SPECMURT ANALYSIS OF MULTI-PITCH MUSIC SIGNALS WITH ADAPTIVE ESTIMATION OF COMMON HARMONIC STRUCTURE

Shoichiro Saito, Hirokazu Kameoka, Takuya Nishimoto and Shigeki Sagayama
Graduate School of Information Science and Technology
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
E-mail: {saito,kameoka,nishi,sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

This paper describes a multi-pitch analysis method using *specmurt analysis* with iterative estimation of the quasi-optimal common harmonic structure function. Specmurt analysis (Sagayama et al., 2004) is based upon the idea that superimposed harmonic structure pattern can be expressed as a convolution of two components, a fundamental frequency distribution and a 'common harmonic structure' function if each underlying tone component has similar harmonic structure pattern. As proved in our previous work (Sagayama et al., 2004) inappropriate common structure function leads to inaccurate analysis results. The iterative algorithm proposed in this paper automatically chooses a proper structure, which results in finding concurrent multiple fundamental frequencies and reduces the dependency on heuristically chosen initial common harmonic structure. The experimental evaluation showed promising results.

Keywords: audio feature extraction, specmurt analysis, visualization of the fundamental frequency.

1 INTRODUCTION

In audio feature extraction, the fundamental frequency is one of the useful features characterizing music structure. Accurately extracted fundamental frequencies can be translated into a musical score by applying some recently developed rhythm recognition techniques as post-processing. What makes it difficult to extract fundamental frequencies from multi-pitch signals is the existence of harmonics. In general, harmonic structure patterns differ among instruments or fundamental frequencies, and also vary along time. Top-down approaches using Graphical models such as Bayesian networks and Hidden Markov Models (HMM) for detecting pitch class of note event (for music transcription use) were recently proposed by Kashino et al. (1995), Raphael (2002), Cemgil et al. (2003). Feder and Weinstein (1988), Goto (2004), Godsill and Davy (2002) and Kameoka et al. (2005) proposed parametric harmonic structure pattern models, that allow flexible estimation. Such parametric models have strong advantages especially for the purpose of precisely analyzing real-world audio signals. Meanwhile, we have proposed *specmurt analysis*, another simple, yet efficient approach to multi-pitch audio signals(Sagayama et al., 2004).

In this method, a *common* harmonic structure is assumed and fundamental frequency distribution is obtained by deconvolution of the power spectrum in log-frequency domain by the common harmonic structure. This is a new method in that fundamental frequency distribution can be obtained as an analytic solution, but at the same time it has the drawback that the accuracy of the calculated fundamental frequency depends on the initial envelope of the common harmonic structure.

This paper proposes an adaptive estimation of the common harmonic structure pattern for each frame which maximizes the resolution between significant fundamentals and other unnecessary components in specmurt analysis through iterative non-linear mapping.

The rest of the paper is organized as follows. After reviewing *specmurt analysis* in section 2, we discuss the algorithm for quasi-optimizing the harmonic structure in section 3 followed by evaluation of the accuracy of the method in section 4 and conclusion in section 5.

2 SPECMURT ANALYSIS

Harmonic signals such as single tones in music contain an energy component of fundamental frequency, f_1, as well as multiple overtones whose frequencies are integral number multiples, $nf_1, n = 2, 3, 4, \ldots$, of the fundamental frequency. In linear frequency scale f, the distance between fundamental frequency and the n-th harmonic frequency is $(n-1)f_1$ and depends on the fundamental frequency. In logarithmic frequency (log-frequency) scale $x = \log f$, on the other hand, the harmonic frequencies are located log2, $\log 3$, ..., $\log n$ away from the fundamental log-frequency, $x_1 = \log f_1$; the relative positions remain constant no matter how the fundamental frequency fluctuates and is an overall parallel shift depending on the degree of fluctuation (see Fig. 1).

We now define a *common* harmonic structure pattern, $h(x)$, as a function of log-frequency, x, choosing the origin $x = 0$ at the fundamental frequency, and $h(0) = 1$. We assume that the relative powers at harmonic frequencies are constant and independent of the fundamental frequency. If $h(x)$ is shifted by Δx to the direction in which x increases, this pattern represents the harmonic structure pattern whose fundamental frequency is Δx. Therefore, if power spectrum is additive[1], the multi-pitch power spectrum, $v(x)$, is represented by the convolution of the common harmonic structure, $h(x)$, and the power distribution function, $u(x)$, representing the power of the fundamental frequency component at log-frequency x, i.e.,

$$v(x) = h(x) * u(x) \ . \tag{1}$$

Conversely, if a multi-pitch spectrum $\tilde{v}(x)$ is observed, we can calculate the distribution of fundamental frequencies, $u(x)$, by deconvolution:

$$u(x) = h(x)^{-1} * \tilde{v}(x) \ . \tag{2}$$

According to the convolution theorem, convolution becomes multiplication in the frequency domain by Fourier transform. Suppose that $U(y)$, $H(y)$ and $\tilde{V}(y)$ are the inverse Fourier-transformed function of $u(x)$, $h(x)$ and $\tilde{v}(x)$, respectively. $U(y)$ is obtained by dividing $\tilde{V}(y)$ by $H(y)$:

$$U(y) = \frac{\tilde{V}(y)}{H(y)} \ , \tag{3}$$

where y means Fourier transformed log-frequency. Therefore, we can calculate $u(x)$ as the Fourier transform of $U(y)$:

$$u(x) = \mathcal{F}\left[U(y)\right] \tag{4}$$

In this way, we can estimate the fundamental frequency distribution from the multi-pitch spectrum. We call this method "*specmurt analysis*" (Sagayama et al., 2004).

