Sinusoidal modeling for piano tones

Wai Man SZETO
Office of University General Education
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
wmszeto@cuhk.edu.hk

Kin Hong WONG
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
khwong@cse.cuhk.edu.hk

***Abstract*—In sinusoidal modeling, a musical sound is represented by a sum of time-varying sinusoidals. Here, we propose a sinusoidal model specifically designed for piano tones and develop an iterative method to estimate the parameters of the model. Our model assumes the following: (1) the input signals are isolated piano tones, and (2) there may be more than one instances of the same pitch. We also have designed a spectral pick-peaking method to estimate the number of partials and the partial frequencies of piano tones. Experiments on real piano signals show that our model gives high modeling quality and yields better quality in modeling than those reported in the previous work.**

***Index Terms*—Music signal processing, sinusoidal modeling, piano sound.**

I. Introduction

In sinusoidal modeling, a musical sound is represented by a sum of time-varying sinusoidals. Sinusoidal modeling is effective for the sounds generated from pitched musical instruments such as piano because the vibrating system of a pitched instrument vibrates at the resonant frequencies. The goal of sinusoidal modeling is to estimate the parameter values of each sinusoidal. Sinusoidal modeling is widely used in the applications of signal analysis, music synthesis, audio effects, audio coding, music transcription and music source separation [1], [2]. Here, we propose a sinusoidal model specifically designed for piano tones and develop an iterative method to estimate the parameters of the model. In particular, we are interested in applying sinusoidal modeling to address the problems of music source separation of piano music signals. This paper is our preliminary work towards building a complete music source separation system for extracting individual tones from a mixture of piano tones.

There is numerous work using sinusoidal modeling to address the problems of monaural music source separation [3], [4], [5], [6], [7], [8], [9]. In our project, we make use of the fact that the input signals are piano sounds. This allows us to design a piano-specific sinusoidal model with high modeling quality. Moreover, in piano music, a particular pitch rarely appears only once. The tones of the same pitch share the same frequency structure which can be captured by our sinusoidal model. In this paper, we work on the following problem: given instances of isolated piano tones with the same pitch, the task is to estimate the parameters in our sinusoidal model which represents these tones. The pitch of these tones has been known in advance. It can be obtained by using pitch detection algorithms [2].

The rest of the paper is organized as follows. Section II-A gives a brief overview of sinusoidal modeling of music signals. The properties of piano sounds, which is essential in designing a piano-specific sinusoidal model, will be covered in Section II-B. Our proposed sinusoidal model, will be presented in Section III. Then, the parameter estimation will be examined in Section IV. In Section V, we will present the experimental results of our sinusoidal model on real piano signals and compare our model to another model. A conclusion will be given in Section VI.

II. Background

A. Sinusoidal modeling

For a music signal, due to its time-varying property, it is commonly analyzed in short-time segments called *frames* [2]. The duration of a frame is usually from 10 ms to 100 ms. Each segment is multiplied by a window function to smooth the boundaries across frames. A musical sound x is segmented into frames as below:

$$x_r[l] = w[l]\mathsf{x}[(r-1)D + l] \tag{1}$$

where $x_r[l]$ is the rth frame at the local time index l where $l = 0, 1, \ldots L-1$ and L is the window length, $w[l]$ is the window function, and D is the hop size. The typeface x denotes the entire piano tone while the typeface x refers to the windowed segment of a frame.

The signal of a frame can be represented by sinusoidal modeling which uses a sum of sinusoidals to represent the signal. Sinusoidal modeling is a well-established technique to model audio signals including speech signals [10] and music signals [1]. A frame-wise sinusoidal model of a musical sound $\widehat{\mathsf{x}}_r$ at the rth frame can be written as below:

$$\widehat{x}_r[l] = \sum_{m=1}^{M_r} w[l]\left(\alpha_{m,r}\cos(2\pi f_{m,r}t_l) + \beta_{m,r}\sin(2\pi f_{m,r}t_l)\right) \tag{2}$$

where M_r is the number of sinusoidals, $\alpha_{m,r}$ is the amplitude of the cosine component, $\beta_{m,r}$ is the amplitude of the sine component, $f_{m,r}$ is the frequency, t_l is the time in seconds at the index l so $t_l = l/f_s$ and f_s is the sampling frequency in Hz. In sinusoidal modeling, the parameters M_r, $\alpha_{m,r}$,

$\beta_{m,r}$ and $f_{m,r}$ are usually fixed within a frame but they can be different across frames. This models the time-varying properties of music signals. To reconstruct or resynthesize the entire signal from the sinusoidal model, the parameter values between two frames can be estimated by some interpolation methods such as [10]. Another reconstruction approach is to overlap and add all estimated signals in the frames [2].

