Source Separation and Analysis of Piano Music Signals

SZETO, Wai Man

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in
Computer Science & Engineering

The Chinese University of Hong Kong
August, 2010

Thesis/Assessment Committee

Professor CHAN Lai Wan (Chair)

Professor WONG Kin Hong (Thesis Supervisor)

Professor BOGDANOV Andrej (Committee Member)

Professor HORNER Andrew (External Examiner)

Source Separation and Analysis of Piano Music Signals

submitted by

SZETO, Wai Man

for the degree of Doctor of Philosophy
at the Chinese University of Hong Kong

Abstract

What makes a good piano performance? An expressive piano performance owes its emotive power to the performer's skills in shaping the music with nuances. For the purpose of performance analysis, nuance can be defined as any subtle manipulation of sound parameters including attack, timing, pitch, loudness and timbre. A major obstacle to a systematic computational analysis of musical nuances is that it is often difficult to uncover relevant sound parameters from the complex audio signal of a piano music performance. A piano piece invariably involves simultaneous striking of multiple keys, and it is not obvious how one may extract the parameters of individual keys from the combined mixed signal. This problem of parameter extraction can be formulated as a source separation problem. Our research goal is to extract individual tones (frequencies, amplitudes and phases) from a mixture of piano tones.

We propose a Bayesian monaural source separation system to extract each individual tone from mixture signals of piano music performance. Specifically, tone extractions can be facilitated by model-based inference. Two signal models based on summation of sinusoidal waves were employed to represent piano tones. The first model is the traditional General Model, which is a variant of sinusoidal modeling, for representing a tone for high modeling quality; but this

model often fails for mixtures of tones. The second model is an instrument-specific model tailored for the piano sound; its modeling quality is not as high as the traditional General Model, but its structure makes source separation easier. To exploit the benefits offered by both the traditional General Model and our proposed Piano Model, we used the hierarchical Bayesian framework to combine both models in the source separation process. These procedures allowed us to recover suitable parameters (frequencies, amplitudes, phases, intensities and fine-tuned onsets) for thorough analyses and characterizations of musical nuances. Isolated tones from a target recording were used to train the Piano Model, and the timing and pitch of individual music notes in the target recording were supplied to our proposed system for different experiments. Our results show that our proposed system gives robust and accurate separation of signal mixtures, and yields a separation quality significantly better than those reported in previous works.

摘要

優秀的鋼琴演奏為何優秀？情感豐富的鋼琴演奏之所以具感染力，全賴演奏者以細微差別呈現音樂。為了將演奏加以分析，「細微差別」可定義為細微地操縱任何聲音參數，包括起奏、時間的安排、音高、音量和音色。以計算機有系統地分析音樂的細微差別時，最大障礙在於鋼琴音樂演奏的音頻信號十分複雜，因此往往很難找出相關的聲音參數。鋼琴作品毫無例外地要求同時彈出多個琴鍵，而怎樣從多個琴鍵的混合信號提取個別琴鍵的參數，方法也並非一目了然。這個提取參數的難題可闡述為「源分離問題」。本研究的目標是從混合了鋼琴多個樂音的聲音裡抽取個別樂音（頻率、振幅和相位）。

我們提議採用「貝葉斯單聲道源分離系統」從鋼琴音樂演奏的混合信號中提取每個樂音。具體來說，基於模型的推理有助提取樂音，而我們則採用兩個根據正弦波相加的信號模型來代表鋼琴樂音。「模型一」是傳統的一般模型，是正弦建模的變種，代表樂音的高質素建模，但這個模型往往不能應付混合的樂音。「模型二」是專為鋼琴聲音而設的特定樂器模型。相較於傳統的一般模型，「模型二」的建模質素較低，但其結構使源分離更容易。為了盡量利用傳統的一般模型和我們提出的鋼琴模型的優點，我們在源分離過程使用分層貝葉斯框架把這兩種模型合二為一。這樣便能取得適當的參數（頻率、振幅、相位、強度和微調聲母）以深入分析和呈現音樂的細微差別。「模型二」利用抽取自將要被分析的錄音裡的孤立樂音來訓練鋼琴模型，並根據個別樂音的時間安排和音高資料，作為實驗的基礎。結果表明我們提出的系統能清楚準確地將混合的信號分離，其分離質素明顯較過去的研究成果為佳。

To Eos

Acknowledgments

First of all, I wish to express my profound indebtedness to Professor Kin Hong Wong, my supervisor, without whose guidance and encouragement I would not have been able to complete this work.

Besides, I would like to thank Professor Andrej Bogdanov, Professor Laiwan Chan and Professor Andrew Horner, who are members of my thesis committee, for their support and advice.

I wish to thank Dr. Yipeng Li, Mr. John Woodruff and Professor DeLiang Wang for providing me with the implementation of their research work.

I have also profited from the invaluable comments from Professor Philip H. W. Leong, Dr. Vincent C. K. Cheung, Mr. Chi Hang Wong, Mr. Ken K. L. Law, Dr. Chris Ng, Dr. Kun Zhang, Professor Yiu Tong Chan, Mr. Kai Tong Sun and Miss Eos H. T. Cheng.

Finally, I would also like to express special gratitude to Dr. Koon Kit Lam and Mr. Man Chun Kam for their support.

Contents

Abstract ii

Acknowledgements vi

1 Introduction 1

1.1 Applications 2

1.2 Problem definition (mixtures, tones and partials) 3

1.3 Overview of our proposed system 7

1.4 Thesis organization and contributions 10

2 Literature review 14

2.1 Research on music performance 14

2.2 Music source separation 15

2.2.1 Monaural source separation - solving the underdetermined problem 15

2.2.2 Resolving overlapping partials 17

3 Signal model representations 21

3.1 Properties of piano tones 22

3.2 Traditional General Model 23

3.3 Proposed Piano Model 31

3.4 Estimation of the number of partials 36

4 Bayesian framework for source separation **39**

4.1 Summary of our main idea 40

4.2 Motivation: problems of parameter estimation in the General Model 41

4.2.1 Example to illustrate rank deficient $\mathbf{H}$ 45

4.3 Bayesian analysis for source separation 48

4.3.1 Example to illustrate how the Bayesian framework works 51

4.4 Problem formulation for source separation 53

5 Training: parameter estimation **61**

5.1 Problem formulation for training 61

5.2 Frequency estimation by peak-picking 67

5.3 Extraction of partials with the General Model 67

5.3.1 Step 1: update the amplitude matrix $\mathbf{G}$ 73

5.3.2 Step 2: update the noise variances σ_V^2 73

5.3.3 Step 3: update the frequencies $\mathbf{f}$ 73

5.3.4 Summary of the partial extraction 76

5.4 Finding the initial guess for the Piano Model 77

5.4.1 Finding the initial guess $\varphi_m^{(0)}$ 78

5.4.2 Finding the initial guess $\phi_m^{(0)}$ 79

5.5 Parameter estimation of the Piano Model 79

6 Source separation: parameter estimation **81**

6.1 Stage 1: source separation with the Piano Model 82

6.2 Stage 2: source separation with the General Model 84

6.2.1 Bayesian analysis for the General Model 84

6.2.1.1 Step 1: update the amplitude matrix $\mathbf{G}$ 85

6.2.1.2 Step 2: update the frequencies $\mathbf{f}$ 87

6.2.2 Estimation of the hyperparameters 89

6.2.2.1 Estimation of the noise variance $\sigma_{v_r}^2$ 89

6.2.2.2 Estimation of the prior 91

7 Experiments **95**

7.1 Databases of piano tones 95

7.2 Generation of mixtures 97

7.3 Results 100

7.3.1 Evaluation criteria 100

7.3.2 Evaluation on modeling quality 101

7.3.3 Evaluation on separation quality 102

7.3.3.1 Overall analysis 103

7.3.3.2 Effect of window length 106

7.3.3.3 Comparison with other method 107

7.3.4 Computation time 109

8 Conclusion and discussion **110**

A Notation **112**

B Derivation **115**

B.1 Derivation of the normalization coefficient in the Piano Model . 115

C List of piano pieces and mixtures **117**

List of Tables

2.1 Comparison of different monaural source separation methods on recovery of amplitude and phase. 18

3.1 Invariant PM parameters and varying PM parameters. 35

7.1 The average SNR of all estimated individual tones of the 25 mixtures before mixing. 101

7.2 The average SNR of the 25 mixtures. The number of tones in a mixture is denoted by K. The column of $\overline{\text{SNR}}$ is the average SNR in dB. The column of $\overline{\Delta\text{SNR}}$ is the average SNR difference between modeling and source separation. 104

7.3 The average absolute error ratio of intensity ER_c 106

7.4 The absolute error of the estimated time shift Err_τ in PM. . . . 106

7.5 Comparison of Li's system and our proposed system PM and GM. 108

7.6 Average computation for one mixture. 109

C.1 Piano pieces from RWC database [34] for generation of mixtures. 118

C.2 List of the 25 mixtures. Loudness: "S" is soft; "M" is medium; and "L" is loud. 119

List of Figures

1.1 The generation of a mixture signal and the goal of source separation. 4

1.2 (a) The magnitude spectrum of tone 1 (C4). (b) The magnitude spectrum of tone 2 (C5). (c) The spectrum of the mixture of tone 1 (C4) and tone 2 (C5). This mixture was obtained by adding the C4 and the C5 in the time domain. The duration is 372 ms and Hamming window was applied. Note that the magnitude spectrum of the mixture C4 and C5 is not equal to the addition of those of individual C4 and C5. 6

1.3 Example of illustrating the pre-processing steps (music transcription and segmentation). 11

1.4 The musical score of the opening of Beethoven's Piano Sonata in E♭, Op. 81a "Les Adieux". 11

1.5 The whole source separation process from a target recording to the estimated individual tones. 12

1.6 The main steps of our source separation process. See the text for the explanation. 12

3.1 Spectrogram of a C4 piano tone played moderately loud. 23

3.2 Extracted partials of the C4 piano tone in Figure 3.1. The partials are extracted by using the method in Section 5.3 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. 24
3.3 The matrix form of $\widehat{\mathbf{x}}_{k,r}$. 29
3.4 Envelope surface against peak amplitude of the time-domain signal and time for the first four partials. (a) The first partial (fundamental frequency). (b) The second partial. (c) The third partial. (d) The fourth partial. 32
3.5 The number of partials M_k of each pitch. 38

4.1 Comparison between the General Model and the Piano Model. (a) Mixture without overlapping partials. (b) Mixture with overlapping partials. The invariant PM parameters are estimated from training data. 40
4.2 Flow of our source separation process. 41
4.3 Rank deficient $\mathbf{H}$. 44
4.4 The likelihood functions of the four data sets. The plots in the left column show the data space. The red line is the line generated by $\mathbf{g}_{\text{true}}$. The plots in the right column show the corresponding likelihood in the parameter space. The white crosses show the true parameter values $\mathbf{g}_{\text{true}}$. 49

4.5 (a) The likelihood function. (b) The prior distribution. (c) The posterior distribution. This schematic diagram shows that an appropriate prior gives the desirable MAP solution. The vertical line shows the true value of Θ. 50

4.6 Illustration of Bayesian analysis for the data set in Figure 4.4(d) with an appropriate prior. (a) Data space. The blue circles are the observed data. The red solid line is the true line. The blue dotted line is generated from the prior mean by setting $\mathbf{g} = \boldsymbol{\mu}_g$. The dashed black line is generated from the posterior mean by setting $\mathbf{g} = \mathbf{m}_g$. (b) The prior distribution. (c) The likelihood function. (d) The posterior distribution. All the white crosses in (b), (c) and (d) are the location of $\mathbf{g}_{\text{true}}$. 54

4.7 Illustration of Bayesian analysis for the data set in Figure 4.4(d) with an inappropriate prior. (a) Data space. The blue circles are the observed data. The red solid line is the true line. The blue dotted line is generated from the prior mean by setting $\mathbf{g} = \boldsymbol{\mu}_g$. The dashed black line is generated from the posterior mean by setting $\mathbf{g} = \mathbf{m}_g$. (b) The prior distribution. (c) The likelihood function. (d) The posterior distribution. All the white crosses in (b), (c) and (d) are the location of $\mathbf{g}_{\text{true}}$. 55

4.8 Bayesian framework for the source separation. 60

5.1 The system flow of the training stage. 66

5.2 The procedures for extracting partials with GM. 72

6.1 The two stages in source separation. 82

7.1 Recording setup of our recorded piano database. 97

7.2 The counts of the mixtures with number of tones K for the experiments. The total number of mixtures is 25. 99

7.3 Generation of mixtures. 99
7.4 The procedures of evaluation on modeling quality. 102
7.5 The procedures of evaluation on separation quality for a mixture. 103
7.6 Average SNR against the number of tones K. 105
7.7 Average SNR difference against the number of tones K. 105
7.8 Average SNR against the window length. 107
7.9 Average SNR against the number of tones K for our system and Li's system. 108

Chapter 1

Introduction

Why do some music performances sound more expressive and alluring than others? Why does a particular music performance exert greater emotional impact to the audience? One approach to answering these age-old questions is to compare the audio signal of an expressive performance with that of an unexpressive one, and analyze how they differ in their nuances. Nuance may 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 rather than dead and mechanical [45]. In recent years, researchers from various disciplines including musicology, psychology, neuroscience and computer science have tried to quantify musical nuances through these sound parameters in order to unveil the mystery behind expressive music performances.

A major obstacle to a systematic computational analysis of musical nuances is that it is often difficult to uncover relevant sound parameters from the complex audio signal of a music performance. For instance, a piano piece invariably involves simultaneous striking of multiple keys, and it is not obvious how one may extract the parameters of individual keys from the combined mixed signal. This problem can be formulated as a source separation problem. Source separation of music signals commonly refers to the challenging problem

of separating the signal of an individual instrument from the mixed signal containing sources from various musical instruments, or of extracting individual tones from a mixture of musical tones.

Here, we propose a Bayesian monaural source separation system to extract each individual tone from mixture signals of piano music performance. Specifically, tone extraction can be facilitated by model-based inference. In this research, two signal models based on the summation of sinusoidal waves are employed to represent piano tones. (1) We use a traditional General Model, which is a variant of sinusoidal modeling, to represent a tone for high modeling quality; but the model often fail for mixtures of tones. (2) We propose an instrument-specific model tailored for the piano sound. Although its modeling quality is not as high as the traditional General Model, it makes source separation easier. (3) To exploit the benefits of both the traditional General Model and our proposed Piano Model, we use the hierarchical Bayesian framework to combine both models in the source separation process. These procedures will allow us to recover suitable parameters for thorough analyzes and characterizations of musical nuances, which, in turn, will open up new avenues in many applications.

1.1 Applications

Performance analysis and manipulation If each individual tone in music recordings can be extracted by our source separation method, we can analyze the styles of various artists by comparing the similarities and differences in the nuances in their performances. In particular, the elements of nuance we are most interested in are intensity and timbre. In piano playing, dynamic shaping for chords is an important means of expression [31, p. 148]. It is delivered by delicately controlling the intensity of the tones in a chord and in the melody represented by the chord sequence. After tone extraction, each extracted tone

can also be individually manipulated for different musical effects.

Regeneration of high sound quality recordings One can analyze old recordings and quantify the nuances of a past master's performing style. Such knowledge may enable a computer-controlled instrument (e.g., electric piano) to remake high-quality sound track based on the genuine style of the past master. The company Zenph Studio has issued commercial recordings based on this concept, but the sound parameters of the nuances they used were acquired by intensively and carefully tuning the sound parameters of each tone by professional musicians with the help of computers [35]. Our source separation method may automate this whole labor intensive process.

Object-based audio coding The extracted information of the tones can be stored as an object in a compressed file structure [63]. Under this coding scheme, high compression rate with high sound quality can be achieved because only the estimated parameters of the tones are stored and transmitted.

Music tutorial system When a piano student practicing a piece, the playing of the piano student can be recorded. If each tone in the playing can be extracted by our source separation method, the extracted information can be used as an input to a music tutorial system which analyzes the playing and then gives feedback to the learner.

1.2 Problem definition (mixtures, tones and partials)

When a piano key is pressed, the hammer hits the strings and the strings vibrate. Then the energy is transferred from the strings to the soundboard, and the sound radiates from here [2]. The sound can be recorded by a microphone

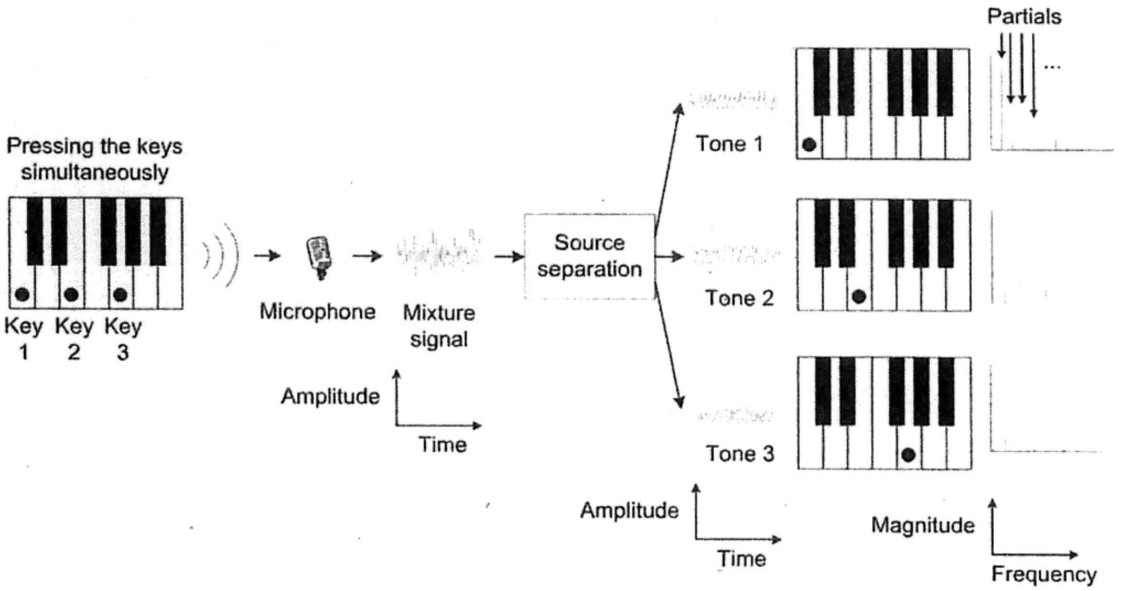

Figure 1.1: The generation of a mixture signal and the goal of source separation.

and the recorded sound can be stored as a time-domain acoustic signal. The signal generated by pressing a piano key is called a piano tone. When multiple piano keys are pressed, a mixture signal is generated as shown in Figure 1.1. The goal of source separation is to extract the individual tones from the mixture signal. Here, we tackle the problem of monaural source separation in which multiple sources were prerecorded by a single microphone or mixed into a single audio channel.

In this research, an individual tone in a mixture is considered as a particular sound source of the corresponding key. We model a mixture signal as a linear superposition of its corresponding individual tones:

$$\mathsf{y}(t_n) = \sum_{k=1}^{K} \mathsf{x}_k(t_n) \tag{1.1}$$

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, t_n is time in second, and n is an index for discrete time. In this mixture model,

the mixture signal is a sum of the individual tones. The problem of source separation is to recover the signal of each individual tone, $x_k(t)$, from the observed mixture signal $y(t)$.

Recovering individual tones from a mixture is challenging. In fact, the problem is under-determined if no assumption of the signals is made. The number of unknown variables (all $x_k(t_n)$) in (1.1) is more than that of the knowns (all $y(t_n)$) for more than one tone. This means that appropriate assumptions are essential for solving the problem. Some researchers are using more than one channel for separation [4, 14, 68]. In this dissertation, we only focus on the separation of monaural signals.

Another difficulty arises from the nature of piano tones. When a piano key is pressed, the hammer hits the strings. The strings vibrate at their resonance frequencies and the result is a set of simultaneously sounding sinusoidals, called *partials*, forming a piano tone. Hence, a piano tone is the sum of its partials. The partial with the lowest frequency is called the first partial or fundamental frequency. The second lowest frequency is called the second partial, the third lowest frequency is called the third partial, and so on. For a piano tone, the partials are quasi-harmonic, meaning that the partials are approximate multiples of the fundamental frequency. When multiple keys are pressed, a mixture signal is formed. The mixture is the sum of all partials from each individual tone. As music is usually not entirely dissonant, it is common that some partials from different tones may overlap with each other. If two partials are overlapping, their frequencies have very close values so that they are effectively identical. For example, octave intervals often appear in piano music; two tones are in octave if the fundamental frequency of the higher tone is twice as that of the lower. For an octave mixture, the second partial of the upper tone overlaps with the first partial of the lower tone, the fourth partial of the upper tone overlaps with the second partial of the lower tone, and so on. Hence, the frequencies of the upper tone are totally immersed within those

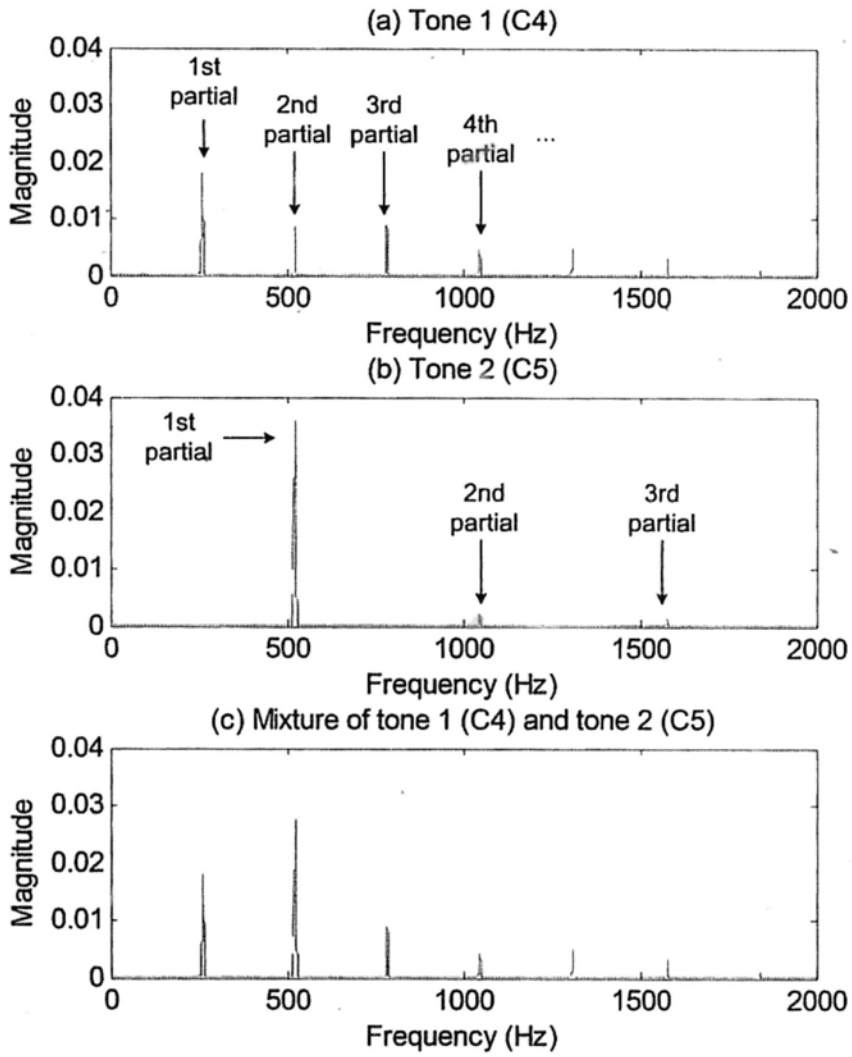

Figure 1.2: (a) The magnitude spectrum of tone 1 (C4). (b) The magnitude spectrum of tone 2 (C5). (c) The spectrum of the mixture of tone 1 (C4) and tone 2 (C5). This mixture was obtained by adding the C4 and the C5 in the time domain. The duration is 372 ms and Hamming window was applied. Note that the magnitude spectrum of the mixture C4 and C5 is not equal to the addition of those of individual C4 and C5.

of the lower (Figure 1.2). 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.

1.3 Overview of our proposed system

Many existing monaural source separation systems use sinusoidal modeling to model musical sounds [71, 46, 32, 69, 20, 14]. 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. The existing systems approach the under-determined problem and the overlapping partial problem by constraining or favoring certain regions of the parameter values in the sum-of-sine model. This can be imposed by the following ways:

1. Given a target music recording, the whole source separation process is divided into two main stages. In the first stage, the timing and sequence of music notes played in the piece are first analyzed. This information can be obtained by using music transcription systems [41] which find the onsets, duration, and pitch or fundamental frequency of each tone in the recording. The transcribed result can be further corrected manually with the help of an annotation tool [15]. Based on the transcribed results, the parameter space of the sum-of-sinusoidal model can be delineated. In the second stage, source separation is performed in this parameter space.

2. According to the general properties of pitched musical sounds, assumptions are made to resolve overlapping partials. For example, the spectral envelope of tones is assumed to be smooth (as in [71, 28]). The information of neighboring non-overlapping partials can also be utilized to estimate the parameters of an overlapping partial. Another assumption is that amplitude of partials from the same source are similar [46]. This is known as common amplitude modulation (CAM).

In our source separation system, sinusoidal modeling is used to represent piano tones. We also follow the two-stage method so that the onset, duration and pitch of each tone are fed into the source separation stage for limiting the parameter search space. The source separation stage is the focus of our research and we propose a source separation system for this stage. Our system, however, does not make assumption on the spectral envelope and CAM because these assumptions are often violated in piano music signals. 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 as discussed in the previous section. Usually CAM performs well in separating mixtures of different musical instruments. However, individual tones in a mixture of piano signals have much more contrasting timbre than tones from different musical instruments. Thus, CAM cannot 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 common characteristics which can be captured by our Piano Model. 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 the Piano Model. The common characteristics captured by the Piano Model do not vary when no pedaling is applied. 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 possibilities of training the Piano Model with mixtures and separating mixtures generated with pedaling will be discussed in Chapter 8. The procedures of the whole source separation process are presented below.

The required input of our source separation algorithm is a set of mixture signals. The mixture signals are obtained from a given target recording of a piano piece via two pre-processing steps as shown in the example in Figure 1.3. The figure shows the signal of the target recording which is a performance of the opening of Beethoven's Piano Sonata in E♭, Op. 81a "Les Adieux" (Figure 1.4). Note that the musical score is provided here only for clarifying the example; the music transcription process does not require the score. In the first pre-processing step, the onset, duration and pitch of each tone in the recording are found by a music transcription system. In the second pre-processing step, the signal of the recording is first segmented into a sequence of segmented signals according to the transcribed result. The pitches within each segment do not change. Each segment may contain different numbers of tone and different pitches. For example in Figure 1.3, the first segment contains E♭4 and G4, the second, B♭3 and F4, and the third, C2, C3, G3 and E♭4.

The whole source separation process is summarized in Figure 1.5. The segmented signals from the pre-processing steps are divided into two sets. These two sets are the inputs to our source separation system. One set contains only the isolated tones which are used for training. Another set contains the mixture signals. The information gained from training helps to separate the mixtures into their individual tones. The outputs of the whole process are the

estimated tones, the estimated intensities and the fine-tuned onsets.

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 3.3. The main steps in our source separation system are depicted in Figure 1.6. 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 the Piano Model are estimated. The Piano Model (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. Estimates of these invariant and varying parameters are then fed into the procedure for estimating the parameters in the General Model (GM). Given the estimated PM parameters, we can favor certain regions of values of the GM parameters under the Bayesian framework. The outputs of GM parameter estimation procedure are the estimated GM parameters and the estimated signals of the individual tones in the mixtures.

1.4 Thesis organization and contributions

The rest of the thesis is organized as follows. Chapter 2 gives a literature review of music performance research and music source separation. The signal models, including the traditional General Model and our proposed Piano Model, will

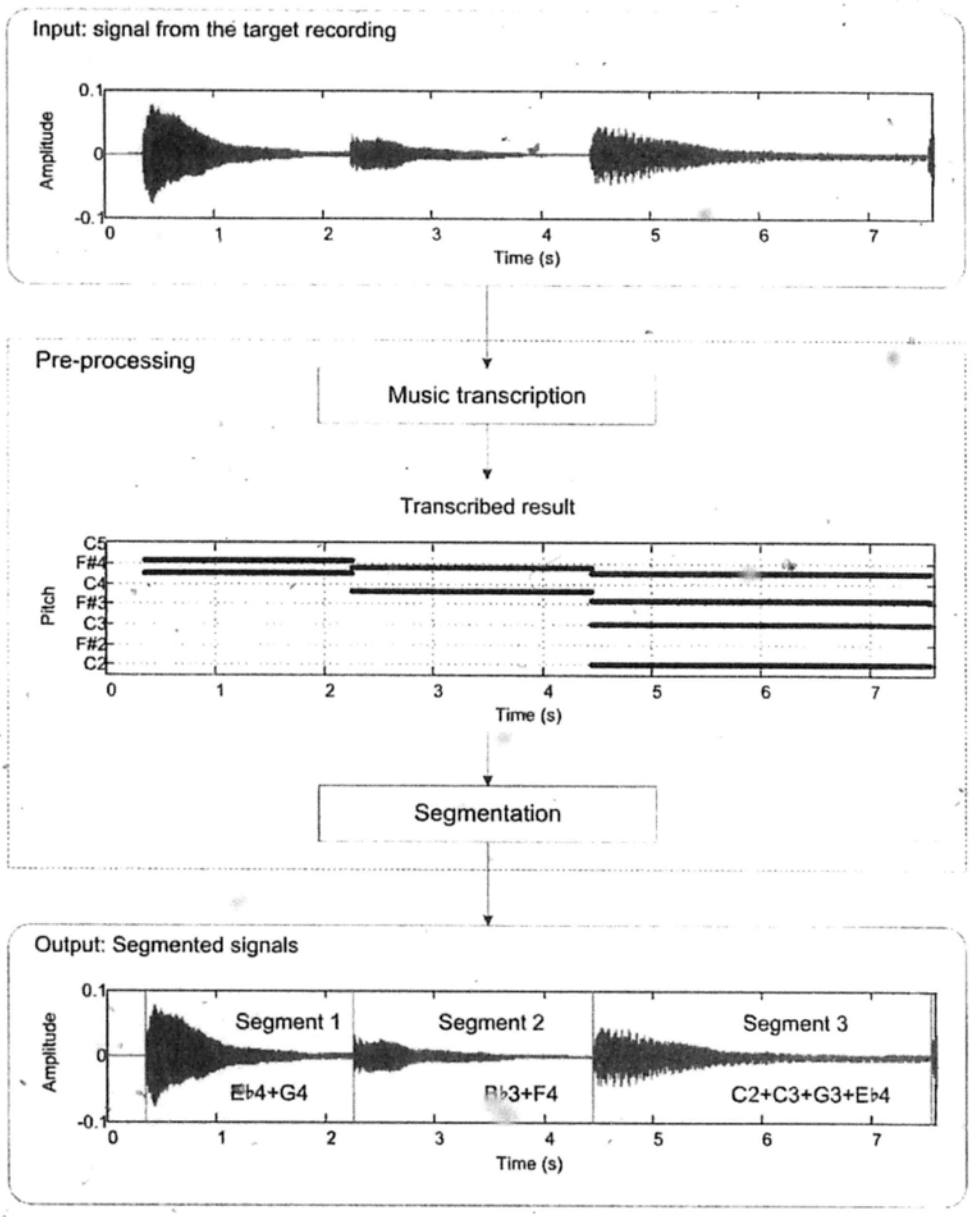

Figure 1.3: Example of illustrating the pre-processing steps (music transcription and segmentation).

