# A HIERARCHICAL BAYESIAN FRAMEWORK FOR SCORE-INFORMED SOURCE SEPARATION OF PIANO MUSIC SIGNALS

**Wai Man SZETO**
Office of University General Education
The Chinese University of Hong Kong
wmszeto@cuhk.edu.hk

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

## ABSTRACT

Here we propose a score-informed monaural source separation system to extract every tone from a mixture of piano tone signals. Two sinusoidal models in our earlier work are employed in the above-mentioned system to represent piano tones: the General Model and the Piano Model. The General Model, a variant of sinusoidal modeling, can represent a single tone with high modeling quality, yet it fails to separate mixtures of tones due to the overlapping partials. The Piano Model, on the other hand, is an instrument-specific model tailored for piano. Its modeling quality is lower but it can learn from training data (consisting entirely of isolated tones), resolve the overlapping partials and thus separate the mixtures. We formulate a new hierarchical Bayesian framework to run both Models in the source separation process so that the mixtures with overlapping partials can be separated with high quality. The results show that our proposed system gives robust and accurate separation of piano tone signal mixtures (including octaves) while achieving significantly better quality than those reported in related work done previously.

## 1. INTRODUCTION

Here we propose a score-informed monaural source separation system under a new hierarchical Bayesian framework to extract every tone from a mixture of piano tone signals with high separation quality. Two sinusoidal models in our earlier work in [14, 15] are employed in the above mentioned system to represent piano tones. Sinusoidal modeling is commonly used in many existing monaural source separation systems to model pitched musical sounds [6, 7, 9, 11, 16]. The major difficulty of source separation (SS) is to resolve overlapping partials.

Existing systems are based on assumptions on the general properties of pitched musical sounds. For example, the spectral envelope of tones is assumed to be smooth (as in [7, 16]), or that the amplitude envelope of each partial from the same note tends to be similar [11] (known as common amplitude modulation (CAM)), or that the amplitude envelope of a partial evolves similarly among different notes of the same musical instrument in [9]. Yet these assumptions may not be suitable for SS of piano mixtures as explained in [15]. A very recent work in [17] can resolve two closed partials but it may not work on octaves, in which the partials of the upper tone are totally immersed within the frequencies of the lower tone. Moreover, it assumes that partials are exact multiples of the fundamental frequency. This assumption is not valid for piano because piano tones are only quasi-harmonic [1].

Instead of formulating similar assumptions, we limit input mixtures to piano music signals. This allows us to design a piano-specific model called the Piano Model (PM) to resolve overlapping partials in [15]. In piano music, a particular pitch tends to appear more than once. The tones of the same pitch share some common characteristics which can be captured by PM. Our system is based on two requirements. First, the pitches in the mixtures should reappear as isolated tones in the target recording. Second, the piano music is performed without pedaling. Then the isolated tones can be used as the training data for PM to resolve the overlapping partials even for octaves.

Although PM can resolve the overlapping partials, its modeling quality of single piano tones is lower than our General Model (GM) in [14]. However, GM cannot be directly applied to SS because it fails to separate mixtures of tones due to the overlapping partials. Here we formulate a new hierarchical Bayesian framework to run both PM and GM in the SS process so that the mixtures with overlapping partials can be separated with high quality. The separation process is divided into the training stage and the SS stage. Given the estimated PM parameters and the training data, we can, in the SS stage, set the prior distributions of the GM parameters to favor the proper regions of values under the Bayesian framework, estimate the GM parameters successfully even the case of overlapping partials, and reconstruct the individual tones in the mixtures with high quality. We hope that our system could shed some light on the empirical study of expressiveness in music performance [5] by comparing the subtleties of various artists' performances, based on individual tones extracted by SS.

## 2. SIGNAL MODELS

Here 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 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 $k$th individual tone in the mixture, and $t$ is the time in seconds. We assume that the score has been known so that the pitch and the duration of each $\mathsf{x}_k(t)$ are given (music transcription systems [2, 10] can be used here). The goal of our research is to recover the signal of each individual tone $\mathsf{x}_k(t)$ from the mixture signal $\mathsf{y}(t)$ via the signal models GM and PM.

### 2.1 General Model (GM)

In [14], we present a frame-wise sinusoidal model called GM to represent a piano tone. For a piano tone, the frequencies of the partials are stable so the frequencies can be fixed across frames. The number of partials can also be fixed for a tone. In GM, the estimated tone $\widehat{\mathsf{x}}_{k,r}$, which is the estimate of the $k$th tone in a mixture at the $r$th frame, can be written as:

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

where $M_k$ is the number of partials, $\alpha_{k,m,r}$ is the amplitude of the cosine component, $\beta_{k,m,r}$ is the amplitude of the sine component, $f_{k,m}$ is the frequency, $w[l]$ is the window function with the window length $L$ and $l=0,...,L-1$, and $t_l$ is the time in second at the index $l$ so $t_l = l/f_s$ and $f_s$ is the sampling frequency in Hz. The overlap-and-add method in [18] can be used to reconstruct the entire signal from GM.

Based on the above model, the estimated mixture $\widehat{y}_r[l]$ at the $r$th frame is the sum of each estimated tone $\widehat{x}_{k,r}[l]$ such that $\widehat{y}_r[l] = \sum_{k=1}^{K} \widehat{x}_{k,r}[l]$. The observed mixture is the sum of the estimated mixture and the noise term so $y_r[l] = \widehat{y}_r[l] + v_r[l]$ where $v_r[l]$ is the noise component. To estimate the parameters in each frame, it is convenient to rewrite the model in (1) into the matrix form. Let $\mathbf{H}_k$ be the frequency matrix of the $k$th tone in the form of

$$H_k[l,u] = \begin{cases} w[l]\cos(2\pi f_{k,u} t_l) & \text{if } 1 \le u \le M_k, \\ w[l]\sin(2\pi f_{k,u-M_k} t_l) & \text{if } M_k+1 \le u \le 2M_k \end{cases} \quad (2)$$

and we also let $\mathbf{f}_k$ be the frequency vector containing all $f_{k,u}$.

The amplitudes of the cosine and sine terms of the $k$th tone at the $r$th frame can be expressed as a column vector $\mathbf{g}_{k,r}$ defined by

$$g_{k,r}[u] = \begin{cases} \alpha_{k,u,r} & \text{if } 1 \le u \le M_k, \\ \beta_{k,u-M_k,r} & \text{if } M_k+1 \le u \le 2M_k \end{cases}. \quad (3)$$

For the mixture, the frequency matrices from each tone are concatenated into the matrix $\mathbf{H} = [\mathbf{H}_1 \cdots \mathbf{H}_K]$ and all $\mathbf{f}_k$ are concatenated into the column vector $\mathbf{f} = \left[\mathbf{f}_1^{\mathrm{T}} \cdots \mathbf{f}_K^{\mathrm{T}}\right]^{\mathrm{T}}$. The amplitude vectors of each tone can also be concatenated into a column vector $\mathbf{g}_r = \left[\mathbf{g}_{1,r}^{\mathrm{T}} \cdots \mathbf{g}_{K,r}^{\mathrm{T}}\right]^{\mathrm{T}}$. The estimated mixture at $r$th frame can be expressed as $\widehat{\mathbf{y}}_r = \mathbf{H}\mathbf{g}_r$ and the estimated mixture is related to the observed mixture as below:

$$\mathbf{y}_r = \mathbf{H}\mathbf{g}_r + \mathbf{v}_r \quad (4)$$

where $\mathbf{v}_r$ is the noise term. It is modeled as the zero-mean Gaussian noise with the variance $\sigma_{v_r}^2$.

The observed mixture signal can be expressed in the form of $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_R]$. Then the estimated mixture for all frames can be written as

$$\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{G} \quad (5)$$

where $\widehat{\mathbf{Y}} = [\widehat{\mathbf{y}}_1 \cdots \widehat{\mathbf{y}}_R]$, $\mathbf{G} = [\mathbf{g}_1 \cdots \mathbf{g}_R]$ and $R$ is the number of frames. All GM parameters can be grouped into $\mathbf{\Theta} = \{\mathbf{f}, \mathbf{G}\}$. The goal of our SS is to estimate both the frequency matrix $\mathbf{H}$ and the amplitude matrix $\mathbf{G}$ so that each individual tone can be reconstructed. However, $\mathbf{H}$ is often rank deficient. This happens when some of the partials from different tones in the mixture are overlapping. This implies that if only the mixture in such case is given, it is impossible to separate the mixture into its individual tones unless more information is provided. This problem can be solved by using the training data as the prior information under the Bayesian framework in Section 3.