B. Properties of piano tones

When a piano key is pressed, the hammer hits the strings of the corresponding key. Then the strings vibrate and the energy transfers from strings to the soundboard, and the sound radiates from the soundboard. The resulting sound can be analyzed by using the spectrogram which shows how the spectrum changes along the time. The spectrogram of a C4 piano tone is depicted in Figure 1. The spectrogram shows that the piano tone consists of its frequency components and noise. The frequency components, also called partials, correspond to the resonance frequencies of the strings. The frequency values of the partials in piano tones are stable against time and also instances. In piano sound, the partials of a tone are usually not exactly harmonic. If the partials are exactly harmonic, the frequencies of the partials are exact multiples of the fundamental frequency, and the frequency ratios between the partials are $1:2:3:4:5$ and so on. For piano tones, the frequency ratios are slightly stretched. The frequency ratios of the first five partials in Figure 1 are $1.0000:2.0000:3.0033:4.0075:5.0163$. This phenomenon is called *inharmonicity* and it is caused by the bending stiffness of the strings [11]. Inharmonicity is perceptually significant for the sound quality of pianos [11] so harmonicity cannot be assumed for modeling piano tones. The spectrogram also shows that the amplitude of each partial is time-varying. It generally follows a rapid rise and then a slow decay.

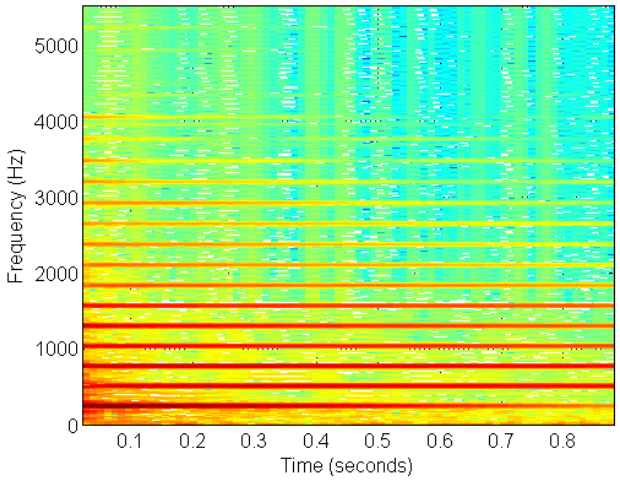

Figure 1. Spectrogram of a C4 piano tone played moderately loud.

III. Signal Model

According to previous section, the frequencies of the partials for a piano tone are stable so the frequencies can be fixed across frames. The number of partials can also be fixed for a tone. Moreover, in piano music, a particular pitch rarely appears only once. The tones of the same pitch share the same partial frequencies. Then the problem of sinusoidal modeling is formulated as follows: given multiple instances of isolated piano tones with the same pitch, the task is to estimate the parameters in our sinusoidal model which represents isolated piano tones. We introduce the instance index i and rewrite the model in (2) with fixed partial frequencies across frames and instances:

$$\widehat{x}_r^i[l] = \sum_{m=1}^{M} w[l]\left(\alpha_{m,r}^i \cos(2\pi f_m t_l) + \beta_{m,r}^i \sin(2\pi f_m t_l)\right) \tag{3}$$

where M is the number of partials, f_m is the frequency of the mth partial, and $\alpha_{m,r}^i$ and $\beta_{m,r}^i$ are the amplitudes of the mth partial of the ith instance at the rth frame. For the window function w, the hamming window is used in this research. The number of partials M of each pitch is assumed to be fixed. The estimation of the number of partials will be discussed in Section IV-A.

Then the observed tone is the sum of the estimated tone and the noise term:

$$x_r^i[l] \quad = \quad \widehat{x}_r^i[l] + v_r^i[l] \tag{4}$$

where $v_r^i[l]$ is the noise component.

To estimate the parameters in each frame of an instance, it is convenient to rewrite the model in (3) as the matrix form. Let $\mathbf{H}$ be the frequency matrix of the piano tone and it is an L-by-$2M$ matrix in the form of

$$H[l,u] = \begin{cases} w[l]\cos(2\pi f_u t_l) & \text{if } 1 \le u \le M, \\ w[l]\sin(2\pi f_{u-M} t_l) & \text{if } M+1 \le u \le 2M \end{cases} \tag{5}$$

so the matrix $\mathbf{H}$ contains two blocks. The left and right blocks contain the cosine and sine terms respectively.

The amplitudes of the cosine and sine terms of the ith instance at the rth frame can be expressed as a $2M$-dimensional vector $\mathbf{g}_r^i$ as below

$$g_r^i[u] = \begin{cases} \alpha_{u,r}^i & \text{if } 1 \le u \le M, \\ \beta_{u-M,r}^i & \text{if } M+1 \le u \le 2M \end{cases} \tag{6}$$

which gives $\mathbf{g}_r^i = \left[\alpha_{1,r}^i \ \cdots \ \alpha_{M,r}^i \ \beta_{1,r}^i \ \cdots \ \beta_{M,r}^i\right]^{\mathrm{T}}$.

