

BAYESIAN MUSIC ALIGNMENT

Akira MAEZAWA

Abstract

This thesis addresses temporal alignment of music audio signals with a symbolic music score and alignment of multiple music audio signals, each of which represents a common piece of music (*music alignment*). This is an important task in many fields such as music production, musicological analysis, and informed sound source separation. This thesis focuses on alignment among multiple audio signals (*audio-to-audio alignment*), and that between a symbolic score and one or more audio signals (*audio-to-score alignment*). Moreover, the thesis focuses on Western music, which is polyphonic, consists mostly of harmonic sounds, and is played according to a music score.

Music alignment is difficult because there are variations among different performances, even though they all play a same music score. Such a variation of music audio signals calls for a probabilistic treatment of music audio signals. Furthermore, the fact that different audio signals have some aspects in common calls for a framework for encoding different constraints that are known in advance. These requirements can be satisfied through Bayesian inference. This thesis approaches music alignment from a Bayesian perspective. Chapter 1 presents an overview of music alignment and its relevance. Chapter 2 reviews background on alignment techniques.

Chapter 3 presents a Bayesian audio-to-score alignment method. Here, the symbolic music score is given, and the main goal is to infer the alignment, taking into account variations of audio signals in timbre, volume, and tempo. To deal with variations in timbre and dynamics, polyphonic pitched sounds in the spectral domain are modeled as a probabilistically weighted sum of spikes. These spikes are placed at frequencies where prominent energy is expected to be observed, via analyzing which notes are present in the music score at a given position. Robustness to variations in timbre and dynamics is attained by assuming a weakly informative prior on the weights. Furthermore, the model represents a smoothly changing tempo trajectory shared among different parts, while allowing for slight asynchronies between different parts. An experimental evaluation demonstrated that the proposed method achieved a median alignment error of about 30 ms on real-world ensemble pieces, and about 50 ms on synthetic orchestral recordings.

Chapter 4 presents a Bayesian audio-to-audio alignment method. The problem is difficult because, unlike audio-to-score alignment, the underlying music score is not given. To tackle this issue, the underlying musical piece is expressed probabilistically. Then, it is inferred from the input audio signals, assuming that each audio signal plays

the same musical piece. Since a music score consists of a relatively few note combinations that get reused throughout the piece, the musical piece is represented as a Markov chain with sparse transition probabilities. Furthermore, the audio signals are represented so that they have different but interrelated tempo trajectories. An experimental evaluation on real-world piano music collection showed a median alignment error of 60 ms, outperforming existing probabilistic methods.

In a real-world use scenario, music alignment is further deteriorated for two more reasons. First, some audio signals may play only a subset of the music score. For example, a user might be interested in aligning between an orchestral piece and a hummed melodic line to the piece. This new kind of alignment is defined as *subset music alignment*. Second, the room acoustics may vary significantly. For example, an orchestral recording may be recorded in a highly reverberant hall, while a hummed melodic line to the piece may be recorded in a bedroom. Next two chapters address these two problems.

Chapter 5 presents a subset music alignment method. It is difficult because the input signals play different sequences of note combinations, with some notes that are played in common. To tackle this issue, the proposed method decomposes the input audio into components common to different recordings and those unique to each recording, and simultaneously aligning the audio signals based on the common components. Namely, a hierarchical point process called the hierarchical Dirichlet process (HDP) is used to represent a subset-like relationship in two layers. On the upper layer, one HDP represents each note combination in a piece of music as a subset of all possible notes that are used in the piece. Then, on the lower layer, the other HDP represents the note combinations inside each audio signal as subsets of the note combinations of the top layer HDP. HMMs are used to encode the order of possible notes to play. Evaluations showed that when aligning audio signals that had only one part in common, the proposed method achieved a median alignment of about 200 ms, outperforming existing methods.