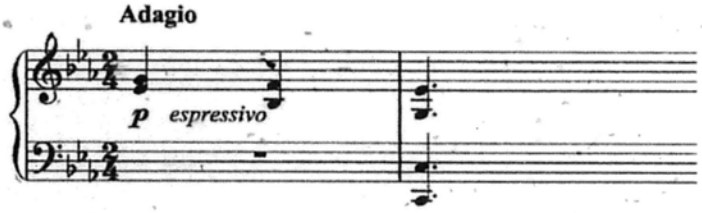

Figure 1.4: The musical score of the opening of Beethoven's Piano Sonata in E♭, Op. 81a "Les Adieux".

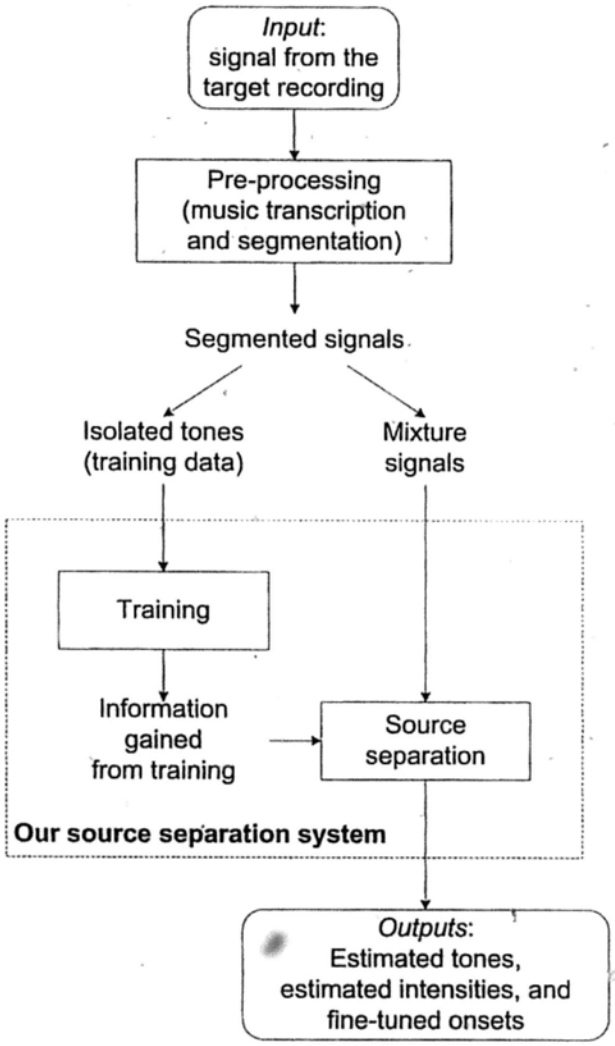

Figure 1.5: The whole source separation process from a target recording to the estimated individual tones.

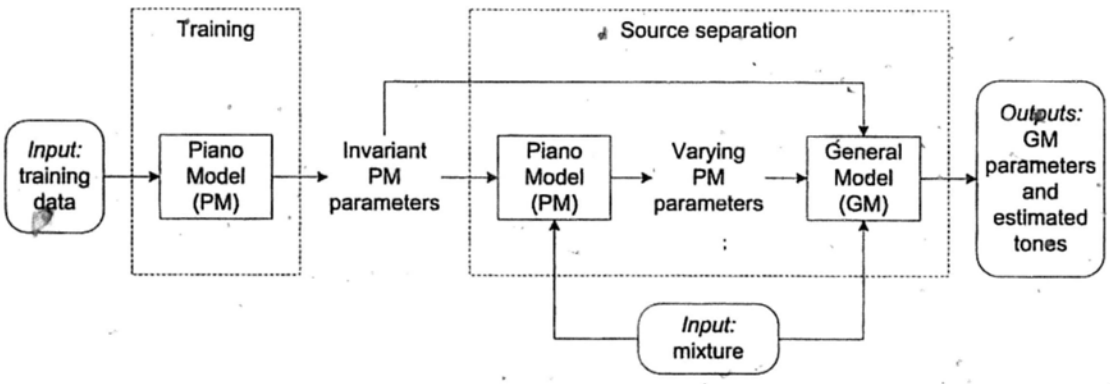

Figure 1.6: The main steps of our source separation process. See the text for the explanation.

be presented in Chapter 3. The Bayesian framework integrating both signal models will be explained in Chapter 4. Then, parameter estimations in the training stage and the source separation will be examined in Chapters 5 and 6 respectively. We will also present the experimental results of our source separation process on real piano signals and compare our system to another system. A conclusion will be given in Chapter 8.

The main contributions of this dissertation are

- Formulating a mathematical model for representing different sounds of the piano in a parametric form using a set of parameters. The complexity of our proposed Piano Model will be optimized for tractable parameter estimation while not sacrificing the quality of subsequent source separation too much. The Piano Model captures the common properties of multiple instances of the same pitch, enabling us to resolve overlapping partials.

- Developing an efficient method for estimating the proposed Piano Model which is both high dimensional and nonlinear.

- Establishing a hierarchical Bayesian framework for the source separation problem for the piano. In order to achieve good separation quality, a tone will be represented by the traditional General Model which is a frame-wise sinusoidal model. The problem's solution will be based on constraining the parameters in the frame-wise model by the estimated parameters in the Piano Model via a Bayesian framework.

- Recovering the amplitude, phase and frequency of each partial in a mixture so that the overlapping partials can be resolved.

- Designing a spectral pick-peaking method for piano tones for estimating the frequencies of partials in a piano tone.

Chapter 2

Literature review

The goal of our research is to address the monaural source separation of piano music signals and to find the nuance of piano music performance. In this chapter, some related work is reviewed. In Section 2.1, some empirical studies on music performances are presented. The work on the monaural music source separation will be presented in Section 2.2.

2.1 Research on music performance

In recent years, there have been many empirical studies on music performances in the aspects of measurement of performance, performing styles, models of performance, and performance planning and practice [30, 22, 55, 51]. One limitation of many published studies in analyzing nuances in music performances is that most of them focus on timing including analysis of note inter-onset interval, rhythm, tempo and rubato [58]. Research on other sound parameters including attack, pitch, loudness, and timbre is relatively rare. Many of them obtain data from specific instruments through specialized sensors for detecting how the instruments were controlled and manipulated by the artists (such as sensors in digital pianos or computer-monitored pianos) [33]. There are other studies which focus on analysis of commercially available music recordings instead of audio data recorded in experimental conditions, but these body of

work usually only focuses on analysis of the overall intensity of the mixture signals [59]. These limitations motivate our research on source separation of piano music signals.

2.2 Music source separation

Source separation of music signals commonly refer to the challenging problem of separating the mixed signals from various musical instruments or extracting the individual tones from a mixture of musical tones. The two major challenges are that monaural source separation is an underdetermined problem and overlapping partials often appear in piano music. Different approaches to tackle these two challenges are reviewed below.

2.2.1 Monaural source separation - solving the underdetermined problem

Independent Component Analysis (ICA) [36, 7] is a widely-used technique to separate mixture signals into source signals. It can be used to address the blind source separation problem. The term "blind" is used in the sense that very little information is known about the source signals [66]. In our research, we have already known that the source signals are piano tones and we will make use of this important information for the source separation process. Therefore, we refer to our problem as "source separation" instead of "blind source separation". ICA aims to extract the source signals from the mixture signals based on some general criteria such as by maximizing the statistical independence between the estimated source signals. The standard ICA demands that the number of observed variable, i.e., the number of microphones or channels, is equal to or larger than the number of source signals. In our context, if ICA is applied to the time-domain mixture signals, ICA requires that the number of channels is

equal to or larger than the number of tones in a mixture. As music signals are often available in one channel (monaural) or two channels (stereo), and it is common to have more than two tones sounding at the same time, the requirement of standard ICA cannot be fulfilled. Hence, the standard ICA cannot directly be applied to the monaural source separation problem.

To deal with the underdetermined problem in the case of monaural music source separation, one of the major approaches is to assume that the source signals exhibit some statistical properties. In [18], Casey and Westner extend ICA into Independent Subspace Analysis (ISA) which works with spectrogram instead of time-domain signals and separates mixtures of monaural audio signals. ISA assumes the statistical independence of source signals when they are represented by a spectrogram. Another technique is sparse coding [70, 1, 12] which represents a mixture signal by a weighted sum of bases from a larger set. It is assumed that most of the weights are zero. This means that only a few bases are active at a time. The third technique is Nonnegative Matrix Factorization (NMF) [43] which factorizes a mixture signal representation into the product of the basis function matrix and the gain matrix. Each element in both matrices is assumed to be non-negative. NMF has been applied in music source separation and music transcription [74, 9, 19, 65, 72, 73, 57]

Another major approach of the underdetermined problem is to use a parametric model to represent the source signals. Formulating a parametric model for representing the sounds of a musical instrument requires a good understanding of how musical sounds can be analyzed and then, artificially synthesized [6, 60]. Existing methods for synthesizing musical sounds, especially the piano sounds, include additive synthesis [38, 61], FM synthesis [21], group synthesis [44], physical modeling [3, 37], and sinusoidal modeling [64]. The most common model in representing music signals for the source separation problem is sinusoidal modeling which has briefly mentioned in Section 1.3. In sinusoidal modeling, a musical sound is represented by a sum of time-varying sinusoidals.

The mixture is usually segmented into short-time segments called *frames*. The parameters of a sinusoidal, including frequency, amplitude and phase, are assumed to be stationary in the frame. The goal of source separation is to estimate the sinusoidal parameters of each source signal. There are numerous work using sinusoidal modeling to address the problems of monaural music source separation and music transcription [71, 24, 69, 14, 23, 28, 46, 27, 5].

2.2.2 Resolving overlapping partials

In Section 1.3, we discuss that overlapping partials often appear in piano music and resolving overlapping partials is essential for maintaining high separation quality. The techniques of ISA, sparse coding and NMF work with the magnitude spectrum of the mixture signals, so phase information is generally ignored but the phase information is essential to resolve the overlapping partials. Moreover, these techniques often assume that the sum of magnitude spectrum of sources is equal to the magnitude spectrum of the mixture. This assumption is only valid for overlapping partials when they are completely in phase. Results in [76] shows that this assumption deteriorates the performance of amplitude estimation. Hence, ISA, sparse coding and NMF cannot give robust performance of resolving overlapping partials.

Sinusoidal modeling gives the possibility to retain the phase information. A comparison of ISA, sparse coding, NMF and sinusoidal modeling on recovery of amplitude and phase is shown in Table 2.1. As mentioned in Section 1.3, assumptions of general musical sound properties can be made to resolve overlapping partials. For example, spectral envelope is assumed to be smooth in [71, 28]. The information of neighboring non-overlapping partials can be utilized to estimate the parameters of an overlapping partial. This can be done by interpolating the non-overlapping partials to estimate the overlapping partial. In [71], Virtanen uses a two-stage method in his monaural source separation

Method	Recovery of amplitude	Recovering of phase
ISA, sparse coding and NMF	Limited	No
Sinusoidal modeling	Possible	Possible

Table 2.1: Comparison of different monaural source separation methods on recovery of amplitude and phase.

system. In the first stage, the number of tones and their rough frequencies are estimated by the multiple frequency estimation method in [42]. In the second stage, the parameters in the sinusoidal model are estimated and the source separation is performed under the least squares criterion. The overlapping partials are resolved by considering the smoothness of spectral envelope. It is done by modeling the partial amplitudes by a weighted sum of fixed basis functions that does not allow large changes between the amplitudes of adjacent partials. However, for a piano tone, spectral envelope may not be smooth.

Another assumption of musical sound properties is that amplitude of partials from the same source is similar. It is known as common amplitude modulation (CAM) [13, 46]. The CAM-based method in [46], which is based on a least squares estimation framework, is performed well in separating the mixtures from different musical instruments. To recover the phase of an overlapping partial, the system in [46] uses the information that the change in phase of a partial is related to the pitch of the tone. However, for recovering amplitudes, the performance of CAM-based method may be significant affected when there are many overlapping partials. For example, the partials of the upper tone in an octave are totally overlapping with those of the lower tone. It is difficult to obtain any non-overlapping partial in the upper tone to estimate the overlapping partials by the property of CAM. Moreover, each individual tone in a mixture of piano signals exhibits much less contrasting timbre comparing to the tones from different musical instruments. Hence, a CAM-based method is unlikely to be able to resolve the overlapping partials of piano tones accurately.

In [69], the continuity of partial amplitude is assumed and it is modeled under a Bayesian setting. A tone in [69] is represented as a sum of exactly harmonic sinusoidal partials. The noise of the tone is modeled by a weighted Gaussian residual prior to minimize the perpetual difference between the original signal and the synthesized signal. The weighting of the frequency bands follows the distortion measures proposed in [70]. The inference is done by using a Maximum A Posteriori (MAP) criterion with an approximate staged inference procedure. The duration and continuity priors are incorporated to avoid perceptually annoying discontinuities. Note that the continuity of partial amplitude is imposed by the priors. The prior of partial amplitude is constructed by a linear scaling of a fixed spectral envelope. The prior of the scaling factor depends on the value of the factor in the previous frame so that the prior ensures a smooth change of the factor. However, the fixed spectral envelope, which is treated as a point estimate of the hyperparameters of the model, is learned from a music tone database of various instruments and it is not adaptive to for the mixture signals. Moreover, the uniform prior is assigned to the phase of a partial. For the case of overlapping partials, the phases cannot be estimated accurately.

Another work using the Bayesian approach for monaural music source separation is proposed in [24]. In [24], a tone is represented as a sum of quasi-harmonic sinusoidals and Gaussian noise. The time-varying amplitudes of sinusoidals are achieved by successive frame-by-frame windowing. Inharmonicity is also modeled to allow the partials which are not exact multiples of the fundamental frequency. The frequencies within a tone are assumed to be fixed and the onsets of the tones are already known. The parameters of the model are estimated by a fully Bayesian inference via Markov chain Monte Carlo (MCMC) method. Some of the parameters controlling the distributions of the parameters in the model, i.e., the hyperparameters, are estimated by using the incoming musical signals; while some of them are estimated by inspecting

a database of musical tones. A partial in an analysis frame is written in the form of $\alpha\cos(2\pi ft) + \beta\sin(2\pi ft)$ where f is the frequency, α and β contain the information of the amplitude and the phase of a partial. The priors of α and β are assigned to be the zero-mean Gaussian which do not supply any information to resolve the overlapping partials. Hence, overlapping partials may not be able to resolve. As we will show in Chapter 4, an appropriate prior is essential for resolving overlapping partials.

Chapter 3

Signal model representations

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 below:

$$\mathsf{y}(t) = \sum_{k=1}^{K} \mathsf{x}_k(t) \tag{3.1}$$

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 second. This model is called *instantaneous linear mixing* in the literature of general source separation [52, 14]. The notation of this dissertation can be found in Appendix A.

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 called *partials*. We use two sum-of-sinusoidal models to represent $\mathsf{x}_k(t)$ - a traditional General Model (GM) and our proposed Piano Model (PM). GM is more flexible to represent an isolated tone for better quality; while PM can capture the common properties across the reappearance of pitches that helps to separate the mixtures. Before discussing these two models, properties of piano tones will be introduced first.

3.1 Properties of piano tones

When a piano key is pressed, the hammer hits the strings of the corresponding key. Then the strings vibrate and the energy transfers from strings to the soundboard, and the sound radiates from the soundboard. The resulting sound can be analyzed by using the spectrogram which shows how the spectrum changes along the time. The spectrogram of a C4 piano tone is depicted in Figure 3.1. The spectrogram shows that the piano tone consists of the frequency components and the noise. The frequency components, also called partials, correspond to the resonance frequencies of the strings. The relation among mixtures, tones and partials can be found in Figure 1.1. The frequency values of the partials in piano tones are stable against time. In piano sound, the partials of a tone are usually not exactly harmonic. If the partials are exactly harmonic, the frequencies of the partials are exact multiples of the fundamental frequency, and the frequency ratios between the partials are $1 : 2 : 3 : 4 : 5$ and so on. For piano tones, the frequency ratios are slightly stretched. The frequency ratios of the first five partials in Figure 3.1 are $1.0000 : 2.0000 : 3.0033 : 4.0075 : 5.0163$. This phenomenon is called *inharmonicity* and it is caused by the bending stiffness of the strings [2]. Inharmonicity is perceptually significant for the sound quality of pianos [2].

Figure 3.2 shows the extracted partials of the C4 piano tone in Figure 3.1. 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 [54]. 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 3.2, the fundamental frequency has the highest peak. The third partial is stronger than the second and the fifth is stronger than the fourth. Figure 3.2 (d) shows the unwrapped

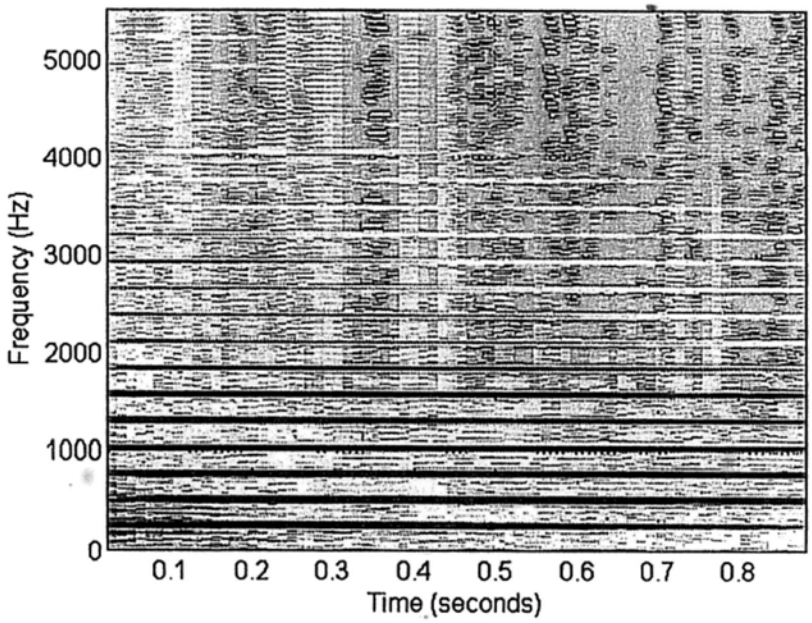

Figure 3.1: Spectrogram of a C4 piano tone played moderately loud.

phase against time. The unwrapped phase is linear and the partials can be considered as linear-phase signals.

3.2 Traditional General Model

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

$$y_r[l] = w[l]\mathrm{y}[(r-1)D + l] \tag{3.2}$$

where $y_r[l]$ is the rth frame at the local time index l where $l = 0, 1, \ldots L-1$ and L is the window length, $w[l]$ is the window function, and D is the hop size. The hamming window is used in this research. The typeface y denotes the entire mixture while the typeface y refers to the windowed segment of a

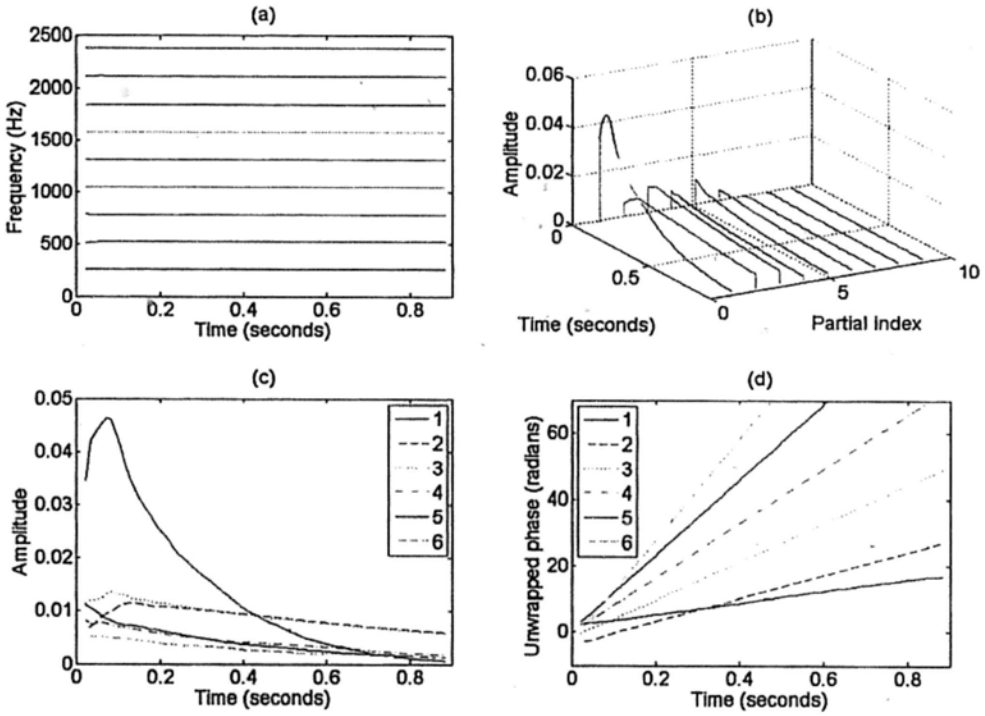

Figure 3.2: Extracted partials of the C4 piano tone in Figure 3.1. The partials are extracted by using the method in Section 5.3 with time-varying frequencies. (a) The frequencies of the first nine partials against time. (b) The amplitude of the first nine partials against time. The partial index one 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.

frame.

The signal of a frame can be represented by sinusoidal modeling which uses a sum of sinusoidals to represent the signal. Sinusoidal modeling is a well-established technique to model audio signals including speech signals [49] and music signals [64]. A frame-wise sinusoidal model of a piano tone $\widehat{\mathrm{x}}_{k,r}$ of the kth tone at the rth frame can be written as below:

$$\widehat{x}_{k,r}[l] = \sum_{m=1}^{M_{k,r}} w[l] \left(\alpha_{k,m,r} \cos(2\pi f_{k,m,r} t_l) + \beta_{k,m,r} \sin(2\pi f_{k,m,r} t_l)\right) \tag{3.3}$$

where $M_{k,r}$ 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, $f_{k,m,r}$ is the frequency, 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. In sinusoidal modeling, the parameters $M_{k,r}$, $\alpha_{k,m,r}$, $\beta_{k,m,r}$ and $f_{k,m,r}$ are fixed within a frame but they can be different across frames. This models the time-varying properties of music signals. To reconstruct or resynthesize the entire signal from the sinusoidal model, the parameter values between two frames can be estimated by some interpolation methods such as [49]. Another reconstruction approach is to overlap and add all estimated signals in the frames [77]. The overlap-and-add method will be used in this research for its simplicity.

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. Then the model in (3.3) can be rewritten as

$$\widehat{x}_{k,r}[l] = \sum_{m=1}^{M_k} w[l] \left(\alpha_{k,m,r} \cos(2\pi f_{k,m} t_l) + \beta_{k,m,r} \sin(2\pi f_{k,m} t_l)\right) \tag{3.4}$$

where M_k is the number of partials of the kth tone and $f_{k,m}$ is the frequency of the mth partial in the kth tone. We refer this model as the *traditional General Model* (GM).

Then the estimated mixture $\widehat{y}_r[l]$ at the rth frame is the sum of each estimated tone $\widehat{x}_{k,r}[l]$ as below

$$\widehat{y}_r[l] = \sum_{k=1}^{K} \widehat{x}_{k,r}[l] \tag{3.5}$$

where K is the number of tones in the mixture. The observed mixture is the sum of the estimated mixture and the noise term:

$$\begin{aligned} y_r[l] &= \widehat{y}_r[l] + v_r[l] && (3.6) \\ &= \sum_{k=1}^{K} \widehat{x}_{k,r}[l] + v_r[l] && (3.7) \end{aligned}$$

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

To estimate the parameters in each frame, it is convenient to rewrite the model in (3.4) into the matrix form. Let $\mathbf{H}_k$ be the frequency matrix of the kth tone and it is an L-by-$2M_k$ matrix in the form of

$$H_k[l,u] = \begin{cases} w[l]\cos(2\pi f_{k,u} t_l) & \text{if } 1 \le u \le M_k, \\ w[l]\sin(2\pi f_{k,u-M_k} t_l) & \text{if } M_k + 1 \le u \le 2M_k \end{cases} \tag{3.8}$$

so the matrix $\mathbf{H}_k$ contains two blocks

$$\mathbf{H}_k = \begin{bmatrix} \mathbf{H}_k^{\cos} & \mathbf{H}_k^{\sin} \end{bmatrix} \tag{3.9}$$

where

$$\mathbf{H}_k^{\cos} = \begin{bmatrix} w[0]\cos(2\pi f_{k,1} t_0) & \cdots & w[0]\cos(2\pi f_{k,M_k} t_0) \\ w[1]\cos(2\pi f_{k,1} t_1) & \cdots & w[1]\cos(2\pi f_{k,M_k} t_1) \\ \vdots & & \vdots \\ w[L-1]\cos(2\pi f_{k,1} t_{L-1}) & \cdots & w[L-1]\cos(2\pi f_{k,M_k} t_{L-1}) \end{bmatrix} \tag{3.10}$$

and

$$\mathbf{H}_k^{\sin} = \begin{bmatrix} w[0]\sin(2\pi f_{k,1}t_0) & \cdots & w[0]\sin(2\pi f_{k,M_k}t_0) \\ w[1]\sin(2\pi f_{k,1}t_1) & \cdots & w[1]\sin(2\pi f_{k,M_k}t_1) \\ \vdots & & \vdots \\ w[L-1]\sin(2\pi f_{k,1}t_{L-1}) & \cdots & w[L-1]\sin(2\pi f_{k,M_k}t_{L-1}) \end{bmatrix}. \tag{3.11}$$

Then

$$H_k^{\cos}[l,m] = w[l]\cos(2\pi f_{k,m}t_l) \tag{3.12}$$

$$H_k^{\sin}[l,m] = w[l]\sin(2\pi f_{k,m}t_l). \tag{3.13}$$

The amplitudes of the cosine and the sine terms of the kth tone at the rth frame can be expressed as a $2M_k$-dimensional vector $\mathbf{g}_{k,r}$ as below

$$g_{k,r}[u] = \begin{cases} \alpha_{k,u,r} & \text{if } 1 \le u \le M_k, \\ \beta_{k,u-M_k,r} & \text{if } M_k+1 \le u \le 2M_k \end{cases} \tag{3.14}$$

which gives

$$\mathbf{g}_{k,r} = [\alpha_{k,1,r}\ \alpha_{k,2,r}\ \cdots\ \alpha_{k,M_k,r}\ \beta_{k,1,r}\ \beta_{k,2,r}\ \cdots\ \beta_{k,M_k,r}]^{\mathrm{T}}. \tag{3.15}$$

The estimated tone $\widehat{\mathbf{x}}_{k,r}$ can be written as

$$\widehat{\mathbf{x}}_{k,r} = \mathbf{H}_k\mathbf{g}_{k,r}. \tag{3.16}$$

To illustrate the matrix form of $\widehat{\mathbf{x}}_{k,r}$ in (3.16), substituting (3.12) and (3.13)

into (3.4) gives

$$\begin{aligned}\widehat{x}_{k,r}[l] &= \sum_{m=1}^{M_k} w[l]\left(\alpha_{k,m,r}\cos(2\pi f_{k,m}t_l) + \beta_{k,m,r}\sin(2\pi f_{k,m}t_l)\right) \\ &= \sum_{m=1}^{M_k}\left(\alpha_{k,m,r}H_k^{\cos}[l,m] + \beta_{k,m,r}H_k^{\sin}[l,m]\right). \end{aligned} \tag{3.17}$$

Figure 3.3 depicts the matrix form of $\widehat{\mathbf{x}}_{k,r}$ with (3.16) and (3.17).

For the mixture, the frequency matrices from each tone are concatenated into the matrix $\mathbf{H}$:

$$\mathbf{H} = \left[\begin{array}{cccc}\mathbf{H}_1 & \mathbf{H}_2 & \cdots & \mathbf{H}_K\end{array}\right] \tag{3.18}$$

where $\mathbf{H}$ is an L-by-$2M$ matrix, M is the total number of partials and $M = \sum_{k=1}^{K} M_k$. The amplitude vectors from of each tone can also be concatenated into a vector $\mathbf{g}_r$:

$$\mathbf{g}_r = \left[\begin{array}{c}\mathbf{g}_{1,r}\\ \mathbf{g}_{2,r}\\ \vdots \\ \mathbf{g}_{K,r}\end{array}\right] \tag{3.19}$$

where $\mathbf{g}_r$ is a $2M$-by-1 vector. The estimated mixture at rth frame can be expressed as:

$$\widehat{\mathbf{y}}_r = \mathbf{H}\mathbf{g}_r \tag{3.20}$$

and the estimated mixture is related to the observed mixture as below:

$$\begin{aligned}\mathbf{y}_r &= \widehat{\mathbf{y}}_r + \mathbf{v}_r \\ &= \mathbf{H}\mathbf{g}_r + \mathbf{v}_r \end{aligned} \tag{3.21}$$

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

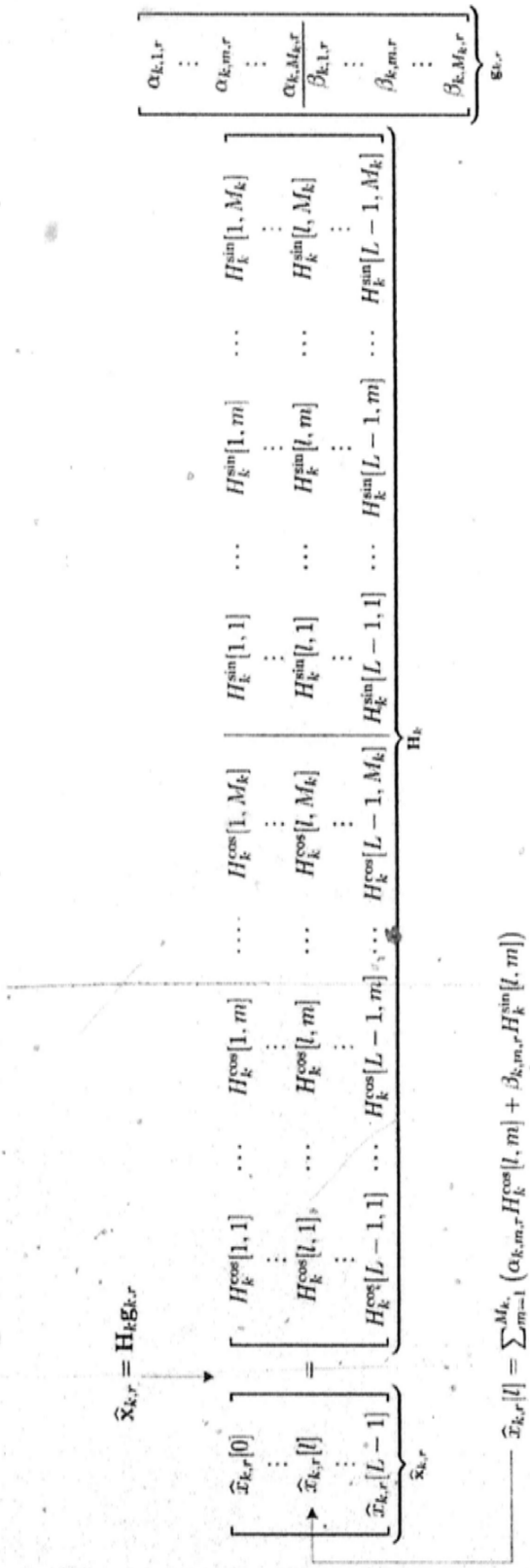

Figure 3.3: The matrix form of $\widehat{\mathbf{x}}_{k,r}$.