The estimated tone $\widehat{\mathbf{x}}_r^i$ of the ith instance at the rth frame can be written as

$$\widehat{\mathbf{x}}_r^i = \mathbf{H}\mathbf{g}_r^i. \tag{7}$$

A series of $\widehat{\mathbf{x}}_r^i$ at different frames can be written in the matrix form $\widehat{\mathbf{X}}^i$ which is the concatenation of the column vectors $\widehat{\mathbf{x}}_r^i$ so that $\widehat{\mathbf{X}}^i = \left[\widehat{\mathbf{x}}_1^i \ \cdots \ \widehat{\mathbf{x}}_{R^i}^i\right]$ where $\widehat{\mathbf{x}}_r^i$ is the rth column of $\widehat{\mathbf{X}}^i$, $\widehat{\mathbf{X}}^i$ is an $L \times R^i$ matrix and R^i is the number

of frames in the ith instance. Then the signal model in (7) is rewritten as

$$\widehat{\mathbf{X}}^i = \mathbf{H}\mathbf{G}^i \tag{8}$$

where the matrix $\mathbf{G}^i$ is the amplitude matrix of the ith instance and it is the concatenation of the column vectors $\mathbf{g}_r^i$ so that $\mathbf{G}^i = \left[\widehat{\mathbf{g}}_1^i \ \cdots \ \widehat{\mathbf{g}}_{R^i}^i\right]$ where $\widehat{\mathbf{g}}_r^i$ is the rth column of $\mathbf{G}^i$ and $\mathbf{G}^i$ is a $2M \times R^i$ matrix.

The estimated $\widehat{\mathbf{X}}^i$ is related to observed $\mathbf{X}^i$ in the form of

$$\mathbf{X}^i = \widehat{\mathbf{X}}^i + \mathbf{V}^i \tag{9}$$

where $\mathbf{V}^i$ is the noise matrix of the ith instance and each element in $\mathbf{V}^i$ is modeled as the zero-mean Gaussian noise with the variance $\sigma_{V^i}^2$. The noise variance $\sigma_{V^i}^2$ is modeled to be the same for all frames for simplicity, but each instance has its own noise variance. The noise variances are grouped into the noise variance vector $\boldsymbol{\sigma}_V^2 = [\sigma_{V^1}^2 \ \cdots \ \sigma_{V^I}^2]^{\mathrm{T}}$.

All instances of $\widehat{\mathbf{X}}^i$ in (8) can be written as

$$\widehat{\mathbf{X}} = \mathbf{H}\mathbf{G} \tag{10}$$

where $\widehat{\mathbf{X}} = \left[\widehat{\mathbf{X}}^1 \ \cdots \ \widehat{\mathbf{X}}^I\right]$ and $\mathbf{G} = \left[\mathbf{G}^1 \ \cdots \ \mathbf{G}^I\right]$. The size of the matrix $\widehat{\mathbf{X}}$ is $L \times R$ and that of the matrix $\mathbf{G}$ is $2M \times R$ where $R = \sum_{i=1}^{I} R^i$.

The goal of the parameter estimation is to estimate both the frequency matrix $\mathbf{H}$ and the amplitude matrix $\mathbf{G}$ so that $\widehat{\mathbf{X}}$ can be found and each instance can be reconstructed.

IV. Parameter estimation

A. Estimation of the number of partials

In the sinusoidal model, we have assumed that the number of partials M of each tone is known. In this section, we will show how M can be found. The values of M are different for different pitches. Lower pitch usually has more partials than the higher. In some research such as [4], [12], M is dynamically estimated. However, this estimation is very computationally intensive. As we have already known that the signals are piano sounds, we estimate the average number of partials required for each pitch from different pianos. Once M for each pitch is determined, it will be fixed for all experiments in Section V.

For each instance of tones with the same pitch, we estimate the frequency values of partials up to $f_s/2$ where f_s is the sampling frequency. The frequency estimation is done by our proposed spectral peak-pick method tailored for piano tones. Then we choose the number of the partials that contains 99.5% of the power of all partials picked.