### 2.2 Piano Model (PM)

In [15], we propose PM to resolve the overlapping partials by exploring the common properties of recurring tones. PM employs a time-varying sum-of-sinusoidal signal model for piano tones, and it describes a tone in an entire duration instead of a single analysis frame as

$$\widehat{\mathsf{x}}_k(t_n) = \sum_{m=1}^{M_k} a(t_n; c_k, \boldsymbol{\varphi}_{k,m}) \cdot \cos(2\pi f_{k,m} t_n + \phi_{k,m}) \quad (6)$$

where $M_k$ is the number of partials of the $k$th tone, $f_{k,m}$ and $\phi_{k,m}$ are the frequency and the phase respectively, and $a(t_n; c_k, \boldsymbol{\varphi}_{k,m})$ is the time-varying amplitude of the partial stated in [15] where the envelope parameters $\boldsymbol{\varphi}_{k,m}$ control the envelope surface against the intensity $c_k$ and the time $t_n$. The intensity $c_k$ is assigned to be the peak amplitude of the observed time-domain signal of the tone. 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 $\widehat{\mathsf{y}}(t_n) = \sum_{k=1}^{M_k} \widehat{\mathsf{x}}_k(t_n - \tau_k)$ where $\tau_k$ is the time shift in seconds.

All parameters in PM for the $k$th tone can be grouped into a parameter set $\boldsymbol{\psi}_k$ so $\boldsymbol{\psi}_k = \{\boldsymbol{\varphi}_{k,m}, f_{k,m}, \phi_{k,m}, c_k, \tau_k\}$ and $\mathbf{\Psi} = \{\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_K\}$. The PM parameters $\boldsymbol{\psi}_k$ can be divided into two groups: 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 and they are estimated from the training data. The varying PM parameters consist of parameters which may vary across instances. Given a mixture, only the varying PM parameters of the mixture are required to be estimated if the invariant PM parameters have been estimated from the training data.

In both GM and PM, we have assumed that the number of partials $M_k$ of each tone is known. The number of partials $M_k$ is fixed for all experiments. The details of finding $M_k$ can be found in [14].

## 3. BAYESIAN FRAMEWORK FOR SS

This section will explain how the Bayesian framework integrates the two models in the previous section and incorporates the training data to resolve overlapping partials. Given the mixture $\mathbf{y}$ and the training data $\boldsymbol{\mathcal{X}}$, the goal of Bayesian SS with GM is to find the *Maximum A Posterior* (MAP) solution $\widehat{\mathbf{\Theta}}_y$ that maximizes the posterior $p(\mathbf{\Theta}_y | \mathbf{y}, \boldsymbol{\mathcal{X}})$ where $\mathbf{\Theta}_y$ is the

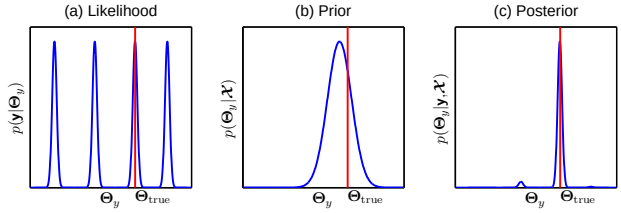


**Figure 1**. (a) The likelihood function. (b) The prior. (c) The posterior. This schematic diagram shows that an appropriate prior gives the desirable MAP solution. The vertical line shows the true value of $\boldsymbol{\Theta}_y$.

GM parameter set for $\mathbf{y}$. By Bayes' theorem, the posterior can be written in the form $p(\boldsymbol{\Theta}_y|\mathbf{y},\boldsymbol{\mathcal{X}}) \propto p(\mathbf{y}|\boldsymbol{\Theta}_y)p(\boldsymbol{\Theta}_y|\boldsymbol{\mathcal{X}})$.

The key issue of Bayesian SS is how to set up the prior $p(\boldsymbol{\Theta}_y|\boldsymbol{\mathcal{X}})$. If overlapping partials are present, the frequency matrix $\mathbf{H}$ is rank deficient and many choices of $\boldsymbol{\Theta}$ can give similar values of the likelihood $p(\mathbf{y}|\boldsymbol{\Theta}_y)$. Hence, there are many peaks in the likelihood function as shown in the schematic diagram (Figure 1(a)). In order to find the desirable MAP solution, it is advantageous that the prior distribution has a high density around the correct value of $\boldsymbol{\Theta}_y$. In Figure 1(b), the prior is appropriate so that the MAP solution, i.e. the peak of the posterior, can be located correctly as depicted in Figure 1(c). In short, an appropriate prior of the GM parameters is crucial for resolving the overlapping partials. It can be found by using the training data and the estimated PM parameters.

The prior $p(\boldsymbol{\Theta}_y|\boldsymbol{\mathcal{X}})$ expresses the probability distribution of $\boldsymbol{\Theta}_y$ given the training data $\boldsymbol{\mathcal{X}}$ and before the mixture $\mathbf{y}$ is observed. The functional form of $p(\boldsymbol{\Theta}_y|\boldsymbol{\mathcal{X}})$ can be formulated in terms of PM. The PM parameter set $\boldsymbol{\Psi}_y$ of the mixture $\mathbf{y}$ is divided into two sets: the invariant PM parameter set $\boldsymbol{\Psi}_{y,\mathbb{I}}$ and the varying PM parameter set $\boldsymbol{\Psi}_{y,\mathbb{V}}$. For the training data $\boldsymbol{\mathcal{X}}$, the PM parameter set $\boldsymbol{\Psi}_{\mathcal{X}}$ is divided into the invariant PM parameter set $\boldsymbol{\Psi}_{\mathcal{X},\mathbb{I}}$ and the varying PM parameter set $\boldsymbol{\Psi}_{\mathcal{X},\mathbb{V}}$. Note that both the mixture and the training data share the same set of the invariant PM parameters. The subscripts $y$ and $\mathcal{X}$ for the invariant PM parameters can be omitted for clarity so $\boldsymbol{\Psi}_{\mathbb{I}} = \boldsymbol{\Psi}_{y,\mathbb{I}} = \boldsymbol{\Psi}_{\mathcal{X},\mathbb{I}}$.

The posterior $p(\boldsymbol{\Theta}_y|\mathbf{y},\boldsymbol{\mathcal{X}})$ of the GM parameters can be linked up with the PM parameters by using marginalization:

$$p(\boldsymbol{\Theta}_y|\mathbf{y},\boldsymbol{\mathcal{X}}) = \iint p(\boldsymbol{\Theta}_y,\boldsymbol{\Psi}_{y,\mathbb{V}},\boldsymbol{\Psi}_{\mathbb{I}}|\mathbf{y},\boldsymbol{\mathcal{X}})d\boldsymbol{\Psi}_{y,\mathbb{V}}d\boldsymbol{\Psi}_{\mathbb{I}}. \quad (7)$$

Note that the noise variance $\sigma_{v_r}^2$ of the mixture in (4) is omitted in the derivation for clarity. Then by the product rule of probability, (7) can be put into

$$p(\boldsymbol{\Theta}_y|\mathbf{y},\boldsymbol{\mathcal{X}}) = \iint p(\boldsymbol{\Theta}_y|\mathbf{y},\boldsymbol{\mathcal{X}},\boldsymbol{\Psi}_{y,\mathbb{V}},\boldsymbol{\Psi}_{\mathbb{I}}) p(\boldsymbol{\Psi}_{y,\mathbb{V}},\boldsymbol{\Psi}_{\mathbb{I}}|\mathbf{y},\boldsymbol{\mathcal{X}})d\boldsymbol{\Psi}_{y,\mathbb{V}}d\boldsymbol{\Psi}_{\mathbb{I}} \quad (8)$$

where the first term is the posterior of $\boldsymbol{\Theta}_y$ and the second is the posterior of $\boldsymbol{\Psi}_{y,\mathbb{V}}$ and $\boldsymbol{\Psi}_{\mathbb{I}}$.

