SOURCE SEPARATION AND ANALYSIS OF PIANO MUSIC SIGNALS USING INSTRUMENT-SPECIFIC SINUSOIDAL MODEL

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

Many existing monaural source separation systems use sinusoidal modeling to represent pitched musical sounds during the separation process. In these sinusoidal modeling systems, a musical sound is represented by a sum of time-varying sinusoidal components, and the goal of source separation is to estimate the parameters of each component. Here, we propose an instrument-specific sinusoidal model tailored for a piano tone. Based on our proposed Piano Model, we develop a monaural source separation system to extract each individual tone from mixture signals of piano tones and at the same time, to identify the intensity and adjust the onset of each tone for characterizing the nuance of the music performance. The major difficulty of the source separation problem is to resolve overlapping partials. Our solution collects the training data from isolated tones to train our Piano Model which can capture the common properties across the reappearance of pitches that helps to separate the mixtures. This approach enables high separation quality even for the case of octaves in which the partials of the upper tone completely overlap with those of the lower tone. The results show that our proposed system gives robust and accurate separation of piano tone signal mixtures (including octaves), with the quality significantly better than those reported in the previous work.

1. INTRODUCTION

Many existing monaural source separation systems use sinusoidal modeling to model pitched musical sounds [1, 2, 3, 4, 5]. 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 source separation based on sinusoidal modeling is to estimate the parameter values of each sinusoidal. Here, we propose an instrument-specific sinusoidal model tailored for a piano tone. Based on our proposed Piano Model (PM), we develop a monaural source separation system to extract each individual tone from mixture signals of piano tones. Specifically, tone extraction can be facilitated by estimating the parameters in PM. In addition to source separation, PM can facilitate the analysis of nuance in an expressive piano performance. Nuance can be defined as the subtle differences in manipulation of sound parameters including attack, timing, pitch, loudness and timbre that makes the music sound alive and human [6]. A major obstacle to a computational analysis of musical nuances is that it is often difficult to uncover relevant sound parameters from mixture signals. This problem can be formulated as a source separation problem.

The major difficulty of the source separation problem is to resolve overlapping partials. As music is usually not entirely dissonant, it is common that some partials from different tones may overlap with each other. For example, octave intervals often appear in piano music. For an octave mixture, the frequencies of the upper tone are totally immersed within those of the lower. Overlapping partials cause a serious problem in separation because a sum of two partials with the same frequency also gives a sinusoidal with that same frequency; there are infinite ways to generate the resulting sinusoidal, so the amplitude and the phase of an overlapped partial cannot be uniquely determined and the overlapping partials cannot be resolved. Hence, we cannot recover the original two partials if only the resulting sinusoidal is given.

In the existing systems, assumptions are made to resolve overlapping partials according to the general properties of pitched musical sounds. For example, the spectral envelope of tones is assumed to be smooth (as in [1, 3]). The information of neighboring non-overlapping partials can also be utilized to estimate the parameters of an overlapping partial. Another assumption is that the amplitude envelope of each partial from the same note tends to be similar [4]. This is known as common amplitude modulation (CAM). Non-overlapping partials are used to estimate the overlapping partials of the same note by the property of CAM. However, these assumptions may not be suitable for the source separation of piano mixtures. For a piano tone, the spectral envelope may not be smooth. Moreover, there may be lack of neighboring non-overlapping partials. For example, the partials of the upper tone in an octave are totally immersed within the frequencies of the lower tone. In such cases, spectral smoothness and CAM cannot be applied. Moreover, the assumption in CAM may not be applied to piano sounds. In Figure 2 (c), the amplitude envelopes of the same note are not similar. Harmonic Temporal Envelope Similarity (HTES) tries to these problems by assuming that the amplitude envelope of a partial evolves similarly among different notes of the same musical instrument [5]. Overlapping partials of a note are reconstructed by the non-overlapping partials of another note. However, the amplitude envelopes can vary significantly across pitches in a piano [7]. Thus, HTES may not resolve the overlapping partials of piano tones accurately.

Instead of formulating assumptions from the general properties of musical sounds, we make use of the fact that the input mixtures in question are piano music signals. This allows us to design an instrument-specific model for the piano sound to accurately resolve overlapping partials. In piano music, a particular pitch rarely appears only once. The tones of the same pitch share some com-

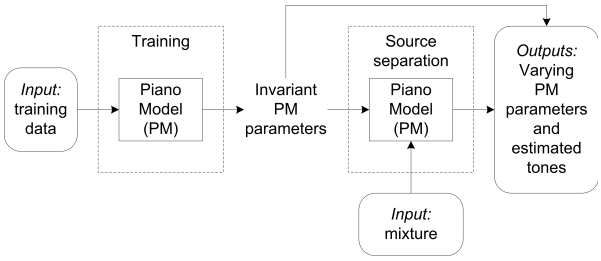

Figure 1: *The main steps of our source separation process.*

mon characteristics which can be captured by PM. In particular, we consider the case when the pitches in the mixtures reappear as isolated tones in the target recording, and when the piano music is performed without pedaling. The isolated tones are used as the training data to train PM. This approach enables high separation quality even for the case of octaves in which the partials of the upper tone completely overlap with those of the lower tone.

The goals of our source separation system are to separate each individual tone from the mixture signal and at the same time, to identify the intensity and adjust the onset of each tone for characterizing the nuance of the music performance. The intensity and fine-tuned onset of a tone will be defined in Section 2.2. The main steps in our source separation system are depicted in Figure 1. The whole separation process is divided into the training stage and the source separation stage. In the training stage, the inputs are the isolated tones from the target recording being investigated. The parameters in PM are estimated. PM contains two sets of parameters. (i) One set contains parameters invariant to instances of the same pitch in the recording. (ii) Another set consists of parameters which may vary across instances. The goal of the training stage is to estimate the invariant model parameters so that they can be used in the source separation stage. If the invariant PM parameters of a mixture are known, only the varying PM parameters are required to be estimated. In the source separation, the varying PM parameters, which include the intensity and fine-tuned onsets, are estimated. Signals of the individual tones in the mixtures can be reconstructed by PM.