The illustration of this process is briefly shown in Fig. 2, 3. The process is done over every short-time analysis frame and thus we finally have a time series of fundamental frequency components, i.e., a piano-roll-like visual representation, with a small amount of computation.

In short-time analysis, spectrum does not look like an impulse (Dirac's delta-function) sequence, but has some broadening coming from the influence of window-function, etc. . .. When we use specmurt method, it is necessary to even up the width of broadening of the spectrum at every (log) frequency. This problem is solved by using wavelet transform and constant-Q filter in calculating $v(x)$.

We created the word "specmurt" for the Fourier transform of linear-scaled spectrum along log-frequency axis by reversing the last four letters in "spectrum" following

[1]Strictly speaking, this assumption is true only in the expectation sense. The power of the sum of sinusoids of identical frequency is not always equal to the sum of the powers of the sinusoids, but depends on the phases of the sinusoids.

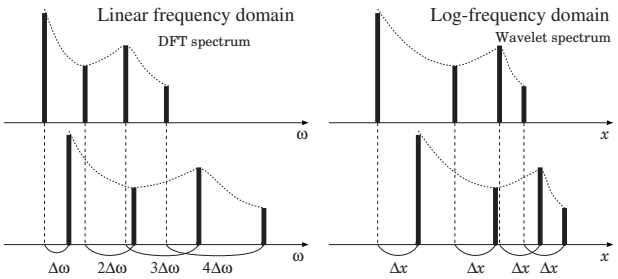


Figure 1: Linear- and log-scaled harmonic structures

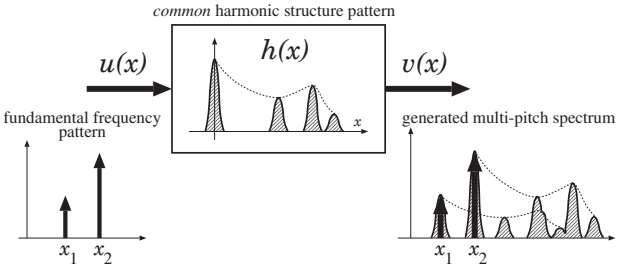


Figure 2: Conceptual diagram of convolution with $h(x)$

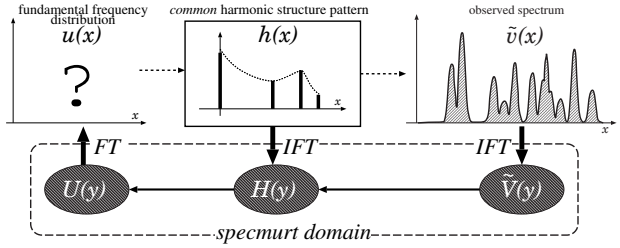


Figure 3: Outline of *specmurt analysis* for multi-pitch signals

the terminology of *cepstrum* which was created by reversing the first four letters of "spectrum"and represents the Fourier transform of log-scaled spectrum with linear frequency. It should be noted that spectrum logarithmically scaled both in frequency and in magnitude is identical to Bode diagram often used in the automatic control theory. Its Fourier transform has no specific name, while it is essentially similar to mel-scaled frequency cepstrum coefficients (MFCC) and is very often used in the feature analysis in speech recognition.

3 OPTIMIZATION OF HARMONIC STRUCTURE PATTERN

3.1 The Role of Common Harmonic Structure

The common harmonic structure pattern, which is described in Section 2, is based on the assumption that the pattern is constant regardless of the sound source, but actually the harmonic structure pattern depends on the sound source and the fundamental frequency. In addition, there is a quite low possibility that the default harmonic structure pattern corresponds with the optimal harmonic structure pattern of the sound source, and it is unrealistic to

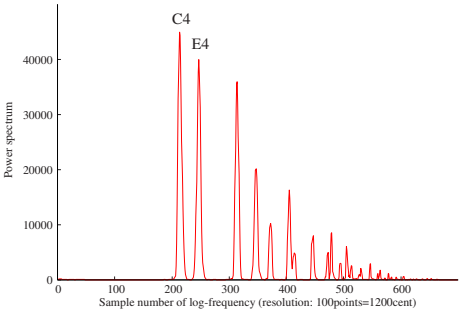
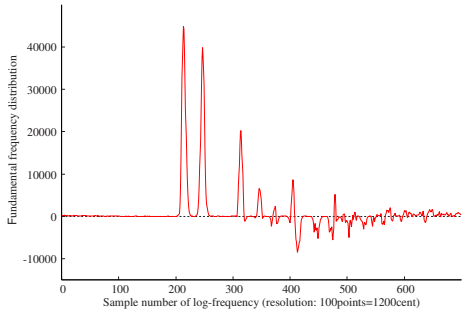
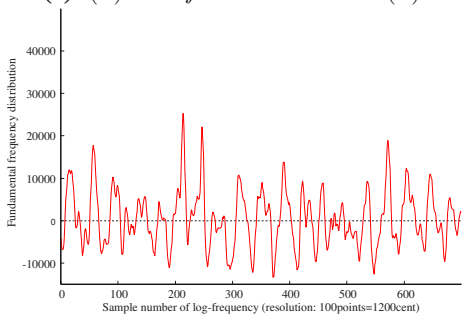


Figure 4: Power spectrum (linear scale) of multi-pitch audio signal of violin's C4 and E4 with log-scaled frequency.