However, finding the MAP solution involves evaluating the integration over all possible values of $\boldsymbol{\Psi}_{y,\mathbb{V}}$ and $\boldsymbol{\Psi}_{\mathbb{I}}$ in (8). PM is a highly dimensional and nonlinear model that makes the integration analytically infeasible. Different approximation techniques can be used to find the MAP solution.

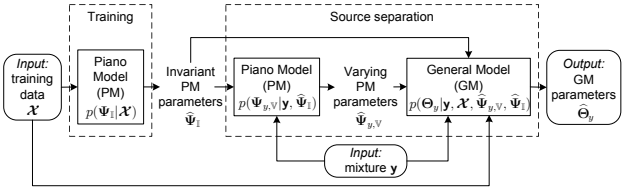


**Figure 2**. Bayesian framework for SS.

For computational efficiency, here we have used the evidence approximation [12, 13]. Following the derivation of the evidence approximation in [3, p. 408], we assume that the posterior $p(\boldsymbol{\Psi}_{y,\mathbb{V}},\boldsymbol{\Psi}_{\mathbb{I}}|\mathbf{y},\boldsymbol{\mathcal{X}})$ is sharply peaked around their most probable values $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ and $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$. Then (8) can be written as $p(\boldsymbol{\Theta}_y|\mathbf{y},\boldsymbol{\mathcal{X}}) \approx p(\boldsymbol{\Theta}_y|\mathbf{y},\boldsymbol{\mathcal{X}},\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}},\widehat{\boldsymbol{\Psi}}_{\mathbb{I}})$.

Hence, the MAP solution $\widehat{\boldsymbol{\Theta}}_y$ is the maximum of the posterior $p(\boldsymbol{\Theta}_y|\mathbf{y},\boldsymbol{\mathcal{X}},\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}},\widehat{\boldsymbol{\Psi}}_{\mathbb{I}})$. The estimation of $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ and $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ can be done as follows: (i) $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ is estimated by maximizing the posterior $p(\boldsymbol{\Psi}_{y,\mathbb{V}}|\mathbf{y},\boldsymbol{\mathcal{X}})$ via the evidence approximation which gives $p(\boldsymbol{\Psi}_{y,\mathbb{V}}|\mathbf{y},\boldsymbol{\mathcal{X}}) \approx p(\boldsymbol{\Psi}_{y,\mathbb{V}}|\mathbf{y},\widehat{\boldsymbol{\Psi}}_{\mathbb{I}})$ (note that $\boldsymbol{\mathcal{X}}$ is omitted because $\boldsymbol{\Psi}_{y,\mathbb{V}}$ is independent of $\boldsymbol{\mathcal{X}}$ if $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ is given); (ii) $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ is estimated by maximizing the posterior $p(\boldsymbol{\Psi}_{\mathbb{I}}|\mathbf{y},\boldsymbol{\mathcal{X}})$ that can be approximated by using the training data only so $p(\boldsymbol{\Psi}_{\mathbb{I}}|\mathbf{y},\boldsymbol{\mathcal{X}}) \approx p(\boldsymbol{\Psi}_{\mathbb{I}}|\boldsymbol{\mathcal{X}})$.

According to these results, the whole SS process is summarized in Figure 2. The whole process is divided into the following two stages:

1. *Training*. Given the training data $\boldsymbol{\mathcal{X}}$, find the most probable value of the invariant PM parameters $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ of $p(\boldsymbol{\Psi}_{\mathbb{I}}|\boldsymbol{\mathcal{X}})$.

2. *SS*. Given the mixture $\mathbf{y}$, the training data $\boldsymbol{\mathcal{X}}$ and the invariant PM parameters $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$, SS functions in two steps:

   (a) *SS with PM*. Given $\mathbf{y}$ and $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$, find the most probable value of the varying PM parameters $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ of $p(\boldsymbol{\Psi}_{y,\mathbb{V}}|\mathbf{y},\widehat{\boldsymbol{\Psi}}_{\mathbb{I}})$.

   (b) *SS with GM*. Given $\mathbf{y}$, $\boldsymbol{\mathcal{X}}$, $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ and $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$, find the MAP solution $\widehat{\boldsymbol{\Theta}}_y$ of $p(\boldsymbol{\Theta}_y|\mathbf{y},\boldsymbol{\mathcal{X}},\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}},\widehat{\boldsymbol{\Psi}}_{\mathbb{I}})$.

## 4. TRAINING AND SS WITH PM

The goal of the training stage is to find the most probable invariant PM parameters $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ that maximize the posterior of the invariant PM parameters $p(\boldsymbol{\Psi}_{\mathbb{I}}|\boldsymbol{\mathcal{X}})$ given the training data $\boldsymbol{\mathcal{X}}$. By Bayes' theorem, the posterior can be rewritten as $p(\boldsymbol{\Psi}_{\mathbb{I}}|\boldsymbol{\mathcal{X}}) \propto p(\boldsymbol{\mathcal{X}}|\boldsymbol{\Psi}_{\mathbb{I}})p(\boldsymbol{\Psi}_{\mathbb{I}})$. The prior $p(\boldsymbol{\Psi}_{\mathbb{I}})$ reflects our prior knowledge of the invariant PM parameters $\boldsymbol{\Psi}_{\mathbb{I}}$. The values of $\boldsymbol{\Psi}_{\mathbb{I}}$ greatly vary from different pitches and pianos. If we have little idea on suitable values for a parameter, it is safe to assign a prior which is insensitive to the values of that parameter [4]. Then maximizing the posterior $p(\boldsymbol{\Psi}_{\mathbb{I}}|\boldsymbol{\mathcal{X}})$ is effectively equivalent to maximize the likelihood $p(\boldsymbol{\mathcal{X}}|\boldsymbol{\Psi}_{\mathbb{I}})$. The details of finding the solution $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ can be found in [15].

Given the invariant PM parameters $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ and the mixture $\mathbf{y}$, we perform SS with PM as shown in Figure 2. The goal of SS

with PM is to find the most probable varying PM parameters $\widehat{\mathbf{\Psi}}_{y,\mathbb{V}}$ that maximize the posterior of the varying PM parameters $p(\mathbf{\Psi}_{y,\mathbb{V}}|\mathbf{y},\widehat{\mathbf{\Psi}}_{\mathbb{I}})$. By Bayes' theorem, the posterior can be rewritten as $p(\mathbf{\Psi}_{y,\mathbb{V}}|\mathbf{y},\widehat{\mathbf{\Psi}}_{\mathbb{I}}) \propto p(\mathbf{y}|\mathbf{\Psi}_{y,\mathbb{V}},\widehat{\mathbf{\Psi}}_{\mathbb{I}})p(\mathbf{\Psi}_{y,\mathbb{V}})$. The prior $p(\mathbf{\Psi}_{y,\mathbb{V}})$ reflects our prior knowledge of the invariant PM parameters $\mathbf{\Psi}_{y,\mathbb{V}}$. The values of $\mathbf{\Psi}_{y,\mathbb{V}}$ greatly vary from different playings. Hence, we choose an insensitive prior for $\mathbf{\Psi}_{y,\mathbb{V}}$ as $\mathbf{\Psi}_{\mathbb{I}}$. Then maximizing the posterior $p(\mathbf{\Psi}_{y,\mathbb{V}}|\mathbf{y},\widehat{\mathbf{\Psi}}_{\mathbb{I}})$ is again effectively equivalent to maximize the likelihood $p(\mathbf{y}|\mathbf{\Psi}_{y,\mathbb{V}},\widehat{\mathbf{\Psi}}_{\mathbb{I}})$. The details of finding $\widehat{\mathbf{\Psi}}_{y,\mathbb{V}}$ are also presented in [15].

## 5. SS WITH GM

The process of SS with GM is divided into the following two steps: (1) estimate the hyperparameters, and (2) given the hyperparameters, find the MAP solution $\widehat{\mathbf{\Theta}}_y$. We will focus on the second step first.