The rest of the paper is organized as follows. Our proposed Piano Model with the properties of piano tones will be presented in Section 2. Then, parameter estimations in the training stage and the source separation will be examined in Sections 3 and 4 respectively. Section 5 will show the experimental results of our source separation process on real piano signals including octaves and compare our system to another system. A conclusion will be given in Section 6.

2. SIGNAL MODEL

In this research, an individual tone (the sound of hitting one piano key) is considered as a particular sound source of the corresponding pitch. When multiple piano keys are pressed, a mixture signal is generated. We model a mixture signal as a sum of its corresponding individual tones that can be expressed as $\mathsf{y}(t) = \sum_{k=1}^{K} \mathsf{x}_k(t)$ where $\mathsf{y}(t)$ is the observed mixture signal in the time domain, K is the number of tones in the mixture, $\mathsf{x}_k(t)$ is the kth individual tone in the mixture, and t is the time in seconds. This model is called instantaneous linear mixing in the literature of general source separation. The pitch of each $\mathsf{x}_k(t)$ is given. This

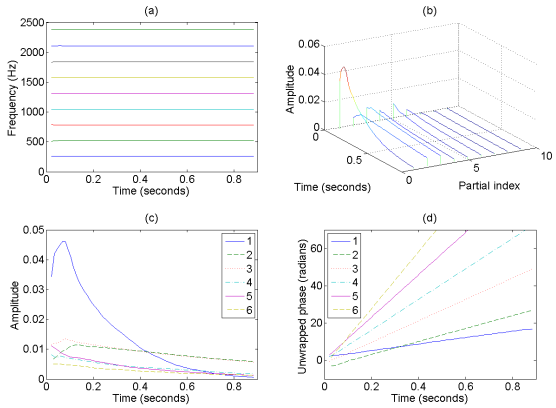

Figure 2: *Extracted partials of a C4 piano tone. The partials are extracted by using the method in [11] with time-varying frequencies. (a) The frequencies of the first nine partials against time. (b) The amplitude of the first nine partials against time. The partial index one corresponds to the fundamental frequency. (c) The amplitude of the first six partials against time. (d) The unwrapped phase of the first six partials against time.*

information can be obtained by using music transcription systems [8]. The goal of our research is to recover the signal of each individual tone $\mathsf{x}_k(t)$ from the mixture signal $\mathsf{y}(t)$. For a piano tone, it consists of a set of time-varying sinusoidals. We use a sum-of-sinusoidal model to represent $\mathsf{x}_k(t)$ - our proposed Piano Model (PM). PM can capture the common properties across the reappearance of pitches that helps to separate the mixtures. The model is formulated according to the properties of piano tones.

2.1. Properties of piano tones

A piano tone consists of its frequency components and noise. The frequency components, also called partials, are usually dominating over the noise and are stable against time. In piano sound, the partials of a tone are usually not exactly harmonic. This phenomenon is called *inharmonicity* and it is perceptually significant for the sound quality of pianos [9]. Hence, the assumption of harmonicity cannot be taken for modeling piano tones. The amplitude of each partial generally follows a rapid rise and then a slow decay. The rapid rise is the building up of the sound. The slow decay is the damping of the sound and it is exponential-like [10]. Note that each partial has its own rate of rising and decaying. The peaks of the partials exhibit a general trend that a higher partial has a weaker peak than a lower partial but there are irregularities. For the piano tone in Figure 2 (b), the fundamental frequency has the highest peak. The third partial is stronger than the second and the fifth is stronger than the fourth. Figure 2 (d) shows the unwrapped phase against time. The unwrapped phase is linear and the partials can be considered as linear-phase signals.

2.2. Proposed Piano Model

Here, we propose PM to resolve the overlapping partials by exploring the common properties of recurring tones. PM employs a time-varying sum-of-sinusoid signal model for piano tones, and it describes a tone in an entire duration instead of a single analy-

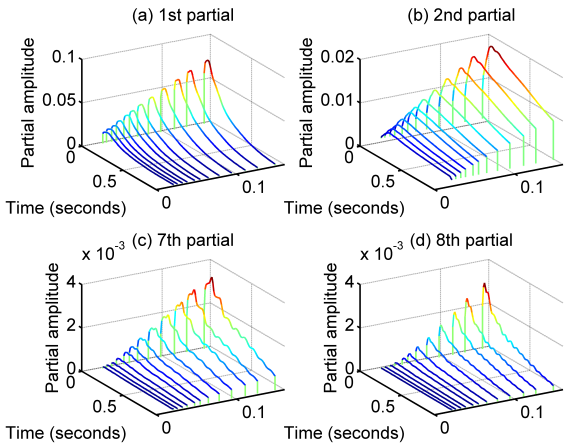

Figure 3: *Envelope surface against peak amplitude of the time-domain signal and time.*

sis frame. For each partial, we aim to model the envelope surface against intensity and time. The intensity of a tone can be measured by the peak amplitude of its time-domain signal. When the key pressing velocity increases, the peak amplitude also increases up to the physical limit of the piano [12]. The envelope surfaces of the first, second, seventh and eighth partials are plotted in Figure 3. The surface is constructed from the extracted partials of the C4 tones from the same piano played with 12 hitting strengths. The partial amplitude and the peak amplitude of the time-domain signal are plotted in the scale that the maximum possible peak amplitude of all input wave files is one.

It is observed that the same partial from various instances of the pitch exhibits a similar shape of rising and decay. When the peak amplitude of the signal increases, the whole partial is also scaled up smoothly. However, this scaling is not the same for all partials. The fact is that a loud note is not a linear amplification of a soft note. High frequency partials are boosted significantly when the key is hit heavily due to nonlinear material property of the piano hammer [9, 7].