The steps for picking the spectral speaks are described below:

1) Perform Discrete Fourier Transform (DFT) of a tone.
2) Find the first partial (fundamental frequency) f_1:
   a) Set f_1^{mid} to the equal-tempered fundamental frequency of the pitch. For example, f_1^{mid} of the pitch A4 is 440 Hz.
   b) Set f_1 to the frequency corresponding to the peak of the magnitude spectrum in the frequency range $[2^{-1/48} f_1^{\text{mid}}, 2^{1/48} f_1^{\text{mid}}]$.
3) Set the inharmonicity coefficient $B^{(0)} = 0$ which is defined in (11) and (12).
4) Find f_m for $m \geq 2$ where f_m is the frequency of the mth partial:
   a) Find f_m^{mid} by
   $$f_m^{\text{mid}} = m f_1 \sqrt{\frac{1 + m^2 B^{(q)}}{1 + B^{(q)}}} \tag{11}$$
   which is the general formula to model the inharmonicity effect for pianos [13, p. 363]. A typical value for the inharmonicity coefficient B is 0.0004 in the middle range of piano keys.
   b) Set f_m to the frequency corresponding to the peak of the magnitude spectrum in the frequency range $[2^{-1/48} f_m^{\text{mid}}, 2^{1/48} f_m^{\text{mid}}]$. If $2^{1/48} f_m^{\text{mid}} > f_s/2$, set the upper bound to $f_s/2$.
5) Update B
   $$B_u = \frac{(f_u/u f_1)^2 - 1}{u^2 - (f_u/u f_1)^2} \tag{12}$$
   Set $B^{(q+1)}$ to the median of all B_u for $1 \leq u \leq m$.
6) Repeat the steps 4-5 until $f_m^{\text{mid}} > f_s/2$ so the frequencies of all partials can be estimated.

B. Finding the initial guess of frequencies by peak-picking

The method of estimating both the frequency matrix $\mathbf{H}$ and the amplitude matrix $\mathbf{G}$, which will be discussed in the next section, starts with an initial guess of the partial frequencies. This initial guess can be found by using the frequency estimation method in Section IV-A. Given an instance $\mathbf{x}^i$, we first find the frequency spectrum by DFT, the peaks are chosen by the iterative method described in Section IV-A. The locations of the peaks are a set of frequencies $\{f_{m,\text{PP}}^i\}$ where $m = 1, \ldots, M$. The initial guess of a partial frequency for extracting a partial is the average of the frequency from peak-picking of all instances. This gives the initial guess in the form

$$f_m^{(0)} = \frac{1}{I} \sum_{i=1}^{I} f_{m,\text{PP}}^i \tag{13}$$

which will be used as the input for the extraction of partials described in the next section.

C. Estimation of the frequency vector $\mathbf{f}$ and the amplitude matrix $\mathbf{G}$

The matrix $\widehat{\mathbf{X}}$ is governed by the amplitude matrix $\mathbf{G}$ and the frequency matrix $\mathbf{H}$ which depends on the frequency vector $\mathbf{f}$ where $\mathbf{f} = [f_1 \ \cdots \ f_M]^{\mathrm{T}}$. The goal of the parameter estimation is to estimate $\mathbf{f}$, $\mathbf{G}$ and $\boldsymbol{\sigma}_V^2$. The weighted least-squares method is used to estimate these parameters. The weights are the inverse of the noise variances $\sigma_{V^i}^2$. The objective function to be minimized is written as

$$E(\mathbf{f}, \mathbf{G}, \boldsymbol{\sigma}_V^2) = \sum_{i=1}^{I} \frac{1}{\sigma_{V^i}^2} \left|\left|\mathbf{X}^i - \widehat{\mathbf{X}}^i\right|\right|_F^2 \tag{14}$$

where $||\cdot||_F$ denotes the Frobenius norm. The summation operation in (14) can be expressed in matrix form. Let $\mathbf{\Sigma}_V$ be the covariance matrix so that

$$\mathbf{\Sigma}_V = \mathrm{diag}(\sigma_{V^1}^2 \mathbf{1}_{LR^1}, \ldots, \sigma_{V^I}^2 \mathbf{1}_{LR^I}) \tag{15}$$

where $\mathbf{1}_{LR^i}$ denotes the LR^i-dimensional column vector filled with 1's. Then (14) can be presented as

$$E(\mathbf{f}, \mathbf{G}, \boldsymbol{\sigma}_V^2) = \left|\left| \mathbf{\Sigma}_V^{-1/2} \left(\mathbf{X}_{\mathrm{vec}} - \widehat{\mathbf{X}}_{\mathrm{vec}} \right) \right|\right|^2. \tag{16}$$

In [14], an iterative least-squares scheme is developed to alternatively update the frequencies and the amplitudes in the general sinusoidal model in (2) for one single frame. Based on this scheme, we propose a scheme to handle the frames of all instances together by using iterative-reweighted least-squares in [15], [16] to minimize the objective function in (16). The minimization is achieved via the procedures summarized in Figure 2:

1) Given $\mathbf{f}$, update $\mathbf{G}$.
2) Given $\mathbf{f}$ and $\mathbf{G}$, update $\boldsymbol{\sigma}_V^2$.
3) Given $\mathbf{G}$ and $\boldsymbol{\sigma}_V^2$, update $\mathbf{f}$.
4) Repeats steps 1 to 3 until convergence.