(a)$u(x)$ with $f^{-1.5}$ chosen for $h(x)$



(b)$u(x)$ with $f^{-0.5}$ chosen for $h(x)$

Figure 5: Estimated fundamental frequency distribution, $u(x)$, obtained by specmurt analysis. Figs. (a) and (b) chose $f^{-1.5}$ and $f^{-0.5}$, respectively, along linear frequency f as the envelope of the common harmonic structure $h(x)$.

change the default pattern little by little until finding the optimal pattern.

Fig. 4 shows the multi-pitch spectrum obtained by adding the real audio signal of violin's C4 and E4. Horizontal axis is in log-frequency domain. The fundamental frequency distribution $u(x)$, calculated from $v(x)$ of Fig. 4, is shown in Fig. 5. In case of (a), the envelope of the common harmonic structure pattern $h(x)$ is assumed to be $f^{-1.5}$ (the n-th harmonic component has a power ratio of $1/\sqrt{n^3}$ relative to the fundamental frequency component) and in the case of (b) it is assumed to be $f^{-0.5}$. In Fig. 5 (a), the first harmonic overtone still has a large power because $f^{-1.5}$ envelope has small effect for the attenuation of harmonic overtones. On the other hand, in Fig. 5(b), though the second harmonic overtone is relatively reduced in power, there arise heavily fluctuating power and many unwanted components as well as negative components in the entire range of frequency.

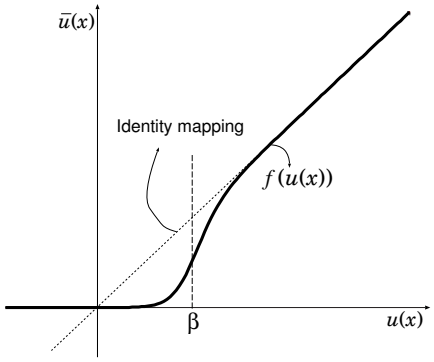


Figure 6: Concept of mapping from $u(x)$ to $\bar{u}(x)$

For the purpose of solving this problem, we propose an adaptive estimation algorithm. This algorithm consists of two steps and generates quasi-optimal $h(x)$ through iterations. First, in Section 3.2 (step I), we renew $u(x)$ calculated by specmurt analysis to a more preferable (or more accurate) distribution $\bar{u}(x)$, and in Section 3.3 (step II) we parameterize $h(x)$ in $\bar{h}(x, \mathbf{\Theta})$ and estimate $\mathbf{\Theta}$ to optimize $\bar{h}(x, \mathbf{\Theta})$. This estimated harmonic structure pattern $\bar{h}(x)$ is quasi-optimal, and we can generate a more accurate fundamental frequency distribution $u(x)$ by applying $\bar{h}(x)$ to specmurt analysis. The following describes these steps in detail.

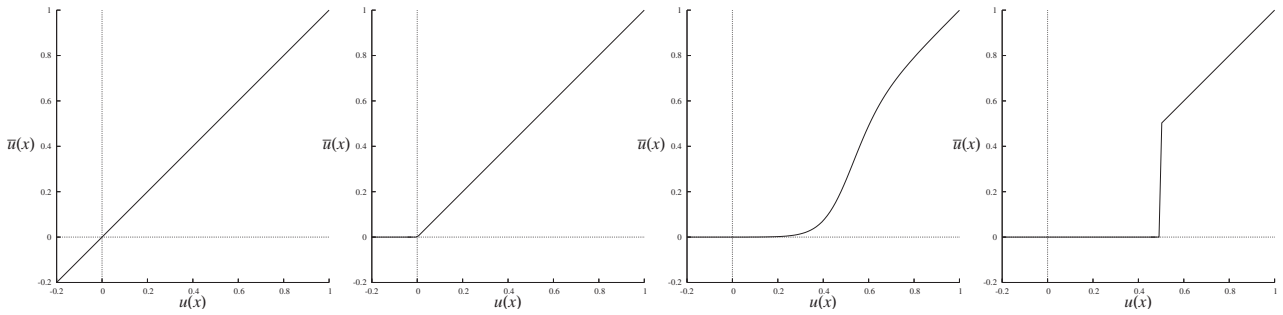
3.2 Step I: Non-Linear Mapping of Fundamental Frequency Distribution

It is difficult to distinguish definitely between true fundamental frequency components and unwanted frequency components in $u(x)$ obtained through specmurt analysis. In consideration of this problem, we introduce a non-linear mapping function to update fundamental frequency distribution, which has a fuzziness parameter α, and a threshold magnitude parameter β. β means the value under which frequency components are assumed to be unwanted, and α represents the degree of fuzziness of the boundary ($\alpha > 0$).