### 5.1 Find the MAP solution

The MAP solution $\widehat{\mathbf{\Theta}}_y$ is found by maximizing the posterior $p(\mathbf{\Theta}_y|\mathbf{y},\boldsymbol{\mathcal{X}},\widehat{\mathbf{\Psi}}_{y,\mathbb{V}},\widehat{\mathbf{\Psi}}_{\mathbb{I}})$. The GM parameters $\mathbf{\Theta}_y$ include the amplitude matrix $\mathbf{G}$ and the frequencies $\mathbf{f}$. An iterative update scheme is designed to find the MAP solution: (1) given $\mathbf{f}$, update $\mathbf{G}$, and (2) given $\mathbf{G}$, update $\mathbf{f}$. Steps 1 to 2 are repeated until convergence. The iterative update starts with the input frequencies from the estimated frequencies in PM in Section 4. The frequencies in PM are close to those in GM. We find that 10 iterations are enough for convergence. In the followings, the iterative update scheme will be discussed in details.

#### *5.1.1 Step 1: update the amplitude matrix* $\mathbf{G}$

Each $\mathbf{g}_r$ in the amplitude matrix $\mathbf{G}$ can be estimated independently. Given the estimated frequencies $\widehat{\mathbf{f}}$, now we rewrite the posterior of $\mathbf{g}_r$ into $p(\mathbf{g}_r|\mathbf{y}_r,\boldsymbol{\mathcal{X}},\widehat{\mathbf{f}},\widehat{\mathbf{\Psi}}_{y,\mathbb{V}},\widehat{\mathbf{\Psi}}_{\mathbb{I}},\widehat{\sigma}^2_{v_r})$. The goal of this step is to find the MAP solution $\widehat{\mathbf{g}}_r$ which maximizes the posterior of $\mathbf{g}_r$. By Bayes' theorem, the posterior of $\mathbf{g}_r$ can be expressed in the form of

$$p(\mathbf{g}_r|\mathbf{y}_r,\boldsymbol{\mathcal{X}},\widehat{\mathbf{f}},\widehat{\mathbf{\Psi}}_{y,\mathbb{V}},\widehat{\mathbf{\Psi}}_{\mathbb{I}},\widehat{\sigma}^2_{v_r}) \propto p(\mathbf{y}_r|\mathbf{g}_r,\widehat{\mathbf{f}},\widehat{\sigma}^2_{v_r})p(\mathbf{g}_r|\boldsymbol{\mathcal{X}},\widehat{\mathbf{\Psi}}_{y,\mathbb{V}},\widehat{\mathbf{\Psi}}_{\mathbb{I}}) \quad (9)$$

where $\widehat{\sigma}^2_{v_r}$ represents the estimated variance of the zero-mean Gaussian noise in (4).

The prior $p(\mathbf{g}_r|\boldsymbol{\mathcal{X}},\widehat{\mathbf{\Psi}}_{y,\mathbb{V}},\widehat{\mathbf{\Psi}}_{\mathbb{I}})$ in (9) represents the prior distribution of $\mathbf{g}_r$ conditioned on the training data $\boldsymbol{\mathcal{X}}$ and the PM parameters $\widehat{\mathbf{\Psi}}_{y,\mathbb{V}}$ and $\widehat{\mathbf{\Psi}}_{\mathbb{I}}$. It is modeled as a Gaussian with the mean $\widehat{\boldsymbol{\mu}}_{g_r}$ and the covariance matrix $\widehat{\mathbf{\Sigma}}_{g_r}$. In this section, it is assumed that the hyperparameters $\widehat{\sigma}^2_{v_r}$, $\widehat{\boldsymbol{\mu}}_{g_r}$ and $\widehat{\mathbf{\Sigma}}_{g_r}$ have been estimated and their values are known. The estimation of these hyperparameters from $\boldsymbol{\mathcal{X}}$, $\widehat{\mathbf{\Psi}}_{y,\mathbb{V}}$ and $\widehat{\mathbf{\Psi}}_{\mathbb{I}}$ will be discussed in Section 5.2. Note that each $\mathbf{g}_r$ has its own set of $\widehat{\boldsymbol{\mu}}_{g_r}$ and $\widehat{\mathbf{\Sigma}}_{g_r}$ so the MAP solution of each $\mathbf{g}_r$ can be found independently.

As $\widehat{\mathbf{y}}_r = \mathbf{H}\mathbf{g}_r$ is a linear model for given $\mathbf{H}$, and both the noise and the prior are Gaussian, the resulting posterior of $\mathbf{g}_r$ is also Gaussian. Therefore, the MAP solution $\widehat{\mathbf{g}}_r$ is equal to the posterior mean. By using the result in [4, p. 153], the MAP solution is

$$\widehat{\mathbf{g}}_r = \left(\widehat{\mathbf{\Sigma}}^{-1}_{g_r} + \widehat{\sigma}^{-2}_{v_r}\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\left(\widehat{\mathbf{\Sigma}}^{-1}_{g_r}\widehat{\boldsymbol{\mu}}_{g_r} + \widehat{\sigma}^{-2}_{v_r}\mathbf{H}^{\mathrm{T}}\mathbf{y}_r\right). \quad (10)$$

#### *5.1.2 Step 2: update the frequencies* $\mathbf{f}$

Given the estimated amplitude matrix $\widehat{\mathbf{G}}$ in Step 1, the goal of Step 2 is to find the MAP solution $\widehat{\mathbf{f}}$ which maximizes the posterior $p(\mathbf{f}|\mathbf{Y},\boldsymbol{\mathcal{X}},\widehat{\mathbf{G}},\widehat{\mathbf{\Psi}}_{y,\mathbb{V}},\widehat{\mathbf{\Psi}}_{\mathbb{I}},\widehat{\boldsymbol{\sigma}}^2_v)$. However, the model $\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{G}$ in (5) is nonlinear with $\mathbf{f}$. Based on our work in [14], we vectorize the matrix $\widehat{\mathbf{Y}}$ into $\widehat{\mathbf{Y}}_{\mathrm{vec}}$ and then linearize $\widehat{\mathbf{Y}}_{\mathrm{vec}}$ by using Taylor's expansion so

$$\widehat{\mathbf{Y}}_{\mathrm{vec}}(\mathbf{f}) \approx \widehat{\mathbf{Y}}_{\mathrm{vec}}(\mathbf{f}^{\mathrm{cur}}) + \mathbf{Z}(\mathbf{f}^{\mathrm{cur}})(\mathbf{f}-\mathbf{f}^{\mathrm{cur}}) \quad (11)$$

where $\widehat{\mathbf{Y}}_{\mathrm{vec}}(\mathbf{f})$ is the estimate depending on the new frequency vector $\mathbf{f}$ which is to be updated, and $\widehat{\mathbf{Y}}_{\mathrm{vec}}(\mathbf{f}^{\mathrm{cur}})$ is the estimate depending on the current estimate of $\mathbf{f}^{\mathrm{cur}}$. The matrix $\mathbf{Z} = \mathbf{Z}(\mathbf{f}^{\mathrm{cur}})$ is the Jacobian matrix $\partial\widehat{\mathbf{Y}}_{\mathrm{vec}}/\partial\mathbf{f}$ evaluated at $\mathbf{f}^{\mathrm{cur}}$ and $\mathbf{Z} = \left[\mathbf{Z}_1^{\mathrm{T}} \cdots \mathbf{Z}_r^{\mathrm{T}} \cdots \mathbf{Z}_R^{\mathrm{T}}\right]^{\mathrm{T}}$. The matrix $\mathbf{Z}_r$ is the Jacobian matrix $\partial\widehat{\mathbf{y}}_r/\partial\mathbf{f}$ at $r$th frame for all tones and $\mathbf{Z}_r = [\mathbf{Z}_{1,r} \cdots \mathbf{Z}_{k,r} \cdots \mathbf{Z}_{K,r}]$. Then the Jacobian matrix $\mathbf{Z}_{k,r}$ at $r$th frame for $k$th tone is

$$Z_{k,r}[l,m] = \frac{\partial\hat{y}_r[l]}{\partial f_{k,m}} = 2\pi t_l w[l] \left(-\alpha_{k,m,r}\sin(2\pi f_{k,m}t_l) + \beta_{k,m,r}\cos(2\pi f_{k,m}t_l)\right). \quad (12)$$

Hence, each element in $\mathbf{Z}$ can be computed from (12).