The iterative update starts with the input frequencies $\mathbf{f}^{(0)}$ found in (13) which are estimated by the peak-picking method described in Section IV-B. We find that 100 iterations are good for convergence. In the followings, each step will be discussed in details.

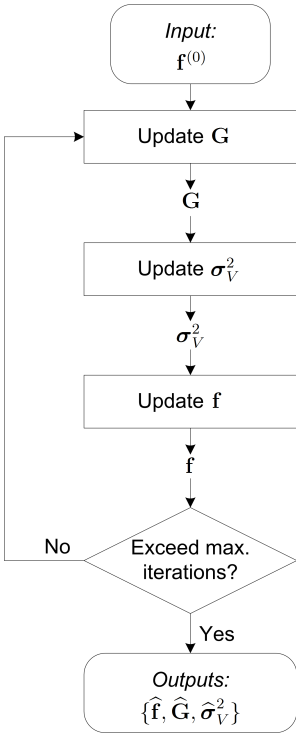

Figure 2. The procedures of the parameter estimation.

1) Step 1: update the amplitude matrix $\mathbf{G}$*:* In Step 1, the frequency matrix $\mathbf{H}$ is calculated from $\mathbf{f}$ by (5). Given $\mathbf{H}$ and the observed tones $\mathbf{X}$, the sinusoidal model becomes a linear model. Then the solution to (14) for updating $\mathbf{G}$ is

$$\mathbf{G} \leftarrow (\mathbf{H}^{\mathrm{T}}\mathbf{H})\mathbf{H}^{\mathrm{T}}\mathbf{X}. \tag{17}$$

Note that the noise variances $\boldsymbol{\sigma}_V^2$ are not involved in updating $\mathbf{G}$ because given $\mathbf{H}$, each $\widehat{\mathbf{X}}^i$ has its independent $\mathbf{G}^i$.

2) Step 2: update the noise variances $\boldsymbol{\sigma}_V^2$*:* Given the updated $\mathbf{G}$ in Step 1, the new estimate $\widehat{\mathbf{X}}$ can be calculated by

$$\widehat{\mathbf{X}} \leftarrow \mathbf{HG}. \tag{18}$$

Then each noise variance $\sigma_{V^i}^2$ is estimated as follows

$$\sigma_{V^i}^2 \leftarrow \frac{1}{LR^i} \left|\left| \mathbf{X}^i - \widehat{\mathbf{X}}^i \right|\right|_F^2. \tag{19}$$

3) Step 3: update the frequencies $\mathbf{f}$*:* Given the updated $\mathbf{G}$ and $\boldsymbol{\sigma}_V^2$, the aim is to update the frequency vector $\mathbf{f}$. However, the sinusoidal model is nonlinear with $\mathbf{f}$. The nonlinear model can be linearized by using Taylor's expansion. In [14], [17], a single frame of the sinusoidal model, in which $\mathbf{X}$, $\widehat{\mathbf{X}}$ and $\mathbf{G}$ are only vectors instead of matrices, is linearized by Taylor's expansion. The Gauss-Newton method is used to update $\mathbf{f}$. Based on these work, we derive the update equation using the weighted least-squares for $\mathbf{f}$. The derivation involves two steps. The first step is to vectorize the matrix $\widehat{\mathbf{X}}$ and the second step is to linearize the vectorized $\widehat{\mathbf{X}}$.

The matrix equation $\widehat{\mathbf{X}} = \mathbf{HG}$ can be converted into a vector equation by the vec operator [18, p. 428] and the Kronecker product [18, p. 422]. We rewrite $\widehat{\mathbf{X}} = \mathbf{HG}$ as

$$\widehat{\mathbf{X}} = \mathbf{HGI}_R \tag{20}$$

where $\mathbf{I}_R$ is an $R \times R$ identity matrix. Vectorizing both sides of (20) gives

$$\widehat{\mathbf{X}}_{\mathrm{vec}} = \mathrm{vec}(\mathbf{HGI}_R). \tag{21}$$

Using the identity of the Kronecker product in [18, p. 429], (21) can be written as a vector equation

$$\widehat{\mathbf{X}}_{\mathrm{vec}} \quad = \quad (\mathbf{I}_R \otimes \mathbf{H})\mathbf{G}_{\mathrm{vec}} \tag{22}$$

which is equivalent to

$$\underbrace{\begin{bmatrix} \widehat{\mathbf{x}}_1^1 \\ \widehat{\mathbf{x}}_2^1 \\ \vdots \\ \widehat{\mathbf{x}}_r^i \\ \vdots \\ \widehat{\mathbf{x}}_{R^I}^I \end{bmatrix}}_{\widehat{\mathbf{X}}_{\mathrm{vec}}} = \underbrace{\begin{bmatrix} \mathbf{H} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{H} & & & & \mathbf{0} \\ \vdots & & \ddots & & & \vdots \\ \mathbf{0} & & & \mathbf{H} & & \mathbf{0} \\ \vdots & & & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{H} \end{bmatrix}}_{(\mathbf{I}_R \otimes \mathbf{H})} \underbrace{\begin{bmatrix} \mathbf{g}_1^1 \\ \mathbf{g}_2^1 \\ \vdots \\ \mathbf{g}_r^i \\ \vdots \\ \mathbf{g}_{R^I}^I \end{bmatrix}}_{\mathbf{G}_{\mathrm{vec}}}$$