As an example of this function, we propose a mapping based on the sigmoid function as follows:

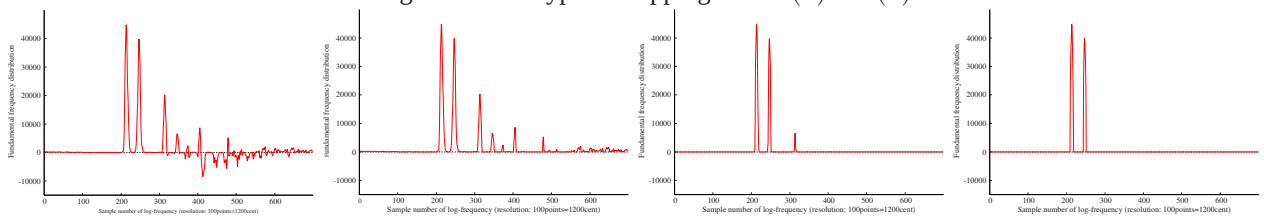
$$\bar{u}(x) = \frac{u(x)}{1 + \exp\{-\alpha(u(x) - \beta)\}} \tag{5}$$

This mapping returns almost the same value when $u(x)$ is significantly larger than β, and a much smaller value when $u(x)$ is smaller than or near β. Fig. 6 shows a sketch of mapping from $u(x)$ to $\bar{u}(x)$. $\bar{u}(x)$ is 50% of $u(x)$ when $u(x) = \beta$, and for other values of $u(x)$, the suppression effect of this mapping depends on α. When α is small, the suppression effect is small and suppression range is wide. If α is large enough, the mapping is similar to a threshold model. In addition, if β is a large negative value, this mapping is approximately equal to the identity mapping. Fig. 7 shows four typical mappings of $u(x)$ to $\bar{u}(x)$ and Fig. 8 shows the experimental results of applying mappings in Fig. 7 to fundamental frequency distribution of Fig. 5 (a) ($f^{-1.5}$ envelope).



(a) $\alpha = 15000,\ \beta = -10000$ (b) $\alpha = 15000,\ \beta = 0$ (c) $\alpha = 15,\ \beta = 0.5$ (d) $\alpha = 15000,\ \beta = 0.5$

Figure 7: Four typical mapping from $u(x)$ to $\bar{u}(x)$.



(a) $\alpha = 15000,\ \beta = -10000$ (b) $\alpha = 15000,\ \beta = 0$ (c) $\alpha = 15,\ \beta = 0.5$ (d) $\alpha = 15000,\ \beta = 0.5$

Figure 8: Sigmoid-mapped fundamental frequency distribution of Fig. 5(a).

By using this mapping, the components of $u(x)$ that have small or negative power are brought close to zero, and the middle power components are still left place for remaining as major fundamental frequency components, unlike a simple threshold mapping.

3.3 Step II: Optimization of the Common Harmonic Structure Pattern

$\bar{u}(x)$ generated in step I is more accurate than $u(x)$ in that the unwanted components are suppressed, therefore a supposed function $\bar{h}(x)$, which is generated by deconvolution of observed multi-pitch spectrum $\tilde{v}(x)$ with $\bar{u}(x)$, will be closer to the actual harmonic structure pattern.

We now define $\bar{h}(x, \boldsymbol{\Theta})$, a function parameterized by Θ_n, which is the amplitude ratio of the n-th harmonic overtone relative to fundamental frequency (shown in Fig. 9). Thus $\bar{h}(x, \boldsymbol{\Theta})$ is described as follows:

$$\begin{aligned} \bar{h}(x, \boldsymbol{\Theta}) &= \Theta_0 \delta(x - \Omega_o) + \cdots + \Theta_N \delta(x - \Omega_N) \\ &= \sum_{n=0}^{N} \Theta_n \delta(x - \Omega_n) \end{aligned} \tag{6}$$

where Ω_n is the x-coordinate of the n-th harmonic overtone on log-frequency scale, and so $\Theta_0 = 1$ and $\Omega_0 = 0$. Then an ideal multi-pitch spectrum $\bar{v}(x)$, generated by convolution of $\bar{h}(x, \boldsymbol{\Theta})$ and $\bar{u}(x)$ calculated in step I, is also parameterized by $\boldsymbol{\Theta}$, and we write it $\bar{v}(x, \boldsymbol{\Theta})$. Actually, $\bar{v}(x, \boldsymbol{\Theta})$ cannot be completely matched observed spectrum $\tilde{v}(x)$, hence we want to know the parameter $\bar{\boldsymbol{\Theta}} = \{\bar{\Theta}_1, \cdots, \bar{\Theta}_N\}$ that minimizes the integral square error between $\tilde{v}(x)$ and $\bar{v}(x, \boldsymbol{\Theta})$. The integral square error is

$$\int_{-\infty}^{\infty} \{\tilde{v}(x) - \bar{v}(x, \boldsymbol{\Theta})\}^2 \, \mathrm{d}x \ , \tag{7}$$

which is rewritten in discrete calculation by:

$$\sum_{i=0}^{I-1} \left\{ \tilde{v}(x_i) - \bar{v}(x_i, \boldsymbol{\Theta}) \right\}^2 \tag{8}$$

where I indicates the number of log-frequency samples. Differentiating eq.8 partially in $\boldsymbol{\Theta}$ and making it equal to zero, the equation below is obtained:

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{\Theta}} \sum_{i=0}^{I-1} \left\{ \sum_{n=0}^{N} \Theta_n u(x_i - \Omega_n) \right\}^2 \\ &= 2 \frac{\partial}{\partial \boldsymbol{\Theta}} \sum_{i=0}^{I-1} \tilde{v}(x_i) \left\{ \sum_{n=0}^{N} \Theta_n u(x_i - \Omega_n) \right\} \end{aligned} \tag{9}$$