Chapter 6 presents a dereverberation method that can be used as a front-end to any alignment method. It is designed to attenuate long-term reverberation (called the late reverberation), which causes past musical notes to smear into the current time instance. Dereverberation is formulated as a deconvolution problem of non-negative autoregressive (AR) process in the power spectrogram domain. Since the order of the AR process is unknown, the Dirichlet process is used to encode a varying number of coefficients that are used. The model is non-conjugate, making inference difficult. To allow inference, a novel inference algorithm based on minorization-maximization is derived. An experimental evaluation showed that even though audio-to-score alignment accuracy degrades under a reverberant environment, the proposed method is capable of recovering the audio-to-score alignment accuracy comparable to that under no reverberation.

Chapter 7 discusses the thesis and presents directions for future research.

Acknowledgements

I would like to express my special thanks and appreciation to my advisor Prof. Hiroshi G. Okuno (now at Waseda University). He has allowed me to pursue a research topic that matters to me, and vigorously advised my research through his broad insights and keen comments. I cannot stress how important it was for me to be researching independently, while receiving the support and advising for conducting and presenting my research. This section would not end if I listed the things I have learned from Prof. Okuno, but, in particular, what I have learned through his style of presentation and writing has become an invaluable asset of my life.

I would also like to express my special thanks and appreciation to Prof. Tatsuya Kawahara, who has supervised my thesis after Prof. Okuno has left Kyoto University, and helped me complete the thesis. His comments were essential for organizing the significance of my research from a broad perspective and presenting it as a unified work. Moreover, he has provided numerous valuable comments that helped me improve the quality of the thesis.

Furthermore, I would like to express my warm thanks and appreciation to Dr. Kazuyoshi Yoshii, who has provided suggestions from the early stage of my research, when my interest in Bayesian inference and music alignment were growing. It has simply been a joy to be working with a researcher whose works I admire. I have learned a lot in Bayesian inference, writing and presentation from him. My interest in Bayesian inference has sparked through conversations with Dr. Yoshii, so I cannot emphasize enough that without him, I would have not written the dissertation.

I also express my special thanks and appreciation to Prof. Toshiyuki Tanaka, the member of my dissertation committee. His numerous astute comments were imperative for improving the quality of the dissertation.

The works presented here were conducted mostly at Yamaha Corporation. This dissertation would have not been written without the support of many staffs from Yamaha. First and foremost, I cannot thank my boss Mr. Takuya Fujishima enough for supporting my pursuing of the Ph.D. program. The various pointers he has offered and the numerous negotiations that he had with the various stakeholders were mandatory for both starting and completing the dissertation. I also express my warm thanks to the division managers Mr. Yukihiro Kawaguchi and Mr. Motoichi Tamura, and the staffs at the human resources division, for allowing me to pursue the Ph.D. program. I would also like to especially thank Mr. Yoshinari Nakamura, Mr. Naoki

Yasuraoka, Dr. Kazunobu Kondo, Dr. Yu Takahashi, Mr. Norihiro Uemura and Mr. Kouhei Sumi for inspiring conversations on statistical signal processing and music alignment.

I also express my special thanks and appreciation to Prof. Masataka Goto, the Prime Senior Researcher of the Information Technology Research Institute at the National Institute of Advanced Industrial Science and Technology. The interest in music alignment sparked during my intern at Prof. Goto's lab. I cannot express enough how valuable the intern was, and how much pleasure and honor it was to work under Prof. Goto and the members of his lab. I learned a lot from his keen insights, often years ahead of the time.

I also express my thanks to former members of Okuno lab, especially Prof. Tetsuya Ogata (now at Waseda University), Prof. Kazunori Komatani (now at Osaka University), Prof. Toru Takahashi (now at Osaka Sangyo University) and Dr. Katsutoshi Itoyama. I have learned a lot from them during the master's program. In particular, I was lucky to co-author papers with Dr. Katsutoshi Itoyama during my Ph.D. program; his sharp comments were invaluable to me. I also warmly thank Ms. Hiromi Okazaki, who has provided secretarial support at our lab. I also thank my colleagues at the lab for the intellectually stimulating conversations.

I also thank my parents for the support that I have received. I especially thank them for allowing me, during my high school years, to spend a great deal of time performing music, composing music, and tinkering with computers – I would not be doing research in music information retrieval if it were not for them.