Note that each subvector $\widehat{\mathbf{x}}_r^i$ in $\widehat{\mathbf{X}}_{\mathrm{vec}}$ is in the form $\widehat{\mathbf{x}}_r^i = \mathbf{H}\mathbf{g}_r^i$ which matches (7).

In the next step of the derivation, $\widehat{\mathbf{X}}_{\mathrm{vec}}$ is linearized by using Taylor's expansion so that

$$\widehat{\mathbf{X}}_{\mathrm{vec}}(\mathbf{f}) \approx \widehat{X}_{\mathrm{vec}}(\mathbf{f}^{\mathrm{cur}}) + \mathbf{Z}(\mathbf{f}^{\mathrm{cur}})\,(\mathbf{f} - \mathbf{f}^{\mathrm{cur}}) \tag{23}$$

where $\mathbf{f}^{\text{cur}}$ is the current estimate of the frequencies, $\mathbf{f}$ is the vector of new frequencies to be estimated, and $\mathbf{Z}(\mathbf{f}^{\text{cur}})$ is Jacobian matrix evaluated at $\mathbf{f}^{\text{cur}}$ and $\mathbf{Z}$ is in the form

$$\begin{aligned}\mathbf{Z} &= \frac{\partial \widehat{\mathbf{X}}_{\text{vec}}}{\partial \mathbf{f}} \\ &= \left[\frac{\partial \widehat{\mathbf{x}}_1^1}{\partial \mathbf{f}} \; \frac{\partial \widehat{\mathbf{x}}_2^1}{\partial \mathbf{f}} \; \cdots \; \frac{\partial \widehat{\mathbf{x}}_r^i}{\partial \mathbf{f}} \; \cdots \; \frac{\partial \widehat{\mathbf{x}}_{R^I}^I}{\partial \mathbf{f}}\right]^{\mathrm{T}} \\ &= \left[\mathbf{Z}_1^1 \; \mathbf{Z}_2^1 \; \cdots \; \mathbf{Z}_r^i \; \cdots \; \mathbf{Z}_{R^I}^I\right]^{\mathrm{T}} \end{aligned} \tag{24}$$

where we let $\mathbf{Z}_r^i = \partial \widehat{\mathbf{x}}_r^i / \partial \mathbf{f}$ and $\mathbf{Z}_r^i$ is the $L \times M$ Jacobian matrix at the rth frame of the ith instance. An element $Z_r^i[l,m]$ in $\mathbf{Z}_r^i$ is

$$\begin{aligned} Z_r^i[l,m] &= \frac{\partial \widehat{X}^i[l,r]}{\partial f_m} \\ &= 2\pi t_l w[l](-\alpha_{m,r}^i \sin(2\pi f_m t_l) + \beta_{m,r}^i \cos(2\pi f_m t_l)). \end{aligned} \tag{25}$$

Then $\mathbf{Z}$ can be computed from (25).

Using the results in [17, pp. 226, 260], the update equation of $\mathbf{f}$ is

$$\mathbf{f} \leftarrow \mathbf{f} + (\mathbf{Z}^{\mathrm{T}} \boldsymbol{\Sigma}_V^{-1} \mathbf{Z})^{-1} \mathbf{Z}^{\mathrm{T}} \boldsymbol{\Sigma}_V^{-1} (\mathbf{X}_{\text{vec}} - \widehat{\mathbf{X}}_{\text{vec}}) \tag{26}$$

in which Gauss-Newton method is used and where $\boldsymbol{\Sigma}_V$ is the covariance matrix in (15).

4) Summary of the parameter estimation: Here is the summary of all update equations. The update starts with the input frequencies $\mathbf{f}^{(0)}$ defined in (13).