In PM, the values of certain parameters do not change across instances of the same pitch. Parameters in the model are divided into two sets: the invariant PM parameters (such as frequencies of partials) and the varying PM parameters (such as the strength of striking a piano key). The invariant PM parameters can be learned from recurring occurrences of the same pitch. The learning process will be fully discussed in Section 3. PM is expressed as below:

$$\widehat{\mathrm{x}}_k(t_n) = \sum_{m=1}^{M_k} a_{k,m}(t_n) \cdot \cos(2\pi f_{k,m} t_n + \phi_{k,m}) \quad (1)$$

where M_k is the number of partials of the kth tone, $f_{k,m}$ and $\phi_{k,m}$ are the frequency and the phase of the mth partial in the kth tone respectively, and $a_{k,m}(t_n)$ is the time-varying amplitude of the partial and it is modeled as a bi-exponential mixture with a nonlinear scaling factor:

$$\begin{aligned} a_{k,m}(t_n) =& a(t_n; c_k, \boldsymbol{\varphi}_{k,m}) \quad (2) \\ =& b_{k,m} \cdot (c_k)^{d_{k,m}} \\ & \cdot \zeta_{k,m} \cdot (\exp\{-\lambda_{k,m} t_n\} - \exp\{-\gamma_{k,m} t_n\}) \quad (3) \end{aligned}$$

where $b_{k,m}$ is the relative amplitude of the mth partial; $d_{k,m}$ controls the significance of the intensity factor c_k; $\lambda_{k,m}$ is the decay rate; $\gamma_{k,m}$ is the rising rate and $\gamma_{k,m} > \lambda_{k,m}$. These envelope parameters are grouped into the parameter set $\boldsymbol{\varphi}_{k,m} = \{b_{k,m}, d_{k,m}, \lambda_{k,m}, \gamma_{k,m}\}$. The intensity factor c_k is assigned to be the peak amplitude of the observed time-domain signal of the tone. All $\alpha_{k,m}, \beta_{k,m}, \gamma_{k,m}, \lambda_{k,m}$ are positive. The term $\zeta_{k,m}$ is the coefficient to normalize the peak of the bi-exponential function $(\exp\{-\lambda_{k,m} t_n\} - \exp\{-\gamma_{k,m} t_n\})$ to one in order to stabilize the optimization process in the parameter estimation. The term $\zeta_{k,m}$ depends on $\lambda_{k,m}$ and $\gamma_{k,m}$:

$$\zeta_{k,m} = \left[\left(\frac{\lambda_{k,m}}{\gamma_{k,m}} \right)^{\frac{\lambda_{k,m}}{\gamma_{k,m} - \lambda_{k,m}}} - \left(\frac{\lambda_{k,m}}{\gamma_{k,m}} \right)^{\frac{\gamma_{k,m}}{\gamma_{k,m} - \lambda_{k,m}}} \right]^{-1} \quad (4)$$

where the derivation of the normalization coefficient is shown in Appendix.

Substituting (2) into (1), we write the estimated signal of a tone in the form

$$\widehat{\mathrm{x}}_k(t_n) = \sum_{m=1}^{M_k} a(t_n; c_k, \boldsymbol{\varphi}_{k,m}) \cdot \cos(2\pi f_{k,m} t_n + \phi_{k,m}). \quad (5)$$

The onset of each tone in the mixture may not be exactly the same so a time-shift factor is introduced for each tone in the estimated mixture $\widehat{\mathrm{y}}(t_n)$:

$$\widehat{\mathrm{y}}(t_n) = \sum_{k=1}^{M_k} \widehat{\mathrm{x}}_k(t_n - \tau_k) \quad (6)$$

where τ_k is the time shift in seconds. The estimated mixture is related to the observed mixture as below:

$$\mathrm{y}(t_n) = \widehat{\mathrm{y}}(t_n) + \epsilon(t_n) \quad (7)$$

where $\epsilon(t_n)$ is the noise term.

All parameters of PM for the kth tone can be grouped into a parameter set $\boldsymbol{\psi}_k$ so $\boldsymbol{\psi}_k = \{\boldsymbol{\varphi}_{k,m}, f_{k,m}, \phi_{k,m}, c_k, \tau_k\}$ which can be divided into two sets: the invariant PM parameters $\boldsymbol{\psi}_{k,\mathbb{I}} = \{\boldsymbol{\varphi}_{k,m}, f_{k,m}, \phi_{k,m}\}$ and the varying PM parameters $\boldsymbol{\psi}_{k,\mathbb{V}} = \{c_k, \tau_k\}$. The invariant PM parameters contain parameters invariant to instances of the same pitch in the recording. The varying PM parameters consist of parameters which may vary across instances. All $\boldsymbol{\psi}_k$ can be grouped into $\boldsymbol{\Psi} = \{\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_K\}$.

The role of the invariant PM parameters $\boldsymbol{\psi}_{k,\mathbb{I}}$ and the varying PM parameters $\boldsymbol{\psi}_{k,\mathbb{V}}$ is shown in Table 1. The key idea is that the invariant PM parameters are estimated from the training data. Given a mixture, only the varying PM parameters of the mixture are required to be estimated. Note that the varying PM parameters including the intensity and the time shift are significant for characterization of musical nuance. As mentioned before, when the key pressing velocity increases, the peak amplitude of the tone in the time domain also increases. Hence, the peak amplitude of the tone can be used as the intensity factor so that the intensity of a tone can be found. The inputs of our source separation system are the mixtures with the onsets detected by a music transcription system. However, existing music transcription systems may not be able to estimate the onsets accurately, and the individual tones in a mixture may not start to sound exactly at the same time. The time shift can be used to obtain the fine-tuned onsets by adding the time shift to the detected onset.

		Training	Source separation
Invariant PM parameters $\boldsymbol{\psi}_{k,\mathbb{I}}$	Envelope parameters $\boldsymbol{\varphi}_{k,m} = \{b_{k,m}, d_{k,m}, \lambda_{k,m}, \gamma_{k,m}\}$	To be estimated	Given
	Frequencies $f_{k,m}$		
	Phases $\phi_{k,m}$		
Varying PM parameters $\boldsymbol{\psi}_{k,\mathbb{V}}$	Intensity c_k	Given	To be estimated
	Time shift τ_k		

Table 1: *Invariant PM parameters and varying PM parameters.*

In PM, we have assumed that the number of partials M_k of each tone is known. The values of M_k are different for different pitches. Lower pitched tones usually have more partials than the higher pitched tones. In some research such as [2, 13], M_k is dynamically estimated. However, this estimation is very computationally intensive. As we know that the mixtures are piano signals, we predetermine M_k. For each pitch in a piano database, we have chosen the number of the partials that contains 99.5% of the power of all partials on average. Note that this database will only be used in estimating M_k and it will not be used in evaluating the performance of our source separation system described in Section 5. The number of partials M_k is fixed for all experiments. The details of finding M_k can be found in [11].