Hence, we obtain the following system of linear equations:

$$\begin{pmatrix} A_{1,1} & \cdots & A_{1,n} & \cdots & A_{1,N} \\ \vdots & & \vdots & & \vdots \\ A_{n,1} & \cdots & A_{n,n} & \cdots & A_{n,N} \\ \vdots & & \vdots & & \vdots \\ A_{N,1} & \cdots & A_{N,n} & \cdots & A_{N,N} \end{pmatrix} \begin{pmatrix} \Theta_1 \\ \vdots \\ \Theta_n \\ \vdots \\ \Theta_N \end{pmatrix} = \begin{pmatrix} B_1 \\ \vdots \\ B_n \\ \vdots \\ B_N \end{pmatrix} , \tag{10}$$

where

$$A_{j,k} = \sum_{i=0}^{I-1} u(x_i - \Omega_j) u(x_i - \Omega_k) \ , \tag{11}$$

$$B_j = \sum_{i=0}^{I-1} \left\{ \tilde{v}(x_i) - u(x_i) \right\} u(x_i - \Omega_j) \ . \tag{12}$$

Now we can work out the system by calculating the inverse of the matrix on the left side, for example with the use of LU decomposition. Then updating $h(x)$ using the

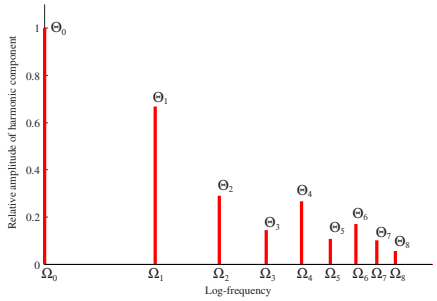
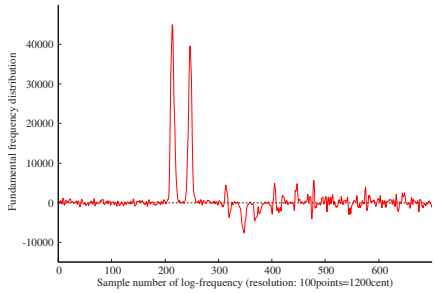
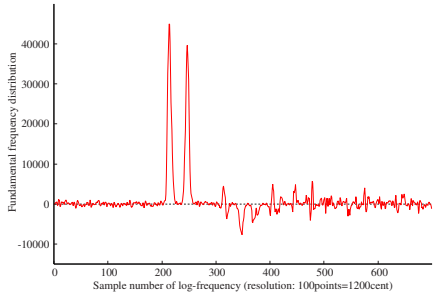


Figure 9: Parameterized harmonic structure pattern $\bar{h}(x, \mathbf{\Theta})$



(a) Starting from $f^{-1.5}$ for $h(x)$



(b) Starting from $f^{-0.5}$ for $h(x)$

Figure 10: Improved fundamental frequency distribution of Fig. 5 after 5 iterations ($\alpha = 15, \beta = 0.5$) starting from Fig. 5(a) and (b), respectively.

optimal parameter $\bar{\mathbf{\Theta}}$ and applying it to specmurt analysis, we obtain the fundamental frequency distribution again. We get back to step I, and iterate.

3.4 Results of Iteration on a Certain Frame

Fig. 10 shows the fundamental frequency distributions which are generated from the distribution of Fig. 5(a) and (b) through 5 iterations of proposed algorithm. In both distribution C4 and E4 are properly emphasized and unwanted components are strongly suppressed.

4 EXPERIMENTAL EVALUATION

4.1 Visualization

Another aspect of specmurt analysis is that the fundamental frequency distribution resulting from the process can be easily visualized. That is, unlike the estimation of various parameters, the result is much comprehensible for

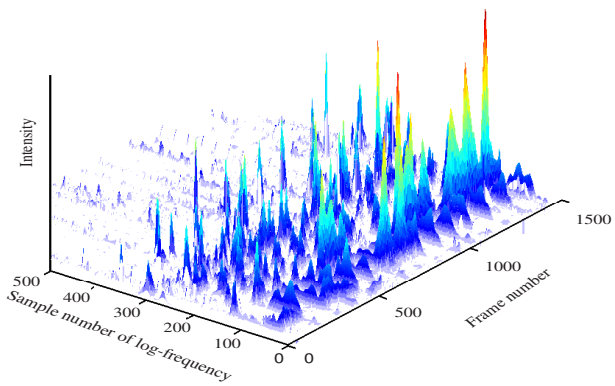


Figure 11: 3D-view of fundamental frequency distribution of data8 (piano) obtained by specmurt analysis through 5 iterations ($\alpha = 15, \beta = 0.5$).

users. An example is shown in Fig. 11. The figure is generated from the $u(x)$ of data8 in Table 2 after 5 iterations where α =15 and $\beta = 0.5$, In this figure, time-frequency-intensity space is considered. Note that intensity does not mean whether a frequency is fundamental or not, but gives a hint about the possibility that a frequency is fundamental. By looking at this figure the user can understand the distribution intuitively and if estimation error occurs, for example of the tone at a certain pitch actually sounds but specmurt analysis drop it out, they can look for the next candidate easily.

The 2-dimensional views of the fundamental frequency distribution are shown in Fig. 12 – 15. Fig. 12 displays the contrast density of the power spectrum at each short frame. Fig. 14 shows the fundamental frequency distribution $u(x)$ without using iteration and Fig. 15 with 5 iterations. Looking at Fig. 12, one can see that there are a lot of overtones, but in Fig. 14 many of overtones disappear or strongly attenuated, and In Fig. 15 much more overtones are attenuated. By means of iterative estimation, Fig. 15 comes closer to Fig. 13, the correct fundamental frequency distribution.