1) Given $\mathbf{f}$, update $\mathbf{G}$. Calculate $\mathbf{H}$ from $\mathbf{f}$. Then
$$\mathbf{G} \leftarrow (\mathbf{H}^{\mathrm{T}}\mathbf{H})\mathbf{H}^{\mathrm{T}}\mathbf{X}. \tag{27}$$
2) Given $\mathbf{f}$ and $\mathbf{G}$, update $\widehat{\mathbf{X}}$ and $\boldsymbol{\sigma}_V^2$
$$\widehat{\mathbf{X}} \leftarrow \mathbf{H}\mathbf{G} \tag{28}$$
$$\sigma_{V^i}^2 \leftarrow \frac{1}{LR^i} \left|\left|\mathbf{X}^i - \widehat{\mathbf{X}}^i\right|\right|_F^2. \tag{29}$$
3) Given $\mathbf{G}$ and $\boldsymbol{\sigma}_V^2$, update $\mathbf{f}$
$$\mathbf{f} \leftarrow \mathbf{f} + (\mathbf{Z}^{\mathrm{T}} \boldsymbol{\Sigma}_V^{-1} \mathbf{Z})^{-1} \mathbf{Z}^{\mathrm{T}} \boldsymbol{\Sigma}_V^{-1} (\mathbf{X}_{\text{vec}} - \widehat{\mathbf{X}}_{\text{vec}}) \tag{30}$$
where $\mathbf{Z}$ is the Jacobian matrix and $\boldsymbol{\Sigma}_V$ is the covariance matrix.
4) Repeats steps 1 to 3 for 100 iterations. The outputs of the extraction of partials are the frequencies $\widehat{\mathbf{f}}$, the amplitude matrix $\widehat{\mathbf{G}}$ and the noise variances $\widehat{\boldsymbol{\sigma}}_V^2$.

V. Experiments

A. Estimation of the number of partials

The piano tone database in [19] is used for estimating the number of partials M as described in Section IV-A. The database contains piano tones from 7 different pianos. Note that this database will only be used in estimating M and it will not be used in evaluating the performance of our sinusoidal model in the next section. After picking all the partials, we choose the number of the partials that contains 99.5% of the power of all partials picked on average. The result is shown in Figure 3.

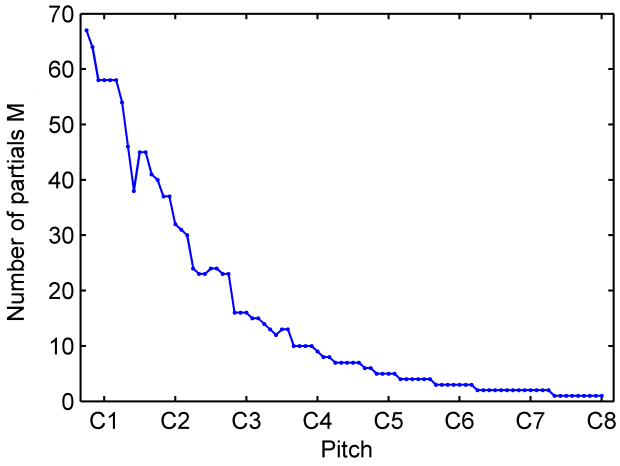

Figure 3. The number of partials M of each pitch.

B. Evaluation on modeling quality

Experiments were performed to test the modeling quality of our sinusoidal model. All data used in the experiments are real signals of piano tones and they are not synthetic. Another piano tone database from four different pianos were used in our experiments. Three of the pianos are from the RWC musical instrument sound database [20] including the grand pianos of Steinway & Sons, Bösendorfer and Yamaha. The remaining piano is a Yamaha Disklavier DU1A upright piano, Mark III series of which we created a piano tone database. Each piano key was played at three different levels of loudness (soft, medium and loud) for each piano. Hence, three instances of each pitch were obtained for each piano. Before performing our experiments, we aligned the instances of a pitch from the same piano in phase by using the cross-correlation method in the following steps:

1) The instance with the medium loudness was selected to be a reference. The onset of the instance was detected by the onset algorithm in [21] and was fine-tuned in our user interface developed in Matlab©. This made the instance to start at time zero.
2) Other instances of the pitch were aligned in phase by time shifting the instances to maximize the cross-correlation between the instances and the reference instance. This alignment also made all instances to start at time zero.

All tones in the piano tone database, including the RWC database and our database, were downsampled from 44.1 kHz to 11.025 kHz for faster processing. The first 0.5 second of the tones were used in the experiments.

In the first experiment, there were 62 pitches randomly selected from 11 piano pieces in the RWC music database including the databases of classical music, jazz music and music genre [20]. For each of these pitches, two instances of different loudness levels were randomly selected from one of the four pianos in the piano tone database described previously. Hence, the number of instances I is equal to 2 and there

were 124 tones for evaluation. The parameter setting for our sinusoidal model is that the window length is 11.61 ms ($L = 128$) with 50% overlapping window.