We can further concatenate the parameter vectors from all frames into matrices. The observed mixture signal can be expressed in the form

$$\mathbf{Y} = [\mathbf{y}_1 \ \mathbf{y}_2 \ \cdots \ \mathbf{y}_R] . \tag{3.22}$$

Similarly, the estimated mixture can be written as

$$\widehat{\mathbf{Y}} = \mathbf{HG} \tag{3.23}$$

where

$$\widehat{\mathbf{Y}} = [\widehat{\mathbf{y}}_1 \ \cdots \widehat{\mathbf{y}}_R] \tag{3.24}$$

$$\mathbf{G} = [\mathbf{g}_1 \ \cdots \ \mathbf{g}_R] \tag{3.25}$$

and R is the number of frames. The goal of our source separation is to estimate both the frequency matrix $\mathbf{H}$ and the amplitude matrix $\mathbf{G}$ so that each individual tone can be reconstructed. Determining the number of partials M_k will be discussed in Section 3.4.

GM of a tone can also be written in the matrix form:

$$\widehat{\mathbf{X}}_k = \mathbf{H}_k \mathbf{G}_k \tag{3.26}$$

where

$$\mathbf{G}_k = [\mathbf{g}_{k,1} \ \cdots \ \mathbf{g}_{k,R}] . \tag{3.27}$$

The frequency vector $\mathbf{f}_k$ and the amplitude matrix $\mathbf{G}_k$ can be grouped into a parameter set $\boldsymbol{\theta}_k = \{\mathbf{f}_k, \mathbf{G}_k\}$. For notational convenience, all $\mathbf{f}_k$ are concatenated into the column vector $\mathbf{f}$ where

$$\mathbf{f} = [\mathbf{f}_1 \ \cdots \ \mathbf{f}_K]^{\mathrm{T}} . \tag{3.28}$$

All parameters of kth tone can be grouped into $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\} = \{\mathbf{f}, \mathbf{G}\}$.

3.3 Proposed Piano Model

The traditional General Model (GM) gives high quality of resynthesis but the parameters of the overlapping partials cannot be accurately estimated. Here, we propose the Piano Model (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 analysis frame. For each partial, we aim to model the surface 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 [53]. The envelope surfaces of the first four partials are plotted in Figure 3.4. 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 each partial has its own rate of rising and decay but 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 [2, 29].

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

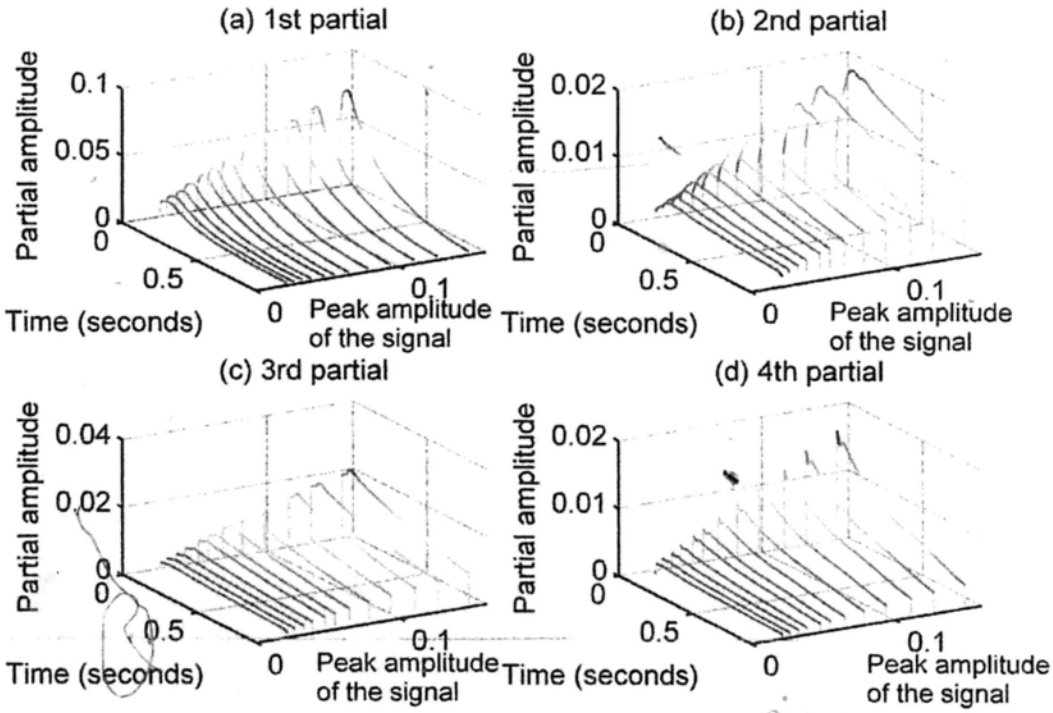

Figure 3.4: Envelope surface against peak amplitude of the time-domain signal and time for the first four partials. (a) The first partial (fundamental frequency). (b) The second partial. (c) The third partial. (d) The fourth partial.

learning process will be fully discussed in Chapter 5. PM is expressed as below:

$$\widehat{\mathrm{x}}_k(t_n) = \sum_{m=1}^{M_k} a_{k,m}(t_n) \cdot \cos(2\pi f_{k,m} t_n + \phi_{k,m}) \tag{3.29}$$

where $f_{k,m}$ is the frequency, $\phi_{k,m}$ is the phase, 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:

$$\begin{aligned} a_{k,m}(t_n) &= a(t_n; c_k, \varphi_{k,m}) && (3.30) \\ &= b_{k,m} (c_k)^{d_{k,m}} \zeta_{k,m} (\exp\{-\lambda_{k,m} t_n\} - \exp\{-\gamma_{k,m} t_n\}) && (3.31) \end{aligned}$$

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 $\varphi_{k,m} = \{b_{k,m}, d_{k,m}, \lambda_{k,m}, \gamma_{k,m}\}$. The intensity factor c_k is assigned to be the peak amplitude of the observed time-domain signal of the tone. The rising and decay of the partial magnitude is modeled by the bi-exponential function $\zeta_{k,m} (\exp\{-\lambda_{k,m} t_n\} - \exp\{-\gamma_{k,m} t_n\})$. The term $\exp\{-\lambda_{k,m} t_n\}$, commonly used in the synthesis of musical sounds, models the slow decay. The term $-\exp\{-\gamma_{k,m} t_n\}$ models the rapid rising. 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 to one, and $\zeta_{k,m}$ depends on $\lambda_{k,m}$ and $\gamma_{k,m}$:

$$\zeta_{k,m} = \left[\left(\frac{\lambda_{k,m}}{\gamma_{k,m}} \right)^{\frac{\lambda_{k,m}}{\gamma_{k,m} - \lambda_{k,m}}} - \left(\frac{\lambda_{k,m}}{\gamma_{k,m}} \right)^{\frac{\gamma_{k,m}}{\gamma_{k,m} - \lambda_{k,m}}} \right]^{-1} \tag{3.32}$$

where the proof the normalization coefficient is shown in Appendix B.

Substituting (3.30) into (3.29), we write the estimated signal of a tone in

the form

$$\widehat{\mathrm{x}}_k(t_n) = \sum_{m=1}^{M_k} a(t_n; c_k, \varphi_{k,m}) \cdot \cos(2\pi f_{k,m} t_n + \phi_{k,m}). \tag{3.33}$$

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 $\widehat{\mathrm{y}}(t_n)$:

$$\widehat{\mathrm{y}}(t_n) = \sum_{k=1}^{M_k} \widehat{\mathrm{x}}_k(t_n - \tau_k) \tag{3.34}$$

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

$$\begin{aligned} \mathrm{y}(t_n) &= \widehat{\mathrm{y}}(t_n) + \epsilon(t_n) \\ &= \sum_{k=1}^{M_k} \widehat{\mathrm{x}}_k(t_n - \tau_k) + \epsilon(t_n) \end{aligned} \tag{3.35}$$

where $\epsilon(t_n)$ is the noise term. This noise is modeled as the zero-mean Gaussian noise with the variance σ_ϵ^2.

All parameters of PM for the kth tone can be grouped into a parameter set $\boldsymbol{\psi}_k$ so

$$\boldsymbol{\psi}_k = \{\varphi_{k,m}, f_{k,m}, \phi_{k,m}, c_k, \tau_k\} \tag{3.36}$$

which can be divided into two sets: the invariant PM parameters $\boldsymbol{\psi}_{k,\mathrm{I}}$ and the varying PM parameters $\boldsymbol{\psi}_{k,\mathrm{V}}$. 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. The invariant PM parameters $\boldsymbol{\psi}_{k,\mathrm{I}}$ contain the envelope parameters $\varphi_{k,m}$, the frequency $f_{k,m}$ and the phase $\phi_{k,m}$, and the invariant PM parameters $\boldsymbol{\psi}_{k,\mathrm{I}}$ are defined by

$$\boldsymbol{\psi}_{k,\mathrm{I}} = \{\varphi_{k,m}, f_{k,m}, \phi_{k,m}\} \tag{3.37}$$

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

Table 3.1: Invariant PM parameters and varying PM parameters.

for $m = 1, 2, \ldots, M_k$.

The varying PM parameters $\psi_{k,\mathrm{V}}$ include the intensity c_k and the time shift τ_k and they are defined by

$$\psi_{k,\mathrm{V}} = \{c_k, \tau_k\}. \tag{3.38}$$

These two sets of the parameters give

$$\psi_k = \{\psi_{k,\mathrm{I}}, \psi_{k,\mathrm{V}}\} \tag{3.39}$$

and all ψ_k can be grouped into

$$\Psi = \{\psi_1, \ldots, \psi_K\}. \tag{3.40}$$

The role of the invariant PM parameters $\psi_{k,\mathrm{I}}$ and the varying PM parameters $\psi_{k,\mathrm{V}}$ is shown in Table 3.1. 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. It is expected that the overlapping partials can be resolved. This will be verified by the experiments presented in Chapter 7. The details of training and source separation will be explained in the next three chapters.

Note that the varying PM parameters $\psi_{k,\mathrm{V}} = \{c_k, \tau_k\}$ including the intensity and the time shift are significant for characterization of musical nuance.

As mentioned in the opening of this section, 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.

3.4 Estimation of the number of partials

In both GM and PM, we have assumed that the number of partials M_k of each tone is known. In this section, we will show how M_k can be found. The values of M_k are different from pitches. Lower pitch usually has more partials than the higher pitch. In some research such as [23, 24, 27], M_k is dynamically estimated. However, this estimation is very computationally intensive. As we have already known that the mixtures are piano signals, we estimate the average number of partials required for each pitch from different pianos. The piano tone database in [39] is used for estimating M_k. The database contains piano tones from 7 different pianos. 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 Chapter 7. Once M_k is determined, it will be fixed for all experiments in Chapter 7.

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

The steps for picking the spectral speaks are described below:

1. Perform Discrete Fourier Transform (DFT) of a tone.

2. Find the first partial (fundamental frequency) f_1:

   (a) Set f_1^{mid} to the equal-tempered fundamental frequency of the pitch. For example, the pitch A4 is with $f_1^{\text{mid}} = 440$ Hz.

   (b) Set f_1 to the frequency corresponding to the peak of the magnitude spectrum in the frequency range $[2^{-1/48} f_1^{\text{mid}}, 2^{1/48} f_1^{\text{mid}}]$.

3. Set the inharmonicity coefficient $B^{(0)} = 0$ which is defined in (3.41) and (3.42).

4. Find f_m for $m \geq 2$ where f_m is the frequency of the mth partial:

   (a) Find f_m^{mid} by

   $$f_m^{\text{mid}} = m f_1 \sqrt{\frac{1 + m^2 B^{(q)}}{1 + B^{(q)}}} \tag{3.41}$$

   which is the general formula to model the inharmonicity effect for pianos [29, p. 363]. A typical value for the inharmonicity coefficient B is 0.0004 in the middle range of piano keys.

   (b) Set f_m to the frequency corresponding to the peak of the magnitude spectrum in the frequency range $[2^{-1/48} f_m^{\text{mid}}, 2^{1/48} f_m^{\text{mid}}]$. If $2^{1/48} f_m^{\text{mid}} > f_s/2$, set the upper bound to $f_s/2$.

5. Update B

   $$B_u = \frac{(f_u/u f_1)^2 - 1}{u^2 - (f_u/u f_1)^2} \tag{3.42}$$

   Set $B^{(q+1)}$ to the median of all B_u for $1 \leq u \leq m$.

6. Repeat the steps 4-5 until $f_m^{\text{mid}} > f_s/2$ so the frequencies of all partials can be estimated.

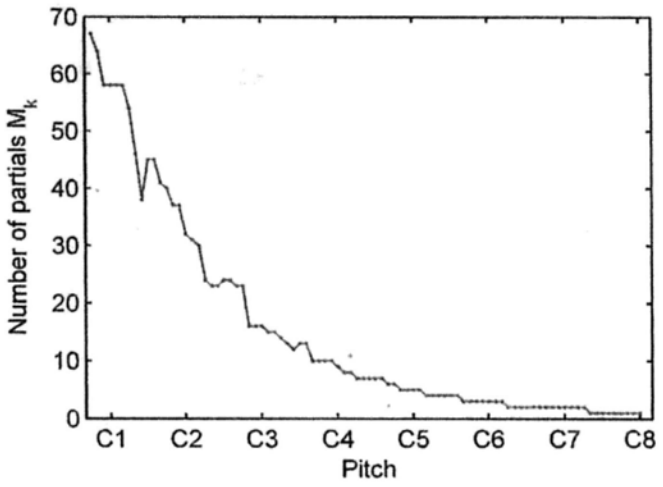

Figure 3.5: The number of partials M_k of each pitch.

After picking all the partials, we choose the number of the partials that contains 99.5% of the power of all partials picked on average. The result is shown in Figure 3.5. Then the numbers of parameters in GM and PM can be calculated from the number of partials. For the estimated mixture $\widehat{\mathbf{Y}}$ in (3.23), the number of parameters in GM is $M(2R+1)$ where M is the total number of partials and $M = \sum_{k=1}^{K} M_k$, and R is the number of frames. For the estimated mixture $\widehat{\mathbf{y}}$ in (3.35), the number of parameters in PM is $6M+1$. For instance, if a mixture contains the tones C4 and G4, the total number of partials in the estimated mixture is $9+7=16$ according to Figure 3.5. If the duration of the mixture is 0.5 second and the window length is 11.61 ms ($L = 128$) with 50% overlapping window, the number of frames R is 87. Then the number of parameters in GM is 2800. For PM, the number of parameters is 97. Although, the number of parameters for GM is much greater than that for PM, parameter estimation in GM is more computationally efficient than that in PM because PM is highly nonlinear. The experiment of investigating the computation time for both GM and PM will be discussed in Section 7.3.4.

Chapter 4

Bayesian framework for source separation

In the previous chapter, two signal model representations are introduced but how do these two models link together to separate a mixture into its individual tones and to resolve overlapping partials? This chapter will explain how the Bayesian framework integrates these two models and incorporates the training data to resolve overlapping partials. The two models have their merits and shortcomings as shown in Figure 4.1. The traditional General Model (GM) is more flexible and has better modeling quality, comparing to our proposed Piano Model (PM). If the mixture does not contain overlapping partials, GM gives higher separation quality. If the mixture contains overlapping partials, unless more information is provided, GM cannot resolve the overlapping partials and it fails to separate the mixture. On the other hand, PM is able to resolve the overlapping partials and output the estimated tones, provided that the values of the invariant PM parameters estimated from the training data given. Estimation of the invariant PM parameters will be discussed shortly.

A simple solution to the source separation problem is that if there is no overlapping partial, GM is used; otherwise, PM is used instead. However, GM has better modeling quality and it is highly desirable that GM can be used even if there are overlapping partials. In other words, the ultimate goal of our

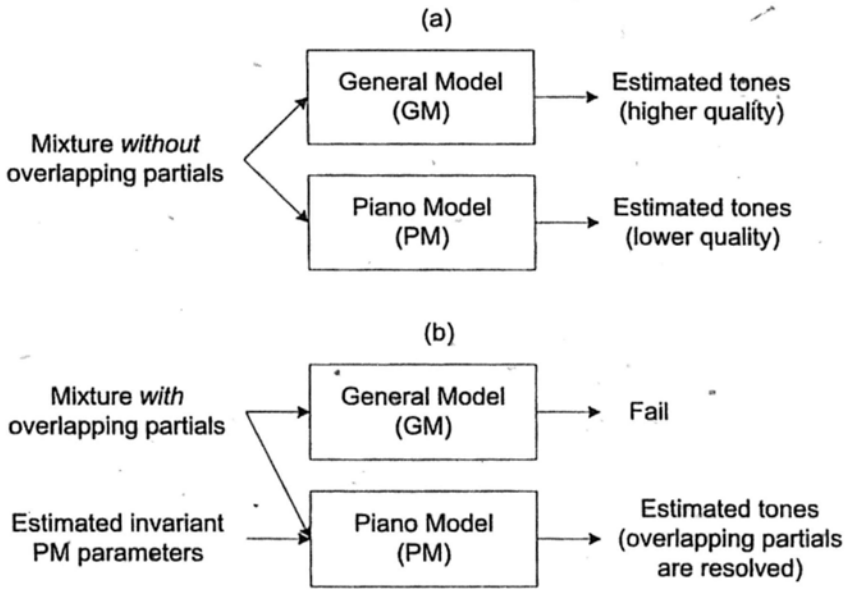

Figure 4.1: Comparison between the General Model and the Piano Model. (a) Mixture without overlapping partials. (b) Mixture with overlapping partials. The invariant PM parameters are estimated from training data.

source separation process is to estimate the parameters in GM for any mixtures no matter there are overlapping partials or not. The goal can be achieved by our proposed source process illustrated in Figure 4.2 which will be explained in the next section.

4.1 Summary of our main idea

The main idea is to use the training data to estimate the invariant PM parameters as shown in Figure 4.2. As the invariant PM parameters capture the common properties across instances of tones, they can be used to resolve the overlapping partials in GM. The process is divided into two stages: the training stage and the source separation stage. Here are the major steps:

1. In the training stage, the invariant PM parameters $\Psi_{\mathbf{I}}$ are estimated by using the training data $\boldsymbol{\mathcal{X}}$.

2. The source separation stage contains two steps:

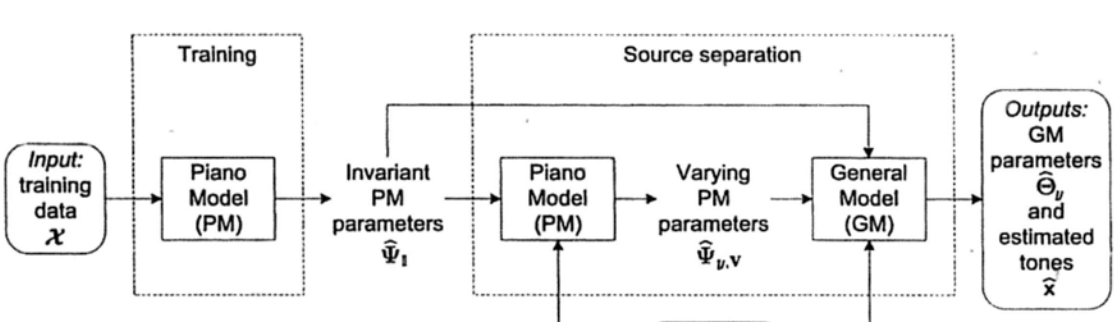

Figure 4.2: Flow of our source separation process.

(a) Given the estimated invariant PM parameters $\widehat{\Psi}_{\mathbb{I}}$ and the mixture $\mathbf{y}$, the varying PM parameters $\Psi_{y,\mathbb{V}}$ in PM for $\mathbf{y}$ are estimated.

(b) Given $\widehat{\Psi}_{\mathbb{I}}$, $\widehat{\Psi}_{y,\mathbb{V}}$ and $\mathbf{y}$, the parameters Θ_y in GM for $\mathbf{y}$ are estimated. The estimated tones can be reconstructed from the estimate $\widehat{\Theta}_y$.

The estimation of all these parameters can be formulated under the Bayesian framework. Before introducing the Bayesian framework, we will discuss how overlapping partials make parameter estimation in GM fail.

4.2 Motivation: problems of parameter estimation in the General Model

Estimating the parameters in GM of a mixture means that both the frequency matrix $\mathbf{H}$ and the amplitude matrix $\mathbf{G}$ in (3.23) are to be estimated in order to find the estimated mixture $\widehat{\mathbf{Y}} = \mathbf{HG}$. To illustrate the difficulties of the estimation, we first consider a simplified case of a frame that the frequency matrix $\mathbf{H}$ is already known and only the amplitude vector $\mathbf{g}$ of the frame is to be estimated. This is written as

$$\mathbf{y} = \mathbf{Hg} + \mathbf{v}. \tag{4.1}$$

If $\mathbf{H}$ is known, GM becomes a linear model. A widely-used estimation method is the least squares which minimizes the sum-of-error squares

$$E = ||\mathbf{y} - \mathbf{H}\mathbf{g}||^2 \tag{4.2}$$

$$= \sum_{l=0}^{L-1} (y[l] - \widehat{y}[l; \mathbf{g}])^2 . \tag{4.3}$$

Then the least-squares solution is

$$\widehat{\mathbf{g}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1} \mathbf{H}^{\mathrm{T}}\mathbf{y} \tag{4.4}$$

which is a unique solution when $\mathbf{H}$ is full column rank.

For music signals, $\mathbf{H}$ is often rank deficient. This happens when some of the partials from different tones in the mixture are overlapping. Overlapping partials frequently occur as discussed in Chapter 1. When there are overlapping partials, these overlapping partials have very close frequencies that cannot be effectively distinguished. If two sinusoidals have an identical frequency, the sum of these two sinusoidals gives a sinusoidal with the same frequency and the two sinusoidals cannot be recovered from their sum. This problem can be further analyzed from the perspective of matrix analysis. Recalling $\mathbf{H}$ is a concatenation of the frequency matrix $\mathbf{H}_k$ of each individual tone in (3.18) so

$$\mathbf{H} = \left[\begin{array}{cccc} \mathbf{H}_1 & \mathbf{H}_2 & \cdots & \mathbf{H}_K \end{array} \right]. \tag{4.5}$$

Following the definitions in (3.9), (3.10) and (3.11), each matrix $\mathbf{H}_k$ contains the cosine block matrix and the sine block matrix

$$\mathbf{H}_k = \left[\begin{array}{cc} \mathbf{H}_k^{\cos} & \mathbf{H}_k^{\sin} \end{array} \right] \tag{4.6}$$

where

$$\mathbf{H}_k^{\cos} = \begin{bmatrix} w[0]\cos(2\pi f_{k,1}t_0) & \cdots & w[0]\cos(2\pi f_{k,M_k}t_0) \\ w[1]\cos(2\pi f_{k,1}t_1) & \cdots & w[1]\cos(2\pi f_{k,M_k}t_1) \\ \vdots & & \vdots \\ w[L-1]\cos(2\pi f_{k,1}t_{L-1}) & \cdots & w[L-1]\cos(2\pi f_{k,M_k}t_{L-1}) \end{bmatrix} \quad (4.7)$$

and

$$\mathbf{H}_k^{\sin} = \begin{bmatrix} w[0]\sin(2\pi f_{k,1}t_0) & \cdots & w[0]\sin(2\pi f_{k,M_k}t_0) \\ w[1]\sin(2\pi f_{k,1}t_1) & \cdots & w[1]\sin(2\pi f_{k,M_k}t_1) \\ \vdots & & \vdots \\ w[L-1]\sin(2\pi f_{k,1}t_{L-1}) & \cdots & w[L-1]\sin(2\pi f_{k,M_k}t_{L-1}) \end{bmatrix}. \quad (4.8)$$

If a pair of partials overlaps, their frequency values are very close so their corresponding columns in the frequency matrix $\mathbf{H}$ are almost identical. Then $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ are nearly singular. If $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ is nearly singular, the resulting solution may greatly depart from the desirable solution. In the case of singular $\mathbf{H}^{\mathrm{T}}\mathbf{H}$, there are infinite number of solutions. An example of rank deficient $\mathbf{H}$ is shown in Figure 4.3. Suppose there are two tones in the mixture. Hence, $\mathbf{H}$ contains the blocks of $\mathbf{H}_1$ and $\mathbf{H}_2$, and each of them contains its cosine and sine blocks. If there exists a pair of overlapping partials, i.e., a partial from the 1st tone has the same frequency of a partial from the 2nd tone, the columns corresponding to the overlapping partials in $\mathbf{H}_1^{\cos}$ and $\mathbf{H}_2^{\cos}$ are identical. Similarly, the columns corresponding to the overlapping partials in $\mathbf{H}_1^{\sin}$ and $\mathbf{H}_2^{\sin}$ are also identical. Hence, a pair of overlapping partials gives two pairs of identical columns in $\mathbf{H}$. As a result, $\mathbf{H}$ is rank deficient.

Another case of singular $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ is that the total number of partials in a mixture is large. This may happens when the mixture contains several tones

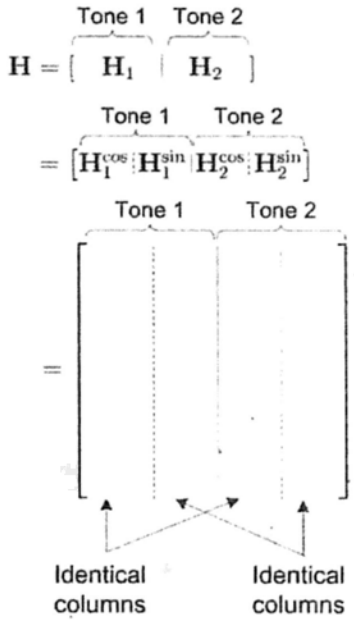

Figure 4.3: Rank deficient $\mathbf{H}$.

and/or the tones are low in pitch so they have many partials. Note that $\mathbf{H}$ is an L-by-$2M$ matrix where L is the window length and M is the total number of partials. If L is less than $2M$, the matrix $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ is also singular. Increasing the window length may solve the problem but this will decrease the time resolution and sacrifice the separation quality.

These two situations give an important implication - if only the mixture is given, and there are overlapping partials and/or L is less than $2M$, the mixture cannot be separated into its individual tones unless more information is provided.

This problem arising from singular $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ can be solved by using the training data as the prior information under the Bayesian framework. The following simple example of straight-line fitting adapted from [11, p. 154] will illustrate how the Bayesian framework works.

4.2.1 Example to illustrate rank deficient H

Suppose we have some data generated from a linear model of the form

$$y(t, \mathbf{g}) = \alpha + \beta t \tag{4.9}$$

where t is the input variable, y is the output variable, α and β are the coefficients of the linear model and they are grouped into $\mathbf{g} = [\alpha\ \beta]^{\mathrm{T}}$. There are N data points observed and they are denoted by the vector $\mathbf{y} = [y_1\ y_2\ \cdots\ y_N]^{\mathrm{T}}$. The input variable of each y_n is t_n. The observed data is contaminated by the noise $\mathbf{v} = [v_1\ v_2\ \cdots\ v_N]^{\mathrm{T}}$. The observed data can be written in the matrix form

$$\mathbf{y} = \mathbf{H}\mathbf{g} + \mathbf{v} \tag{4.10}$$

where

$$\mathbf{H} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_N \end{bmatrix}. \tag{4.11}$$

The goal is to estimate the coefficient vector $\mathbf{g}$ given the noisy data $\mathbf{y}$. In addition to the least squares estimation, another widely-used estimation method is *maximum likelihood* which estimates the value of $\mathbf{g}$ by maximizing the likelihood function $p(\mathbf{y}|\mathbf{g})$. Given that the noise $\mathbf{v}$ is the zero-mean Gaussian with the variance σ_v^2 and v_n is independent and identically distributed, the likelihood function is in the form

$$p(\mathbf{y}|\mathbf{g}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n|\widehat{y}_n, \sigma_v^2\right) \tag{4.12}$$

$$= \frac{1}{(2\pi\sigma_v^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_v^2} \parallel \mathbf{y} - \mathbf{H}\mathbf{g} \parallel^2\right\} \tag{4.13}$$

where $\widehat{y}_n$ is the estimate of y_n in the form

$$\widehat{y}_n(t_n, \mathbf{g}) = \alpha + \beta t_n \tag{4.14}$$

and $\mathcal{N}(\cdot)$ denotes the Gaussian distribution. For the case of a single variable y, the Gaussian distribution is defined by

$$\mathcal{N}\left(y|\mu, \sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y-\mu\right)^2\right\} \tag{4.15}$$

where μ is the mean and σ^2 is the variance. For the case of an N-dimensional vector $\mathbf{y}$, which will appear in later sections, the Gaussian distribution is defined by

$$\mathcal{N}\left(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{(2\pi)^{N/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right\} \tag{4.16}$$

where $\boldsymbol{\mu}$ is the N-dimensional mean vector and $\boldsymbol{\Sigma}$ is the $N \times N$ covariance matrix.

From [11, p. 142], maximization of (4.12) gives the maximum likelihood solution in the form

$$\widehat{\mathbf{g}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{y} \tag{4.17}$$

which is equivalent to the least-squares solution. Note that maximum likelihood and least squares give the same solution when the noise is zero-mean Gaussian [11].