Most importantly, I thank my wife Yumiko for (1) her unconditional support, (2) her infinite patience, and (3) understanding how to handle me when I am coding.

Contents

Contents	vi
List of Figures	viii
List of Tables	ix
1 Introduction	1
1.1 Motivation	1
1.2 Uses of alignment techniques	2
1.3 Problem	4
1.4 Issues in music alignment	4
1.5 Approach	7
1.6 Organization of this thesis	7
2 Review of Music Alignment	9
2.1 Preliminaries on music and musical acoustics	9
2.2 Preliminaries on Bayesian inference	13
2.3 Music alignment	21
2.4 Formulation of Bayesian music alignment	27
3 Bayesian Audio-to-Score Alignment	29
3.1 Formulation	30
3.2 Inference	38
3.3 Evaluation	44
3.4 Summary	50
4 Bayesian Audio-to-Audio Alignment	51
4.1 Conceptual overview	53
4.2 Formulation	56
4.3 Inference	63
4.4 Evaluation	67
4.5 Summary	70

5	Bayesian Subset Music Alignment	71
5.1	Subset music alignment	72
5.2	Formulation	73
5.3	Inference	84
5.4	Evaluation	86
5.5	Summary	89
6	Bayesian Dereverberation for Music Alignment	91
6.1	Formulation	92
6.2	Inference	99
6.3	Evaluation	104
6.4	Summary	113
7	Conclusion	115
7.1	Contribution of this thesis	115
7.2	Directions for future research	116
A	Application: Violin Fingering Inference	119
A.1	Violin fingering: a primer	121
A.2	Formulation	122
A.3	Experiments	133
A.4	Discussion	136
A.5	Application: de-mystifying historical recordings	136
A.6	Conclusion	138
B	List of Distributions	139
	Bibliography	141
	List of Publications	153

List of Figures

1.1	An example of the recording workflow.	2
1.2	An interpretation retrieval system using alignment technique.	4
1.3	The relationship between a music piece, a music score and audio signals.	5
1.4	Aspects conveyed in a music score	6
1.5	Organization of this thesis.	8
2.1	Source-filter model.	10
2.2	Scientific pitch notation and f_0	12
2.3	Aspects conveyed in a music score)	13
2.4	The spectrogram of a Western music	14
2.5	Hierarchical Dirichlet Process.	17
2.6	Illustration of PLCA	20
2.7	Problem statement of music alignment.	21
2.8	Example of inter-part asynchrony.	27
2.9	Unified formulation of music alignment.	28
3.1	Conceptual idea of the proposed method.	31
3.2	Model of the music score as a state sequence.	32
3.3	Model of the state duration pdf.	34
3.4	Graphical model of the proposed audio-to-score alignment method.	37
3.5	SDR and SIR of the separated parts, using score-informed separation based on the proposed method.	49
4.1	An overview of generative audio alignment.	52
4.2	The idea of how a piece is generated.	54
4.3	The idea of the duration modeling.	55
4.4	Similarity matrix computed from Chopin Op. 41-2	57
4.5	Feature sequences (chroma vector) of two performances, overlaid by points where the state of the latent composition changes.	58
4.6	A short phrase played with 24 different interpretations, and the estimated $\mathbf{u}$ and Λ . Dotted arrows indicate the choice of executed phrasing.	60
4.7	The model of state duration pdf ($p(l_{nd})$ in Fig. 4.8).	62
4.8	The model of the performance and composition sequence	62