3. TRAINING: PARAMETER ESTIMATION

This section will show how to use the training data to train our proposed Piano Model (PM). The goal of the training stage is to estimate the invariant PM parameters given the training data. The major difficulty of estimating the invariant PM parameters is that PM in (5) is nonlinear. A good initial guess, which is close to the optimal solution, is crucial for accurately estimating the parameters. The procedures for finding a good initial guess will be discussed in Sections 3.2 and 3.3. The main idea is to extract the partials of each isolated tone in the training data, so that the initial guess for the PM parameters for each partial can be found independently. Before discussing how to find the initial guess, the problem of estimating the invariant PM parameters will be formulated first.

3.1. Problem formulation for training

The goal of the training stage is to estimate the invariant PM parameters $\boldsymbol{\Psi}_{\mathbb{I}}$ from the training data $\boldsymbol{\mathcal{X}}$ by finding $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ that maximizes the likelihood $p(\boldsymbol{\mathcal{X}}|\boldsymbol{\Psi}_{\mathbb{I}})$. The invariant PM parameters $\boldsymbol{\Psi}_{\mathbb{I}}$ are divided into K sets so $\boldsymbol{\Psi}_{\mathbb{I}} = \{\boldsymbol{\psi}_{1,\mathbb{I}}, \ldots, \boldsymbol{\psi}_{K,\mathbb{I}}\}$. Each $\boldsymbol{\psi}_{k,\mathbb{I}}$ corresponds to the invariant PM parameters of the pitch p_k. The maximum likelihood solution of $\boldsymbol{\Psi}_{\mathbb{I}}$ is defined as $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$. Note that each $\mathcal{X}_k$ is generated independently from others and $\mathcal{X}_k$ only depends on $\boldsymbol{\psi}_{k,\mathbb{I}}$. This implies that maximizing $p(\boldsymbol{\mathcal{X}}|\boldsymbol{\Psi}_{\mathbb{I}})$ can be done by maximizing each $p(\mathcal{X}_k|\boldsymbol{\psi}_{k,\mathbb{I}})$ independently. Then the training process is performed pitch-by-pitch and each $\mathcal{X}_k$ is processed independently. Each $\mathcal{X}_k$ may contain more than one instance. We introduce the index i to denote the quantities associated with the ith instance and $i = 1, \ldots, I_k$. In this section, the index k is omitted for brevity. Hence, $\mathcal{X}_k$ is rewritten as $\mathcal{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^I\}$ and $\boldsymbol{\psi}_{k,\mathbb{I}}$ is rewritten as $\boldsymbol{\psi}_{\mathbb{I}}$.

Each tone $\mathbf{x}^i$ is represented by its PM estimate $\widehat{\mathbf{x}}^i$. Adding the instance index i, we rewrite PM in (3) and (5) into

$$\widehat{\mathrm{x}}^i(t_n) = \sum_{m=1}^{M} a(t_n; c^i, \boldsymbol{\varphi}_m) \cdot \cos(2\pi f_m t_n + \phi_m) \qquad (8)$$

where $n = 0, \ldots, N^i - 1$ and N^i is the length of $\widehat{\mathbf{x}}^i$. The variable c^i is the intensity factor which is equal to the peak amplitude of the time-domain signal $\mathbf{x}^i$ so c^i is known. Following (6), the observed tone $\mathbf{x}^i$ and the estimated tone $\widehat{\mathbf{x}}^i$ are related by

$$\mathrm{x}^i(t_n) \quad = \quad \widehat{\mathrm{x}}^i(t) + \epsilon^i(t_n) \qquad (9)$$

where $\epsilon^i(t_n)$ is the noise term which is modeled as the zero-mean Gaussian noise with the variance $\sigma^2_{\epsilon^i}$. Note that the time shift factor τ^i in (6) is omitted by setting $\tau^i = 0$. It is because each $\mathbf{x}^i$ is an isolated tone so its onset can be detected by using onset detection algorithms or manually annotated. Then $\mathbf{x}^i$ can be adjusted to start from the time zero.

In summary, the invariant PM parameters $\boldsymbol{\psi}_{\mathbb{I}} = \{\boldsymbol{\varphi}_m, f_m, \phi_m\}$ are estimated in the training stage. The varying PM parameters $\boldsymbol{\psi}_{\mathbb{V}} = \{c^i, \tau^i\}$ are given. The likelihood $p(\mathcal{X}|\boldsymbol{\psi}_{\mathbb{I}})$ is rewritten as $p(\mathcal{X}|\boldsymbol{\psi}_{\mathbb{I}}, \boldsymbol{\sigma}^2_{\epsilon})$ to include the noise variances $\boldsymbol{\sigma}^2_{\epsilon}$ where $\boldsymbol{\sigma}^2_{\epsilon} = \{\sigma^2_{\epsilon^1}, \ldots, \sigma^2_{\epsilon^I}\}$. The likelihood $p(\mathcal{X}|\boldsymbol{\psi}_{\mathbb{I}}, \boldsymbol{\sigma}^2_{\epsilon})$ is expressed in the form

$$p(\mathcal{X}|\boldsymbol{\psi}_{\mathbb{I}}, \boldsymbol{\sigma}^2_{\epsilon}) = \prod_{i=1}^{I} \frac{1}{\left(2\pi\sigma^2_{\epsilon^i}\right)^{N^i/2}} \exp\left\{-\frac{1}{2\sigma^2_{\epsilon^i}} \parallel \mathbf{x}^i - \widehat{\mathbf{x}}^i \parallel^2\right\}. \qquad (10)$$

The goal of the training stage is to find the optimal solution $\widehat{\boldsymbol{\psi}}_{\mathbb{I}}$. As PM is a nonlinear model, a good initial guess, which is close to the optimal solution, is crucial for accurately estimating the parameters. The initial guess is obtained by the following procedures:

1. Extract the partials from each $\mathbf{x}^i$ by using the method in [11]. (Section 3.2)
2. Given the extracted partials, find the initial guess of $\boldsymbol{\psi}_{\mathbb{I}}$ for PM. (Section 3.3)
3. Given the initial guess of $\boldsymbol{\psi}_{\mathbb{I}}$, find the optimal solution $\widehat{\boldsymbol{\psi}}_{\mathbb{I}}$ for PM. (Section 3.4)

3.2. Extraction of partials with the General Model

In [11], we propose a framewise sinusoidal model, called *General Model* (GM), to represent piano tones by extracting the partials. In GM, a framewise sinusoidal model of the ith instance of a particular pitch $\widehat{\mathrm{x}}^i_r$ at the rth frame is represented by a sum of sinusoidals. Here we converted the polar form in [11] into

$$\widehat{x}^i_r[l] = \sum_{m=1}^{M} w[l] \left(a^i_{m,r,\mathrm{GM}} \cos\left(2\pi f_{m,\mathrm{GM}} t_l + \phi^i_{m,r,\mathrm{GM}}\right)\right) \qquad (11)$$

where M is the number of partials, $f_{m,\mathrm{GM}}$ is the frequency of the mth partial and it is fixed across frames and instances, $a^i_{m,r,\mathrm{GM}}$ and $\phi^i_{m,r,\mathrm{GM}}$ are the amplitude and the phase of the mth partial of the ith instance at the rth frame respectively, t_l is the time in seconds at the index l where $l = 0, \ldots L-1$ and L is the window length,

and $t_l = l/f_s$ and f_s is the sampling frequency in Hz. Note that the typeface x denotes the entire piano tone while the typeface x refers to the windowed segment of a frame. The problem of sinusoidal modeling in GM is formulated as follows: given multiple instances of isolated piano tones with the same pitch, the task is to estimate the parameters in GM which represents isolated piano tones. The observed tone, which is the sum of the estimated tone and the noise term, is expressed as $x_r^i[l] = \widehat{x}_r^i[l] + v_r^i[l]$ where $v_r^i[l]$ is the noise term and it is modeled as the zero-mean Gaussian noise with the variance $\sigma_{V^i}^2$. The noise variance is the same for all frames for simplicity, but each instance has its own noise variance.

The extraction of partials with GM in [11] gives an estimate of the frequency $\widehat{f}_{m,\text{GM}}$ of a partial, the amplitude $\widehat{a}_{m,r,\text{GM}}^i$ and the phase $\widehat{\phi}_{m,r,\text{GM}}^i$ of a partial for each frame, and the noise variance $\widehat{\sigma}_{V^i}^2$. These estimates will be used to find the initial guess of each partial for PM. The initial guess for frequency $\widehat{f}_m^{(0)}$ is $\widehat{f}_{m,\text{GM}}$ while the initial guess for the envelope parameters $\boldsymbol{\varphi}_m$, the phase ϕ_m and the noise variance $\boldsymbol{\sigma}_\epsilon^2$ will be discussed below.

3.3. Finding the initial guess for the Piano Model

3.3.1. Finding the initial guess $\boldsymbol{\varphi}_m^{(0)}$

The initial guess of the envelope parameters in PM is found by fitting the envelope function to the amplitudes of each frame from GM. Let t_r' be the time at the center of the rth frame so that

$$t_r' = \left((r-1)\,D + 0.5L\right)/f_s \tag{12}$$

where D is the hop size in samples. Define the envelope function at the center of the rth frame as $a_{m,r}^i(\boldsymbol{\varphi}_m) = a(t_r'; c^i, \boldsymbol{\varphi}_m)$ where $a(\cdot)$ is the envelope function defined in (3), and the intensity c^i, which is the peak amplitude of observed tone $\mathbf{x}^i$ in the time domain, is already known. Fitting $a_{m,r}^i(\boldsymbol{\varphi}_m)$ with $\widehat{a}_{m,r,\text{GM}}^i$ using weighted least-squares, we have the objective function

$$E_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_m) = \sum_{i=1}^{I}\sum_{r=1}^{R^i} \frac{1}{\widehat{\sigma}_{V^i}^2}\left(\widehat{a}_{m,r,\text{GM}}^i - a_{m,r}^i(\boldsymbol{\varphi}_m)\right)^2 \tag{13}$$

where the weights are the inverse of the variances $\widehat{\sigma}_{V^i}^2$. The objective function $E_{\boldsymbol{\varphi}}$ can be minimized by using the trust-region-reflective algorithm implemented in Matlab©. Ten starting points are randomly generated to minimize $E_{\boldsymbol{\varphi}}$. The best solution which gives the smallest $E_{\boldsymbol{\varphi}}$ will be chosen as the initial guess $\boldsymbol{\varphi}_m^{(0)}$ for estimating the PM parameters.

3.3.2. Finding the initial guess $\phi_m^{(0)}$

The phase $\widehat{\phi}_{m,r,\text{GM}}^i$ in GM is the initial phase at the beginning of a frame. In order to perform fitting as finding the initial guess $\boldsymbol{\varphi}_m^{(0)}$, the phase $\widehat{\phi}_{m,r,\text{GM}}^i$ is shifted to the center of a frame. The centered phase $\widehat{\phi}_{m,r,\text{GM}}^{i,\text{cent}}$ is in the form

$$\widehat{\phi}_{m,r,\text{GM}}^{i,\text{cent}} = \widehat{f}_m L/f_s + \widehat{\phi}_{m,r,\text{GM}}^i. \tag{14}$$

The objective function for finding the initial guess $\phi_m^{(0)}$ is also in the form of weighted least-squares which gives

$$E_\phi(\phi_m) = \sum_{i=1}^{I}\sum_{r=1}^{R^i} \frac{1}{\widehat{\sigma}_{V^i}^2}\cdot \left(\widehat{a}_{m,r,\text{GM}}^i \cos(\widehat{\phi}_{m,r,\text{GM}}^{i,\text{cent}}) - \widehat{a}_{m,r,\text{GM}}^i \cos\left(2\pi\widehat{f}_m t_r' + \phi_m\right)\right)^2 \tag{15}$$

where $\widehat{a}_{m,r,\text{GM}}^i \cos(\widehat{\phi}_{m,r,\text{GM}}^{i,\text{cent}})$ is the partial generated by the GM estimate, and $\widehat{a}_{m,r,\text{GM}}^i \cos\left(2\pi\widehat{f}_m t_r' + \phi_m\right)$ is the partial generated by PM. The weights are also the inverse of the variances $\widehat{\sigma}_{V^i}^2$. The objective function E_ϕ is again minimized by using the trust-region-reflective algorithm. There are 30 starting points randomly generated as E_ϕ is more sensitive to the starting points than $E_{\boldsymbol{\varphi}}$. The best solution will be chosen as the initial guess $\phi_m^{(0)}$.