The performance of our sinusoidal model is evaluated by the signal-to-noise ratio (SNR) which is defined by

$$\text{SNR} = 10 \log_{10} \frac{\sum_n \mathsf{x}(t_n)^2}{\sum_n (\mathsf{x}(t_n) - \widehat{\mathsf{x}}(t_n))^2} \tag{31}$$

where $\mathsf{x}(t_n)$ is the isolated piano tone in the time domain and $\widehat{\mathsf{x}}(t_n)$ is the estimated tone in the time domain. The estimated tone is reconstructed from $\widehat{\mathbf{X}}$ by using the overlap-and-add method in [2]. Higher SNR means higher quality of estimated signals. The result shows that the average SNR of all the 124 tones is 17.62 dB which reflects a high modeling quality.

We also compared our sinusoidal model to a system of monaural source separation (Li's system) in [8] which is also based on sinusoidal modeling. The implementation of Li's system is provided by the authors. All the 124 isolated tones in the previous experiment were inputted to Li's system to test the modeling quality without any mixing. The true fundamental frequency of each tone was supplied to Li's system. The result is shown in Table I. Our model performs better than Li's system for the average SNR.

	Average SNR (dB)	
	Proposed	Li
All 124 tones	17.62	14.84

Table I
COMPARISON OF OUR PROPOSED SINUSOIDAL MODEL AND LI'S SYSTEM.

VI. CONCLUSIONS AND DISCUSSIONS

Here, we have proposed a sinusoidal model specifically designed for piano tones and have developed an iterative method to estimate the parameters of the model. Our model assumes the following: (1) the input signals are isolated piano tones, and (2) there may be more than one instances of the same pitch. Based on this formulation, we have developed a sinusoidal model with fixed partial frequencies across frames and instances. We also have designed a spectral pick-peaking method to estimate the number of partials and the partial frequencies of piano tones. Experiments show that our model performs well and gives better modeling quality than those reported in the previous work.

This paper is our preliminary work towards building a complete music source separation system for extracting individual tones from a mixture of piano tones. One of the challenges in this project is to resolve the overlapping partials. We will try to make use of the input signals known to be piano and multiple instances of the same pitch to resolve the overlapping partials.

ACKNOWLEDGMENT

This work is supported by a direct grant (Project Code: 2050486, project title: Music nuance extraction from audio signals) from the Faculty of Engineering of the Chinese University of Hong Kong.

REFERENCES

[1] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. Poli, Eds. Lisse, the Netherlands: Swets & Zeitlinger, 1997.
[2] U. Zöler, Ed., *DAFX - Digital Audio Effects*, 2nd ed. Wiley, 2011.
[3] T. Virtanen, "Sound source separation in monaural music signals," Ph.D. dissertation, Tampere University of Technology, Finland, November 2006.
[4] M. Davy, S. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2498–2517, April 2006.
[5] E. Vincent and M. D. Plumbley, "Single-channel mixture decomposition using bayesian harmonic models," in *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Source Separation (ICA 2006)*, Charleston, SC, USA, March 2006, pp. 722–730.
[6] M. R. Every and J. E. Szymanski, "Separation of synchronous pitched notes by spectral filtering of harmonics," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 5, pp. 1845–1856, 2006.
[7] J. J. Burred, "From sparse models to timbre learning: New methods for musical source separation," Ph.D. dissertation, Technical University of Berlin, Berlin, Germany, September 2008.
[8] Y. Li, J. Woodruff, and D. Wang, "Monaural musical sound separation based on pitch and common amplitude modulation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1361–1371, 2009.
[9] J. Han and B. Pardo, "Reconstructing completely overlapped notes from musical mixtures," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 249–252.
[10] R. J. McCaulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, August 1986.
[11] A. Askenfelt, Ed., *Five Lectures on the Accoutics of the Piano*. Royal Swedish Academy of Music, 1990, available online at http://www.speech.kth.se/music/5_lectures/.
[12] C. Dubois and M. Davy, "Joint detection and tracking of time-varying harmonic components: A flexible bayesian approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1283–1295, May 2007.
[13] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, 2nd ed. Springer Verlag, 1998.
[14] P. Depalle and L. Tromp, "An improved additive analysis method using parametric modelling of the short-time Fourier transform," in *Proceedings of International Computer Music Conference*, Hong Kong, 1996, pp. 297–300.
[15] R. J. Carroll and D. Ruppert, *Transformation and Weighting in Regression*. New York; London: Chapman and Hall, 1988.
[16] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
[17] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, N.J.: Prentice-Hall, 1993.
[18] T. K. Moon and W. C. Stirling, *Mathematical methods and algorithms for signal processing*. Upper Saddle River, N.J.: Prentice Hall, 2000.
[19] P. M. Instruments, "Piano magic: The PMI piano sample collection," 2005.
[20] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Music genre database and musical instrument sound database," in *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003)*, October 2003, pp. 229–230.
[21] C. H. Wong, W. M. Szeto, and K. H. Wong, "Automatic lyrics alignment for cantonese popular music," *Multimedia Systems*, vol. 12, no. 4-5, pp. 307–323, March 2007.