Here, let us have a closer look on two segments of the Fig. 15, which are between the frame approximately 470 –560 and approximately 1200 – 1300. In the first segment, a double-F0 error is made in estimation. However, double-F0 is also much stronger in Fig. 12. That is to say, the harmonic structure in this segment has a kind of missing fundamental feature, and specmurt analysis cannot estimate it correctly in this case in principle. In the second segment, the fundamental frequency near sample number 300 disappears, while in Fig. 14 it exists. This is because the frequency at 300 is the fourth harmonic of the frequency at 100 whose amplitude is largest in these frames. As estimation is iterated, the harmonic structure of the lower and stronger fundamental frequency takes in the harmonic structure of the higher and weaker fundamental frequency and it seems to be optimal. These are weak points of specmurt analysis and we need to examine

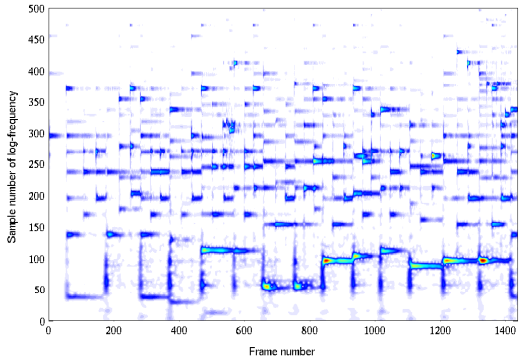


Figure 12: Observed spectrogram of data8 (piano) where overlapping harmonics make it difficult to follow multiple fundamentals. (Used as input to specmurt analysis)

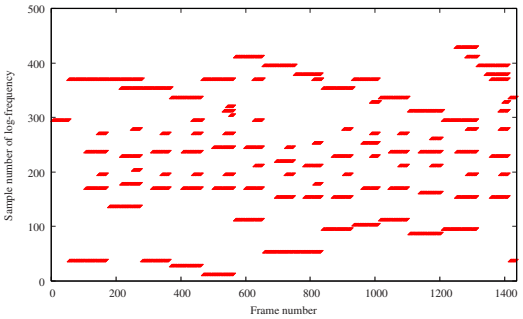


Figure 13: Handcrafted MIDI reference of fundamental frequency time pattern (piano-roll display) for data8, each red line indicating a single note event activation.

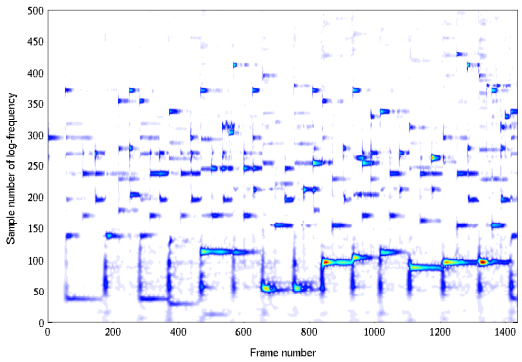


Figure 14: Fundamental frequency of data8 obtained through specmurt analysis with a fixed common harmonic structure.

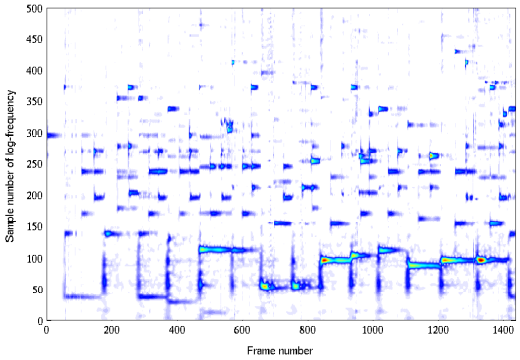


Figure 15: Estimated fundamental frequency of data8 obtained through specmurt analysis with adaptive estimation of common harmonic structure after 5 iterations.

the issues.

4.2 Preparation

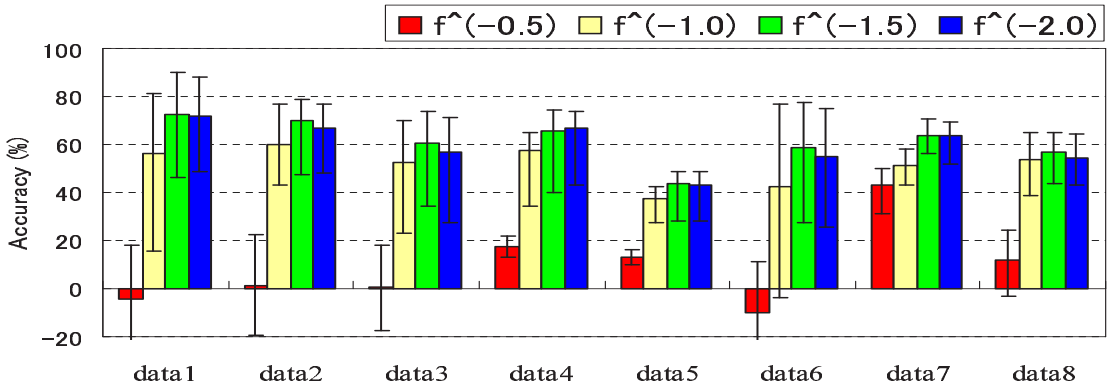
In this section, we applied the specmurt analysis with iterative algorithm to several music signals to obtain their fundamental frequency distributions and evaluate their accuracy. Specmurt analysis is not a method to estimate discrete parameters or states but, so to speak, to emphasize the (continuous) fundamental frequency distribution. Therefore, there is no obvious criterion for evaluation of the accuracy, but as one strategy, we considered a certain frequency as being in "ON" state if the amplitude was over a certain threshold intensity. To evaluate the accuracy, we first divided the fundamental frequency distribution at each time and frequency into two states ("ON" and "OFF") by comparing the peak of distribution with the threshold. After that we compared the obtained two-state table with correct two-state table and calculated the accuracy.