Synthetic data is generated to visualize the likelihood function and study how $\mathbf{H}$ affects the likelihood function. Four data sets are generated from the function $y(t, \mathbf{g}_{\text{true}})$ with the same $\mathbf{g}_{\text{true}} = [-0.2\ 0.5]^{\mathrm{T}}$ but different sets of t. The zero-mean Gaussian noise v_n with the standard deviation σ_v of 0.1 is added to $y(t, \mathbf{g}_{\text{true}})$ to obtain the observed value y_n. Each data set contains three data points and the results are shown in Figure 4.4. In Figure 4.4(a), the input

values t_n are -0.5, 0 and 0.5 so

$$\mathbf{H} = \begin{bmatrix} 1 & -0.5 \\ 1 & 0 \\ 1 & 0.5 \end{bmatrix}. \tag{4.18}$$

The input values are well separated. The likelihood is sharply peaked and the location of the peak is close to the true values $\mathbf{g}_{\text{true}}$.

In Figure 4.4(b), the input values t_n are 0.1, 0.3 and 0.5 so

$$\mathbf{H} = \begin{bmatrix} 1 & 0.1 \\ 1 & 0.3 \\ 1 & 0.5 \end{bmatrix}. \tag{4.19}$$

The input values are closer. Although the likelihood is elliptical and the likelihood is less sharply peaked, the location of the peak is still close to the true values.

In Figure 4.4(c), the input values t_n are 0.4, 0.45 and 0.50 so

$$\mathbf{H} = \begin{bmatrix} 1 & 0.4 \\ 1 & 0.45 \\ 1 & 0.5 \end{bmatrix}. \tag{4.20}$$

The input values are very close to each other so $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ is nearly singular. The likelihood is highly elliptical and the likelihood of the true values has very similar values at many locations.

In Figure 4.4(d), all inputs are set to 0.5 so

$$\mathbf{H} = \begin{bmatrix} 1 & 0.5 \\ 1 & 0.5 \\ 1 & 0.5 \end{bmatrix} \tag{4.21}$$

and $\mathbf{H}^T\mathbf{H}$ is singular and there are infinite number of solutions. In such case, if only the data is given, the true line cannot be found unless some prior knowledge is given. The next section will show how the Bayesian framework makes use of the prior knowledge to solve the problem.

4.3 Bayesian analysis for source separation

The example of straight-line fitting illustrates that maximum likelihood or least squares cannot handle the case that $\mathbf{H}$ is not full column rank. The Bayesian framework does not require $\mathbf{H}$ to be full column rank by incorporating the prior knowledge of the parameters. Back to the source separation problem, given a mixture $\mathbf{y}$, our aim is that to estimate both the frequency matrix $\mathbf{H}$ and the amplitude matrix $\mathbf{G}$. The frequency matrix $\mathbf{H}$ is generated by the frequencies $\mathbf{f}$ of all partials from (3.8). Let Θ be the parameter set containing all parameters such that $\Theta = \{\mathbf{f}, \mathbf{G}\}$. The prior knowledge of Θ can be quantified by the prior probability distribution $p(\Theta)$. The prior $p(\Theta)$ represents our prior knowledge of Θ before observing the mixture $\mathbf{y}$. The likelihood function $p(\mathbf{y}|\Theta)$ describes how likely the observed mixture $\mathbf{y}$ is generated by the parameter set Θ. The prior and the likelihood can be linked up by Bayes' theorem in the form of

$$p(\Theta|\mathbf{y}) = \frac{p(\mathbf{y}|\Theta)p(\Theta)}{p(\mathbf{y})} \tag{4.22}$$

where $p(\Theta|\mathbf{y})$ is the posterior probability distribution that expresses the probability distribution of Θ after observing the mixture $\mathbf{y}$. The denominator $p(\mathbf{y})$ in (4.22) is a normalization constant. It makes the integral of $p(\Theta|\mathbf{y})$ with respect to Θ equal to one so $p(\Theta|\mathbf{y})$ a valid probability distribution. As $p(\mathbf{y})$ is a normalization constant, the Bayes' theorem can be rewritten as

$$\underbrace{p(\Theta|\mathbf{y})}_{\text{posterior}} \propto \underbrace{p(\mathbf{y}|\Theta)}_{\text{likelihood}} \underbrace{p(\Theta)}_{\text{prior}} \tag{4.23}$$

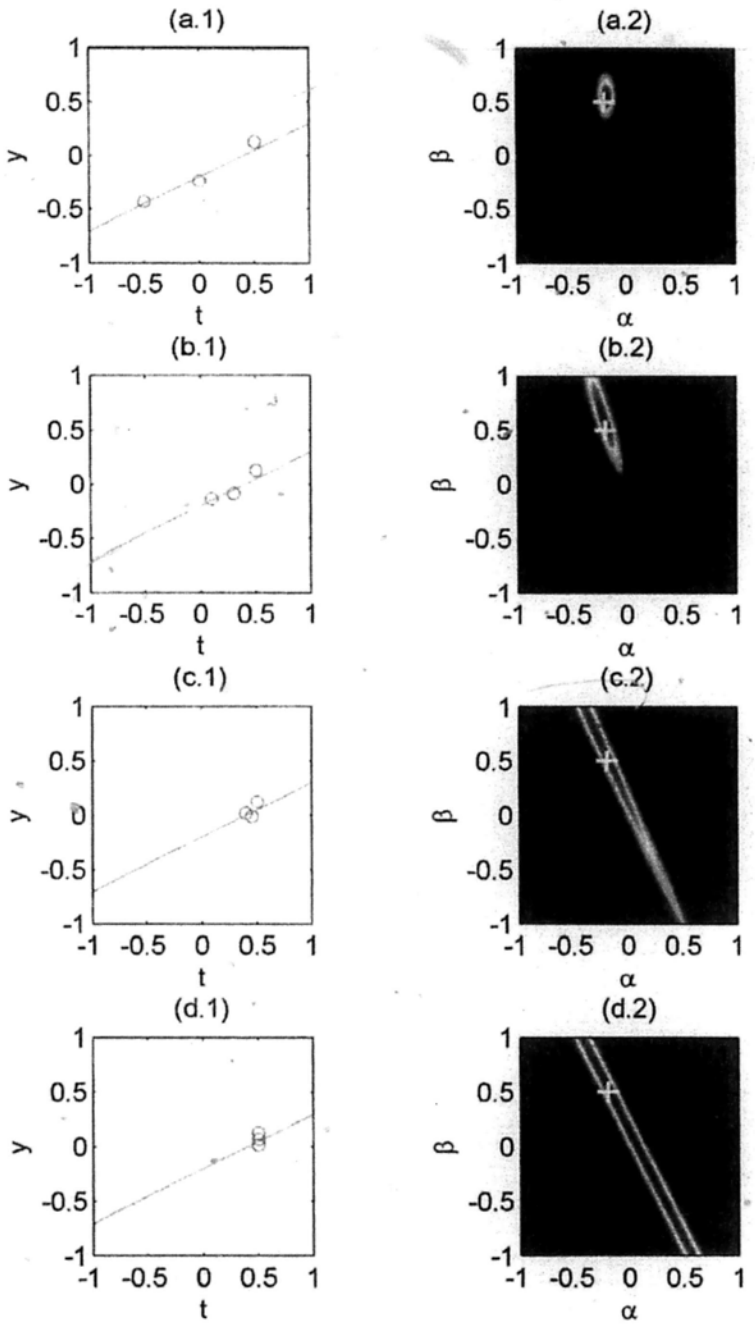

Figure 4.4: The likelihood functions of the four data sets. The plots in the left column show the data space. The red line is the line generated by $\mathbf{g}_{\text{true}}$. The plots in the right column show the corresponding likelihood in the parameter space. The white crosses show the true parameter values $\mathbf{g}_{\text{true}}$.

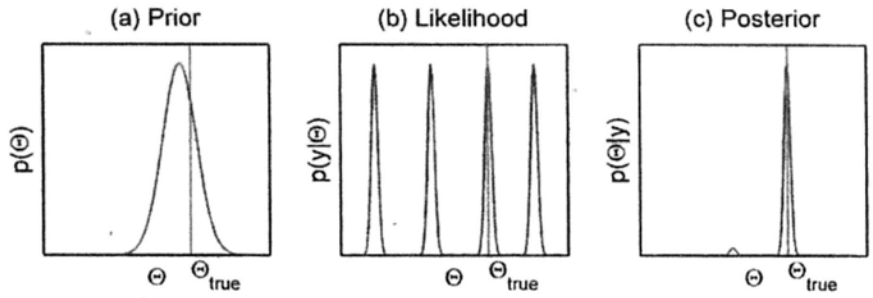

Figure 4.5: (a) The likelihood function. (b) The prior distribution. (c) The posterior distribution. This schematic diagram shows that an appropriate prior gives the desirable MAP solution. The vertical line shows the true value of Θ.

which means the posterior is directly proportional to the product of the likelihood and the prior. The goal of Bayesian analysis in the source separation problem is to find $\widehat{\Theta}$ that maximizes the posterior $p(\Theta|\mathbf{y})$ so

$$\widehat{\Theta} = \arg\max_{\Theta} p(\Theta|\mathbf{y}) \tag{4.24}$$

where $\widehat{\Theta}$ is called the *Maximum A Posterior* (MAP) solution.

The key issue of Bayesian source separation is how to set up the prior $p(\Theta)$. If overlapping partials are present, the matrix $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ is nearly singular, many choices of Θ can give similar values of the likelihood. Hence, there are many peaks in the likelihood function as shown in the schematic diagram (Figure 4.5(b)). In order to find the desirable MAP solution, it is desirable that the prior distribution has a high density around the correct value of Θ. In Figure 4.5(a), the prior is appropriate so that the MAP solution, i.e. the peak of the posterior, can be located correctly as depicted in Figure 4.5(c). Before discussing how to find the appropriate prior, we illustrate how the prior affects the posterior in the straight-line fitting example.

4.3.1 Example to illustrate how the Bayesian framework works

In Section 4.2.1, we show that maximum likelihood fails to handle the case that $\mathbf{H}$ is not full column rank. In this section, we continue to discuss the example of straight-line fitting and show how Bayesian analysis handles such case. In the example of straight-line fitting, the goal is to estimate the coefficient vector $\mathbf{g} = [\alpha \ \beta]^{\mathrm{T}}$ given the observed data $\mathbf{y}$ and its input vector $\mathbf{t}$. Let the prior $p(\mathbf{g})$ be a Gaussian with the mean $\boldsymbol{\mu}_g$ and the covariance $\boldsymbol{\Sigma}_g$ so

$$\begin{aligned} p(\mathbf{g}) &= \mathcal{N}(\mathbf{g}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) && (4.25) \\ &= \frac{1}{2\pi|\boldsymbol{\Sigma}_g|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{g}-\boldsymbol{\mu}_{\mathbf{g}})^{\mathrm{T}} \boldsymbol{\Sigma}_g^{-1}(\mathbf{g}-\boldsymbol{\mu}_{\mathbf{g}})\right\}. && (4.26) \end{aligned}$$

The likelihood $p(\mathbf{y}|\mathbf{g})$ is the same as (4.12) and it is restated here for convenience

$$p(\mathbf{y}|\mathbf{g}) = \frac{1}{(2\pi\sigma_v^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_v^2} \| \mathbf{y} - \mathbf{H}\mathbf{g} \|^2\right\}. \tag{4.27}$$

The resulting posterior $p(\mathbf{g}|\mathbf{y})$ is also a Gaussian as shown in [11, p. 153]. It is in the form

$$p(\mathbf{g}|\mathbf{y}) = \mathcal{N}(\mathbf{g}|\mathbf{m}_g, \mathbf{S}_g) \tag{4.28}$$

where

$$\begin{aligned} \mathbf{m}_g &= \mathbf{S}_g\left(\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\mu}_g + \sigma_v^{-2}\mathbf{H}^{\mathrm{T}}\mathbf{y}\right) && (4.29) \\ \mathbf{S}_g &= \left(\boldsymbol{\Sigma}_g^{-1} + \sigma_v^{-2}\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}. && (4.30) \end{aligned}$$

The mode of a Gaussian distribution coincides with its mean. Therefore, the MAP solution $\widehat{\mathbf{g}}$ is equal to the posterior mean $\mathbf{m}_g$. Substituting (4.30) into (4.29) gives

$$\widehat{\mathbf{g}} = \left(\boldsymbol{\Sigma}_g^{-1} + \sigma_v^{-2}\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\left(\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\mu}_g + \sigma_v^{-2}\mathbf{H}^{\mathrm{T}}\mathbf{y}\right). \tag{4.31}$$

Note that finding a unique $\widehat{\mathbf{g}}$ does not require that $\mathbf{H}$ is full column rank. The relationship between MAP and maximum likelihood can be studied through the prior. Here we consider $\boldsymbol{\Sigma}_g$ is a diagonal matrix in the form

$$\boldsymbol{\Sigma}_g = \begin{bmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\beta^2 \end{bmatrix} \tag{4.32}$$

where σ_α^2 and σ_β^2 are the variances of the prior for α and β respectively. If the prior is infinitely broad, both σ_α^2 and σ_β^2 tend to infinity then $\boldsymbol{\Sigma}_g^{-1} = \mathbf{0}$ and $\widehat{\mathbf{g}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{y}$. Thus the MAP solution is the same as the maximum likelihood solution given in (4.17). An infinite broad prior indicates that prior knowledge of $\mathbf{g}$ is unavailable. In such case, $\mathbf{H}$ is required to be full column rank for the unique estimate of $\widehat{\mathbf{g}}$.

We illustrate the importance of the prior when $\mathbf{H}$ is not full column. This is the case in Figure 4.4(d) when all inputs are set to 0.5. This case is further investigated in Figure 4.6. Suppose we have the prior knowledge of $\mathbf{g}$ that $\boldsymbol{\mu}_g = [0\ 0.7]^{\mathrm{T}}$ and $\sigma_\alpha = \sigma_\beta = 0.2$. The prior is plotted in Figure 4.6(b). This prior is appropriate in the sense that its center is not far away from the true value of $\mathbf{g}$ ($\mathbf{g}_{\text{true}} = [-0.3\ 0.5]^{\mathrm{T}}$). The MAP solution shown in Figure 4.6(d) is near to the true values. In Figure 4.6(a), the line generated by the prior mean does not pass through the data points. After observing the data, the likelihood can be calculated to quantify the influence of the data. Both the prior and the likelihood affect the posterior as shown in (4.23). The line generated by the posterior mean, i.e. the MAP solution, passes close to the data points and its direction is also close to the line generated by the prior mean as well as the true line.

The importance of the prior can be further illustrated by an inappropriate prior. Suppose the prior mean is changed to $\boldsymbol{\mu}_g = [0\ -0.5]^{\mathrm{T}}$ with the same standard derivation $\sigma_\alpha = \sigma_\beta = 0.2$. This prior is plotted in Figure 4.7(b). The

likelihood in Figure 4.7(c) is the identical to that in 4.6(c). The posterior in Figure 4.7(d) does not give a good MAP solution. In Figure 4.7(a), the line generated by the MAP solution passes close to the data points but its direction is close to the line of the prior mean. Both the lines of the posterior mean and the prior mean are wrongly directed. The result is not surprising because the MAP solution comes from the likelihood and the prior. The likelihood is only able to provide a soft constraint as in Figures 4.7(c) and 4.6(c). If $\mathbf{H}$ is not full column and the prior is wrongly specified, no good MAP solution can be found. Therefore, in the case of overlapping partials, an appropriate prior is crucial for resolving the overlapping partials. How to find an appropriate prior will be discussed in the next section.

4.4 Problem formulation for source separation

In the previous section, we show that an appropriate prior is crucial for resolving the overlapping partials. The prior can be found by using the training data. In piano music, a particular pitch rarely appears only once. The tones of the same pitch share some common characteristics which are captured by our proposed Piano Model (PM) to resolve the overlapping partials. In our research, we focus on the case that given a mixture $\mathbf{y}$, all the pitches in the mixture reappear as isolated tones in the target recording. These isolated tones are extracted from the target recording as discussed in Section 1.3 to form the training data $\boldsymbol{\mathcal{X}}$. If no training data is available, source separation may be performed with other methods discussed in Chapter 8.

To include the training data in the Bayesian framework, the posterior $p(\boldsymbol{\Theta}|\mathbf{y})$ in (4.24) is rewritten as $p(\boldsymbol{\Theta}_y|\mathbf{y}, \boldsymbol{\mathcal{X}})$ where $\boldsymbol{\Theta}_y$ is the parameter set of the traditional General Model (GM) for the mixture $\mathbf{y}$. Then the goal of source separation is to find the MAP solution, i.e. to find $\widehat{\boldsymbol{\Theta}}_y$ that maximizes

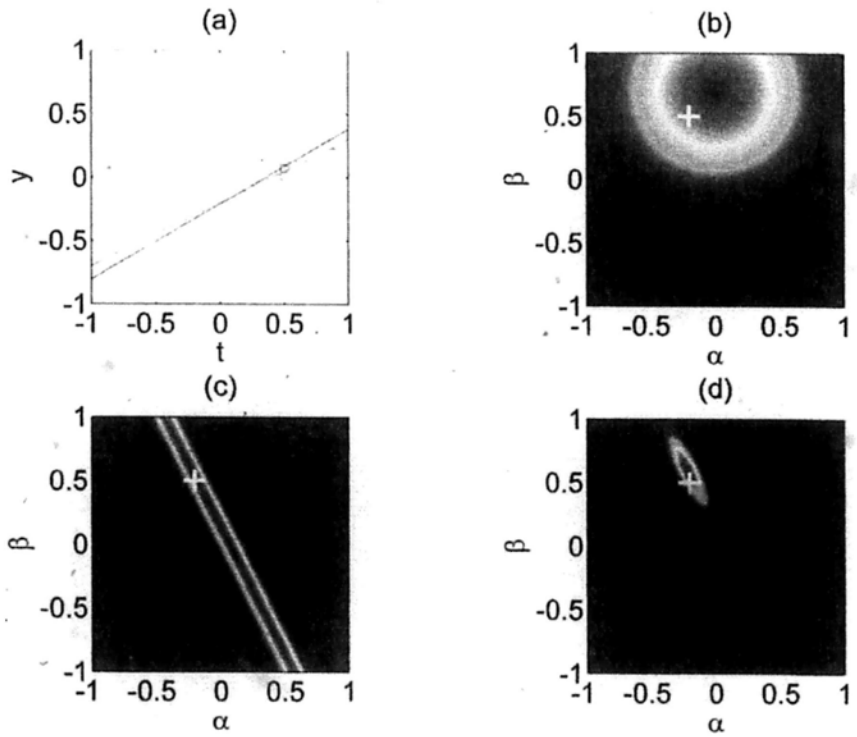

Figure 4.6: Illustration of Bayesian analysis for the data set in Figure 4.4(d) with an appropriate prior. (a) Data space. The blue circles are the observed data. The red solid line is the true line. The blue dotted line is generated from the prior mean by setting $\mathbf{g} = \boldsymbol{\mu}_g$. The dashed black line is generated from the posterior mean by setting $\mathbf{g} = \mathbf{m}_g$. (b) The prior distribution. (c) The likelihood function. (d) The posterior distribution. All the white crosses in (b), (c) and (d) are the location of $\mathbf{g}_{\text{true}}$.

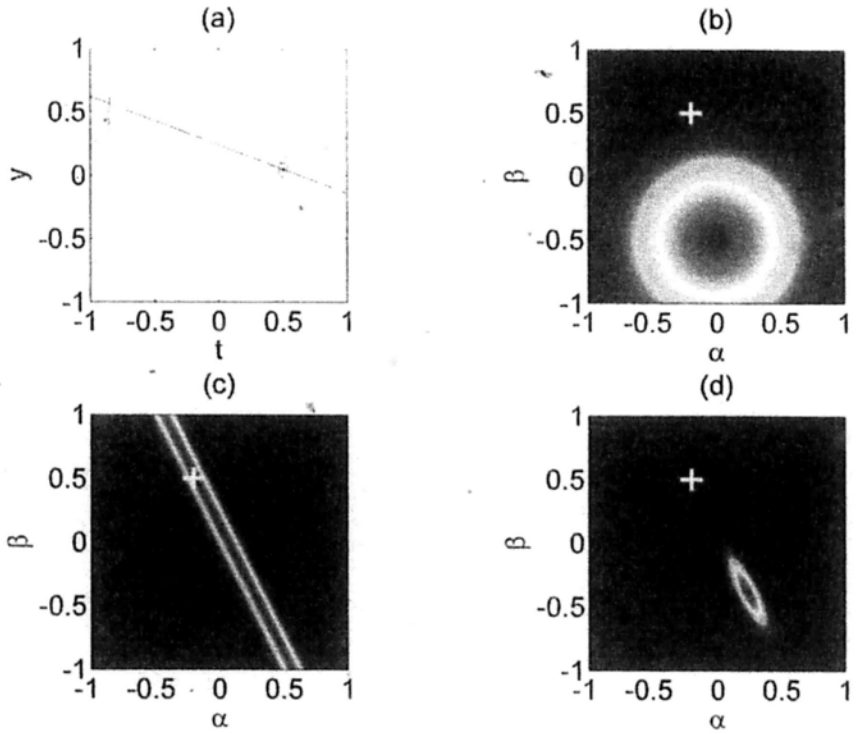

Figure 4.7: Illustration of Bayesian analysis for the data set in Figure 4.4(d) with an inappropriate prior. (a) Data space. The blue circles are the observed data. The red solid line is the true line. The blue dotted line is generated from the prior mean by setting $\mathbf{g} = \boldsymbol{\mu}_g$. The dashed black line is generated from the posterior mean by setting $\mathbf{g} = \mathbf{m}_g$. (b) The prior distribution. (c) The likelihood function. (d) The posterior distribution. All the white crosses in (b), (c) and (d) are the location of $\mathbf{g}_{\text{true}}$.

the posterior $p(\Theta_y|\mathbf{y}, \mathcal{X})$ so

$$\widehat{\Theta}_y = \arg\max_{\Theta_y} p(\Theta_y|\mathbf{y}, \mathcal{X}) \tag{4.33}$$

and by Bayes' theorem, the posterior can be written in the form

$$\begin{aligned} p(\Theta_y|\mathbf{y}, \mathcal{X}) &\propto p(\mathbf{y}|\Theta_y, \mathcal{X})p(\Theta_y|\mathcal{X}) && (4.34) \\ &= p(\mathbf{y}|\Theta_y)p(\Theta_y|\mathcal{X}). && (4.35) \end{aligned}$$

The last step makes use of the likelihood depending on the parameter set Θ_y and independent of $\mathcal{X}$. The functional form of $p(\mathbf{y}|\Theta_y)$ will be discussed in Chapter 5. The remaining section will focus on how to find the prior $p(\Theta_y|\mathcal{X})$.

The prior $p(\Theta_y|\mathcal{X})$ expresses the probability distribution of the parameter set Θ_y of the mixture $\mathbf{y}$ given the training data $\mathcal{X}$ and before the mixture $\mathbf{y}$ is observed. The key problem is that how to make use of the training data $\mathcal{X}$ and to find a functional form for the prior $p(\Theta_y|\mathcal{X})$. This comes our PM. In PM for a mixture $\mathbf{y}$, the parameter set Ψ_y is divided into two sets: the invariant PM parameter set $\Psi_{y,\mathbb{I}}$ and the varying PM parameter set $\Psi_{y,\mathbb{V}}$. For the training data $\mathcal{X}$, the parameter set $\Psi_{\mathcal{X}}$ is divided into the invariant PM parameter set $\Psi_{\mathcal{X},\mathbb{I}}$ and the varying PM parameter set $\Psi_{\mathcal{X},\mathbb{V}}$. Note that both the mixture and the training 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 $\Psi_{\mathbb{I}} = \Psi_{y,\mathbb{I}} = \Psi_{\mathcal{X},\mathbb{I}}$.

Before deriving the prior $p(\Theta_y|\mathcal{X})$, the sum and product rules of probability will be discussed first. In addition to Bayes' theorem, the sum and product rules of probability will be extensively used in the Bayesian framework. It is worth to restate Bayes' theorem and define these basic rules of probability. If A and B and two real continuous random variables, the Bayes' theorem can

be stated in the form

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}. \tag{4.36}$$

The sum rule is

$$p(A) = \int p(A, B)dB. \tag{4.37}$$

The product rule is

$$p(A, B) = p(B|A)p(A). \tag{4.38}$$

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

$$p(\Theta_y|\mathbf{y}, \boldsymbol{\mathcal{X}}) = \iint p(\Theta_y, \Psi_{y,\mathrm{V}}, \Psi_{\mathrm{I}}|\mathbf{y}, \boldsymbol{\mathcal{X}})d\Psi_{y,\mathrm{V}}d\Psi_{\mathrm{I}}. \tag{4.39}$$

Integrating out $\Psi_{y,\mathrm{V}}$ and Ψ_{I} from $p(\Theta_y, \Psi_{y,\mathrm{V}}, \Psi_{\mathrm{I}}|\mathbf{y}, \boldsymbol{\mathcal{X}})$ is called *marginalization* in Bayesian analysis [11]. The noise variance $\sigma_{v_r}^2$ of the mixture is omitted in the derivation for clarity. The estimation of the noise variance will be discussed in Chapter 6. Then by the product rule, (4.39) can be put into

$$\begin{aligned} p(\Theta_y|\mathbf{y}, \boldsymbol{\mathcal{X}}) &= \iint p(\Theta_y|\mathbf{y}, \boldsymbol{\mathcal{X}}, \Psi_{y,\mathrm{V}}, \Psi_{\mathrm{I}})p(\Psi_{y,\mathrm{V}}, \Psi_{\mathrm{I}}|\mathbf{y}, \boldsymbol{\mathcal{X}})d\Psi_{y,\mathrm{V}}d\Psi_{\mathrm{I}} \\ &= \iint p(\Theta_y|\mathbf{y}, \Psi_{y,\mathrm{V}}, \Psi_{\mathrm{I}})p(\Psi_{y,\mathrm{V}}, \Psi_{\mathrm{I}}|\mathbf{y}, \boldsymbol{\mathcal{X}})d\Psi_{y,\mathrm{V}}d\Psi_{\mathrm{I}} \end{aligned} \tag{4.40}$$

where $p(\Theta_y|\mathbf{y}, \Psi_{y,\mathrm{V}}, \Psi_{\mathrm{I}})$ is the posterior of the GM parameters Θ_y given the mixture $\mathbf{y}$, and the PM parameters $\Psi_{y,\mathrm{V}}$ and Ψ_{I}; while $p(\Psi_{y,\mathrm{V}}, \Psi_{\mathrm{I}}|\mathbf{y}, \boldsymbol{\mathcal{X}})$ is the posterior of the PM parameters $\Psi_{y,\mathrm{V}}$ and Ψ_{I} given the mixture $\mathbf{y}$ and the training data $\boldsymbol{\mathcal{X}}$. Here we have omitted $\boldsymbol{\mathcal{X}}$ in $p(\Theta_y|\mathbf{y}, \Psi_{y,\mathrm{V}}, \Psi_{\mathrm{I}})$ because the dependence of $\boldsymbol{\mathcal{X}}$ has been expressed via Ψ_{I}.

Changing the posterior from $p(\Theta_y|\mathbf{y}, \mathcal{X})$ to $p(\Theta_y|\mathbf{y}, \Psi_{y,\mathrm{v}}, \Psi_{\mathrm{I}})$ is crucial because now the posterior $p(\Theta_y|\mathbf{y}, \Psi_{y,\mathrm{v}}, \Psi_{\mathrm{I}})$ depends on the invariant PM parameters Ψ_{I} which is conditioned on the training data $\mathcal{X}$ in the form of $p(\Psi_{y,\mathrm{v}}, \Psi_{\mathrm{I}}|\mathbf{y}, \mathcal{X})$. Hence, the posterior $p(\Theta_y|\mathbf{y}, \Psi_{y,\mathrm{v}}, \Psi_{\mathrm{I}})$ is able to use the training data to resolve the overlapping partials.

Finding the MAP solution involves evaluating the integration over all possible values of $\Psi_{y,\mathrm{v}}$ and Ψ_{I} in (4.40). PM is a highly dimensional and nonlinear model that makes the integration analytically infeasible. Different approximation methods can be used to find the MAP solution including the deterministic methods (e.g. evidence approximation [47, 48] and variational approximation [67]) and the probabilistic methods (e.g. Markov chain Monte Carlo [62] and particle filtering [26]). For computational efficiency, here we will use the evidence approximation. It is also called type-II maximum likelihood [8] or empirical Bayes [16].

The evidence approximation has two main steps. The first step is to evaluate the posterior $p(\Theta_y|\mathbf{y}, \Psi_{y,\mathrm{v}}, \Psi_{\mathrm{I}})$ at the most probable values of $\Psi_{y,\mathrm{v}}$ and Ψ_{I} to avoid from performing the integration[1] . The second step is to find the most probable values of $\Psi_{y,\mathrm{v}}$ and Ψ_{I}. The main idea of these two steps will be discussed below.

Following the derivation of the evidence approximation in [10, p. 408], let us suppose that the posterior $p(\Psi_{y,\mathrm{v}}, \Psi_{\mathrm{I}}|\mathbf{y}, \mathcal{X})$ is sharply peaked around their most probable values $\widehat{\Psi}_{y,\mathrm{v}}$ and $\widehat{\Psi}_{\mathrm{I}}$. Then (4.40) can be written

$$\begin{aligned} p(\Theta_y|\mathbf{y}, \mathcal{X}) &\approx p(\Theta_y|\mathbf{y}, \widehat{\Psi}_{y,\mathrm{v}}, \widehat{\Psi}_{\mathrm{I}}) \iint p(\Psi_{y,\mathrm{v}}, \Psi_{\mathrm{I}}|\mathbf{y}, \mathcal{X}) d\Psi_{y,\mathrm{v}} d\Psi_{\mathrm{I}} \\ &= p(\Theta_y|\mathbf{y}, \widehat{\Psi}_{y,\mathrm{v}}, \widehat{\Psi}_{\mathrm{I}}). \end{aligned} \tag{4.41}$$

[1] In the evidence approximation, it is usually to be said that the posterior is evaluated at the most probable values of the *hyperparameters*. The concept of hyperparameters will be introduced in Chapter 6.