Following the prior distribution of $\mathbf{g}_r$, the prior distribution of $\mathbf{f}$ is also modeled as a Gaussian with the mean $\widehat{\boldsymbol{\mu}}_f$ and the covariance matrix $\widehat{\mathbf{\Sigma}}_f$. By applying (11) to the result in [4, p. 93], the MAP solution $\widehat{\mathbf{f}}$ is

$$\widehat{\mathbf{f}} = \left(\widehat{\mathbf{\Sigma}}_f^{-1} + \mathbf{Z}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_v^{-1}\mathbf{Z}\right)^{-1} \left(\widehat{\mathbf{\Sigma}}_f^{-1}\widehat{\boldsymbol{\mu}}_f + \mathbf{Z}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_v^{-1}\left(\mathbf{Y}_{\mathrm{vec}} - \widehat{\mathbf{Y}}_{\mathrm{vec}} + \mathbf{Z}\mathbf{f}^{\mathrm{cur}}\right)\right) \quad (13)$$

where $\mathbf{Z} = \mathbf{Z}(\mathbf{f}^{\mathrm{cur}})$, $\widehat{\mathbf{Y}}_{\mathrm{vec}} = \widehat{\mathbf{Y}}_{\mathrm{vec}}(\mathbf{f}^{\mathrm{cur}})$, and the covariance matrix $\widehat{\mathbf{\Sigma}}_v = \mathrm{diag}(\widehat{\sigma}^2_{v_1}\mathbf{1}_L, \ldots, \widehat{\sigma}^2_{v_R}\mathbf{1}_L)$ and $\mathbf{1}_L$ denotes the $L$-dimensional column vector filled with 1's. In the next section, we will show how to find the hyperparameters which are crucial for resolving overlapping partials.

### 5.2 Estimation of the hyperparameters

Given the training data $\boldsymbol{\mathcal{X}}$, we first estimate the GM parameters for each isolated tone in $\boldsymbol{\mathcal{X}}$ by the method in [14]. Together with the estimated PM parameters $\widehat{\mathbf{\Psi}}_{y,\mathbb{V}}$ and $\widehat{\mathbf{\Psi}}_{\mathbb{I}}$ found in Section 4, we will estimate the hyperparameters $\widehat{\sigma}^2_{v_r}$, $\widehat{\boldsymbol{\mu}}_{g_r}$, $\widehat{\mathbf{\Sigma}}_{g_r}$, $\widehat{\boldsymbol{\mu}}_f$ and $\widehat{\mathbf{\Sigma}}_f$ .

#### *5.2.1 Estimation of the noise variance* $\sigma^2_{v_r}$

To estimate the noise variance $\sigma^2_{v_r}$ of $\mathbf{y}_r$ in (4), we model the noise variance of an isolated tone $\mathbf{x}_{k,r}$ at a frame is directly proportional to the signal power. Then the noise variance of $\mathbf{x}_{k,r}$ is $\sigma^2_{v_{k,r}} = \bar{\sigma}^2_{v_k}\left\|\mathbf{x}_{k,r}\right\|^2$ where $\bar{\sigma}^2_{v_k}$ is the proportionality constant for pitch $p_k$ and it can be determined by the training data $\boldsymbol{\mathcal{X}}$ which may contain multiple instances of the same pitch. Let $\mathbf{x}^i_{k,r,\mathcal{X}}$ be a frame of an isolated tone in $\boldsymbol{\mathcal{X}}$ where the index $i$ denotes the $i$th instance of the pitch $p_k$. Then $\bar{\sigma}^2_{v_k}$ can be estimated by

$$\bar{\sigma}^2_{v_k} = \frac{1}{I_k R_k L}\sum_{i=1}^{I_k}\sum_{r=1}^{R^i_k}\sum_{l=0}^{L-1}\left(\frac{x^i_{k,r,\mathcal{X}}[l] - \hat{x}^i_{k,r,\mathcal{X}}[l]}{\left\|\mathbf{x}^i_{k,r,\mathcal{X}}\right\|}\right)^2 \quad (14)$$

where $I_k$ is the number of instances of pitch $p_k$ in $\boldsymbol{\mathcal{X}}$, $R_k^i$ is the number of frames in the $i$th instance of the pitch $p_k$ and $R_k = \sum_{i=1}^{I_k} R_k^i$, and $\hat{\mathbf{x}}_{k,r,\mathcal{X}}^i$ is the estimate of $\mathbf{x}_{k,r,\mathcal{X}}^i$ and is found by using the method in [14].

The noise variance of the mixture $\mathbf{y}_r$ is

$$\sigma_{v_r}^2 = \sum_{k=1}^{K} \sigma_{v_{k,r,y}}^2 = \sum_{k=1}^{K} \bar{\sigma}_{v_k}^2 ||\mathbf{x}_{k,r,y}||^2 \tag{15}$$

where $\mathbf{x}_{k,r,y}$ is the $k$th individual tone in the mixture, and $\sigma_{v_{k,r,y}}^2$ is its noise variance. However, $\mathbf{x}_{k,r,y}$ is not known. In order to estimate $\sigma_{v_r}^2$, we approximate $||\mathbf{x}_{k,r,y}||^2$ into

$$||\mathbf{x}_{k,r,y}||^2 \approx \left( \frac{\widehat{c}_k}{\sum_{k=1}^{K} \widehat{c}_k} \right) ||\mathbf{y}_r||^2 \tag{16}$$

where the estimated intensity $\widehat{c}_k$ in PM determines the proportion of $||\mathbf{x}_{k,r,y}||^2$ in $||\mathbf{y}_r||^2$. Substituting (16) into (15), we estimate the noise variance $\widehat{\sigma}_{v_r}^2$ in the mixture $\mathbf{y}_r$ in the form of

$$\widehat{\sigma}_{v_r}^2 = \sum_{k=1}^{K} \left( \frac{\widehat{c}_k \bar{\sigma}_{v_k}^2}{\sum_{k=1}^{K} \widehat{c}_k} \right) ||\mathbf{y}_r||^2. \tag{17}$$

#### *5.2.2 Estimation of the prior distribution of the amplitudes* $\mathbf{g}_r$

The prior distribution $p(\mathbf{g}_r|\widehat{\boldsymbol{\mu}}_{g_r}, \widehat{\boldsymbol{\Sigma}}_{g_r})$ of $\mathbf{g}_r$ is modeled as the Gaussian with the mean $\widehat{\boldsymbol{\mu}}_{g_r}$ and the covariance $\widehat{\boldsymbol{\Sigma}}_{g_r}$. Both $\widehat{\boldsymbol{\mu}}_{g_r}$ and $\widehat{\boldsymbol{\Sigma}}_{g_r}$ depend on $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ and $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$. This dependence can be formulated by converting the PM parameters $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ and $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ into the GM parameters. Let $t_r'$ be the time at the center of the $r$th frame so that $t_r' = ((r-1)D + 0.5L)/f_s$ where $D$ is the hop size in samples. Evaluating the envelope function of PM in (6) at the center of the $r$th frame, we can find the estimated amplitude $\widehat{a}_{k,m,r,y,\text{PM}} = a(t_r'; \widehat{c}_k, \widehat{\boldsymbol{\varphi}}_m)$ where $\widehat{c}_k$ and $\widehat{\boldsymbol{\varphi}}_m$ are included in $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ and $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ respectively.