3.4. Parameter estimation of the Piano Model

For efficient computation, the maximum likelihood solution of $\boldsymbol{\psi}_{\mathbb{I}}$ will be approximated by the weighted least-squares solution. Assuming that the noise variance in PM is directly proportional to that in GM, this means that $\sigma_{\epsilon^i}^2 \propto \sigma_{V^i}^2$ so the noise variance $\sigma_{\epsilon^i}^2$ in PM can be replaced by the noise variance $\widehat{\sigma}_{V^i}^2$ in GM. Note that the value of $\widehat{\sigma}_{V^i}^2$ is fixed for finding $\widehat{\boldsymbol{\psi}}_{\mathbb{I}}$. Replacing $\sigma_{\epsilon^i}^2$ by $\widehat{\sigma}_{V^i}^2$ and omitting the constant terms, we can rewrite (10) as the following objective function

$$E_{\text{train}}(\boldsymbol{\psi}_{\mathbb{I}}) = \sum_{i=1}^{I}\left(\frac{1}{2\widehat{\sigma}_{V^i}^2}\left|\left|\mathbf{x}^i - \widehat{\mathbf{x}}^i\right|\right|^2\right). \tag{16}$$

Given the initial guess $\boldsymbol{\psi}_{\mathbb{I}}^{(0)} = \{\boldsymbol{\varphi}_m^{(0)}, \widehat{f}_m^{(0)}, \phi_m^{(0)}\}$ for all m in PM, parameter estimation of PM can be done by minimizing the objective function E_{train} in (16) by using the trust-region-reflective algorithm. The outputs are the estimated invariant PM parameters $\widehat{\boldsymbol{\psi}}_{\mathbb{I}}$ which will be used in the source separation process.

4. SOURCE SEPARATION: PARAMETER ESTIMATION

Given the invariant PM parameters $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ estimated in the previous section and the mixture $\mathbf{y}$, we perform the source separation by estimating the varying PM parameters $\boldsymbol{\Psi}_{y,\mathbb{V}}$ for the mixture $\mathbf{y}$. The varying PM parameters $\boldsymbol{\Psi}_{y,\mathbb{V}}$ include the intensity c_k and the time shift τ_k for each kth tone in the mixture. The output of this stage is the estimated varying PM parameters $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ which maximize the likelihood function of $\boldsymbol{\Psi}_{y,\mathbb{V}}$. With $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ and $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ in PM, the signals of each individual tone in the mixture can be reconstructed by using PM.

The noise term $\epsilon(t_n)$ in (7) is modeled as the zero-mean Gaussian noise. Hence, the maximization of the likelihood is equivalent to the minimization of the least-squares errors. Then given the mixture $\mathbf{y}$ and the estimated invariant PM parameters $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$, the objective function for source separation with PM is

$$E_{\text{sep}}(\boldsymbol{\Psi}_{y,\mathbb{V}}) = ||\mathbf{y} - \widehat{\mathbf{y}}(\boldsymbol{\Psi}_{y,\mathbb{V}})||^2. \tag{17}$$

The goal of source separation with PM is to find the varying PM parameters $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ which minimize E_{sep} in (17). The objective function $E_{\text{sep,PM}}$ can be minimized by using the trust-region-reflective algorithm. There are 100 starting points randomly generated to

minimize E_{sep}. The best solution, which gives the smallest E_{sep}, will be chosen as the estimated varying PM parameters $\widehat{\mathbf{\Psi}}_{y,\mathbb{V}}$.

5. EXPERIMENTS

Experiments were performed to test the modeling and separation qualities of PM. All data used in the experiments are real signals of piano tones and they are not synthetic. The piano tones were used to generate mixtures from musical chords which include octaves. The generation of mixtures will be discussed in Section 5.1. In Section 5.2, the experimental results will be presented. The input mixtures of our experiments were generated by mixing isolated tones from the recorded piano databases. So the ground truth of these testing mixtures is known. Then our separation method was applied to these mixtures to separate them into the individual tones. The estimated tones were compared with the input isolated tones for evaluation.

5.1. Piano tone database and generation of mixtures

Piano tones from four different pianos were used in our experiments. Three of the pianos are from the RWC musical instrument sound database [14] 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 [11]. All tones, including our database and the RWC database, were downsampled to 11.025 kHz for faster processing.

In the experiments, there are 25 mixtures randomly selected from 11 piano pieces in the RWC music database including the databases of classical music, jazz music and music genre [14]. The lists of all the piano pieces and mixtures are shown in Appendix. The RWC database provides the MIDI files of the transcribed performance of these pieces. We extracted all chords from the MIDI files. A chord is a set of simultaneous pitches. These chords provide the pitch information for the mixtures. In order to measure the performance of our proposed system in real music, we randomly selected the 25 mixtures from the extracted chords according to the distribution of the number of pitches in these chords. The number of tones K in our selected mixtures are ranging from 1 to 6 with the counts 8, 6, 5, 4, 1 and 1. The 25 mixtures consist of 62 tones. There are 9 mixtures containing at least one pair of octaves. Two of them ($K = 5$ and $K = 6$) contain 2 pairs of octaves.