Hence, the MAP solution of Θ_y becomes

$$\widehat{\Theta}_y = \arg\max_{\Theta_y} p(\Theta_y|\mathbf{y}, \widehat{\Psi}_{y,\mathrm{v}}, \widehat{\Psi}_{\mathrm{I}}) \tag{4.42}$$

The most probable value $\widehat{\Psi}_{y,\mathrm{v}}$ can be estimated by maximizing the posterior $p(\Psi_{y,\mathrm{v}}|\mathbf{y}, \boldsymbol{\mathcal{X}})$

$$\widehat{\Psi}_{y,\mathrm{v}} = \arg\max_{\Psi_{y,\mathrm{v}}} p(\Psi_{y,\mathrm{v}}|\mathbf{y}, \boldsymbol{\mathcal{X}}). \tag{4.43}$$

Then the sum rule and the evidence approximation can be applied in the same fashion

$$\begin{aligned} p(\Psi_{y,\mathrm{v}}|\mathbf{y}, \boldsymbol{\mathcal{X}}) &= \int p(\Psi_{y,\mathrm{v}}, \Psi_{\mathrm{I}}|\mathbf{y}, \boldsymbol{\mathcal{X}}) d\Psi_{\mathrm{I}} \\ &= \int p(\Psi_{y,\mathrm{v}}|\mathbf{y}, \boldsymbol{\mathcal{X}}, \Psi_{\mathrm{I}}) p(\Psi_{\mathrm{I}}|\mathbf{y}, \boldsymbol{\mathcal{X}}) d\Psi_{\mathrm{I}} \\ &= \int p(\Psi_{y,\mathrm{v}}|\mathbf{y}, \Psi_{\mathrm{I}}) p(\Psi_{\mathrm{I}}|\mathbf{y}, \boldsymbol{\mathcal{X}}) d\Psi_{\mathrm{I}} \\ &\approx p(\Psi_{y,\mathrm{v}}|\mathbf{y}, \widehat{\Psi}_{\mathrm{I}}) \int p(\Psi_{\mathrm{I}}|\mathbf{y}, \boldsymbol{\mathcal{X}}) d\Psi_{\mathrm{I}} \\ &= p(\Psi_{y,\mathrm{v}}|\mathbf{y}, \widehat{\Psi}_{\mathrm{I}}). \end{aligned} \tag{4.44}$$

The most probable value $\widehat{\Psi}_{\mathrm{I}}$ can be estimated by maximizing the posterior $p(\Psi_{\mathrm{I}}|\mathbf{y}, \boldsymbol{\mathcal{X}})$

$$\widehat{\Psi}_{\mathrm{I}} = \arg\max_{\Psi_{\mathrm{I}}} p(\Psi_{\mathrm{I}}|\mathbf{y}, \boldsymbol{\mathcal{X}}). \tag{4.45}$$

This can be approximated by finding the posterior only given the training data so

$$p(\Psi_{\mathrm{I}}|\mathbf{y}, \boldsymbol{\mathcal{X}}) \approx p(\Psi_{\mathrm{I}}|\boldsymbol{\mathcal{X}}). \tag{4.46}$$

According to these results, the whole source separation process is summarized in Figure 4.8. The whole process is divided into the following steps:

1. Given the training data, find the most probable value of the invariant PM parameters $\widehat{\Psi}_{\mathrm{I}}$ in (4.46).

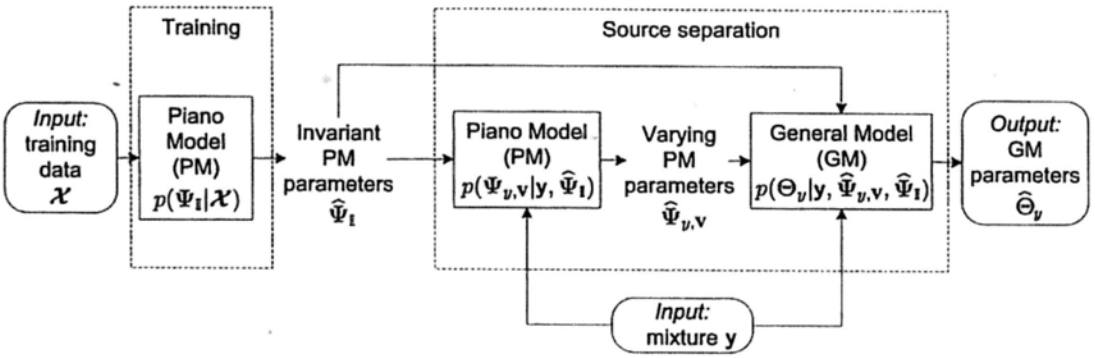

Figure 4.8: Bayesian framework for the source separation.

2. Given the most probable value $\widehat{\Psi}_{\mathrm{I}}$ and the mixture $\mathbf{y}$, find the most probable value of the varying PM parameters $\Psi_{y,\mathrm{V}}$ of the mixture $\mathbf{y}$ in (4.44).

3. Given the most probable values $\widehat{\Psi}_{y,\mathrm{V}}$ and $\widehat{\Psi}_{\mathrm{I}}$, find the MAP solution $\widehat{\Theta}_y$ in (4.41).

The first step is the training stage which will be discussed in Chapter 5. The second and third steps perform the source separation with PM and GM respectively. These two steps will be explained in Chapter 6.

Chapter 5

Training: parameter estimation

This chapter 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 (3.33) 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 5.2 to 5.4. 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.

5.1 Problem formulation for training

The goal of the training stage is to estimate the invariant PM parameters $\Psi_{\mathbb{I}}$, which contain a set of common parameters defined in Section 3.3 for the training data $\boldsymbol{\mathcal{X}}$ and the mixture $\mathbf{y}$. The training data $\boldsymbol{\mathcal{X}}$ is used to estimate $\Psi_{\mathbb{I}}$ so that source separation of $\mathbf{y}$ can be performed. The stage of source separation will be discussed in Chapter 6. Before formulating the training problem, we will introduce the notation first.

A mixture signal $\mathbf{y}$ consists of K individual tones. The index k denotes the index of kth tone in the mixture and $k = 1, 2, \ldots, K$. Let p_k be the pitch of the kth tone. The training data contains the isolated tones of each p_k. Hence, the training data can be divided into K sets, i.e., $\boldsymbol{\mathcal{X}} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_K\}$, and each $\mathcal{X}_k$ consists of the isolated tones of pitch p_k. Moreover, each $\mathcal{X}_k$ may contain more than one instance of the pitch p_k. We introduce the index i to denote the quantities associated with the ith instance. The time-domain signal of the ith instance of the pitch p_k is written as $\mathbf{x}_k^i$ so $\mathcal{X}_k = \{\mathbf{x}_k^1, \mathbf{x}_k^2, \ldots, \mathbf{x}_k^{I_k}\}$ where I_k is the number of instances of pitch p_k in the training data. For example, a mixture $\mathbf{y}$ contains two simultaneous tones. The first tone is C4 and the second is G5 so p_1 = C4 and p_2 = G5. In the training data $\boldsymbol{\mathcal{X}}$, there are three isolated tones for C4 and two for G4 so $I_1 = 3$ and $I_2 = 2$. The training data $\boldsymbol{\mathcal{X}}$ can be written as $\boldsymbol{\mathcal{X}} = \{\mathcal{X}_1, \mathcal{X}_2\}$. The data set $\mathcal{X}_1$ contains the three isolated tones of C4 and $\mathcal{X}_1 = \{\mathbf{x}_1^1, \mathbf{x}_1^2, \mathbf{x}_1^3\}$. The data set $\mathcal{X}_2$ contains the two isolated tones of G4 and $\mathcal{X}_2 = \{\mathbf{x}_2^1, \mathbf{x}_2^2\}$.

In the previous chapter, it states that the goal of the training stage is to find the most probable invariant PM parameters $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ that maximizes the posterior of the invariant PM parameters $p(\boldsymbol{\Psi}_{\mathbb{I}}|\boldsymbol{\mathcal{X}})$. By Bayes' theorem, the posterior can be rewritten as

$$p(\boldsymbol{\Psi}_{\mathbb{I}}|\boldsymbol{\mathcal{X}}) \propto p(\boldsymbol{\mathcal{X}}|\boldsymbol{\Psi}_{\mathbb{I}})p(\boldsymbol{\Psi}_{\mathbb{I}}). \tag{5.1}$$

The prior of the invariant PM parameters $p(\boldsymbol{\Psi}_{\mathbb{I}})$ reflects our prior knowledge of $\boldsymbol{\Psi}_{\mathbb{I}}$. The values of $\boldsymbol{\Psi}_{\mathbb{I}}$ can 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 very insensitive to the values of that parameter [11]. Therefore, we choose a very insensitive prior for $\boldsymbol{\Psi}_{\mathbb{I}}$. Then maximizing the posterior $p(\boldsymbol{\Psi}_{\mathbb{I}}|\boldsymbol{\mathcal{X}})$ is effectively equivalent to maximize the likelihood $p(\boldsymbol{\mathcal{X}}|\boldsymbol{\Psi}_{\mathbb{I}})$ so the most probable parameters $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ are equivalent to the maximum likelihood solution. Hence, the goal of the training stage becomes to find $\widehat{\boldsymbol{\Psi}}_{\mathbb{I}}$ that maximize the likelihood

$p(\boldsymbol{\mathcal{X}}|\Psi_{\mathbb{I}})$ so

$$\widehat{\Psi}_{\mathbb{I}} = \arg\max_{\Psi_{\mathbb{I}}} p(\boldsymbol{\mathcal{X}}|\Psi_{\mathbb{I}}). \tag{5.2}$$

The invariant PM parameters $\Psi_{\mathbb{I}}$ are also divided into K sets as the training data $\boldsymbol{\mathcal{X}}$ so $\Psi_{\mathbb{I}} = \{\psi_{1,\mathbb{I}}, \psi_{2,\mathbb{I}}, \ldots, \psi_{K,\mathbb{I}}\}$. Each $\psi_{k,\mathbb{I}}$ corresponds to the invariant PM parameters of the pitch p_k. The maximum likelihood solution of $\Psi_{\mathbb{I}}$ is defined by $\widehat{\Psi}_{\mathbb{I}} = \{\widehat{\psi}_{1,\mathbb{I}}, \widehat{\psi}_{2,\mathbb{I}}, \ldots, \widehat{\psi}_{K,\mathbb{I}}\}$. Note that each pair of $\mathcal{X}_k$ is independently from each other and $\mathcal{X}_k$ only depends on $\psi_{k,\mathbb{I}}$ so the likelihood $p(\boldsymbol{\mathcal{X}}|\Psi_{\mathbb{I}})$ can be factorized into the product of $p(\mathcal{X}_k|\psi_{k,\mathbb{I}})$

$$p(\boldsymbol{\mathcal{X}}|\Psi_{\mathbb{I}}) = \prod_{k=1}^{K} p(\mathcal{X}_k|\psi_{k,\mathbb{I}}) \tag{5.3}$$

where $p(\mathcal{X}_k|\psi_{k,\mathbb{I}})$ is the likelihood for $\mathcal{X}_k$ given $\psi_{k,\mathbb{I}}$. This implies that maximizing $p(\boldsymbol{\mathcal{X}}|\Psi_{\mathbb{I}})$ can be done by maximizing each $p(\mathcal{X}_k|\psi_{k,\mathbb{I}})$ independently. Then the maximum likelihood solution of $\psi_{k,\mathbb{I}}$ is

$$\widehat{\psi}_{k,\mathbb{I}} = \arg\max_{\psi_{k,\mathbb{I}}} p(\mathcal{X}_k|\psi_{k,\mathbb{I}}). \tag{5.4}$$

This means that the training process is performed pitch-by-pitch and each $\mathcal{X}_k$ is processed independently. In this chapter, the index k is omitted for brevity. Hence, $\mathcal{X}_k$ is rewritten as $\mathcal{X} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^I\}$ and $\psi_{k,\mathbb{I}}$ is rewritten as $\psi_{\mathbb{I}}$.

Each tone $\mathbf{x}^i$ is represented by its PM $\widehat{\mathbf{x}}^i$. Adding the instance index i to (3.31) and (3.33), we rewrite PM in (3.31) and (3.33) into

$$\widehat{\mathrm{x}}^i(t_n) = \sum_{m=1}^{M} a(t_n; c^i, \varphi_m) \cdot \cos(2\pi f_m t_n + \phi_m) \tag{5.5}$$

and

$$a(t_n; c^i, \varphi_m) \;=\; b_m \left(c^i\right)^{d_m} \zeta_m \left(\exp\{-\lambda_m t_n\} - \exp\{-\gamma_m t_n\}\right) \tag{5.6}$$

where

$$\zeta_m = \left[\left(\frac{\lambda_m}{\gamma_m} \right)^{\frac{\lambda_m}{\gamma_m - \lambda_m}} - \left(\frac{\lambda_m}{\gamma_m} \right)^{\frac{\gamma_m}{\gamma_m - \lambda_m}} \right]^{-1} . \tag{5.7}$$

The definition of each parameter is restated for convenience. The function $a(\cdot)$ represents the envelope of the partials. The variable t_n is the time in second, n is the time index and $n = 0, 1, \ldots, N^i - 1$ where 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. The variables φ_m are the envelope parameters of the mth partial. They include the relative amplitude b_m , the decay rate $\lambda_{k,m}$, the rising rate $\gamma_{k,m}$ and the control of the intensity factor $d_{k,m}$. Thus, the variables φ_m can be written as $\varphi_m = \{b_m, d_m, \lambda_m, \gamma_m\}$. The term $\zeta_{k,m}$ is the normalization coefficient governed by λ_m and γ_m. The variables f_m and ϕ_m are the frequency and the phase of the mth partial respectively. The index m is from 1 to M where M is the number of partials given in Section 3.4.

Following (3.34), the observed tone $\mathbf{x}^i$ and the estimated tone $\widehat{\mathbf{x}}^i$ are related by

$$\mathrm{x}^i(t_n) \;=\; \widehat{\mathrm{x}}^i(t) + \epsilon^i(t_n) \tag{5.8}$$

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 (3.34) 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 such as the algorithm in [75] or manually annotated via a graphical interface (such as [15]). Then $\mathbf{x}^i$ can be adjusted to start from the time zero.

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

variances $\boldsymbol{\sigma}_\epsilon^2$ where $\boldsymbol{\sigma}_\epsilon^2 = \{\sigma_{\epsilon^1}^2, \sigma_{\epsilon^2}^2, \ldots, \sigma_{\epsilon^I}^2\}$. The likelihood $p(\mathcal{X}|\boldsymbol{\psi}_{\mathbb{I}}, \boldsymbol{\sigma}_\epsilon^2)$ is expressed in the form

$$\begin{aligned} p(\mathcal{X}|\boldsymbol{\psi}_{\mathbb{I}}, \boldsymbol{\sigma}_\epsilon^2) &= \prod_{i=1}^{I} p(\mathbf{x}^i|\boldsymbol{\psi}^i, \sigma_{\epsilon^i}^2) \\ &= \prod_{i=1}^{I} \frac{1}{\left(2\pi\sigma_{\epsilon^i}^2\right)^{N^i/2}} \exp\left\{-\frac{1}{2\sigma_{\epsilon^i}^2} \parallel \mathbf{x}^i - \widehat{\mathbf{x}}^i \parallel^2\right\}. \end{aligned} \tag{5.9}$$

The goal of the training stage is to find the optimal solution $\widehat{\boldsymbol{\psi}}_{\mathbb{I}}$. For efficient computation, the maximum likelihood solution of $\boldsymbol{\psi}_{\mathbb{I}}$ will be approximated by the weighted least-squares solution. The noise variance $\boldsymbol{\sigma}_\epsilon^2$ is replaced by a fixed value which will be estimated before finding $\widehat{\boldsymbol{\psi}}_{\mathbb{I}}$. The details will be explained in Section 5.5. As mentioned in the beginning of this chapter, PM is nonlinear. 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. Estimate the frequencies of the partials for each $\mathbf{x}^i$. (Section 5.2)

2. Given the estimated frequencies, extract the partials from each $\mathbf{x}^i$ by using GM. (Section 5.3)

3. Given the extracted partials, find the initial guess of $\boldsymbol{\psi}_{\mathbb{I}}$ for PM. (Section 5.4)

4. Given the initial guess of $\boldsymbol{\psi}_{\mathbb{I}}$, find the optimal solution $\widehat{\boldsymbol{\psi}}_{\mathbb{I}}$ for PM. (Section 5.5)

The procedures are summarized in Figure 5.1. The parameters will be defined and explained in the later sections.

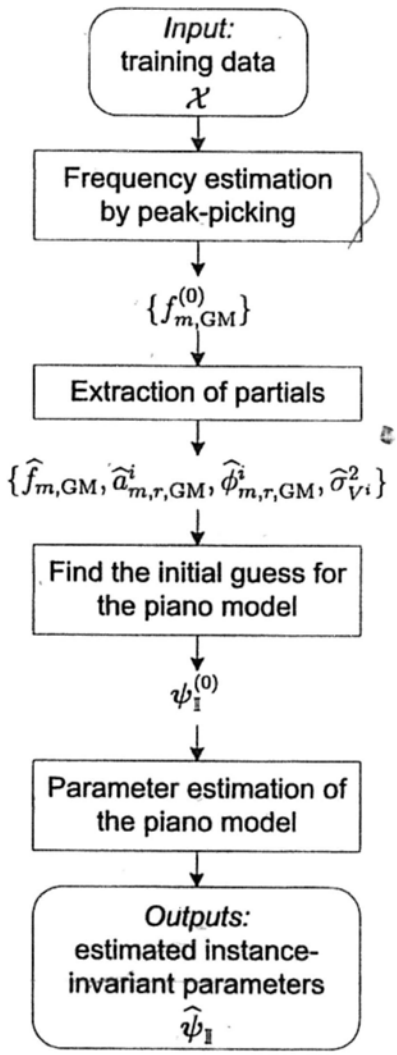

Figure 5.1: The system flow of the training stage.

5.2 Frequency estimation by peak-picking

The method of extracting partials in each training tone, which will be discussed in the next section, starts with an initial guess of the partial frequencies. This initial guess can be found by using the frequency estimation method in Section 3.4. In Section 3.4, the partial frequencies are estimated by picking the peaks in the frequency spectrum and locating the frequencies of the peaks. Given an isolated tone $\mathbf{x}^i$ in the training data $\mathcal{X}$, we first find the frequency spectrum by discrete Fourier transform (DFT), the peaks are chosen by the iterative method described in Section 3.4. The locations of the peaks are a set of frequencies $\{f^i_{m,\mathrm{PP}}\}$ where $m = 1, 2, \ldots, M$ and M is the number of partials which has been determined in Section 3.4.

The initial guess of a partial frequency for extracting a partial is the average of the frequency from peak-picking of all instances. This gives the initial guess in the form

$$f^{(0)}_{m,\mathrm{GM}} = \frac{1}{I} f^i_{m,\mathrm{PP}} \tag{5.10}$$

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

5.3 Extraction of partials with the General Model

The traditional General Model (GM) in Section 3.2 can be used to extract the partials of the isolated tones in the training data. Based on the extracted partials, we can find the initial guess of each partial for the Piano Model (PM) independently. This section will explain how to extract the partials from the training data with GM. Finding the initial guess for PM will be discussed in the next section.

In the training data $\mathcal{X}$, there are multiple instances of isolated tones with

the same pitch. As frequencies of the partials are the invariant PM parameters in PM, frequencies are also modeled as invariant PM parameters in GM. This means that the instances in $\mathcal{X}$ have the same values of frequencies so they share the same frequency matrix $\mathbf{H}$ in (3.26). Based on the notation in (3.26), we introduce the instance index i and rewrite (3.26) into

$$\widehat{\mathbf{X}}^i = \mathbf{H}\mathbf{G}^i \tag{5.11}$$

where $\widehat{\mathbf{X}}^i$ is an $L \times R^i$ matrix, L is the window length, and R^i is the number of frames for the ith instance. The matrix $\mathbf{H}$ is the frequency matrix defined in (3.9). The matrix $\mathbf{G}^i$ is the amplitude matrix of the ith instance and the size is $2M \times R^i$ where M is the number of partials. Following (3.4), an element in $\widehat{\mathbf{X}}^i$ is in the form

$$\widehat{X}^i[l,r] = \sum_{m=1}^{M} w[l]\left(\alpha^i_{m,r}\cos(2\pi f_m t_l) + \beta^i_{m,r}\sin(2\pi f_m t_l)\right). \tag{5.12}$$

The matrix $\widehat{\mathbf{X}}^i$ can also be expressed as the concatenation of the column vectors $\widehat{\mathbf{x}}^i_r$ so that

$$\widehat{\mathbf{X}}^i = \left[\widehat{\mathbf{x}}^i_1\ \widehat{\mathbf{x}}^i_2\ \cdots\ \widehat{\mathbf{x}}^i_{R^i}\right] \tag{5.13}$$

where $\widehat{\mathbf{x}}^i_r$ is rth column of $\widehat{\mathbf{X}}^i$.

The amplitude matrix $\mathbf{G}^i$ is written as

$$\mathbf{G}^i = \left[\begin{array}{ccccc} \alpha^i_{1,1} & \cdots & \alpha^i_{1,r} & \cdots & \alpha^i_{1,R^i} \\ \vdots & & \vdots & & \vdots \\ \alpha^i_{m,1} & \cdots & \alpha^i_{m,r} & \cdots & \alpha^i_{m,R^i} \\ \vdots & & \vdots & & \vdots \\ \alpha^i_{M,1} & \cdots & \alpha^i_{M,r} & \cdots & \alpha^i_{M,R^i} \\ \hline \beta^i_{1,1} & \cdots & \beta^i_{1,r} & \cdots & \beta^i_{1,R^i} \\ \vdots & & \vdots & & \vdots \\ \beta^i_{m,1} & \cdots & \beta^i_{m,r} & \cdots & \beta^i_{m,R^i} \\ \vdots & & \vdots & & \vdots \\ \beta^i_{M,1} & \cdots & \beta^i_{M,r} & \cdots & \beta^i_{M,R^i} \end{array}\right] \tag{5.14}$$

and $\mathbf{G}^i$ can be viewed as the concatenation of the column vectors $\mathbf{g}^i_r$ so that

$$\mathbf{G}^i = [\widehat{\mathbf{g}}^i_1 \ \widehat{\mathbf{g}}^i_2 \ \cdots \ \widehat{\mathbf{g}}^i_{R^i}] \tag{5.15}$$

where $\widehat{\mathbf{g}}^i_r$ is rth column of $\mathbf{G}^i$.

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

$$\mathbf{X}^i = \widehat{\mathbf{X}}^i + \mathbf{V}^i \tag{5.16}$$

where each element in $\mathbf{V}^i$ is the zero-mean Gaussian noise with the variance $\sigma^2_{V^i}$. Note that in (3.21), the noise variance can be different from frames. In applying GM in extracting the partials in this section, the noise variance $\sigma^2_{V^i}$ is the same for all frames for simplicity, but each instance has its own noise variance.

All instances in (5.11) can be written as

$$\widehat{\mathbf{X}} = \mathbf{HG} \tag{5.17}$$

where

$$\widehat{\mathbf{X}} = \left[\widehat{\mathbf{X}}^1\ \widehat{\mathbf{X}}^2\ \cdots\ \widehat{\mathbf{X}}^I\right] \tag{5.18}$$

$$= \left[\widehat{\mathbf{x}}^1_1\ \widehat{\mathbf{x}}^1_2\ \cdots\ \widehat{\mathbf{x}}^1_{R^1}\middle|\widehat{\mathbf{x}}^2_1\ \widehat{\mathbf{x}}^2_2\ \cdots\ \widehat{\mathbf{x}}^2_{R^2}\middle|\ \cdots\ \middle|\widehat{\mathbf{x}}^I_1\ \widehat{\mathbf{x}}^I_2\ \cdots\widehat{\mathbf{x}}^I_{R^I}\right] \tag{5.19}$$

and

$$\mathbf{G} = \left[\mathbf{G}^1\ \mathbf{G}^2\ \cdots\ \mathbf{G}^I\right]$$

$$= \left[\widehat{\mathbf{g}}^1_1\ \widehat{\mathbf{g}}^1_2\ \cdots\ \widehat{\mathbf{g}}^1_{R^1}\middle|\widehat{\mathbf{g}}^2_1\ \widehat{\mathbf{g}}^2_2\ \cdots\ \widehat{\mathbf{g}}^2_{R^2}\middle|\ \cdots\ \middle|\widehat{\mathbf{g}}^I_1\ \widehat{\mathbf{g}}^I_2\ \cdots\widehat{\mathbf{g}}^I_{R^I}\right].$$

The size of the matrix $\widehat{\mathbf{X}}$ is $L \times R$ and that of the matrix $\mathbf{G}$ is $2M \times R$ where $R = \sum_{i=1}^{I} R^i$. The matrix $\widehat{\mathbf{X}}$ is governed by $\boldsymbol{\theta}_{\mathcal{X}} = \{\mathbf{f}, \mathbf{G}\}$ where $\mathbf{f}$ is the frequency vector and $\mathbf{f} = [f_1\ f_2\ \cdots\ f_M]^{\mathrm{T}}$. The frequency matrix $\mathbf{H}$ depends on $\mathbf{f}$. The noise variances are grouped into $\boldsymbol{\sigma}_V^2 = [\sigma_{V^1}^2\ \sigma_{V^2}^2\ \cdots\ \sigma_{V^I}^2]^{\mathrm{T}}$.

The goal of the extraction of partials is to estimate $\mathbf{f}$, $\mathbf{G}$ and $\boldsymbol{\sigma}_V^2$. Weighted least-squares method is used to estimate these parameters. The weights are the inverse of the noise variances $\sigma_{V^i}^2$. The objective function to be minimized is written as

$$E_{\mathrm{GM}}(\mathbf{f}, \mathbf{G}, \boldsymbol{\sigma}_V^2) = \sum_{i=1}^{I} \frac{1}{\sigma_{V^i}^2} \left|\left|\mathbf{X}^i - \widehat{\mathbf{X}}^i\right|\right|_F^2 \tag{5.20}$$

$$= \sum_{i=1}^{I} \frac{1}{\sigma_{V^i}^2} \left|\left|\mathbf{X}^i - \mathbf{H}\mathbf{G}^i\right|\right|_F^2 \tag{5.21}$$

where $||\cdot||_F$ denotes the Frobenius norm. For an $m \times n$ matrix $\mathbf{A}$, the Frobenius norm of $\mathbf{A}$ is defined as

$$||\mathbf{A}||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} (A[i,j])^2}. \tag{5.22}$$

Then (5.20) can also be written as

$$E_{\mathrm{GM}}(\mathbf{f}, \mathbf{G}, \boldsymbol{\sigma}_V^2) = \sum_{i=1}^{I} \sum_{l=0}^{L-1} \sum_{r=1}^{R^i} \frac{1}{\sigma_{V^i}^2} \left(X^i[l,r] - \widehat{X}^i[l,r] \right)^2 . \tag{5.23}$$

The summation operation in (5.20) can be expressed in matrix form. Let $\boldsymbol{\Sigma}_V$ be the covariance matrix so that

$$\boldsymbol{\Sigma}_V = \mathrm{diag}(\sigma_{V^1}^2 \mathbf{1}_{LR^1}, \sigma_{V^2}^2 \mathbf{1}_{LR^2}, \ldots, \sigma_{V^I}^2 \mathbf{1}_{LR^I}) \tag{5.24}$$

where $\mathbf{1}_{LR^i}$ denotes the LR^i-dimensional column vector filled with 1's. Then (5.20) can be presented as

$$E_{\mathrm{GM}}(\mathbf{f}, \mathbf{G}, \boldsymbol{\sigma}_V^2) = \left|\left| \boldsymbol{\Sigma}_V^{-1/2} \left(\mathbf{X}_{\mathrm{vec}} - \widehat{\mathbf{X}}_{\mathrm{vec}}(\mathbf{f}) \right) \right|\right|^2 . \tag{5.25}$$

In [25], an iterative least-squares scheme is developed to alternatively update the frequencies and amplitudes of GM for one single frame. Based on this scheme, we propose a scheme to handle the frames of all instances together by using iterative-reweighted least-squares[17, 11]. Here are the procedures summarized in Figure 5.2:

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

The iterative update starts with the input frequencies $\mathbf{f}_{\mathrm{GM}}^{(0)}$ found in (5.10) which are estimated by the peak-picking method described in Section 5.2. We find that 100 iterations are good for convergence. In the followings, each step will be discussed in details.

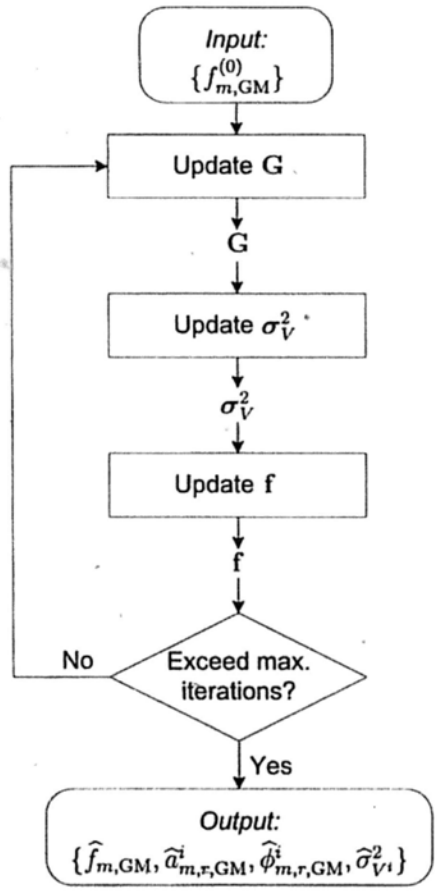

Figure 5.2: The procedures for extracting partials with GM.

5.3.1 Step 1: update the amplitude matrix G

In Step 1, the frequency matrix $\mathbf{H}$ is calculated from $\mathbf{f}$ by (3.8). Given $\mathbf{H}$ and the observed tones $\mathbf{X}$, GM becomes a linear model. Then the solution to (5.20) for updating $\mathbf{G}$ is

$$\mathbf{G} \leftarrow (\mathbf{H}^{\mathrm{T}}\mathbf{H})\mathbf{H}^{\mathrm{T}}\mathbf{X}. \tag{5.26}$$

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

5.3.2 Step 2: update the noise variances $\boldsymbol{\sigma}_V^2$

Given the updated $\mathbf{G}$ in Step 1, the new estimate $\widehat{\mathbf{X}}$ can be calculated

$$\widehat{\mathbf{X}} \leftarrow \mathbf{HG}. \tag{5.27}$$

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

$$\sigma_{V^i}^2 \leftarrow \frac{1}{LR^i}\left\|\mathbf{X}^i - \widehat{\mathbf{X}}^i\right\|_F^2. \tag{5.28}$$

5.3.3 Step 3: update the frequencies f

Given the updated $\mathbf{G}$ and $\boldsymbol{\sigma}_V^2$, the aim of Step 3 is to update the frequency vector $\mathbf{f}$. However, GM is nonlinear with $\mathbf{f}$. The nonlinear GM model can be linearized by using Taylor's expansion. In [25, 40], a single frame of GM, in which $\mathbf{X}$, $\widehat{\mathbf{X}}$ and $\mathbf{G}$ are only vectors instead of matrices, is linearized by Taylor's expansion. The Gauss-Newton method is used to update $\mathbf{f}$. Based on the work in [25, 40], we derive the update equation using the weighted least-squares for $\mathbf{f}$. The derivation involves two steps. The first step is to vectorize the matrix $\widehat{\mathbf{X}}$ and the second step is to linearize the vectorized $\widehat{\mathbf{X}}$.

To vectorize $\widehat{\mathbf{X}}$, we introduce the vec operator and the Kronecker product. For an $m \times n$ matrix $\mathbf{A} = [\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_n]$ where $\mathbf{a}_i$ is the ith column of $\mathbf{A}$, the

vec operator converts the matrix into a column vector by stacking the columns of $\mathbf{A}$,

$$\text{vec}(\mathbf{A}) = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix}, \tag{5.29}$$

to obtain an mn-dimensional column vector [50, p. 428]. For notational convenience, we denote $\text{vec}(\mathbf{A})$ by $\mathbf{A}_{\text{vec}}$.

Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ be a $p \times q$ matrix. The Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} A[1,1]\mathbf{B} & A[1,2]\mathbf{B} & \cdots & A[1,n]\mathbf{B} \\ A[2,1]\mathbf{B} & A[2,2]\mathbf{B} & \cdots & A[2,n]\mathbf{B} \\ \vdots & \vdots & & \vdots \\ A[m,1]\mathbf{B} & A[m,2]\mathbf{B} & \cdots & A[m,n]\mathbf{B} \end{bmatrix} \tag{5.30}$$

where $\mathbf{A} \otimes \mathbf{B}$ is an $mp \times nq$ matrix [50, p. 422].

The matrix equation $\widehat{\mathbf{X}} = \mathbf{HG}$ can be converted into a vector equation by the vec operator and the Kronecker product. We rewrite $\widehat{\mathbf{X}} = \mathbf{HG}$ into

$$\widehat{\mathbf{X}} = \mathbf{HGI}_I \tag{5.31}$$

where $\mathbf{I}_I$ is an $I \times I$ identity matrix. Vectoring both sides of (5.31) gives

$$\widehat{\mathbf{X}}_{\text{vec}} = \text{vec}(\mathbf{HGI}_I). \tag{5.32}$$

Using the identity in [50, p. 429], (5.32) can be written as a vector equation

$$\widehat{\mathbf{X}}_{\text{vec}} = (\mathbf{I}_I \otimes \mathbf{H})\mathbf{G}_{\text{vec}} \tag{5.33}$$

which is equivalent to

$$\underbrace{\begin{bmatrix} \widehat{\mathbf{x}}_1^1 \\ \widehat{\mathbf{x}}_2^1 \\ \vdots \\ \widehat{\mathbf{x}}_r^i \\ \vdots \\ \widehat{\mathbf{x}}_{R^I-1}^I \\ \widehat{\mathbf{x}}_{R^I}^I \end{bmatrix}}_{\widehat{\mathbf{X}}_{\text{vec}}} = \underbrace{\begin{bmatrix} \mathbf{H} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} & & & & & \mathbf{0} \\ \vdots & & \ddots & & & & \vdots \\ \mathbf{0} & & & \mathbf{H} & & & \mathbf{0} \\ \vdots & & & & \ddots & & \vdots \\ \mathbf{0} & & & & & \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{H} \end{bmatrix}}_{(\mathbf{I}_I \otimes \mathbf{H})} \underbrace{\begin{bmatrix} \mathbf{g}_1^1 \\ \mathbf{g}_2^1 \\ \vdots \\ \mathbf{g}_r^i \\ \vdots \\ \mathbf{g}_{R^I-1}^I \\ \mathbf{g}_{R^I}^I \end{bmatrix}}_{\mathbf{G}_{\text{vec}}} = \begin{bmatrix} \mathbf{H}\mathbf{g}_1^1 \\ \mathbf{H}\mathbf{g}_2^1 \\ \vdots \\ \mathbf{H}\mathbf{g}_r^i \\ \vdots \\ \mathbf{H}\mathbf{g}_{R^I-1}^I \\ \mathbf{H}\mathbf{g}_{R^I}^I \end{bmatrix}. \tag{5.34}$$

Note that each subvector $\widehat{\mathbf{x}}_r^i$ in $\widehat{\mathbf{X}}_{\text{vec}}$ is in the form

$$\widehat{\mathbf{x}}_r^i = \mathbf{H}\mathbf{g}_r^i \tag{5.35}$$

so (5.33) gives the correct result.

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

$$\widehat{\mathbf{X}}_{\text{vec}}(\mathbf{f}) \approx \widehat{X}_{\text{vec}}(\mathbf{f}^{\text{cur}}) + \mathbf{Z}(\mathbf{f}^{\text{cur}})\,(\mathbf{f} - \mathbf{f}^{\text{cur}}) \tag{5.36}$$

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

$$\begin{aligned} \mathbf{Z} &= \frac{\partial \widehat{\mathbf{X}}_{\text{vec}}}{\partial \mathbf{f}} \\ &= \left[\frac{\partial \widehat{\mathbf{x}}_1^1}{\partial \mathbf{f}} \; \frac{\partial \widehat{\mathbf{x}}_2^1}{\partial \mathbf{f}} \; \cdots \; \frac{\partial \widehat{\mathbf{x}}_r^i}{\partial \mathbf{f}} \; \cdots \; \frac{\partial \widehat{\mathbf{x}}_{R^I-1}^I}{\partial \mathbf{f}} \; \frac{\partial \widehat{\mathbf{x}}_{R^I}^I}{\partial \mathbf{f}} \right]^{\mathrm{T}} \\ &= \left[\mathbf{Z}_1^1 \; \mathbf{Z}_2^1 \; \cdots \; \mathbf{Z}_r^i \; \cdots \; \mathbf{Z}_{R^I-1}^I \; \mathbf{Z}_{R^I}^I \right]^{\mathrm{T}} \end{aligned} \tag{5.37}$$

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

$$\begin{aligned} Z_r^i[l,m] &= \frac{\partial \widehat{X}^i[l,r]}{\partial f_m} \\ &= \frac{\partial}{\partial f_m} w[l] \sum_{u=1}^{M} \left(\alpha_{u,r}^i \cos(2\pi f_u t_l) + \beta_{u,r}^i \sin(2\pi f_u t_l) \right) \\ &= 2\pi t_l w[l] \left(-\alpha_{m,r}^i \sin(2\pi f_m t_l) + \beta_{m,r}^i \cos(2\pi f_m t_l) \right). \end{aligned} \tag{5.38}$$

Then $\mathbf{Z}$ can be computed from (5.38).

Using the results in [40, pp. 226, 260], the update equation of $\mathbf{f}$ is

$$\mathbf{f} \leftarrow \mathbf{f} + (\mathbf{Z}^{\mathrm{T}} \Sigma_V^{-1} \mathbf{Z})^{-1} \mathbf{Z}^{\mathrm{T}} \Sigma_V^{-1} (\mathbf{X}_{\mathrm{vec}} - \widehat{\mathbf{X}}_{\mathrm{vec}}) \tag{5.39}$$

in which Gauss-Newton method is used.

5.3.4 Summary of the partial extraction

Here is the summary of all update equations. The update starts with the input frequencies $\mathbf{f}_{\mathrm{GM}}^{(0)}$ defined in (5.10).

1. Given $\mathbf{f}$, update $\mathbf{G}$. Calculate $\mathbf{H}$ from $\mathbf{f}$. Then

$$\mathbf{G} \leftarrow (\mathbf{H}^{\mathrm{T}}\mathbf{H})\mathbf{H}^{\mathrm{T}}\mathbf{X}. \tag{5.40}$$

2. Given $\mathbf{f}$ and $\mathbf{G}$, update $\widehat{\mathbf{X}}$ and σ_V^2

$$\widehat{\mathbf{X}} \leftarrow \mathbf{HG} \tag{5.41}$$

$$\sigma_{V^i}^2 \leftarrow \frac{1}{LR^i} \left\| \mathbf{X}^i - \widehat{\mathbf{X}}^i \right\|_F^2. \tag{5.42}$$

3. Given $\mathbf{G}$ and σ_V^2, update $\mathbf{f}$

$$\mathbf{f} \leftarrow \mathbf{f} + (\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}_V^{-1}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}_V^{-1}(\mathbf{X}_{\mathrm{vec}} - \widehat{\mathbf{X}}_{\mathrm{vec}}) \tag{5.43}$$

where $\mathbf{Z}$ is the Jacobian matrix and $\boldsymbol{\Sigma}_V$ is the covariance matrix defined in (6.25).

4. Repeats steps 1 to 3 for 100 iterations. The outputs of the extraction of partials are the frequencies $\widehat{\mathbf{f}}_{\mathrm{GM}}$, the amplitude matrix $\widehat{\mathbf{G}}$ and the noise variances $\widehat{\sigma}_V^2$.

5.4 Finding the initial guess for the Piano Model

The extraction of partials with GM in the previous section gives an estimate of the frequency $\widehat{f}_{m,\mathrm{GM}}$ and the amplitude vector $\widehat{\mathbf{g}}_{m,r}^i = [\widehat{\alpha}_{m,r}^i \ \widehat{\beta}_{m,r}^i]^{\mathrm{T}}$ of a partial for each frame. 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 $\widehat{f}_{m,\mathrm{GM}}$ while the initial guess for the envelope parameters φ_m in (3.31), the phase ϕ_m and the noise variance $\widehat{\sigma}_\epsilon^2$ will be discussed below. Before showing how the initial guess can be found, we first convert the estimated GM parameters into the values for PM. A partial can be expressed in the following forms

$$\widehat{\alpha}_{m,r}^i \cos(2\pi \widehat{f}_{m,\mathrm{GM}} t_l) + \widehat{\beta}_{m,r}^i \sin(2\pi \widehat{f}_{m,\mathrm{GM}} t_l) = \widehat{a}_{m,r,\mathrm{GM}}^i \cos(2\pi \widehat{f}_{m,\mathrm{GM}} t_l + \widehat{\phi}_{m,r,\mathrm{GM}}^i) \tag{5.44}$$

where $\widehat{a}_{m,r,\mathrm{GM}}^i$ is the amplitude in the form

$$\widehat{a}_{m,r,\mathrm{GM}}^i = \sqrt{(\widehat{\alpha}_{m,r}^i)^2 + (\widehat{\beta}_{m,r}^i)^2} \tag{5.45}$$

and $\widehat{\phi}^i_{m,r,\mathrm{GM}}$ is the phase in the form

$$\widehat{\phi}^i_{m,r,\mathrm{GM}} = \tan^{-1}\left(-\frac{\widehat{\beta}^i_{m,r}}{\widehat{\alpha}^i_{m,r}}\right). \tag{5.46}$$

5.4.1 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

$$t'_r = \left((r-1)\,D + 0.5L\right)/f_s \tag{5.47}$$

where D is the hop size in samples, L is the window length and f_s is the sampling frequency in Hz. Define the envelope function at the center of the rth frame as

$$a^i_{m,r}(\boldsymbol{\varphi}_m) = a(t'_r; c^i, \boldsymbol{\varphi}_m) \tag{5.48}$$

where $a(\cdot)$ is the envelope function defined in (3.31), and the intensity c^i, which is the peak amplitude of observed tone $\mathbf{x}^i$ in the time domain, is already known. Fitting $a^i_{m,r}(\boldsymbol{\varphi}_m)$ with $\widehat{a}^i_{m,r,\mathrm{GM}}$ using weighted least-squares, we have the objective function

$$E_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_m) = \sum_{i=1}^{I}\sum_{r=1}^{R^i} \frac{1}{\widehat{\sigma}^2_{V^i}}\left(\widehat{a}^i_{m,r,\mathrm{GM}} - a^i_{m,r}(\boldsymbol{\varphi}_m)\right)^2 \tag{5.49}$$

where the weights are the inverse of the variances $\widehat{\sigma}^2_{V^i}$. 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 discussed in Section 5.5.

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

The phase $\widehat{\phi}^i_{m,r,\mathrm{GM}}$ given in (5.46) is the initial phase at the beginning of a frame. In order to perform fitting as finding the initial guess $\varphi_m^{(0)}$, the phase $\widehat{\phi}^i_{m,r,\mathrm{GM}}$ is shifted to the center of a frame. The centered phase $\widehat{\phi}^{i,\mathrm{cent}}_{m,r,\mathrm{GM}}$ is in the form

$$\begin{aligned} \widehat{\phi}^{i,\mathrm{cent}}_{m,r,\mathrm{GM}} &= 2\pi \widehat{f}_m (L/(2f_s)) + \widehat{\phi}^i_{m,r,\mathrm{GM}} && (5.50) \\ &= \pi \widehat{f}_m L/f_s + \widehat{\phi}^i_{m,r,\mathrm{GM}}. && (5.51) \end{aligned}$$

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

$$E_\phi(\phi_m) = \sum_{i=1}^{I} \sum_{r=1}^{R^i} \frac{1}{\widehat{\sigma}^2_{V^i}} \left(\widehat{a}^i_{m,r,\mathrm{GM}} \cos(\widehat{\phi}^{i,\mathrm{cent}}_{m,r,\mathrm{GM}}) - \widehat{a}^i_{m,r,\mathrm{GM}} \cos\left(2\pi \widehat{f}_m t'_r + \phi_m\right) \right)^2 \tag{5.52}$$

where $\widehat{a}^i_{m,r,\mathrm{GM}} \cos(\widehat{\phi}^{i,\mathrm{cent}}_{m,r,\mathrm{GM}})$ is the partial generated by the GM estimate, and $\widehat{a}^i_{m,r,\mathrm{GM}} \cos\left(2\pi \widehat{f}_m t'_r + \phi_m\right)$ is the partial generated with RM. The weights are also the inverse of the variances $\widehat{\sigma}^2_{V^i}$. The objective function E_ϕ is again minimized by using the trust-region-reflective algorithm implemented in Matlab. There are 30 starting points randomly generated as E_ϕ is more sensitive to the starting points than E_φ. The best solution will be chosen as the initial guess $\phi_m^{(0)}$.

5.5 Parameter estimation of the Piano Model

As mentioned in Section 5.1, the maximum likelihood solution of ψ_{I} is approximated by the weighted least-squares solution for efficient computation. To work towards the weighted least-squares solution, the likelihood $p(\mathcal{X}|\psi_{\mathrm{I}}, \sigma_\epsilon^2)$

in (5.9) is expressed in the form of the negative log-likelihood

$$-\ln p(\mathcal{X}|\psi_{\mathrm{I}}, \sigma_{\epsilon}^2) = \sum_{i=1}^{I} \left(\frac{1}{2\sigma_{\epsilon^i}^2} \left|\left| \mathbf{x}^i - \widehat{\mathbf{x}}^i \right|\right|^2 + N^i \ln \sigma_{\epsilon^i} \right) + \frac{N^i I}{2} \ln(2\pi) . \quad (5.53)$$

Assuming that the noise variance in PM is directly proportional to that in GM for the extraction of partials, this means that

$$\sigma_{\epsilon^i}^2 \propto \sigma_{V^i}^2 \quad (5.54)$$

so the noise variance $\sigma_{\epsilon^i}^2$ in PM can be replaced by the noise variance $\widehat{\sigma}_{V^i}^2$ in GM outputted from the final iteration in (5.42). Note that the value of $\widehat{\sigma}_{V^i}^2$ is fixed for finding $\widehat{\psi}_{\mathrm{I}}$. Replacing $\sigma_{\epsilon^i}^2$ by $\widehat{\sigma}_{V^i}^2$ and omitting the constant terms, we can rewrite the negative log-likelihood in (5.53) into the following objective function

$$E_{\text{train}}(\psi_{\mathrm{I}}) = \sum_{i=1}^{I} \left(\frac{1}{2\widehat{\sigma}_{V^i}^2} \left|\left| \mathbf{x}^i - \widehat{\mathbf{x}}^i \right|\right|^2 \right) . \quad (5.55)$$

Given the initial guess $\psi_{\mathrm{I}}^{(0)} = \{\varphi_m^{(0)}, \widehat{f}_m^{(0)}, \phi_m^{(0)}\}$ for all m in PM, parameter estimation of PM can be done by minimizing the objective function E_{train} in (5.55) by using the trust-region-reflective algorithm implemented in Matlab. The outputs are the estimated invariant PM parameters $\widehat{\psi}_{\mathrm{I}}$. Hence, all of the PM parameters, including the invariant PM parameters, can be estimated. These parameters will be used in the source separation process explained in the next chapter.

Chapter 6

Source separation: parameter estimation

In the previous chapter, we explain how to estimate the invariant PM parameters $\widehat{\Psi}_{\mathrm{I}}$ from the training data $\boldsymbol{\mathcal{X}}$. The invariant PM parameters $\widehat{\Psi}_{\mathrm{I}}$ include the estimate of the envelope parameters $\widehat{\varphi}_k$, the frequencies $\widehat{\mathbf{f}}_k$ and the phases $\widehat{\phi}_k$ for each kth tone in the mixture $\mathbf{y}$. Given $\widehat{\Psi}_{\mathrm{I}}$ and $\mathbf{y}$, we perform the source separation in two stages as shown in Figure 6.1:

Stage 1: source separation with the Piano Model (PM). Given the invariant PM parameters $\widehat{\Psi}_{\mathrm{I}}$ in PM, the goal is to estimate the varying PM parameters $\Psi_{y,\mathrm{V}}$ for the mixture $\mathbf{y}$. The varying PM parameters $\Psi_{y,\mathrm{V}}$ 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 $\widehat{\Psi}_{y,\mathrm{V}}$ which maximize the likelihood function of $\Psi_{y,\mathrm{V}}$. With $\widehat{\Psi}_{\mathrm{I}}$ and $\widehat{\Psi}_{y,\mathrm{V}}$ in PM, the signals of each individual tone in the mixture can be reconstructed by using PM.

Stage 2: source separation with the General Model (GM). After separating the sources with PM, we use GM to further improve the separation quality. Given $\widehat{\Psi}_{\mathrm{I}}$ and $\widehat{\Psi}_{y,\mathrm{V}}$, the goal of Stage 2 is to estimate the GM parameters Θ_y. The GM parameters Θ_y consist of the amplitude matrix $\mathbf{G}$ and the frequencies $\mathbf{f}$. The prior distribution of the GM parameters Θ_y is learned from $\widehat{\Psi}_{\mathrm{I}}$ and $\widehat{\Psi}_{y,\mathrm{V}}$ to facilitate the process of source separation even for the case of overlapping

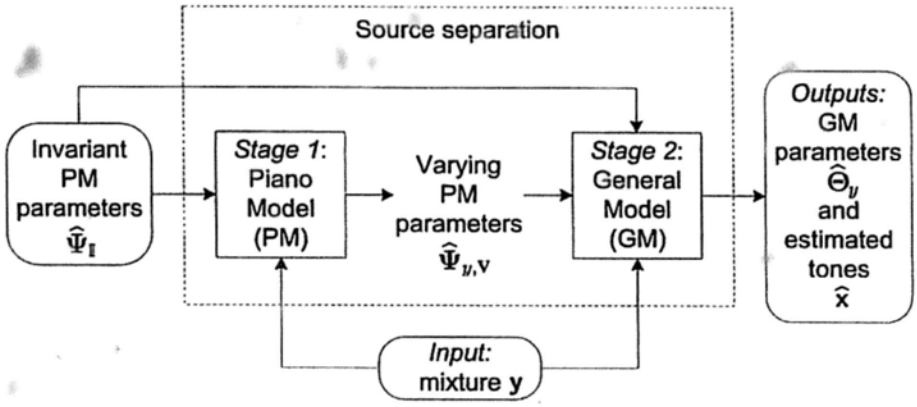

Figure 6.1: The two stages in source separation.

partials. The output of Stage 2 is the estimated GM parameters $\widehat{\Theta}_y$ which maximize the posterior distribution of Θ_y. The signals of each individual tone in the mixture can be reconstructed by using the overlap-and-add method [77] based on $\widehat{\Theta}_y$.

The following section will present Stage 1. Stage 2 will be explained in Section 6.2.

6.1 Stage 1: source separation with the Piano Model

In Chapter 5, it states that given the mixture $\mathbf{y}$ and the estimated invariant PM parameters $\widehat{\Psi}_{\mathbb{I}}$, the goal of source separation with PM is to find the most probable varying PM parameters $\widehat{\Psi}_{y,\mathbf{v}}$ that maximize the posterior of the varying PM parameters $p(\Psi_{y,\mathbf{v}}|\mathbf{y}, \widehat{\Psi}_{\mathbb{I}})$. By Bayes' theorem, the posterior can be rewritten as

$$p(\Psi_{y,\mathbf{v}}|\mathbf{y}, \widehat{\Psi}_{\mathbb{I}}) \propto p(\mathbf{y}|\Psi_{y,\mathbf{v}}, \widehat{\Psi}_{\mathbb{I}})p(\Psi_{y,\mathbf{v}}|\widehat{\Psi}_{\mathbb{I}}) \quad (6.1)$$

$$= p(\mathbf{y}|\Psi_{y,\mathbf{v}}, \widehat{\Psi}_{\mathbb{I}})p(\Psi_{y,\mathbf{v}}). \quad (6.2)$$

where $\Psi_{y,\mathbf{v}}$ are the varying PM parameters including the intensity c_k and the time shift τ_k for all tones k in the mixture $\mathbf{y}$. The prior of the invariant PM parameters $p(\Psi_{y,\mathbf{v}})$ reflects our prior knowledge of $\Psi_{y,\mathbf{v}}$. The values of $\Psi_{y,\mathbf{v}}$ can greatly vary from different playings. If we have little idea on suitable values for a parameter, we choose a very insensitive prior for $\Psi_{y,\mathbf{v}}$ as explained in Section 5.1. Then maximizing the posterior $p(\Psi_{y,\mathbf{v}}|\mathbf{y}, \widehat{\Psi}_{\mathbb{I}})$ is effectively equivalent to maximize the likelihood $p(\mathbf{y}|\Psi_{y,\mathbf{v}}, \widehat{\Psi}_{\mathbb{I}})$ so the most probable parameters $\widehat{\Psi}_{y,\mathbf{v}}$ are equivalent to the maximum likelihood solution. Hence, the goal of the source separation with PM becomes finding $\widehat{\Psi}_{y,\mathbf{v}}$ which maximize the likelihood $p(\mathbf{y}|\Psi_{y,\mathbf{v}}, \widehat{\Psi}_{\mathbb{I}})$ so that

$$\widehat{\Psi}_{y,\mathbf{v}} = \arg\max_{\Psi_{y,\mathbf{v}}} p(\mathbf{y}|\Psi_{y,\mathbf{v}}, \widehat{\Psi}_{\mathbb{I}}). \tag{6.3}$$

Recalling (3.35), the estimated mixture $\widehat{\mathbf{y}}$ is related to the observed mixture $\mathbf{y}$ as below

$$\mathrm{y}(t_n) = \widehat{\mathrm{y}}(t_n) + \epsilon(t_n) \tag{6.4}$$

where $\epsilon(t_n)$ is the zero-mean Gaussian noise with the variance σ_ϵ^2. Given a single noise variance σ_ϵ^2, the maximization of the likelihood is equivalent to the minimization of the least-squares errors [11]. Then the objective function for source separation with PM is

$$E_{\text{sep,PM}}(\Psi_{y,\mathbf{v}}) = ||\mathbf{y} - \widehat{\mathbf{y}}(\Psi_{y,\mathbf{v}})||^2. \tag{6.5}$$

The goal of source separation with PM is to find the varying PM parameters $\widehat{\Psi}_{y,\mathbf{v}}$ which minimize $E_{\text{sep,PM}}$ in (6.5). The objective function $E_{\text{sep,PM}}$ can be minimized by using the trust-region-reflective algorithm implemented in Matlab. There are 100 starting points randomly generated to minimize $E_{\text{sep,PM}}$. The best solution which gives the smallest $E_{\text{sep,PM}}$ will be chosen as the estimated varying PM parameters $\widehat{\Psi}_{y,\mathbf{v}}$.

6.2 Stage 2: source separation with the General Model

As presented in Section 4.4, the goal of source separation with the General Model (GM) is to find the MAP solution of the GM parameters $\widehat{\Theta}_y$ which maximizes the posterior distribution $p(\Theta_y|\mathbf{y}, \widehat{\Psi}_{y,\mathrm{V}}, \widehat{\Psi}_{\mathrm{I}}, \widehat{\sigma}_v^2)$ given the mixture $\mathbf{y}$. The GM parameters Θ_y include the amplitude matrix $\mathbf{G}$ and the frequencies $\mathbf{f}$. Finding the varying PM parameters $\widehat{\Psi}_{y,\mathrm{V}}$ and the invariant PM parameters $\widehat{\Psi}_{\mathrm{I}}$ have been discussed in Section 6.1 and Chapter 5 respectively. The estimation of the noise variances $\widehat{\sigma}_v^2$ will be covered in Section 6.2.2.1.

The process of source separation with GM is divided into the following two steps:

1. Estimate the hyperparameters, i.e. the noise variance $\widehat{\sigma}_v^2$ and the parameters in the prior distribution of Θ_y. (Section 6.2.2)

2. Given the hyperparameters, find the MAP solution $\widehat{\Theta}_y$ and reconstruct the signals of each individual tone by using the overlap-and-add method [77]. (Section 6.2.1)

Finding the MAP solution $\widehat{\Theta}_y$ (Step 2) will be explained in the next section.

6.2.1 Bayesian analysis for the General Model

In GM, the time-domain signal $\mathbf{y}$ is segmented into frames via the operation in (3.2). This gives the matrix $\mathbf{Y}$ defined in (3.22). The posterior distribution of Θ_y can be rewritten as $p(\Theta_y|\mathbf{Y}, \widehat{\Psi}_{y,\mathrm{V}}, \widehat{\Psi}_{\mathrm{I}}, \widehat{\sigma}_v^2)$. The GM parameters Θ_y include the amplitude matrix $\mathbf{G}$ and the frequencies $\mathbf{f}$ so the posterior distribution can also be expressed in the form of $p(\mathbf{f}, \mathbf{G}|\mathbf{Y}, \widehat{\Psi}_{y,\mathrm{V}}, \widehat{\Psi}_{\mathrm{I}}, \widehat{\sigma}_v^2)$. In Section 5.5, an iterative update scheme is designed to find the least-squares solution of the GM parameters. A similar iterative update scheme is also applied to find the MAP solution of the GM parameters:

1. Given $\mathbf{f}$, update $\mathbf{G}$.

2. Given $\mathbf{G}$, update $\mathbf{f}$.

3. Repeats steps 1 to 2 until convergence.

The iterative update starts with the input frequencies from the estimated frequencies of PM in (5.55). The frequencies of PM are close to those of GM. We find that 10 iterations are enough for convergence. Note that unlike the scheme for the least-squares solution, the noise variance is not updated in the scheme for the MAP solution because the noise variance can be estimated from the training data in advance. Estimation of the hyperparameters, including the noise variance, will be explained in Section 6.2.2. In the followings, the iterative update scheme will be discussed in details.

6.2.1.1 Step 1: update the amplitude matrix G

The amplitude matrix $\mathbf{G}$ is a concatenation of the amplitude vector $\mathbf{g}_r$ of each frame defined in (3.25). We will show that each $\mathbf{g}_r$ can be estimated independently. Given the estimated frequencies $\widehat{\mathbf{f}}$, now we rewrite the posterior distribution into $p(\mathbf{g}_r|\mathbf{y}_r, \widehat{\mathbf{f}}, \widehat{\mathbf{\Psi}}_{y,\mathbf{V}}, \widehat{\mathbf{\Psi}}_{\mathbb{I}}, \widehat{\sigma}^2_{v_r})$ where $\mathbf{y}_r$ and $\widehat{\sigma}^2_{v_r}$ are the mixture and the noise variance at the rth frame respectively. The goal of this step is to find the MAP solution $\widehat{\mathbf{g}}_r$ which maximizes the posterior distribution $p(\mathbf{g}_r|\mathbf{y}_r, \widehat{\mathbf{f}}, \widehat{\mathbf{\Psi}}_{y,\mathbf{V}}, \widehat{\mathbf{\Psi}}_{\mathbb{I}}, \widehat{\sigma}^2_{v_r})$ for each frame r.

By Bayes' theorem, the posterior distribution $p(\mathbf{g}_r|\mathbf{y}_r, \widehat{\mathbf{f}}, \widehat{\mathbf{\Psi}}_{y,\mathbf{V}}, \widehat{\mathbf{\Psi}}_{\mathbb{I}}, \widehat{\sigma}^2_{v_r})$ of $\mathbf{g}_r$ can be expressed in the form of

$$\begin{aligned} p(\mathbf{g}_r|\mathbf{y}_r, \widehat{\mathbf{f}}, \widehat{\mathbf{\Psi}}_{y,\mathbf{V}}, \widehat{\mathbf{\Psi}}_{\mathbb{I}}, \widehat{\sigma}^2_{v_r}) &\propto p(\mathbf{y}_r|\mathbf{g}_r, \widehat{\mathbf{f}}, \widehat{\mathbf{\Psi}}_{y,\mathbf{V}}, \widehat{\mathbf{\Psi}}_{\mathbb{I}}, \widehat{\sigma}^2_{v_r}) p(\mathbf{g}_r|\widehat{\mathbf{f}}, \widehat{\mathbf{\Psi}}_{y,\mathbf{V}}, \widehat{\mathbf{\Psi}}_{\mathbb{I}}, \widehat{\sigma}^2_{v_r}) \\ &= p(\mathbf{y}_r|\mathbf{g}_r, \widehat{\mathbf{f}}, \widehat{\sigma}^2_{v_r}) p(\mathbf{g}_r|\widehat{\mathbf{\Psi}}_{y,\mathbf{V}}, \widehat{\mathbf{\Psi}}_{\mathbb{I}}) \end{aligned} \tag{6.6}$$

where $p(\mathbf{y}_r|\mathbf{g}_r, \widehat{\mathbf{f}}, \widehat{\sigma}^2_{v_r})$ is the likelihood function and $p(\mathbf{g}_r|\widehat{\mathbf{\Psi}}_{y,\mathbf{V}}, \widehat{\mathbf{\Psi}}_{\mathbb{I}})$ is the prior distribution of $\mathbf{g}_r$.

As stated in (3.21), the estimated mixture $\widehat{\mathbf{y}}_r$ is related to the observed mixture $\mathbf{y}_r$ as below:

$$\begin{aligned} \mathbf{y}_r &= \widehat{\mathbf{y}}_r + \mathbf{v}_r & (6.7) \\ &= \mathbf{H}\mathbf{g}_r + \mathbf{v}_r. & (6.8) \end{aligned}$$

where $\mathbf{v}_r$ is the zero-mean Gaussian noise with the variance $\sigma^2_{\mathbf{v}_r}$. Following the result in (4.12), the likelihood function is

$$p(\mathbf{y}_r|\mathbf{g}_r, \widehat{\mathbf{f}}, \widehat{\sigma}^2_{v_r}) = \frac{1}{\left(2\pi\widehat{\sigma}^2_{v_r}\right)^{L/2}} \exp\left\{-\frac{1}{2\widehat{\sigma}^2_{v_r}} \parallel \mathbf{y}_r - \mathbf{H}\mathbf{g}_r \parallel^2\right\} \quad (6.9)$$

where $\mathbf{H}$ is the frequency matrix generated from $\widehat{\mathbf{f}}$ in (3.18).