The phase at the center of $r$th frame can be calculated from $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$ and $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ by

$$\widehat{\phi}_{k,m,r,y,\text{PM}} = 2\pi \widehat{f}_{k,m,\text{PM}}(t_r' - \widehat{\tau}_k) + \widehat{\phi}_{k,m,\text{PM}} \tag{18}$$

where the frequency $\widehat{f}_{k,m,y,\text{PM}}$ and the phase $\widehat{\phi}_{k,m,\text{PM}}$ are included in $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$, and the time shift $\widehat{\tau}_k$ is included in $\widehat{\boldsymbol{\Psi}}_{y,\mathbb{V}}$. Then $\widehat{a}_{k,m,r,y,\text{PM}}$ and $\widehat{\phi}_{k,m,r,y,\text{PM}}$ in PM can be transformed into the amplitude of cosine $\widehat{\alpha}_{k,m,r,y,\text{PM}}$ and the amplitude of sine $\widehat{\beta}_{k,m,r,y,\text{PM}}$ in GM. The mean $\widehat{\boldsymbol{\mu}}_{g_r}$ of the prior is assigned to be these estimated amplitudes from PM so that

$$\widehat{\mu}_{\alpha_{k,m,r}} = \widehat{\alpha}_{k,m,r,y,\text{PM}} = \widehat{a}_{k,m,r,y,\text{PM}} \cos\widehat{\phi}_{k,m,r,y,\text{PM}} \tag{19}$$

$$\widehat{\mu}_{\beta_{k,m,r}} = \widehat{\beta}_{k,m,r,y,\text{PM}} = -\widehat{a}_{k,m,r,y,\text{PM}} \sin\widehat{\phi}_{k,m,r,y,\text{PM}} \tag{20}$$

where $\widehat{\mu}_{\alpha_{k,m,r}}$ and $\widehat{\mu}_{\beta_{k,m,r}}$ are the elements in $\widehat{\boldsymbol{\mu}}_{g_r}$ and they follow the ordering in (3).

The covariance $\widehat{\boldsymbol{\Sigma}}_{g_r}$ measures the deviation between the values of $\mathbf{g}_r$ estimated by PM and those estimated by GM. It is modeled as a diagonal matrix of which the diagonal is filled with the variances $\widehat{\sigma}_{\alpha_{k,m,r}}^2$ and $\widehat{\sigma}_{\beta_{k,m,r}}^2$ and follows the ordering in (3). We assume that the variances $\widehat{\sigma}_{\alpha_{k,m,r}}^2$ and $\widehat{\sigma}_{\beta_{k,m,r}}^2$ are identical and they are directly proportional to the power of the partial amplitude. This gives

$$\widehat{\sigma}_{\alpha_{k,m,r}}^2 = \widehat{\sigma}_{\beta_{k,m,r}}^2 = \bar{\sigma}_{G_k}^2 (\widehat{a}_{k,m,r,y,\text{PM}})^2 \tag{21}$$

where $\bar{\sigma}_{G_k}^2$ is the proportionality constant and it can be determined by the training data $\boldsymbol{\mathcal{X}}$ as below.

Let $\widehat{\alpha}_{k,m,r,\mathcal{X},\text{GM}}^i$ and $\widehat{\beta}_{k,m,r,\mathcal{X},\text{GM}}^i$ be the amplitudes in GM for $\boldsymbol{\mathcal{X}}$ and they have been estimated by the method in [14]. Let $\widehat{\alpha}_{k,m,r,\mathcal{X},\text{PM}}^i$ and $\widehat{\beta}_{k,m,r,\mathcal{X},\text{PM}}^i$ be the amplitudes in GM for $\boldsymbol{\mathcal{X}}$ and they are converted from the PM estimate. The conversion from the PM estimate to the GM estimate for $\boldsymbol{\mathcal{X}}$ follows that for the mixture $\mathbf{y}$ in (19) and (20). Let $\widehat{a}_{k,m,r,\mathcal{X},\text{PM}}^i$ be the partial amplitude in PM then

$$\widehat{a}_{k,m,r,\mathcal{X},\text{PM}}^i = \sqrt{\left(\widehat{\alpha}_{k,m,r,\mathcal{X},\text{GM}}^i\right)^2 + \left(\widehat{\beta}_{k,m,r,\mathcal{X},\text{GM}}^i\right)^2}. \tag{22}$$

Following (21), we can estimate $\bar{\sigma}_{G_k}^2$ from $\boldsymbol{\mathcal{X}}$ by

$$\bar{\sigma}_{G_k}^2 = \frac{1}{2I_k M_k R_k} \sum_{i=1}^{I_k} \sum_{m=1}^{M_k} \sum_{r=1}^{R_k^i} \left\{ \left( \frac{\delta\widehat{\alpha}_{k,m,r}^i}{\widehat{a}_{k,m,r,\mathcal{X},\text{PM}}^i} \right)^2 + \left( \frac{\delta\widehat{\beta}_{k,m,r}^i}{\widehat{a}_{k,m,r,\mathcal{X},\text{PM}}^i} \right)^2 \right\} \tag{23}$$

where $\delta\widehat{\alpha}_{k,m,r}^i = \widehat{\alpha}_{k,m,r,\mathcal{X},\text{GM}}^i - \widehat{\alpha}_{k,m,r,\mathcal{X},\text{PM}}^i$ and $\delta\widehat{\beta}_{k,m,r}^i = \widehat{\beta}_{k,m,r,\mathcal{X},\text{GM}}^i - \widehat{\beta}_{k,m,r,\mathcal{X},\text{PM}}^i$.

Note that the prior $p(\mathbf{g}_r|\widehat{\boldsymbol{\mu}}_{g_r}, \widehat{\boldsymbol{\Sigma}}_{g_r})$ reflects the difference between the individual tones estimated by GM and PM. As PM gives satisfactory quality of estimation in [15], the difference should be small enough to make the prior distribution $p(\mathbf{g}_r|\widehat{\boldsymbol{\mu}}_{g_r}, \widehat{\boldsymbol{\Sigma}}_{g_r})$ has a high density around the correct value of $\mathbf{g}_r$ as shown in the schematic diagram in Figure 1. Hence, overlapping partials can be resolved and higher quality of SS can be obtained. It will be verified and explained in the experiments.

#### *5.2.3 Estimation of the prior distribution of frequencies* $\mathbf{f}$

The prior distribution $p(\mathbf{f}|\widehat{\boldsymbol{\mu}}_f, \widehat{\boldsymbol{\Sigma}}_f)$ of $\mathbf{f}$ is modeled as the Gaussian with the mean $\widehat{\boldsymbol{\mu}}_f$ and the covariance $\widehat{\boldsymbol{\Sigma}}_f$. The mean $\widehat{\boldsymbol{\mu}}_f$ is set to the estimated frequencies in PM from $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ so that

$$\widehat{\mu}_{f_{k,m}} = \hat{f}_{k,m,\text{PM}} \tag{24}$$

where $\widehat{\mu}_{f_{k,m}}$ are the elements in $\widehat{\boldsymbol{\mu}}_f$. Following the derivation of $\widehat{\boldsymbol{\Sigma}}_{g_r}$, we also assume that $\widehat{\boldsymbol{\Sigma}}_f$ is a diagonal matrix of which the diagonal is filled with each variance $\widehat{\sigma}_{f_{k,m}}^2$. The variance $\widehat{\sigma}_{f_{k,m}}^2$ is modeled to be directly proportional to the square of the frequency in PM. This gives

$$\widehat{\sigma}_{f_{k,m}}^2 = \bar{\sigma}_{f_k}^2 \left( \widehat{f}_{k,m,\text{PM}} \right)^2 \tag{25}$$

where $\bar{\sigma}_{f_k}^2$ is the proportionality constant which can also be determined by the training data $\boldsymbol{\mathcal{X}}$. The estimate of $\bar{\sigma}_{f_k}^2$ is

$$\bar{\sigma}_{f_k}^2 = \frac{1}{M_k} \sum_{m=1}^{M_k} \left( \frac{\widehat{f}_{k,m,\mathcal{X},\text{GM}} - \widehat{f}_{k,m,\text{PM}}}{\widehat{f}_{k,m,\text{PM}}} \right)^2 \tag{26}$$

where $\widehat{f}_{k,m,\mathcal{X},\text{GM}}$ is the estimated frequency in GM for $\boldsymbol{\mathcal{X}}$ and it can be estimated by using the method in [14]. Note that there is no subscript $\mathcal{X}$ in $\widehat{f}_{k,m,\text{PM}}$ because $\widehat{f}_{k,m,\text{PM}}$ are the invariant PM parameters so the training data and the mixture share the same set of $\widehat{f}_{k,m,\text{PM}}$.