The procedures of generating a mixture are shown below. Each mixture was generated by mixing its individual tones. The pitches of the tones in a mixture correspond to the pitches of a selected chord. All tones in a mixture were randomly selected from the isolated tones in one of the four pianos described before and the individual tones in a mixture come from the same piano. The choices of loudness of a tone in a mixture are soft, medium and loud. The loudness of each tone was assigned according to the MIDI velocity in the MIDI files. When a particular loudness of the tone was selected, the remaining two instances were put in the training data. Hence, the number of instances I_k is equal to 2. Random time shifts were added to the isolated tones in the range of $-10 \leq \tau \leq 10$ ms before mixing to test whether the time shift can be estimated in PM. A mixture was formed by a summation of the selected time-shifted isolated tones. The first 0.5 second of the mixtures and the training data were used in the experiments.

5.2. Results

5.2.1. Evaluation criteria

The performance of our source separation system is evaluated by the signal-to-noise ratio (SNR) which is defined by

$$\text{SNR} = 10 \log_{10} \frac{\sum_n \mathsf{x}(t_n)^2}{\sum_n (\mathsf{x}(t_n) - \widehat{\mathsf{x}}(t_n))^2} \tag{18}$$

where $\mathsf{x}(t_n)$ is the time-shifted isolated tone in the time domain before mixing and $\widehat{\mathsf{x}}(t_n)$ is the estimated tone in the time domain. The estimated tone is reconstructed from PM. Higher SNR means higher quality of estimated signals.

The musical nuance is related to the estimated intensity $\widehat{c}_k$ and the estimated time shift $\widehat{\tau}_k$. These two parameters will also be examined. As intensity is at a relative scale, the accuracy of the estimated intensity is evaluated by the absolute error ratio

$$\text{ER}_c = \left| \frac{c_k - \widehat{c}_k}{c_k} \right| \tag{19}$$

where c_k is intensity of the input isolated tone, and $\widehat{c}_k$ is the estimated intensity of the tone. Lower absolute error ratio means higher accuracy of the estimated intensity.

The accuracy of the estimated time shift $\widehat{\tau}_k$ is evaluated by the absolute error

$$\text{Err}_\tau = |\tau_k - \widehat{\tau}_k| \tag{20}$$

where τ_k is the time shift of the input isolated tone in seconds, and $\widehat{\tau}_k$ is the estimated time shift of the tone. The input time shift τ_k has been added to the isolated tones from the piano databases as described in Section 5.1.

5.2.2. Evaluation on modeling quality

Before evaluating the separation quality, we first evaluate the modeling quality, i.e. the quality of PM to represent an isolated tone before mixing. PM was used to find the estimated signals of the time-shifted isolated tones before the tones were mixed into mixtures. The modeling quality provides a benchmark for the source separation experiments. We will compare the performance difference before and after mixing.

The procedures of evaluation on the modeling quality are shown in Figure 4 (a). For each mixture in the 25 mixtures described in Section 5.1, the individual tone of the mixture was selected from the isolated tone in the piano databases. Then a random time shift was added to each isolated tone, and the shifted tones were inputted into our source separation system. The outputs of our system were the estimated tones reconstructed from PM. The estimated tones were compared to the shifted tones to evaluate the modeling quality. If the parameters obtained in PM are accurate, they can regenerate the original shifted tones in high quality. The average SNR ($\overline{\text{SNR}}$) is 11.15 dB which is satisfactory.

5.2.3. Evaluation on separation quality

After evaluating the modeling quality, we evaluate the separation quality, i.e. the quality of PM to separate a mixture into its individual tones. Figure 4 (b) illustrates the procedures of evaluation

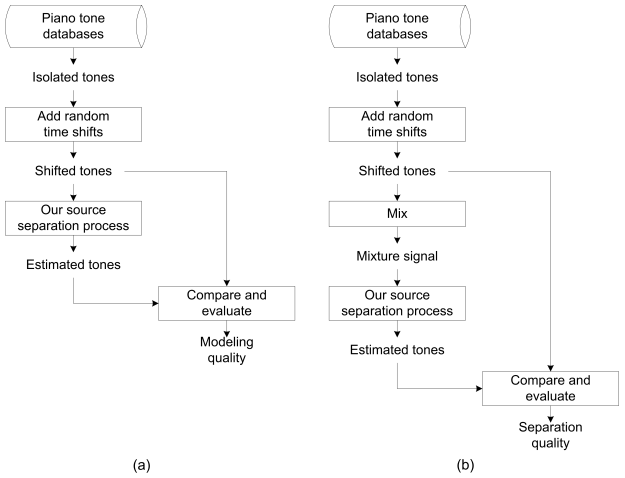

Figure 4: *The procedures of evaluation on (a) modeling quality for a tone, and (b) separation quality for a mixture.*

	$\overline{\text{SNR}}$ (dB)	$\overline{\Delta\text{SNR}}$ (dB)
All mixtures	10.88	-
$2 \leq K \leq 6$	10.97	-0.31
Upper tones in octaves	10.95	-0.37

Table 2: *The average SNR of the 25 mixtures. The number of tones in a mixture is denoted by K. The column of $\overline{SNR}$ is the average SNR in dB. The column of $\overline{\Delta SNR}$ is the average SNR difference between modeling and source separation.*

on the separation quality for one mixture. The quality is evaluated with one mixture at a time. The steps are similar to those in evaluation on the modeling quality. The difference starts from the shifted tones. In evaluating the separation quality, the shifted tones were mixed into a mixture by summing these tones. Then the mixture signal was inputted into our proposed source separation system. The outputs of our system were the estimated tones reconstructed from PM. The estimated tones were compared to the input shifted tones to evaluate the separation quality.

For the 25 mixtures, the average SNR is 10.88 dB. The results are shown in Table 2. PM is able to reconstruct the upper tone in an octave. The partials of the upper tone in an octave are completely overlapping with the lower tone. Hence, the overlapping partials were successfully resolved. The average SNR may not completely illustrate the separation quality because high average SNR may be due to high modeling quality. To evaluate the separation quality effectively, the average SNR difference is used. The SNR difference between the modeling benchmark in Section 5.2.2 and the separation is defined by

$$\Delta\text{SNR}=(\text{SNR from modeling}) - (\text{SNR from separation}) \quad (21)$$

which measures the drop of SNR from the modeling benchmark to the separation result. The average SNR difference, which is the average of ΔSNR of different cases, is shown in Table 2. The average SNR difference is small. This means that PM is robust to overlapping partials.