The prior $p(\mathbf{g}_r|\widehat{\Psi}_{y,\mathbf{v}}, \widehat{\Psi}_{\mathbb{I}})$ represents the prior distribution of $\mathbf{g}_r$ conditioned on the PM parameters $\widehat{\Psi}_{y,\mathbf{v}}$ and $\widehat{\Psi}_{\mathbb{I}}$. It is modeled as a Gaussian so that

$$p(\mathbf{g}_r|\widehat{\Psi}_{y,\mathbf{v}}, \widehat{\Psi}_{\mathbb{I}}) = \mathcal{N}(\mathbf{g}_r|\widehat{\mu}_{g_r}, \widehat{\Sigma}_{g_r}) \quad (6.10)$$

where $\widehat{\mu}_{g_r}$ is the mean and $\widehat{\Sigma}_{g_r}$ is the covariance matrix. Both $\widehat{\mu}_{g_r}$ and $\widehat{\Sigma}_{g_r}$ depend on $\widehat{\Psi}_{y,\mathbf{v}}$ and $\widehat{\Psi}_{\mathbb{I}}$. Finding $\widehat{\mu}_{g_r}$ and $\widehat{\Sigma}_{g_r}$ will be discussed in Section 6.2.2.2. In this section, it is assumed that $\widehat{\sigma}^2_{v_r}$, $\widehat{\mu}_{g_r}$ and $\widehat{\Sigma}_{g_r}$ have been estimated and their values are known. The parameters $\widehat{\sigma}^2_{v_r}$, $\widehat{\mu}_{g_r}$ and $\widehat{\Sigma}_{g_r}$ are called *hyperparameters* in Bayesian analysis [10] because they themselves control the distribution of other parameters (i.e. $\mathbf{g}_r$). Note that each $\mathbf{g}_r$ has its own set of $\widehat{\mu}_{g_r}$ and $\widehat{\Sigma}_{g_r}$ 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 $p(\mathbf{g}|\mathbf{y})$ is also Gaussian as shown in [11, p. 153]. Following (4.28), the posterior is in the form

$$p(\mathbf{g}_r|\mathbf{y}_r) = \mathcal{N}(\mathbf{g}_r|\mathbf{m}_{g_r}, \mathbf{S}_{g_r}) \quad (6.11)$$

where

$$\mathbf{m}_{g_r} = \mathbf{S}_{g_r}\left(\widehat{\boldsymbol{\Sigma}}_{g_r}^{-1}\widehat{\boldsymbol{\mu}}_{g_r} + \widehat{\sigma}_{v_r}^{-2}\mathbf{H}^{\mathrm{T}}\mathbf{y}_r\right) \tag{6.12}$$

$$\mathbf{S}_{g_r} = \left(\widehat{\boldsymbol{\Sigma}}_{g_r}^{-1} + \widehat{\sigma}_{v_r}^{-2}\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}. \tag{6.13}$$

The mode of a Gaussian distribution coincides with its mean. Therefore, the MAP solution $\widehat{\mathbf{g}}_r$ is equal to the posterior mean $\mathbf{m}_{g_r}$. Substituting (6.13) into (6.12) gives

$$\widehat{\mathbf{g}}_r = \left(\widehat{\boldsymbol{\Sigma}}_{g_r}^{-1} + \widehat{\sigma}_{v_r}^{-2}\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\left(\widehat{\boldsymbol{\Sigma}}_{g_r}^{-1}\widehat{\boldsymbol{\mu}}_{g_r} + \widehat{\sigma}_{v_r}^{-2}\mathbf{H}^{\mathrm{T}}\mathbf{y}_r\right). \tag{6.14}$$

Note that as analyzed in Chapter 4, finding a unique $\widehat{\mathbf{g}}_r$ does not require that $\mathbf{H}$ is full column rank. The frequency matrix $\mathbf{H}$ is rank deficient if there are overlapping partials.

6.2.1.2 Step 2: update the frequencies f

Given the estimated amplitude matrix $\widehat{\mathbf{G}}$ concatenated from $\widehat{\mathbf{g}}_r$ in Step 1, the goal of Step 2 is to find the MAP solution $\widehat{\mathbf{f}}$ which maximizes the posterior distribution $p(\mathbf{f}|\mathbf{Y}, \widehat{\mathbf{G}}, \widehat{\boldsymbol{\Psi}}_{y,\mathrm{V}}, \widehat{\boldsymbol{\Psi}}_{\mathbb{I}}, \widehat{\boldsymbol{\sigma}}_v^2)$. The model $\widehat{\mathbf{Y}} = \mathbf{HG}$ is nonlinear with $\mathbf{f}$ where $\mathbf{f} = [\mathbf{f}_1 \ \cdots \ \mathbf{f}_K]^{\mathrm{T}}$ defined in (3.28). Following Section 5.3.3, we vectorize the matrix $\widehat{\mathbf{Y}}$ into $\widehat{\mathbf{Y}}_{\mathrm{vec}}$ and linearize the vectorized $\widehat{\mathbf{Y}}_{\mathrm{vec}}$ as in (5.36) by using Taylor's expansion so

$$\widehat{\mathbf{Y}}_{\mathrm{vec}}(\mathbf{f}) \approx \widehat{\mathbf{Y}}_{\mathrm{vec}}(\mathbf{f}^{\mathrm{cur}}) + \mathbf{Z}(\mathbf{f}^{\mathrm{cur}})\,(\mathbf{f} - \mathbf{f}^{\mathrm{cur}}) \tag{6.15}$$

where $\widehat{\mathbf{Y}}_{\mathrm{vec}}(\mathbf{f})$ is the estimate of the vectorized $\widehat{\mathbf{Y}}$ depending on $\mathbf{f}$ and $\mathbf{f}$ is the vector of new frequencies to be estimated, $\widehat{\mathbf{Y}}_{\mathrm{vec}}(\mathbf{f}^{\mathrm{cur}})$ is the estimate of the vectorized $\widehat{\mathbf{Y}}$ depending on $\mathbf{f}^{\mathrm{cur}}$ and $\mathbf{f}^{\mathrm{cur}}$ is the current estimate of the frequencies, and $\mathbf{Z}(\mathbf{f}^{\mathrm{cur}})$ is the Jacobian matrix evaluated at $\mathbf{f}^{\mathrm{cur}}$ and $\mathbf{Z}$ is an

$LR \times M$ matrix in the form

$$\mathbf{Z} = \frac{\partial \widehat{\mathbf{Y}}_{\text{vec}}}{\partial \mathbf{f}} \tag{6.16}$$

$$= \left[\frac{\partial \widehat{\mathbf{y}}_1}{\partial \mathbf{f}} \frac{\partial \widehat{\mathbf{y}}_2}{\partial \mathbf{f}} \cdots \frac{\partial \widehat{\mathbf{y}}_r}{\partial \mathbf{f}} \cdots \frac{\partial \widehat{\mathbf{y}}_{R-1}}{\partial \mathbf{f}} \frac{\partial \widehat{\mathbf{y}}_R}{\partial \mathbf{f}} \right]^{\mathrm{T}} \tag{6.17}$$

$$= [\mathbf{Z}_1 \, \mathbf{Z}_2 \cdots \mathbf{Z}_r \cdots \mathbf{Z}_{R-1} \, \mathbf{Z}_R]^{\mathrm{T}} \tag{6.18}$$

where $\mathbf{Z}_r$ is the $L \times M$ Jacobian matrix at rth frame for all tones and it is defined by

$$\mathbf{Z}_r = [\mathbf{Z}_{1,r} \, \mathbf{Z}_{2,r} \cdots \mathbf{Z}_{K,r}] \tag{6.19}$$

and $\mathbf{Z}_{k,r}$ is the $L \times M_k$ Jacobian matrix at rth frame for kth tone in the form

$$Z_{k,r}[l,m] = \frac{\partial \hat{y}_r[l]}{\partial f_{k,m}} \tag{6.20}$$

$$= \frac{\partial}{\partial f_{k,m}} \sum_{u=1}^{K} \sum_{v=1}^{M_u} w[l] \left(\alpha_{u,v,r} \cos(2\pi f_{u,v} t_l) + \beta_{u,v,r} \sin(2\pi f_{u,v} t_l) \right) \tag{6.21}$$

$$= 2\pi t_l w[l] \left(-\alpha_{k,m,r} \sin(2\pi f_{k,m} t_l) + \beta_{k,m,r} \cos(2\pi f_{k,m} t_l) \right). \tag{6.22}$$

Then $\mathbf{Z}$ can be computed from (6.22).

Following the prior distribution of $\mathbf{g}_r$ in (6.23), the prior distribution of $\mathbf{f}$ is also modeled as a Gaussian

$$p(\mathbf{f} | \widehat{\mathbf{\Psi}}_{y,\mathbf{V}}, \widehat{\mathbf{\Psi}}_{\mathbf{I}}) = \mathcal{N}(\mathbf{f} | \widehat{\boldsymbol{\mu}}_f, \widehat{\boldsymbol{\Sigma}}_f) \tag{6.23}$$

where $\widehat{\boldsymbol{\mu}}_f$ is the mean and $\widehat{\boldsymbol{\Sigma}}_f$ is the covariance matrix. Using the result of the Gauss-Newton method in (6.14), the MAP solution $\widehat{\mathbf{f}}$ is

$$\widehat{\mathbf{f}} = \left(\boldsymbol{\Sigma}_f^{-1} + \mathbf{Z}^{\mathrm{T}} \boldsymbol{\Sigma}_v^{-1} \mathbf{Z} \right)^{-1} \left(\boldsymbol{\Sigma}_f^{-1} \boldsymbol{\mu}_f + \mathbf{Z}^{\mathrm{T}} \boldsymbol{\Sigma}_v^{-1} \left(\mathbf{Y}_{\text{vec}} - \widehat{\mathbf{Y}}_{\text{vec}} + \mathbf{Z}\mathbf{f} \right) \right) \tag{6.24}$$

where Σ_v is the covariance matrix for all frames in the form

$$\Sigma_v = \text{diag}(\sigma_{v_1}^2 \mathbf{1}_L, \sigma_{v_2}^2 \mathbf{1}_L, \ldots, \sigma_{v_R}^2 \mathbf{1}_L) \tag{6.25}$$

and $\mathbf{1}_L$ denotes the L-dimensional column vector filled with 1's.

In the next section, we will show how to find $\widehat{\sigma}_{v_r}^2$, $\widehat{\mu}_{g_r}$, $\widehat{\Sigma}_{g_r}$, $\widehat{\mu}_f$ and $\widehat{\Sigma}_f$ which are crucial for resolving overlapping partials.

6.2.2 Estimation of the hyperparameters

In the evidence approximation introduced in [47, 48], the noise is modeled as a zero-mean Gaussian and the prior is a zero-mean univariate Gaussian. The noise variance and the prior variance are estimated by maximizing the evidence function. In our context, the evidence function is $p(\mathbf{y}_r|\sigma_{v_r}^2, \Sigma_\mathbf{g}, \Sigma_f)$. However, the mean of the priors in our case is not zero which will be shown shortly. This means the optimization technique for maximizing the evidence function in [47, 48] cannot be directly applied. In our case, we have the training data which are isolated tones. Instead of using the approach of maximizing the evidence, we will make use of the training data to estimate the hyperparameters $\widehat{\sigma}_{v_r}^2$, $\widehat{\mu}_{g_r}$, $\widehat{\Sigma}_{g_r}$, $\widehat{\mu}_f$ and $\widehat{\Sigma}_f$.

6.2.2.1 Estimation of the noise variance $\sigma_{v_r}^2$

The noise variance $\sigma_{v_r}^2$ is the variance of the zero-mean Gaussian noise $\mathbf{v}_r$ so that

$$\mathbf{y}_r = \widehat{\mathbf{y}}_r + \mathbf{v}_r \tag{6.26}$$

where $\mathbf{y}_r$ is the observed mixture and $\widehat{\mathbf{y}}_r$ is the estimated mixture. To estimate $\sigma_{v_r}^2$ from $\mathbf{y}_r$ with the use of the training data $\mathcal{X}$, we model the noise variance of an isolated tone $\mathbf{x}_{k,r}$ of a frame is directly proportional to the signal power.

This gives

$$\sigma^2_{v_{k,r}} = \bar{\sigma}^2_{v_k} ||\mathbf{x}_{k,r}||^2 \tag{6.27}$$

where $\mathbf{x}_{k,r}$ is the isolated tone at rth frame with pitch p_k, the parameter $\sigma^2_{v_{k,r}}$ is the noise variance of $\mathbf{x}_{k,r}$, and $\bar{\sigma}^2_{v_k}$ is the proportionality constant for pitch p_k and it can be determined by the training data $\boldsymbol{\mathcal{X}}$. Let $\mathbf{x}^i_{k,r,\mathcal{X}}$ be an isolated tone in the training data $\boldsymbol{\mathcal{X}}$. The subscript $\mathcal{X}$ in $\mathbf{x}^i_{k,r,\mathcal{X}}$ denotes that $\mathbf{x}^i_{k,r,\mathcal{X}}$ is obtained from the training data. Then $\bar{\sigma}^2_{v_k}$ can be estimated by

$$\bar{\sigma}^2_{v_k} = \frac{1}{I_k RL} \sum_{i=1}^{I_k} \sum_{r=1}^{R^i} \sum_{l=0}^{L-1} \left(\frac{x^i_{k,r,\mathcal{X}}[l] - \hat{x}^i_{k,r,\mathcal{X}}[l]}{||\mathbf{x}^i_{k,r,\mathcal{X}}||} \right)^2 \tag{6.28}$$

where $\hat{\mathbf{x}}^i_{k,r,\mathcal{X}}$ is the estimate of $\mathbf{x}^i_{k,r,\mathcal{X}}$ and it can be found by using the iterative reweighted least-squares explained in Section 5.3.

For the mixture $\mathbf{y}_r$, the noise variance is

$$\begin{aligned} \sigma^2_{v_r} &= \sum_{k=1}^{K} \sigma^2_{v_{k,r,y}} \\ &= \sum_{k=1}^{K} \bar{\sigma}^2_{v_k} ||\mathbf{x}_{k,r,y}||^2 \end{aligned} \tag{6.29}$$

where $\mathbf{x}_{k,r,y}$ is the kth individual tone in the mixture at rth frame, and $\sigma^2_{v_{k,r,y}}$ is its noise variance. The subscript y in $\mathbf{x}_{k,r,y}$ denotes that $\mathbf{x}_{k,r,y}$ is the individual tone in the mixture $\mathbf{y}_r$. However, $\mathbf{x}_{k,r,y}$ is not known. In order to estimate $\sigma^2_{v_r}$, we approximate $||\mathbf{x}_{k,r,y}||^2$ into

$$||\mathbf{x}_{k,r,y}||^2 \approx \left(\frac{\widehat{c}_k}{\sum_{k=1}^{K} \widehat{c}_k} \right) ||\mathbf{y}_r||^2 \tag{6.30}$$

where $\widehat{c}_k$ is the estimated intensity in PM described in Section 6.1. The intensity $\widehat{c}_k$ determines the proportion of $||\mathbf{x}_{k,r,y}||^2$ in $||\mathbf{y}_r||^2$. Substituting (6.30) into (6.29), we obtain the estimate $\widehat{\sigma}^2_{v_r}$ of the noise variance in the mixture $\mathbf{y}_r$

in the form

$$\widehat{\sigma}_{v_r}^2 = \sum_{k=1}^{K} \left(\frac{\widehat{c}_k \bar{\sigma}_{v_k}^2}{\sum_{k=1}^{K} \widehat{c}_k} \right) ||\mathbf{y}_r||^2 . \tag{6.31}$$

6.2.2.2 Estimation of the prior

Prior distribution of the amplitude vector $\mathbf{g}_r$ The prior distribution $p(\mathbf{g}_r|\widehat{\boldsymbol{\mu}}_{g_r}, \widehat{\boldsymbol{\Sigma}}_{g_r})$ of $\mathbf{g}_r$ is 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 $\widehat{\boldsymbol{\Psi}}_{y,\mathrm{V}}$ and $\widehat{\boldsymbol{\Psi}}_{\mathrm{I}}$. The dependence can be defined by converting the PM parameters $\widehat{\boldsymbol{\Psi}}_{y,\mathrm{V}}$ and $\widehat{\boldsymbol{\Psi}}_{\mathrm{I}}$ into the GM parameters. The aim of the conversion is to find the GM parameters at the center of a frame by finding those values from the PM parameters. Let t'_r be the time at the center of the rth frame as in (5.47) so that

$$t'_r = ((r-1)\,D + 0.5L)\,/f_s \tag{6.32}$$

where D is the hop size in samples, L is the window length and f_s is the sampling frequency in Hz. Evaluating the envelope function of PM in (3.30) at the center of the rth frame, the estimated amplitude of the mth partial of the kth tone in the mixture $\mathbf{y}$ is

$$\widehat{a}_{k,m,r,y,\mathrm{PM}} = a(t'_r; \widehat{c}_k, \widehat{\varphi}_m) \tag{6.33}$$

where $\widehat{c}_k$ and $\widehat{\varphi}_m$ are included in $\widehat{\boldsymbol{\Psi}}_{y,\mathrm{V}}$ and $\widehat{\boldsymbol{\Psi}}_{\mathrm{I}}$ respectively.

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

$$\widehat{\phi}_{k,m,r,y,\mathrm{PM}} \;=\; 2\pi \widehat{f}_{k,m,\mathrm{PM}}\,(t'_r - \widehat{\tau}_k) + \widehat{\phi}_{k,m,\mathrm{PM}} \tag{6.34}$$

where the frequency $\widehat{f}_{k,m,y,\mathrm{PM}}$ and the phase $\widehat{\phi}_{k,m,y,\mathrm{PM}}$ are included in $\widehat{\boldsymbol{\Psi}}_{\mathrm{I}}$, and the time shift $\widehat{\tau}_k$ is included in $\widehat{\boldsymbol{\Psi}}_{\mathrm{I}}$. Then $\widehat{a}_{k,m,r,y,\mathrm{PM}}$ and $\widehat{\phi}_{k,m,r,y,\mathrm{PM}}$ can be

transformed into the amplitudes of sine and cosine:

$$\widehat{\alpha}_{k,m,r,y,\mathrm{PM}} = \widehat{a}_{k,m,r,y,\mathrm{PM}} \cos \widehat{\phi}_{k,m,r,y,\mathrm{PM}} \tag{6.35}$$

$$\widehat{\beta}_{k,m,r,y,\mathrm{PM}} = -\widehat{a}_{k,m,r,y,\mathrm{PM}} \sin \widehat{\phi}_{k,m,r,y,\mathrm{PM}}. \tag{6.36}$$

The mean $\widehat{\boldsymbol{\mu}}_{g_r}$ in the prior is assigned to be the estimate from PM so that

$$\widehat{\mu}_{\alpha_{k,m,r}} = \widehat{\alpha}_{k,m,r,y,\mathrm{PM}} \tag{6.37}$$

$$\widehat{\mu}_{\beta_{k,m,r}} = \widehat{\beta}_{k,m,r,y,\mathrm{PM}} \tag{6.38}$$

where $\widehat{\mu}_{\alpha_{k,m,r}}$ and $\widehat{\mu}_{\beta_{k,m,r}}$ are the elements in $\widehat{\boldsymbol{\mu}}_{g_r}$ which follow the definition of $\mathbf{g}_r$ in (3.19).

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 assumed to be a diagonal matrix of which the diagonal is filled with the variances $\widehat{\sigma}^2_{\alpha_{k,m,r}}$ and $\widehat{\sigma}^2_{\beta_{k,m,r}}$. The ordering of $\widehat{\sigma}^2_{\alpha_{k,m,r}}$ and $\widehat{\sigma}^2_{\beta_{k,m,r}}$ in the diagonal also follows the definition of $\mathbf{g}_r$ in (3.19). We model that the variances $\widehat{\sigma}^2_{\alpha_{k,m,r}}$ and $\widehat{\sigma}^2_{\beta_{k,m,r}}$ are identical and they are directly proportional to the power of the partial amplitude at the rth frame. This gives

$$\widehat{\sigma}^2_{\alpha_{k,m,r}} = \widehat{\sigma}^2_{\beta_{k,m,r}} \tag{6.39}$$

$$= \bar{\sigma}^2_{G_k} \left(\widehat{a}_{k,m,r,y,\mathrm{PM}}\right)^2 \tag{6.40}$$

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

Let $\widehat{\alpha}^i_{k,m,r,\mathcal{X},\mathrm{GM}}$ and $\widehat{\beta}^i_{k,m,r,\mathcal{X},\mathrm{GM}}$ be the amplitudes in GM for the training data $\boldsymbol{\mathcal{X}}$ and they have been estimated by using GM from Section 5.3. The subscript $\mathcal{X}$ in $\widehat{\alpha}^i_{k,m,r,\mathcal{X},\mathrm{GM}}$ and $\widehat{\beta}^i_{k,m,r,\mathcal{X},\mathrm{GM}}$ denotes that their values are obtained from the training data. Let $\widehat{\alpha}^i_{k,m,r,\mathcal{X},\mathrm{PM}}$ and $\widehat{\beta}^i_{k,m,r,\mathcal{X},\mathrm{PM}}$ be the amplitudes in

GM for $\boldsymbol{\mathcal{X}}$ but 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 (6.35) and (6.36). Let $\widehat{a}^i_{k,m,r,\mathcal{X},\mathrm{PM}}$ be the partial amplitude in PM then

$$\widehat{a}^i_{k,m,r,\mathcal{X},\mathrm{PM}} = \sqrt{\left(\widehat{\alpha}^i_{k,m,r,\mathcal{X},\mathrm{GM}}\right)^2 + \left(\widehat{\beta}^i_{k,m,r,\mathcal{X},\mathrm{GM}}\right)^2}. \tag{6.41}$$

Following (6.39) and (6.40), we can estimate $\bar{\sigma}^2_{\mathbf{G}_k}$ from $\boldsymbol{\mathcal{X}}$ by

$$\bar{\sigma}^2_{G_k} = \frac{1}{2I_k M_k R^i_k} \sum_{i=1}^{I_k} \sum_{m=1}^{M_k} \sum_{r=1}^{R^i_k} \left\{ \left(\frac{\widehat{\alpha}^i_{k,m,r,\mathcal{X},\mathrm{GM}} - \widehat{\alpha}^i_{k,m,r,\mathcal{X},\mathrm{PM}}}{\widehat{a}^i_{k,m,r,\mathcal{X},\mathrm{PM}}} \right)^2 + \left(\frac{\widehat{\beta}^i_{k,m,r,\mathcal{X},\mathrm{GM}} - \widehat{\beta}^i_{k,m,r,\mathcal{X},\mathrm{PM}}}{\widehat{a}^i_{k,m,r,\mathcal{X},\mathrm{PM}}} \right)^2 \right\}. \tag{6.42}$$

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, 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 4.5. Hence, overlapping partials can be resolved and higher quality of source separation can be obtained. It will be verified and explained in the experiments.

Prior distribution of frequencies f The prior distribution $p(\mathbf{f}|\widehat{\boldsymbol{\mu}}_f, \widehat{\boldsymbol{\Sigma}}_f)$ of $\mathbf{f}$ is 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 estimate of the frequencies in PM from $\widehat{\boldsymbol{\Psi}}_{\mathrm{I}}$ so that

$$\widehat{\mu}_{f_{k,m}} = \hat{f}_{k,m,\mathrm{PM}}$$

where $\widehat{\mu}_{f_{k,m}}$ are the elements in $\widehat{\boldsymbol{\mu}}_f$ which follows the definition of $\mathbf{f}$ in (3.28).

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 $\widehat{\sigma}^2_{f_{k,m}}$. The variance $\widehat{\sigma}^2_{f_{k,m}}$ is

modeled to be directly proportional to the square of the frequency in PM. This gives

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

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

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

where $\widehat{f}_{k,m,\mathcal{X},\mathrm{GM}}$ has been estimated by using GM from Section 5.3. Note that there is no subscript $\mathcal{X}$ in $\widehat{f}_{k,m,\mathrm{PM}}$ because $\widehat{f}_{k,m,\mathrm{PM}}$ are the invariant PM parameters so the training data and the mixture share the same set of $\widehat{f}_{k,m,\mathrm{PM}}$.

In summary, after estimating the hyperparameters $\widehat{\sigma}^2_{v_r}$, $\widehat{\boldsymbol{\mu}}_{g_r}$, $\widehat{\boldsymbol{\Sigma}}_{g_r}$, $\widehat{\boldsymbol{\mu}}_f$ and $\widehat{\boldsymbol{\Sigma}}_f$, we can find the MAP solution $\widehat{\boldsymbol{\Theta}}_y$ of GM discussed in Section 6.2.1. In the next chapter, experimental results will be presented to show the performance of the whole source separation process.

Chapter 7

Experiments

Experiments were performed to test the separation quality of our proposed Piano Model (PM) and the traditional General Model (GM). All data used in the experiments are real signals of piano tones and they are not synthetic. The databases of piano tones will be described in Section 7.1. The piano tones were used to generate mixtures from musical chords which include octaves. The generation of mixtures will be discussed in Section 7.2. In 7.3, 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.

7.1 Databases of piano tones

Piano tones from four different pianos were used in our experiments. Three of the pianos are from the RWC musical instrument sound database [34] 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. The reason for creating our own database

is that each pitch in the RWC database was only played at three different levels of loudness (soft, medium and loud). In order to design our proposed Piano Model (PM) presented in Section 3.3, we created the database with the Yamaha upright piano which is a computer-controlled piano. Each pitch was played at 12 different levels of loudness to enable a detailed study of timbre change with different hitting strengths. The details of the recording setup will be described as follows.

During the recording session, both the top lid and the front face of the Yamaha upright piano were open as shown in Figure 7.1. The sound was recorded with four RØDE NT1000 condenser microphones placed approximately 20 cm above the keyboard and 18 cm in front of the piano strings. This close-miking setup reduced the effect of room acoustics. The microphones were connected to an RME Fireface 800 Audio Interface, which acted as a microphone preamp and an A/D converter, and transferred the signals to a PC digitally through a firewire cable. The signals were stored in WAV format. The sampling frequency was 44.1 kHz and the number of bits per sample was 24. All 88 keys in the piano were played and recorded. Each tone was played at 12 levels of loudness ranging from very soft to very loud and lasted for around 1 second. After listening, we chose the signals recorded by the microphone in front of the C3 piano strings for our monaural source separation experiments because of their balanced sound quality for a wide range of pitches. All tones, including our database and the RWC database, were downsampled to 11.025 kHz for faster processing.

Before performing our experiments, we aligned the instances of a pitch from the same piano in phase by using the cross-correlation method in the following steps:

1. The instance with the medium loudness was selected to be a reference. The onset of the instance was detected by the onset algorithm in [75] and was fine-tuned in our user interface developed in Matlab. This made the

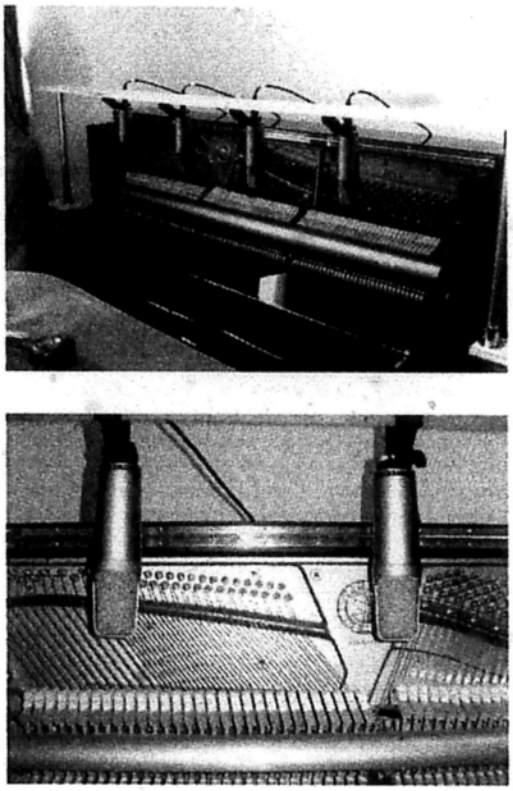

Figure 7.1: Recording setup of our recorded piano database.

instance to start at time zero.

2. Other instances of the pitch were aligned in phase by time shifting the instances to maximize the cross-correlation between the instances and the reference instance. This alignment also made all instances to start at time zero.

7.2 Generation of mixtures

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 [34]. The lists of all the piano pieces and mixtures are shown in Appendix C. 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 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 as shown in Figure 7.2. The 25 mixtures consist of 62 tones. There are 9 mixtures containing at least one pair of octaves. Two of them contain 2 pairs of octaves.

The procedures of generating a mixture is shown in Figure 7.3. 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 in Section 7.1 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.

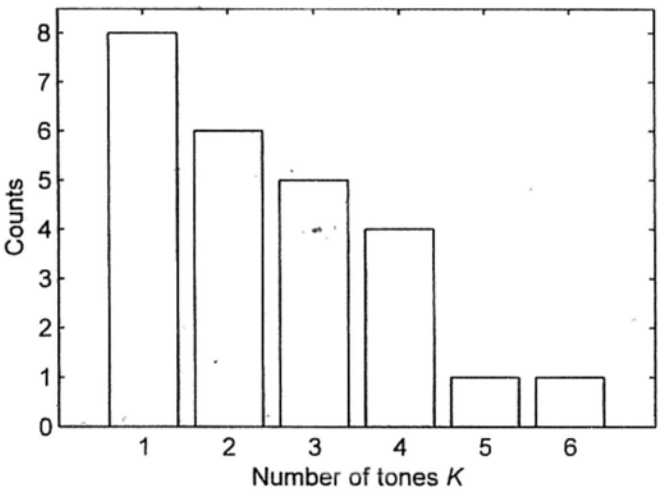

Figure 7.2: The counts of the mixtures with number of tones K for the experiments. The total number of mixtures is 25.

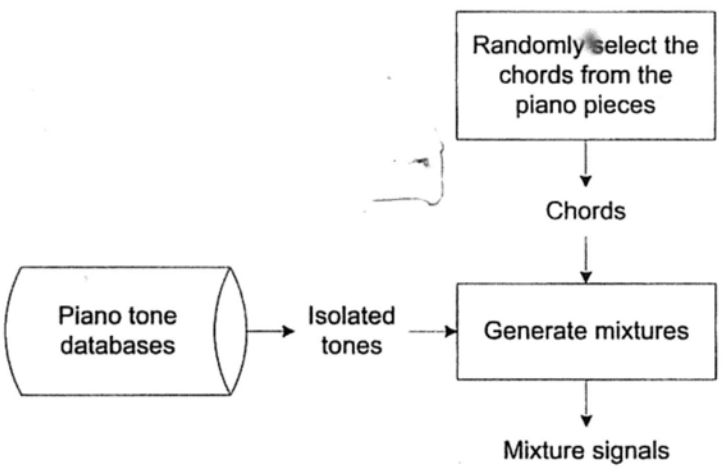

Figure 7.3: Generation of mixtures.