To compare two tables, we used DP matching algorithm for each note number. Matching perfectly in all note numbers the accuracy was 100%. If there were i deletion and j insertion errors, the accuracy was defined as $100 * (N - i - j)/N$, where N is the number of "ON" in the reference table.

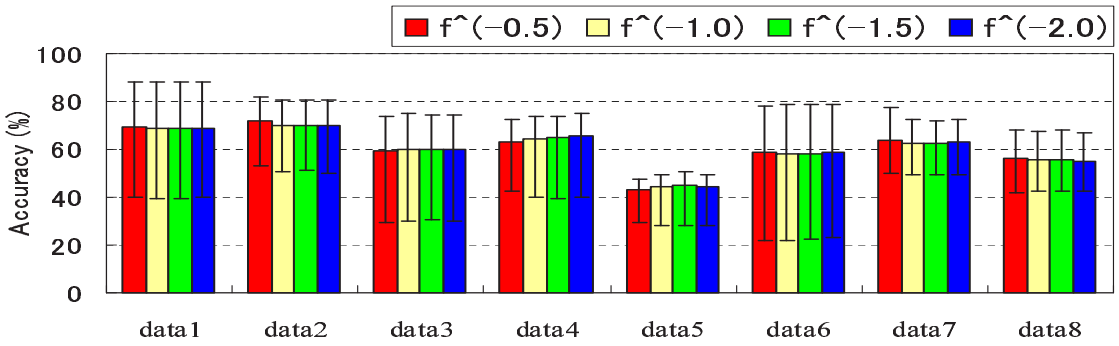
Table 1: Experimental conditions for specmurt analysis.

analysis	sampling rate	16kHz (monaural)
	frame shift	16 ms
	wavelet function	Gabor function
	γ	50
	frequency resolution	12 cent
	frequency bandwidth	60 – 7626.95 Hz
$h(x)$	number of harmonics	8
sigmoid mapping		$\alpha = 15, \beta = 0.5$
number of iterations		5

The strategy to determine threshold amplitude was as follows. First, counting the positive amplitude of $u(x)$ at every time and frequency and making an amplitude histogram, we chose as a threshold candidate the amplitude such that x percent largest amplitudes where higher than it. In other words, this threshold turned only the x percent largest amplitudes into "ON" state. This time, we adopted largest 8 threshold amplitudes, or x is 1, 2, ..., 8. Note that this strategy is temporary and arbitrary, and other policies can be employed.



(a)Results from initial common harmonic structure (no iteration)



(b) 5 iterations to optimize the common harmonic structure

Figure 16: Multi-pitch accuracy of specmurt analysis evaluated over 8 music pieces with errorbars showing the range between the highest and the lowest accuracies. (a) Results with no iteration. (b) Results with 5 iterations.

4.3 Evaluation of Effectiveness of Adaptive Estimation

In this experiment, we used 8 pieces of real music performance data excerpted from RWC music database (listed in Table 2). 23s long WAV files were used in specmurt analysis and MIDI files were used to make correct "ON" table. The power spectrum in log-frequency domain $\tilde{v}(x)$ was calculated using Gabor wavelet transform. Other experiment conditions are shown in Table 1. We calculated the accuracy on both conditions that iteration algorithm is not applied and that it is applied 5 times while changing the initial envelope of $h(x)$ among four types from $f^{-0.5}$ to $f^{-2.0}$.

The results are shown in Fig. 16. In Fig. 16(a) iteration algorithm was not used, and in Fig. 16(b) used 5 times. The height of a box means the average of the accuracy values of 8 thresholds. The bottom of an errorbar means the minimum value among the 8 accuracy values and the top means the maximum value.

When no iterative estimation was used, the accuracy was dependent on the initial envelope of the harmonic structure and, especially in case of $f^{-0.5}$ initial envelope, the accuracy was rather low. On the other hand, in Fig. 16(b), the dependency on the initial envelope disappears and all of the accuracy values were as high as or higher than the maximum accuracy value of Fig. 16(a).

4.4 Evaluation of Sigmoid Mapping Parameter

In this experiment, we evaluated the accuracy depending on different sigmoid parameters, α and β. As shown in Fig. 7, we selected four typical mappings. In addition, we prepared 12 types of mapping, by changing the value of β. The initial envelope of $h(x)$ was $f^{-1.5}$, and other experiment conditions and used music database were the same as in the previous experiment. But unlike the previous experiment, because the average accuracy and whole tendency show little difference, we decided to use in the evaluation the maximum accuracy of 8 data obtained by changing the threshold from 1% to 8%.

The results are shown in Table 3. In each field, we show the maximum accuracy for each data and set of parameters.