In summary, after estimating the hyperparameters $\widehat{\sigma}_{v_r}^2$ in (17), $\widehat{\boldsymbol{\mu}}_{g_r}$ in (19) and (20), $\widehat{\boldsymbol{\Sigma}}_{g_r}$ in (21), $\widehat{\boldsymbol{\mu}}_f$ in (24) and $\widehat{\boldsymbol{\Sigma}}_f$ in (25), we can find the MAP solution $\widehat{\boldsymbol{\Theta}}_y$ of GM by iteratively updating the amplitude matrix $\mathbf{G}$ in (10) and the frequencies $\mathbf{f}$ in (13). In the next section, experimental results will be

presented to show the performance of the whole SS process.

## 6. EXPERIMENTS

### 6.1 Data set and experimental setup

We used the same data set in [15] for comparing the performance. The data set contains 25 mixtures. Each mixture was generated by mixing the isolated tones in the recorded piano databases [8, 15], taken from 4 different pianos. Only tones from the same piano were used to form a mixture. The pitches in each mixture correspond to a chord randomly selected from 11 piano pieces in the RWC database [8]. The number of tones (represented by $K$) in our selected mixtures ranges from 1 to 6: 1 tone (8 mixtures), 2 tones (6), 3 tones (5), 4 tones (4), 5 tones (1) and 6 tones (1). These 25 mixtures consist of 62 tones. 7 mixtures contain one pair of octaves, 2 ($K=5$ and $K=6$) contain 2 pairs of octaves. For the training data, two instances of each pitch are available so $I_k=2$. The first 0.5 second of the mixtures and the training data were used in the experiments. All data were downsampled to 11.025 kHz for faster processing. The window setting in GM is as follows: the window function is the hamming window with length 11.61 ms ($L=128$) and 50% overlap. The titles of the piano pieces used and details of the selected mixtures are available on the website of this paper (link available at the end of this section).

### 6.2 Results

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

$$\text{SNR}=10\log_{10}\frac{\sum_n \mathsf{x}(t_n)^2}{\sum_n (\mathsf{x}(t_n)-\widehat{\mathsf{x}}(t_n))^2} \qquad (27)$$

where $\mathsf{x}(t_n)$ is the isolated tone in the time domain before mixing and $\widehat{\mathsf{x}}(t_n)$ is the estimated tone in the time domain. The isolated tones give the ground truth for evaluation.

#### *6.2.1 Evaluation on modeling quality*

We followed the procedures in [15] to evaluate the modeling quality, i.e. the quality of PM and GM to represent an isolated tone before mixing. The isolated tones of the 25 mixtures were inputted into our proposed SS system including both PM and GM. The outputs of our system were the estimated tones reconstructed from PM and GM. If the parameters obtained in PM and GM are accurate, they can regenerate the original tones in high quality. The result is that the average SNRs of PM and GM are 11.15 dB and 17.38 dB respectively. The average SNR of GM is much higher than that of PM. This is because GM is more flexible to represent piano tones.

#### *6.2.2 Comparing with other systems for separation quality*

The procedures in [15] were followed to evaluate the separation quality, i.e. the quality of PM and GM separating a mixture into its individual tones. We also compared PM and GM with a recent SS system in [11], in which Li, Woodruff and Wang built their system (Li's system) based on CAM mentioned in Section 1. It uses the non-overlapping partials to estimate the overlapping partials of the same note. The implementation of Li's system was provided by the authors. The true fundamental frequency of each tone was supplied to Li's system.

| | $\overline{\text{SNR}}$ (dB) | | |
|---|---|---|---|
| | PM | GM | Li |
| All mixtures | 10.88 | 13.51 | 6.63 |
| $K=2$ | 11.76 | 15.26 | 12.07 |
| $2\leq K\leq 6$ | 10.97 | 13.15 | 5.40 |
| Upper tones in octaves | 10.95 | 12.77 | 1.57 |

**Table 1**. Comparison of Li's system and our PM and GM.

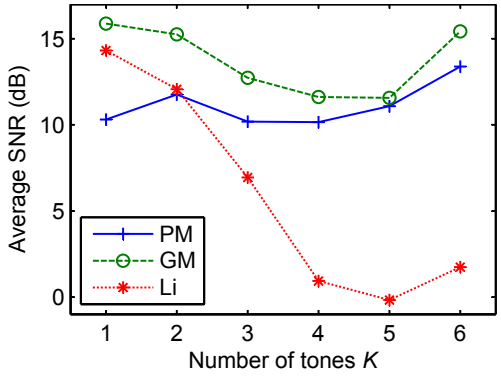


**Figure 3**. Average SNR against the number of tones $K$ for PM, GM, and Li's system.

The results are shown in Table 1. For the 25 mixtures, the average SNRs of PM, GM and Li' system are 10.88 dB, 13.51 dB and 6.63 dB respectively. Both PM and GM outperform Li's system. 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 because non-overlapping partials are not available. On the other hand, both PM and GM are able to reconstruct the upper tone in an octave. The overlapping partials were successfully resolved even for mixtures containing 2 pairs of octaves of C3, G3, C4, E4, G4 ($K=5$) and of F♯3, C4, F4, C5, D5, F5 ($K=6$).

The average SNR against the number of tones $K$ is plotted in Figure 3. The average SNR of Li's system decreases much more rapidly than PM and GM. Our system can make use of the training data to give higher separation quality. Some audio files in the experiments are selected for demonstration purpose. The audio files, titles of piano pieces used, details of the selected mixtures and mathematical notations used in this paper are available at http://www.cse.cuhk.edu.hk/~khwong/www2/conference/ismir2015/ismir2015.html.

## 7. CONCLUSIONS

Here we have proposed a score-informed monaural SS system to extract each tone from a mixture of piano tone signals. Two sinusoidal models, PM and GM, are employed to represent piano tones in the system. We formulate a hierarchical Bayesian framework to run both Models in the SS process so that the mixtures with overlapping partials can be resolved with high quality. Experiments show that our proposed system gives robust and accurate separations of mixtures and improves the separation quality significantly comparing to the previous work.

## 8. REFERENCES


[1] A. Askenfelt, editor. *Five Lectures on the Acoustics of the Piano*. Royal Swedish Academy of Music, 1990. Available online at http://www.speech.kth.se/music/5_lectures/.

[2] T. Berg-Kirkpatrick, J. Andreas, and D. Klein. Unsupervised transcription of piano music. In *Advances in Neural Information Processing Systems*, pages 1538–1546, 2014.

[3] C. M. Bishop. *Neural Network for Pattern Recognition*. Oxford University Press, New York, 1995.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[5] E. Schubert D. Fabian, R. Timmers, editor. *Expressiveness in Music Performance: Empirical Approaches Across Styles and Cultures*. Oxford University Press, 2014.

[6] M. Davy, S. Godsill, and J. Idier. Bayesian analysis of polyphonic western tonal music. *Journal of the Acoustical Society of America*, 119(4):2498–2517, April 2006.

[7] M. R. Every and J. E. Szymanski. Separation of synchronous pitched notes by spectral filtering of harmonics. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1845–1856, 2006.

[8] 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)*, pages 229–230, October 2003.

[9] Jinyu Han and B. Pardo. Reconstructing completely overlapped notes from musical mixtures. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 249–252, 2011.

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

[11] 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*, 17(7):1361–1371, 2009.

[12] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

[13] D. J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4:448–472, 1992.

[14] W. M. Szeto and K. H. Wong. Sinusoidal modeling for piano tones. In *2013 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC)*, pages 1–6, Kunming, China, August 2013. Available online at http://www.cse.cuhk.edu.hk/~khwong/www2/conference/ismir2015/ismir2015.html.

[15] W. M. Szeto and K. H. Wong. Source separation and analysis of piano music signals using instrument-specific sinusoidal model. In *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13)*, pages 109–116, Maynooth, Ireland, September 2013. Available online at http://www.cse.cuhk.edu.hk/~khwong/www2/conference/ismir2015/ismir2015.html.