7.3 Results

7.3.1 Evaluation criteria

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

$$\mathrm{SNR} = 10 \log_{10} \frac{\sum_n \mathsf{x}(t_n)^2}{\sum_n (\mathsf{x}(t_n) - \widehat{\mathsf{x}}(t_n))^2} \tag{7.1}$$

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 either PM or GM. 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

$$\mathrm{ER}_c = \left| \frac{c_k - \widehat{c}_k}{c_k} \right| \tag{7.2}$$

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

$$\mathrm{Err}_\tau = |\tau_k - \widehat{\tau}_k| \tag{7.3}$$

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 7.2.

	$\overline{\text{SNR}}$ (dB)	
	PM	GM
All estimated individual tones	11.15	17.38

Table 7.1: The average SNR of all estimated individual tones of the 25 mixtures before mixing.

7.3.2 Evaluation on modeling quality

Before evaluating the separation quality, we first evaluate the modeling quality, i.e. the quality of PM and GM to represent an isolated tone before mixing. Both PM and GM were 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 of PM and GM before and after mixing.

The procedures of evaluation on the modeling quality is shown in Figure 7.4. For each mixture in the 25 mixtures described in Section 7.2, 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 proposed source separation system including both PM and GM. The parameter setting for GM is that the window length is 11.61 ms ($L = 128$) with 50% overlapping window. The effect of window length will be discussed in Section 7.3.3.2. The outputs of our system were the estimated tones reconstructed from PM and GM. The estimated tones were compared to the shifted tones to evaluate the modeling quality. If the parameters obtained in PM and GM are accurate, they can regenerate the original shifted tones in high quality. Table 7.1 shows the average SNR ($\overline{\text{SNR}}$) which reflects the modeling quality. The average SNR of GM is higher than that of PM. This is because GM is a more flexible model to represent piano tones.

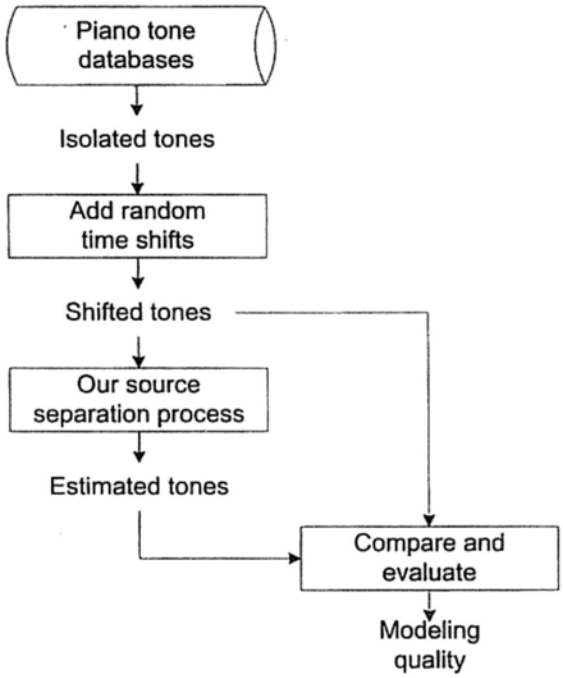

Figure 7.4: The procedures of evaluation on modeling quality.

7.3.3 Evaluation on separation quality

After evaluating the modeling quality, we evaluate the separation quality, i.e. the quality of PM and GM to separate a mixture into its individual tones. Both PM and GM were used to construct the estimated signals of the individual tones in the mixtures. Figure 7.5 illustrates the procedures of evaluation 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 parameter setting of GM was that the window length was set to 11.61 ms ($L = 128$) with 50% overlapping window. The number of instances I_k in the training data was two. This means that there were two isolated tones used in training for each individual tone in a mixture. The outputs of our system were the estimated tones reconstructed from PM and GM.

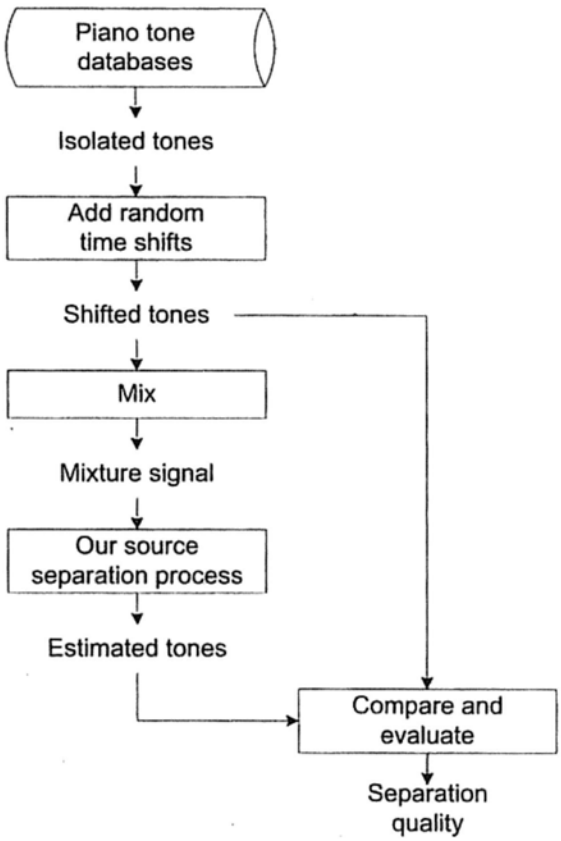

Figure 7.5: The procedures of evaluation on separation quality for a mixture.

The estimated tones were compared to the input shifted tones to evaluate the separation quality.

7.3.3.1 Overall analysis

For the 25 mixtures, the average SNRs of GM and PM are 13.51 dB and 10.88 dB respectively. The results are shown in Table 7.2. Both PM and GM are 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 against the number of tones K is plotted in Figure 7.6. The average SNR of PM decreases slowly when the number of tones K increases. The average SNR of GM decreases more rapidly but it rises significant for the case $K = 6$. The average SNR cannot completely illustrate the separation quality because high average SNR

	$\overline{\text{SNR}}$ (dB)		$\overline{\Delta\text{SNR}}$ (dB)	
	PM	GM	PM	GM
All mixtures	10.88	13.51	-	-
$2 \leq K \leq 6$	10.97	13.15	-0.31	-4.45
Upper tones in octaves	10.95	12.77	-0.37	-3.34

Table 7.2: The average SNR of the 25 mixtures. The number of tones in a mixture is denoted by K. The column of $\overline{\text{SNR}}$ is the average SNR in dB. The column of $\overline{\Delta\text{SNR}}$ is the average SNR difference between modeling and source separation.

may be due to high quality of the modeling. To evaluate the separation quality effectively, the average SNR difference is used. The SNR difference between the modeling benchmark in Section 7.3.2 and the separation is defined by

$$\Delta\text{SNR}=(\text{SNR from modeling}) - (\text{SNR from separation}) \tag{7.4}$$

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 7.2. The average SNR difference has a greater drop for GM than PM because PM is less sensitive to overlapping partials. When the number of tones K increases, the number of overlapping partials generally increases.

The average SNR difference against the number of tones K is plotted in Figure 7.6. The average SNR difference of PM decreases slowly when the number of tones K increases. The average SNR difference of GM decreases more rapidly but it rises significantly when $K = 6$. This may be because the PM parameters are more accurately estimated for the mixture of $K = 6$.

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 7.3. 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

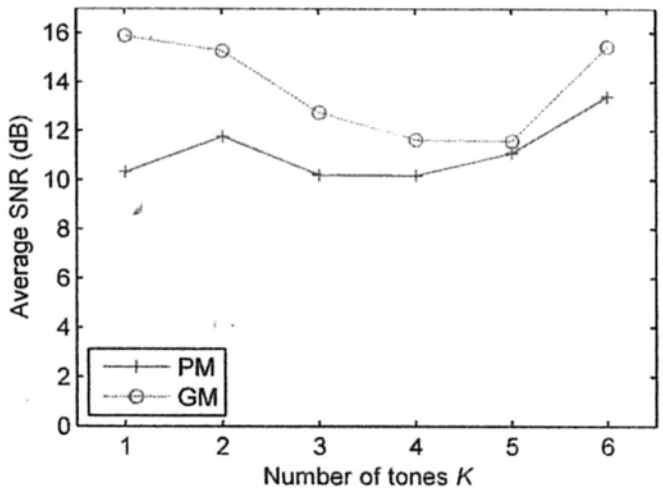

Figure 7.6: Average SNR against the number of tones K.

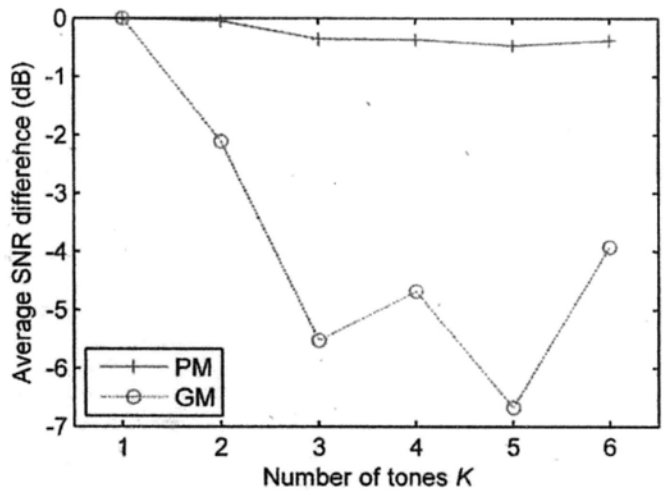

Figure 7.7: Average SNR difference against the number of tones K.

	Average absolute error ratio of intensity ER_c		
	Intensity c_k	Peak from PM	Peak from GM
$2 \leq K \leq 6$	0.074	0.222	0.130

Table 7.3: The average absolute error ratio of intensity ER_c

	Absolute error of the estimated time shift Err_τ in PM (ms)
$2 \leq K \leq 6$	3.16

Table 7.4: The absolute error of the estimated time shift Err_τ in PM.

tone in a mixture, the accuracy of $\widehat{c}_k$ is compared to the peak amplitude of the estimated tones from PM and GM. The average absolute error ratio of $\widehat{c}_k$ is lower than that of PM and GM. This is because the peak amplitude of the estimated tones from PM and GM depend on all estimated parameters. In the other hand, the estimation of $\widehat{c}_k$ is only based on the envelope function defined in (3.31) 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.

The average absolute error of the estimated time shift Err_τ in PM is shown in Table 7.4. The error is only 3.16 ms so the estimated time shift can give an accurate fine-tuned onset.

7.3.3.2 Effect of window length

The effect of window length in GM was studied by changing the window length in the source separation stage with GM. Four window lengths were tested including 5.80, 11.61, 23.22 and 46.44 ms which corresponding to L equal to 64, 128, 256 and 512 samples. The 50% overlapping window were applied in these 4 window lengths. The source separation stage with GM was inputted with the same PM parameters for the tests. The results are shown in Figure 7.8. The best window length from the results is 11.61 ms ($L = 128$).

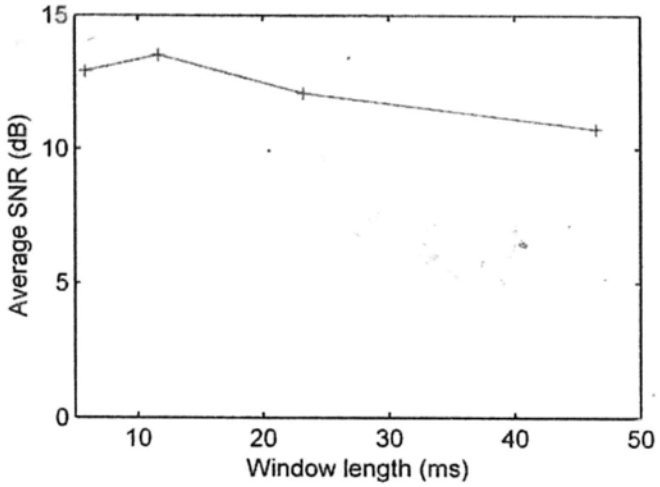

Figure 7.8: Average SNR against the window length.

7.3.3.3 Comparison with other method

In a recent system of monaural source separation in [46], Li, Woodruff and Wang compare their system (Li's system) to the systems in [71] and [56] for the mixtures with the number of tones K equal to 2. They show that Li's system significantly outperforms those two systems. Li's system is based on the principle of common amplitude modulation reviewed in Chapter 2. We compared Li's system to our proposed source system including PM and GM 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 7.5. Our system including both PM and GM perform better than Li's system for the average SNR of all mixtures. A significant improvement is in the octave cases as shown in the table. 7.9. 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 7.9. The average SNR of Li's system decreases more rapidly than our system. Our system can make use of the training data to give higher separation quality.

	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 7.5: Comparison of Li's system and our proposed system PM and GM.

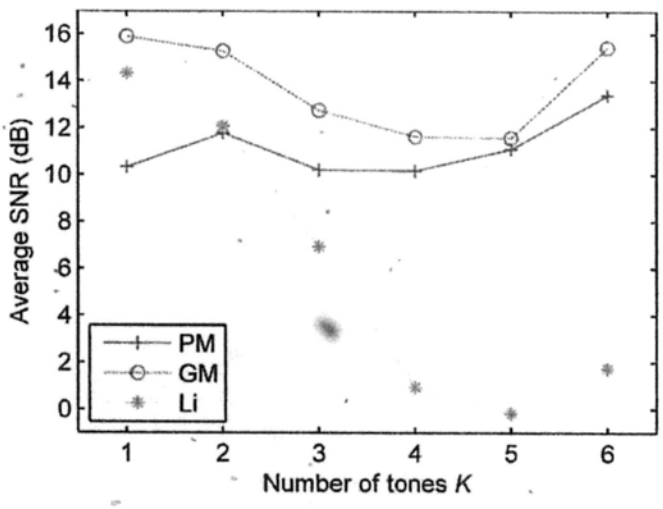

Figure 7.9: Average SNR against the number of tones K for our system and Li's system.

	Average computation time for one mixture (min)
Total (training and source separation)	11.61 (100%)
Training	4.43 (38.15%)
Source separation: PM	7.00 (60.24%)
Source separation: GM	0.19 (1.61%)

Table 7.6: Average computation for one mixture.

7.3.4 Computation time

The experiment in Section 7.3.3.1 was run in a Window Vista PC with an Intel Core2 Quad Q6600 2.4 GHz CPU and 2GB memory. All program codes were written in Matlab. The average computation time for one mixture is shown in Table 7.6. Source separation with PM is much slower than that with GM because the optimization in the source separation with PM is highly nonlinear. For Li's system, the average computation time for one mixture is only 1.38 second which is much lower than that in our system. However, the separation quality of our system is higher than that of Li's system.

Chapter 8

Conclusion and discussion

In this dissertation, we propose a Bayesian monaural source separation system to extract each individual tone from mixture signals of piano music performance. The system incorporates the following in its formulation: (1) the timing and pitch of music notes played in the piece from the target recording are provided, and (2) the isolated tones from the target recording are available. Based on this formulation, we have developed a model-based source separation system. In this research, two signal models based on sum-of-sinusoidal modeling are employed to represent piano tones: (1) We use a traditional General Model (GM), which is a variant of sinusoidal modeling, to represent a tone for high modeling quality but the model often fail for mixtures of tones because of overlapping partials. (2) We propose an instrument-specific model for the piano. Although the modeling quality is not as high as GM, it can resolve the overlapping partials and help to solve the source separation problem. (3) To benefit from the merits of both GM and our proposed Piano Model (PM), we use the hierarchical Bayesian framework to combine both models in the source separation process. Experiments show that our proposed system gives robust and accurate separations of mixtures and improves the separation quality significantly comparing to the previous work. The extension of our current system are discussed below.

If an isolated tone of a particular pitch is not available for training, the

training may be done by pitch-shifting neighboring isolated tones. The mixtures can also be separated by other source separation methods such as [46] which does not require isolated tones. However, this may give poor separation quality if there are a great number of overlapping partials. Hence, our future work is to extend our current system which is able to train PM with mixtures. Possible ways to achieve this goal are to find the MAP solution of GM by using other inference techniques. The techniques include the deterministic methods (e.g. variational approximation [67]) and the probabilistic methods (e.g. Markov chain Monte Carlo [62] and particle filtering [26]).

Another extension is to perform the task of source separation with the piano music signals played with pedaling. GM is able to resynthesize the isolated piano tones with pedaling. However, PM may not be able to give a very accurate estimate of the PM parameters because pedaling affects the time evolution of partials. A possible way to solve this problem is to use the full evidence approximation which learns the variance between GM and PM from the mixtures.

In our research, we only test our model on piano music signals. We may extend the model for other musical instruments, such as the Chinese musical instruments - the Chinese lute *Pipa* or the Chinese hammered dulcimer *Yangqin*. We can use the same approach to analyze these Chinese string-striking musical instruments. These two instruments are usually played without using any damper mechanism so it is expected that there is more than one tone sounding simultaneously. So the problem is very similar to that of a piano. This extension enhances the study, development and preservation of Chinese music tradition.

Appendix A

Notation

Symbol	Meaning
a	partial amplitude
B	inharmonicity factor
b	relative amplitude of a partial in the Piano Model (PM)
c	intensity of a tone in PM
D	hop size
d	control of the intensity significance in PM
E	cost function
f	frequency
$\mathbf{G}$	amplitude matrix in the General Model (GM)
$\mathbf{g}$	amplitude vector in GM
$\mathbf{H}$	frequency matrix in GM
I	number of instances
i	index for ith instance
K	number of tones in a mixture
k	index for kth tone
L	window length in samples
l	discrete time index in a windowed signal
M	number of partials
m	index for mth 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 rth frame
T	transpose of a matrix
t	continuous time value
w	window function
$\mathbf{X}$	matrix of frames of a tone
$\mathcal{X}$	isolated tones in the training data
$\mathbf{x}$	signal vector of a single frame of a tone
$\mathbf{x}$	signal vector of a tone in the time domain
$\mathbf{Y}$	matrix of frames of a mixture
$\mathbf{y}$	signal vector of a single frame of a mixture
$\mathbf{y}$	signal vector of a mixture in the time domain
$\mathbf{Z}$	Jacobian matrix
α	amplitude of cosine term
β	amplitude of sine term
γ	rising rate of a partial in PM
ϵ	noise in PM
ζ	normalization coefficient in PM
$\boldsymbol{\Theta}$	parameter set of a mixture in GM
$\boldsymbol{\theta}$	parameter set of a tone in GM
λ	decay rate of a partial in PM
μ	mean of a Gaussian distribution
$\boldsymbol{\Sigma}$	covariance matrix of a Gaussian distribution
σ	standard deviation of a Gaussian distribution
τ	continuous time shift
ϕ	phase

φ	envelope parameters in PM
Ψ	parameter set of a mixture in PM
$\Psi_{\mathbb{I}}$	invariant PM parameters
$\Psi_{y,\mathbb{V}}$	varying PM parameters of the mixture $\mathbf{y}$
ψ	parameter set of a tone in PM

Appendix B

Derivation

B.1 Derivation of the normalization coefficient in the Piano Model

The peak of the bi-exponential mixture $(\exp\{-\lambda_{k,m}t_n\} - \exp\{-\gamma_{k,m}t_n\})$ is normalized to by the normalization coefficient $\zeta_{k,m}$:

$$\zeta_{k,m}\left(\exp\{-\lambda_{k,m}t_n\} - \exp\{-\gamma_{k,m}t_n\}\right). \tag{B.1}$$

In this appendix, our goal is to find $\zeta_{k,m}$. Let

$$z_{k,m} = \exp\{-\lambda_{k,m}t_n\} - \exp\{-\gamma_{k,m}t_n\}. \tag{B.2}$$

Differentiate both sides and give

$$\frac{dz_{k,m}}{dt_n} = -\lambda_{k,m}\exp\{-\lambda_{k,m}t_n\} + \gamma_{k,m}\exp\{-\gamma_{k,m}t_n\}. \tag{B.3}$$

Set $\frac{dz_{k,m}}{dt_n}$ into zero, then

$$-\lambda_{k,m}\exp\{-\lambda_{k,m}t_n\} + \gamma_{k,m}\exp\{-\gamma_{k,m}t_n\} = 0 \tag{B.4}$$

$$\lambda_{k,m}\exp\{-\lambda_{k,m}t_n\} = \gamma_{k,m}\exp\{-\gamma_{k,m}t_n\}. \tag{B.5}$$

Taking natural logarithms of both sides and rearranging the terms,

$$t_n = \ln \left(\frac{\lambda_{k,m}}{\gamma_{k,m}} \right)^{\frac{1}{\lambda_{k,m} - \gamma_{k,m}}} \tag{B.6}$$

Substitute t_n into $z_{k,m}$ in (B.2),

$$z_{k,m} = \left(\frac{\lambda_{k,m}}{\gamma_{k,m}} \right)^{\frac{\lambda_{k,m}}{\gamma_{k,m} - \lambda_{k,m}}} - \left(\frac{\lambda_{k,m}}{\gamma_{k,m}} \right)^{\frac{\gamma_{k,m}}{\gamma_{k,m} - \lambda_{k,m}}}. \tag{B.7}$$

Then

$$\zeta_{k,m} = \left[\left(\frac{\lambda_{k,m}}{\gamma_{k,m}} \right)^{\frac{\lambda_{k,m}}{\gamma_{k,m} - \lambda_{k,m}}} - \left(\frac{\lambda_{k,m}}{\gamma_{k,m}} \right)^{\frac{\gamma_{k,m}}{\gamma_{k,m} - \lambda_{k,m}}} \right]^{-1} \tag{B.8}$$

The second derivative of $z_{k,m}$ is

$$\frac{d^2 z_{k,m}}{dt_n^2} = \lambda_{k,m}^2 \exp\{-\lambda_{k,m} t_n\} - \gamma_{k,m}^2 \exp\{-\gamma_{k,m} t_n\} \tag{B.9}$$

which implies that the condition for the maximum value also requires $\gamma_{k,m} > \lambda_{k,m}$.

Appendix C

List of piano pieces and mixtures

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 C.1: Piano pieces from RWC database [34] for generation of mixtures.

No.	K	Pitches	Octave	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 C.2: List of the 25 mixtures. Loudness: "S" is soft; "M" is medium; and "L" is loud.

Bibliography

[1] S. A. Abdallah. *Towards Music Perception by Redundancy Reduction and Unsupervised Learning In Probabilistic Models.* PhD thesis, Department of Electronic Engineering, King's College London, London, U.K., 2002.

[2] A. Askenfelt, editor. *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/.

[3] B. Bank. *Physics-based Sound Synthesis of String Instruments Including Geometric Nonlinearities.* PhD thesis, Budapest Univeristy of Technology and Economics, Hungary, February 2006.

[4] D. Barry and B. Lawlor. Sound source separation: Azimuth discrimination and resynthesis. In *Proc. Int. Conf. on Digital Audio Effects (DAFX)*, Naples, Italy, 2004.

[5] M. Bay and J. W. Beauchamp. Harmonic source separation using prestored spectra. In *ICA 2006, LNCS 3889*, pages 561–568, 2006.

[6] J. W. Beauchamp, editor. *Analysis, Synthesis, and Perception of Musical Sounds.* Springer, New York, 2007.

[7] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[8] J. O. Berger. *Statistical decision theory and Bayesian analysis.* Springer, New York, 2nd edition, 1985.

[9] N. Bertin, R. Badeau, and E. Vincent. Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):538 –549, march 2010.

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

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

[12] T. Blumensath. *Bayesian Modelling of Music: Algorithmic Advances and Experimental Studies of Shift-Invariant Sparse Coding.* PhD thesis, University of London, 2006.

[13] A. S. Bregman. *Auditory Scene Analysis: the Perceptual Organization of Sound.* MIT Press, 1990.

[14] J. J. Burred. *From Sparse Models to Timbre Learning: New Methods for Musical Source Separation.* PhD thesis, Technical University of Berlin, Berlin, Germany, September 2008.

[15] C. Cannam, C. Landone, M. Sandler, and J. P. Bello. The sonic visualiser: A visualisation platform for semantic descriptors from musical signals. In *Proceedings of 7th International Conference on Music Information Retrieval (ISMIR 2006)*, pages 324–327, Victoria, Canada, October 2006.

[16] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall/CRC, Boca Raton, 2000.

[17] R. J. Carroll and D. Ruppert. *Transformation and Weighting in Regression.* Chapman and Hall, New York; London, 1988.

[18] M. A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proceedings of the International Computer Music Conference (ICMC)*, Berlin, Germany, August 2000.

[19] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. Technical report, Deparment of Engineering, Universityh of Camibridge, 2008.

[20] A. T. Cemgil, C. Fevotte, and S. J. Godsill. Variational and stochastic inference for bayesian source separation. *Digital Signal Processing*, 17(5):891–913, 2007. Special Issue on Bayesian Source Separation.

[21] J. M. Chowning and D. Bristow. *FM Theory and Applications.* Yamaha Corporation, Tokyo, 1986.

[22] E. Clarke. *Empirical Musicology: Aims, Methods, Prospects*, chapter Empirical methods in the study of performance, pages 77–102. Oxford University Press, Oxford, 2004.

[23] M. Davy and S. Godsill. Bayesian harmonic models for musical signal analysis. In *Bayesian Statistics VII.* Oxford University Press, 2003.

[24] 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.

[25] Ph. Depalle and L. Tromp. An improved additive analysis method using parametric modelling of the short-time Fourier transform. In *Proceedings of International Computer Music Conference*, pages 297–300, Hong Kong, 1996.

[26] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo methods in practice.* Springer, New York, 2001.

[27] 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*, 15(4):1283–1295, May 2007.

[28] 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.

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

[30] A. Gabrielsson. Music performance research at the millennium. *Psychology of Music*, 31(3):221–272, 2003.

[31] J. Gat. *The Technique of Piano Playing.* Collet's (Publishers) Limited, 5th edition, 1980.

[32] S. Godsill and M. Davy. Bayesian harmonic models for musical pitch estimation and analysis. In *IEEE ICASSP*, volume 2, pages 1769–1772, Orlando, USA, May 2002.

[33] W. Goebl. *The Role of Timing and Intensity in the Production and Perception of Melody in Expressive Piano Performance.* PhD thesis, Karl-Franzens-Universitat Graz, Graz, Austria, 2003.

[34] 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.

[35] M. Hamer. Ivory encore for dead piano greats. In *New Scientist*, number 2496, April 22 2005. http://www.newscientist.com/article/mg18624966.700.

[36] A. Hyvärinen and E. Oja. Neural networks. *Independent component analysis: algorithms and applications*, 13(4-5):411–430, 2000.

[37] J. O. Smith III and S. A. Van Duyne. Commuted piano synthesis. In *Proceedings of the 1995 International Computer Music Conference*, pages 319–326, Banff, 1995.

[38] J. O. Smith III and X. Serra. Parshl: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. Technical report, Center for Computer Research in Music and Acoustics (CCRMA), Department of Music, Stanford University, 1987.

[39] Post Musical Instruments. Piano magic: The PMI piano sample collection, 2005.

[40] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Englewood Cliffs, N.J., 1993.

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

[42] A. P. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6):804–815, 2003.

[43] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[44] K. Lee and A. Horner. Modeling piano tones with group synthesis. *Journal of the Audio Engineering Society*, 47(3):101–111, 1999.

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

[46] 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.

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

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

[49] R. J. McCaulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, August 1986.

[50] T. K. Moon and W. C. Stirling. *Mathematical methods and algorithms for signal processing*. Prentice Hall, Upper Saddle River, N.J., 2000.

[51] D. Leech-Wilkinson N. Cook, E. Clarke and J. Rink, editors. *The Cambridge Companion to Recorded Music*. Cambridge University Press, Cambridge, 2009.

[52] P. D. O'Grady, B. A. Peralmutter, and S. T. Rickard. Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology*, 2005. to appear.

[53] C. Palmer and J. C. Brown. Investigations in the amplitude of sounded piano tones. *Journal of the Acoustical Society of America*, 90(1):60–66, July 1991.

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

[55] R. Parncutt and G. E. McPherson, editors. *The Science and Psychology of Music Performance: Creative Strategies for Teaching and Learning.* Oxford University Press, Oxford, 2002.

[56] T. W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *Journal of Acoustical Society of America*, 60(4):911–918, 1976.

[57] P.H. Peeling, A.T. Cemgil, and S.J. Godsill. Generative spectrogram factorization models for polyphonic piano transcription. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):519 –527, march 2010.

[58] B. H. Repp. A microcosm of musical expression: I. quantitative analysis of pianists' timing in the initial measures of chopin's etude in e major. *Journal of the Acoustical Society of America*, 104:1085–1100, 1998.

[59] B. Hermann Repp. A microcosm of musical expression: Ii. quantitative analysis of pianists' dynamics in the initial measures of chopin's etude in e major. *Journal of the Acoustical Society of America*, 105:1972–1988, 1999.

[60] C. Roads, editor. *The Computer Music Tutorial.* MIT Press, 1996.

[61] A. Robel. Adaptive additive synthesis of sound. In *Proc. of the International Computer Music Conference (ICMC)*, pages 256–259, Bejing, China, 1999.

[62] C. P. Robert and G. Casella. *Monte Carlo statistical methods.* Springer, New York, 2nd edition, 2004.

[63] E. D. Scheirer. Structured audio and effects processing in the mpeg-4 multimedia standard. *Multimedia Systems*, 7(1):11–22, January 1999.

[64] X. Serra. Musical sound modeling with sinusoids plus noise. In C. Roads, S. Pope, A. Picialli, and G. Poli, editors, *Musical Signal Processing*. Swets & Zeitlinger, Lisse, the Netherlands, 1997.

[65] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, pages 177–180, New Paltz, NY, October 2003.

[66] J. V. Stone. *Independent Component Analysis: A Tutorial Introduction*. The MIT Press, Cambridge, Massachusetts; London, England, 2004.

[67] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos. The variational approximation for bayesian inference. *IEEE Signal Processing Magazine*, 25(6):131–146, November 2008.

[68] E. Vincent. Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):91098, 2006.

[69] E. Vincent and M. D. Plumbley. Single-channel mixture decomposition using bayesian harmonic models. In *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Source Separation (ICA 2006)*, pages 722–730, Charleston, SC, USA, March 2006.

[70] T. Virtanen. Separation of sound sources by convolutive sparse coding. In *Workshop on Statistical and Perceptual Audio Processing (SAPA)*, Jeju, Korea, October 2004.

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

[72] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions of Audio, Speech and Language Processing*, 15(3):1066–1074, 2007.

[73] T. Virtanen, A. T. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1825–1828, 2008.

[74] B. Wang and M. D. Plumbley. Investigating single-channel audio source separation methods based on non-negative matrix factorization 18-19 sep 2006, pp 17-20, 2006. In *Proceedings of the ICA Research Network International Workshop*, pages 17–20, September 2006.

[75] C. H. Wong, W. M. Szeto, and K. H. Wong. Automatic lyrics alignment for cantonese popular music. *Multimedia Systems*, 12(4-5):307–323, March 2007.

[76] C. Yeh. The expected amplitude of overlapping partials of harmonic sounds. In *ICASSP*, 2009.

[77] Udo Zöler, editor. *DAFX - Digital Audio Effects*. Wiley, 2002.