There were small differences between the results, but the accuracy tended to be higher when using the type (c) of sigmoid mapping. It would be unwise to conclude that this tendency is general, but one can think that if you choose type (c) as the sigmoid mapping, the accuracy will not be small. This needs to be analyzed in more details.

Table 2: List of The Experimental Data Excerpted from RWC Music Database

Symbol	Title (Genre)	Catalog number	Composer/Player	Instruments
data1	Crescent Serenade (Jazz)	RWC-MDB-J-2001 No. 9	S. Yamamoto	Guitar
data2	For Two (Jazz)	RWC-MDB-J-2001 No. 7	H. Chubachi	Guitar
data3	Jive (Jazz)	RWC-MDB-J-2001 No. 1	M. Nakamura	Piano
data4	Lounge Away (Jazz)	RWC-MDB-J-2001 No. 8	S. Yamamoto	Guitar
data5	For Two (Jazz)	RWC-MDB-J-2001 No. 2	M. Nakamura	Piano
data6	Jive (Jazz)	RWC-MDB-J-2001 No. 6	H. Chubachi	Guitar
data7	Three Gimnopedies no. 1 (Classic)	RWC-MDB-C-2001 No. 35	E. Satie	Piano
data8	Nocturne no.2, op.9-2(Classic)	RWC-MDB-C-2001 No. 30	F. F. Chopin	Piano

Table 3: Maximum accuracy for each sigmoid parameter

α	β	**data1**	**data2**	**data3**	**data4**	**data5**	**data6**	**data7**	**data8**
15000	-10000	**90.2%**	78.9%	73.8%	74.1%	49.0%	77.5%	70.8%	65.0%
15000	0	88.6%	76.5%	71.3%	73.9%	47.3%	73.3%	66.5%	62.0%
15	0.2	87.9%	79.1%	72.9%	**76.1%**	49.1%	76.3%	68.6%	65.2%
15	0.3	88.3%	79.7%	73.6%	74.3%	49.4%	77.5%	69.3%	65.6%
15	0.4	88.6%	**80.9%**	74.1%	74.8%	49.4%	78.6%	72.1%	67.2%
15	0.5	87.9%	80.7%	74.3%	73.5%	**50.9%**	**78.9%**	**72.2%**	68.2%
15	0.6	88.1%	79.3%	**75.4%**	75.6%	48.8%	78.4%	71.9%	66.6%
15000	0.2	87.5%	78.8%	73.4%	75.3%	49.1%	75.4%	67.8%	64.6%
15000	0.3	88.1%	79.6%	73.3%	75.4%	49.3%	76.2%	71.2%	65.4%
15000	0.4	88.5%	79.5%	73.9%	75.7%	49.3%	77.4%	71.4%	67.1%
15000	0.5	88.8%	80.4%	74.5%	74.8%	49.2%	78.2%	72.0%	67.0%
15000	0.6	87.9%	80.3%	74.1%	73.4%	50.0%	78.7%	71.5%	**68.6%**

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed a method of iterative estimation of quasi-optimal harmonic structure pattern in specmurt analysis. We have discussed the two steps of the algorithm and the analytical calculation of the quasi-optimal harmonic structure. This method gives a more accurate fundamental distribution which depends less on the initial common harmonic structure pattern. In addition, we could visualize the fundamental frequency distribution. Specmurt analysis is a user-friendly method and the proposed iteration algorithm makes specmurt analysis more accurate and robust.

There is, however, still a room for improving the method. In step I of the iteration, there is no mathematical guarantee that the mapping from $u(x)$ to $\bar{u}(x)$ takes $u(x)$ closer to optimum (and this is why we say "quasi" optimizing). This mapping strategy is based on the assumption that the power of fundamental frequency components should not take negative or low values. Though this rarely harms the convergence of iterative estimation of harmonic structure pattern (actually observed in a few frames out of over 1000 frames), we wish to solve the optimality and stability problems of the iterative estimation.

References

A. T. Cemgil, B. Kappen, and D. Barber. Generative model based polyphonic music transcription. In *Proc. WASPAA2003*, 2003.

M. Feder and E. Weinstein. Parameter estimation of superimposed signals using the em algorithm. *ASSP*, 36 (4):477–489, 1988.

S. Godsill and Manuel Davy. Baysian harmonic models for musical pitch estimation and analysis. In *Proc.IEEE, International Conference on Acoustics, Speech, and Signal Processing (ICASSP2002)*, pages 1769–1772, 2002.

M. Goto. A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *ISCA Journal*, 43(4):311–329, 2004.

H. Kameoka, T. Nishimoto, and S. Sagayama. Minimum bic estimate of harmonic kernel regression model for multi-pitch analysis. *Trans. IEEE.(submitted)*, 2005.

K. Kashino, K. Nakadai, and H. Tanaka. Organization of hierarchical perceptual sounds: Music scene analysis with autonomous processing modules and a quantitive information integration mechanism. In *Proc. IJCAI*, volume 1, pages 158–164, 1995.

C. Raphael. Automatic transcription of piano music. In *Proc.International Conference on Music Information Retrieval(ISMIR2002)*, pages 15–19, 2002.

S. Sagayama, K. Takahashi, H. Kameoka, and T. Nishimoto. "specmurt analysis: A piano-roll-visualization of polyphonic music signal by deconvolution of log-frequency spectrum". In *SAPA2004*, pages in CD–ROM, 2004.