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

[17] M. Zivanovic. Harmonic bandwidth companding for separation of overlapping harmonics in pitched signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(5):898–908, May 2015.

[18] U. Zöler, editor. *DAFX - Digital Audio Effects*. Wiley, 2nd edition, 2011.

# A HIERARCHICAL BAYESIAN FRAMEWORK FOR SCORE-INFORMED SOURCE SEPARATION OF PIANO MUSIC SIGNALS

**Wai Man SZETO**
Office of University General Education
The Chinese University of Hong Kong
wmszeto@cuhk.edu.hk

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

# APPENDIX

## A. LIST OF PIANO PIECES

| No. | Title | Composer | Style |
|---|---|---|---|
| 1 | Piano Sonata in A major, K.331/300i, 1st mvt. | Mozart, Wolfgang Amadeus | Classical |
| 2 | Variations on Ah Vous Dirai-je Maman, K.265/300e | Mozart, Wolfgang Amadeus | Classical |
| 3 | Piano Sonata no. 23 in F minor, op.57 Appassionata, 1st mvt. | Beethoven, Ludwig van | Classical |
| 4 | Traumerei from Suite Kinderszenen, op.15 | Schumann, Robert | Classical |
| 5 | Nocturne no.2 in E♭ major, op.9 no.2 | Chopin, Frederic | Classical |
| 6 | Etude in E major, op.10 no.3 | Chopin, Frederic | Classical |
| 7 | La Campanella from Grandes Etudes de Paganini | Liszt, Franz | Classical |
| 8 | Three Gymnopedies no.1 | Satie, Erik | Classical |
| 9 | Clair de Lune from Suite Bergamasque | Debussy, Claude | Classical |
| 10 | Jive (Piano Solo) | Nakamura, Makoto | Jazz |
| 11 | For Two (Piano Solo) | Nakamura, Makoto | Jazz |
| 12 | Lounge Away (Piano Solo) | Nagai, Takao | Jazz |

**Table 1**. Piano pieces from RWC database [1] for generation of mixtures.



## B. LIST OF MIXTURES

| No. | $K$ | Pitches | Octaves | Loudness |
|---|---|---|---|---|
| 1 | 1 | G2 | - | L |
| 2 | 1 | D♯3 | - | S |
| 3 | 1 | D5 | - | M |
| 4 | 1 | D3 | - | S |
| 5 | 1 | D♯6 | - | M |
| 6 | 1 | E4 | - | L |
| 7 | 1 | F4 | - | M |
| 8 | 1 | C5 | - | L |
| 9 | 2 | D♯4, B4 | 0 | M, M |
| 10 | 2 | G♯4, C5 | 0 | M, M |
| 11 | 2 | C4, C5 | 1 | M, M |
| 12 | 2 | A3, C♯5 | 0 | S, L |
| 13 | 2 | E4, F♯5 | 0 | S, L |
| 14 | 2 | C4, F4 | 0 | M, L |
| 15 | 3 | A♯4, A♯5, C♯6 | 1 | M, M, M |
| 16 | 3 | G4, E5, F5 | 0 | M, L, L |
| 17 | 3 | B2, A♯3, D♯4 | 0 | M, L, M |
| 18 | 3 | B1, D♯4, G♯4 | 0 | S, M, M |
| 19 | 3 | E3, C4, C6 | 1 | M, M, L |
| 20 | 4 | D4, F4, A4, D5 | 1 | L, L, L, L |
| 21 | 4 | C3, G3, E4, G4 | 1 | S, M, M, M |
| 22 | 4 | D3, G3, D4, A♯4 | 1 | S, M, M, L |
| 23 | 4 | A3, C♯4, F♯4, F♯5 | 1 | S, M, M, L |
| 24 | 5 | C3, G3, C4, E4, G4 | 2 | M, M, M, M, M |
| 25 | 6 | F♯3, C4, F4, C5, D5, F5 | 2 | M, M, L, L, M, M |

**Table 2**. List of the 25 mixtures. Loudness: "S" is soft; "M" is medium; and "L" is loud.

## C. BAYESIAN FRAMEWORK

A larger figure of the whole source separation process is shown in Figure 1.

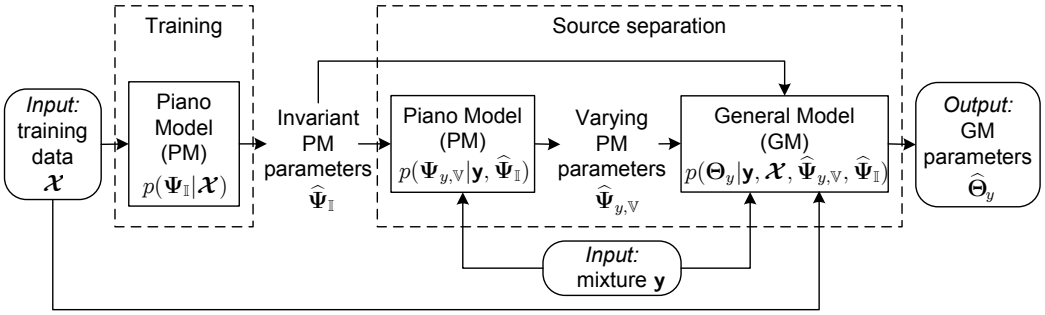


**Figure 1**. Bayesian framework for source separation.

## D. NOTATIONS

| Symbol | Meaning |
|---|---|
| $a$ | Partial amplitude |
| $c$ | Intensity of a tone in PM |
| $D$ | Hop size |
| $\mathbf{f}$ | Frequency vector |
| $f_s$ | Sampling frequency |
| $\mathbf{G}$ | Amplitude matrix in GM |
| $\mathbf{g}$ | Amplitude vector in GM |
| $\mathbf{H}$ | Frequency matrix in GM |
| $I$ | Number of instances |
| $i$ | Index for $i$th instance |
| $K$ | Number of tones in a mixture |
| $k$ | Index for $k$th tone |
| $L$ | Window length in samples |
| $l$ | Discrete time index in a windowed signal |
| $M$ | Number of partials |
| $m$ | Index for $m$th partial |
| $N$ | Time length of a signal in samples |
| $n$ | Discrete time index in a signal |
| $p$ | Pitch |
| $R$ | Number of frames |
| $r$ | Index for $r$th frame |
| $t$ | Time in seconds |
| $v$ | Noise in GM |
| $w$ | Window function |
| $\mathbf{X}$ | Signal matrix of a tone with all frames |
| $\boldsymbol{\mathcal{X}}$ | Isolated tones in the training data |
| $\mathbf{x}$ | Signal vector of a tone in a frame |
| $\mathbf{x}$ | Signal vector of a tone in an entire duration |
| $\mathbf{Y}$ | Signal matrix of a mixture with all frames |
| $\mathbf{y}$ | Signal vector of a mixture in a frame |
| $\mathbf{y}$ | Signal vector of a mixture in an entire duration |
| $\mathbf{Z}$ | Jacobian matrix |
| $\alpha$ | Amplitude of the cosine term |
| $\beta$ | Amplitude of the sine term |
| $\epsilon$ | Noise in PM |
| $\boldsymbol{\Theta}$ | Parameter set of a mixture in GM |
| $\mu$ | Mean of a Gaussian distribution |
| $\boldsymbol{\Sigma}$ | Covariance matrix of a Gaussian distribution |

| Symbol | Meaning |
|---|---|
| $\sigma$ | Standard deviation of a Gaussian distribution |
| $\tau$ | Time shift in seconds |
| $\phi$ | Phase |
| $\boldsymbol{\varphi}$ | Envelope parameters in PM |
| $\boldsymbol{\Psi}$ | Parameter set of a mixture in PM |
| $\boldsymbol{\Psi}_{\mathbb{I}}$ | Invariant PM parameters |
| $\boldsymbol{\Psi}_{y,\mathbb{V}}$ | Varying PM parameters of the mixture $\mathbf{y}$ |
| $\boldsymbol{\psi}$ | Parameter set of a tone in PM |

## E. REFERENCES


[1] 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)*, pages 229–230, October 2003.