In addition to SNR, we also evaluate the separation result by the average absolute error ratio of intensity and the average absolute error of the estimated time shift for $2 \leq K \leq 6$. The average absolute error ratio of intensity ER_c is shown in Table 3. The error ratio is 0.074 for estimating the intensity. As the intensity $\widehat{c}_k$ is used to estimate the peak amplitude of the individual tone in a mixture, the accuracy of $\widehat{c}_k$ is compared to the peak amplitude of the estimated tones from PM. The average absolute error ratio of $\widehat{c}_k$ is lower than that of PM. This is because the peak amplitude of the estimated tones from PM depend on all estimated parameters. In the other hand, the estimation of $\widehat{c}_k$ is only based on the envelope function defined in (3) so the estimation of $\widehat{c}_k$ is less sensitive to the estimation error arisen from phases. As a result, $\widehat{c}_k$ is more robust to estimate the peak amplitude of an individual tone in a mixture.

	Average absolute error ratio of intensity ER_c	
	Intensity c_k	Peak from PM
$2 \leq K \leq 6$	0.074	0.222

Table 3: *The average absolute error ratio of intensity ER_c*

	$\overline{\text{SNR}}$ (dB)	
	PM	Li
All mixtures	10.88	6.63
$K = 2$	11.76	12.07
$2 \leq K \leq 6$	10.97	5.40
Upper tones in octaves	10.95	1.57

Table 4: *Comparison of Li's system and our proposed system PM.*

The average absolute error of the estimated time shift Err_τ for $2 \leq K \leq 6$ is only 3.16 ms so the estimated time shift can give an accurate fine-tuned onset.

5.2.4. Comparison with other system

In a recent system of monaural source separation in [4], Li, Woodruff and Wang built their system (Li's system) based on the principle of common amplitude modulation reviewed in Section 1. We compared Li's system to our proposed source system for all mixtures. The implementation of Li's system is provided by the authors. The true fundamental frequency of each tone was supplied to Li's system. The result is shown in Table 4. Our system performs better than Li's system for the average SNR. A significant improvement is in the octave cases as shown in the table. Li's system is unable to resolve the overlapping partials of the upper tones in octaves. Our system can resolve those overlapping partials. The average SNR against the number of tones K is plotted in Figure 5. Although Li's system performs well for the number of notes equal to 1 or 2, the average SNR of Li's system decreases much more rapidly than our system. Our system can make use of the training data to give higher separation quality.

Some audio files in the experiments are selected for the demonstration purpose. They are available at
`http://www.cse.cuhk.edu.hk/~khwong/www2/conference/dafx13/c2013_dafx13_demo.zip`.

6. CONCLUSIONS

In this paper, we have proposed a monaural source separation system to extract individual tones from mixture signals of piano tones.

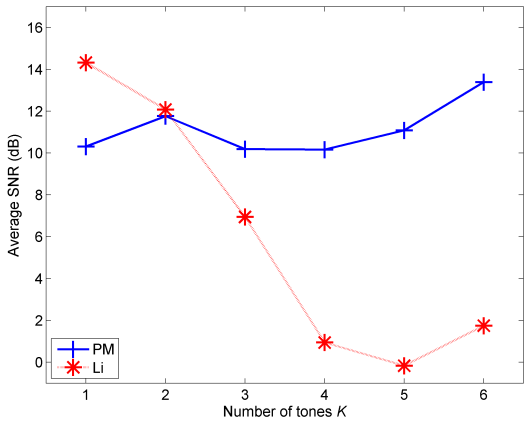

Figure 5: *Average SNR against the number of tones K for our system and Li's system.*

We designed a Piano Model (PM) based on a sum of sinusoidal components to represent piano tones. Based on this PM model, the system is able to resolve overlapping partials in the source separation process. The recovered parameters (frequencies, amplitudes, phases, intensities and fine-tuned onsets) of partials are essential for thorough signal analysis and characterizations of musical nuances. The experiments show that our proposed PM method gives robust and accurate results in separation of signal mixtures even when octaves are included. The separation quality is significantly better than those reported in the previous work. However, when measuring modeling quality used for sound reproduction of isolated tones, our approach is still inferior to other methods such as the framewise model in [4]. Our future direction is to combine these two methods: our PM and our framewise model in [11] by using a hierarchical Bayesian framework to achieve better performances both in source separation and in sound reproduction.

7. ACKNOWLEDGMENTS

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. We are thankful to the reviewers for their valuable comments.

8. REFERENCES

[1] T. Virtanen, *Sound Source Separation in Monaural Music Signals*, Ph.D. thesis, Tampere University of Technology, Finland, November 2006.

[2] 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.

[3] 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.

[4] 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.

[5] Jinyu 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.

[6] A. C. Lehmann, J. A. Sloboda, and R. H. Woody, *Psychology for Musicians: Understanding and Acquiring the Skills*, chapter Expression and interpretation, pp. 85–106, Oxford University Press, 2007.

[7] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, Springer Verlag, 2nd edition, 1998.

[8] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*, Springer, 2006.

[9] 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/.

[10] R. Palmieri, Ed., *Piano: an encyclopedia*, Routledge, London, 2nd edition, 2003.

[11] W. M. Szeto and K. H. Wong, "Sinusoidal modeling for piano tones," in *IEEE International Conference on Signal Processing, Communications and Computing*, Kunming, China, August 2013, Available online at `http://www.cse.cuhk.edu.hk/~khwong/www2/conference/icspcc2013/c2013_icspcc2013_szeto_piano_tones.pdf`.

[12] C. Palmer and J. C. Brown, "Investigations in the amplitude of sounded piano tones," *Journal of the Acoustical Society of America*, vol. 90, no. 1, pp. 60–66, July 1991.

[13] 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.

[14] 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.

Appendix

The appendix is available at
`http://www.cse.cuhk.edu.hk/~khwong/www2/conference/dafx13/c2013_dafx13_appendix.pdf`