4.9	Percentile of absolute alignment error.	69
4.10	Median alignment error against α	70
5.1	The idea behind subset music alignment.	72
5.2	The proposed method considers a latent “master” music score, subsets of which generates the sequence of notes played for each music media, temporally warped by the signal-dependent note durations.	75
5.3	A sample drawn from the model of CQT timeslice.	76
5.4	Illustration of the proposed three-layer HDP-HSMM model.	78
5.5	MAP estimate of the basis $\mathbf{g}$ (top; color in logarithmic scale, larger values are darker) and how frequently each basis is used (bottom).	79
5.6	How the CQT spectrum is created in the proposed model.	80
5.7	An excerpt of the expectation of the spectral basis activation, and the histogram associated with some of the basis (logarithmic scale).	81
5.8	Graphical model of the proposed method.	82
5.9	Absolute alignment error percentile as a function of number of simultaneously shared notes of a four-part ensemble.	88
6.1	Distribution of the effective complexity.	94
6.2	Conceptual diagram of the proposed statistical model.	95
6.3	Graphical model of the proposed method.	98
6.4	Evaluation measures of synthetic impulse response as the reverberation time is varied.	108
6.5	Evaluation measures with different SNR.	109
6.6	Evaluation measures as α is varied.	110
6.7	Percentile of Audio-to-Score Alignment Error.	112
A.1	The essences of violin fingering.	122
A.2	Block diagram of the proposed method.	123
A.3	Block diagram of feature extraction step.	126
A.4	Block diagram of the GMM.	127
A.5	Simplified horizontal-vertical model used in the Viterbi algorithm.	128
A.6	Graphical description of $x_{BN,1}$. Roman numeral indicates the string (IV=lowest, I = highest), and number indicates the pressed finger (1=index finger, 4=little finger).	131
A.7	Graphical descriptions of some features.	132
A.8	Recognition accuracy for different values of ν	134
A.9	Comparison of recognition accuracy for different sequence estimation methods.	135
A.10	Recognition accuracy with different levels of accompaniment.	135
A.11	Estimated fingering from first forty bars of Joachim’s <i>Romanze in C</i> played by Joachim (1903).	137

List of Tables

3.1	Percentile of the absolute error.	45
3.2	True positive rate (TPR) of repeated segment detection. True negative rate was 100%.	47
3.3	Cumulative percentage of the absolute error of the estimated onsets w.r.t. polyphony p	48
4.1	Onset detection rate	68
6.1	List of the parameters used.	105
6.2	Comparison of different methods, averaged over impulse response. . .	106
6.3	Evaluation measures averaged over source signals.	108
6.4	Standard deviation, evaluated over I , of ISD and SRMR for finite and infinite mixtures, averaged over impulse responses.	110
6.5	Average rating (on a 5 point scale) of excerpts averaged over the musical genre. Number following $\pm$ is the standard deviation.	111
A.1	Subjective evaluation of the proposed method.	136

Chapter 1

Introduction

This thesis addresses matching of the time axis of different audio signals that play a same piece of music. This problem is called *music alignment*. Section 1.1 motivates the work. Section 1.2 provides pragmatic motivation for alignment techniques. The goal of the work is provided in Section 1.3 and is outlined in Section 1.6.

1.1 Motivation

Imagine that one day you are walking back from a concert, dazzled from a shockingly vibrant performance of Beethoven’s *Ninth Symphony*. As you walk back home from the concert, you hear a child humming the climactic melody of the same symphony, and are reminded of the vigorous climax of the same piece. The next day, you decide that you liked the piece so much that you buy the recording of the symphony from an online music store. As you listen to it, you feel something is different from the dazzling experience from yesterday – the momentum is lost from a slowly played climax.

What you have heard were three very different acoustic signals, all of which you have associated as audio renditions of a same musical idea – the *Ninth*. We humans listen and associate the input audio signal to past occurrences of the same event, comparing and contrasting the new occurrence with the previous occurrences. This is a nontrivial task because a piece of music contains many notes that are played simultaneously (*polyphony*), meaning that we need to segregate the music audio into relevant auditory streams, while aligning the audio signals based on the segregated streams. The task is further exacerbated by variations in playing: people play with different volumes and timbres, in different acoustic environments.

Humans very naturally associate music media to a same piece of music as the same piece. More specifically, we can identify where a specific location of the piece is in different media, and compare these performances. Unfortunately, this kind of comparative analysis is still difficult to attain with a machine.

This thesis aims to achieve such a comparative machine listening capability.

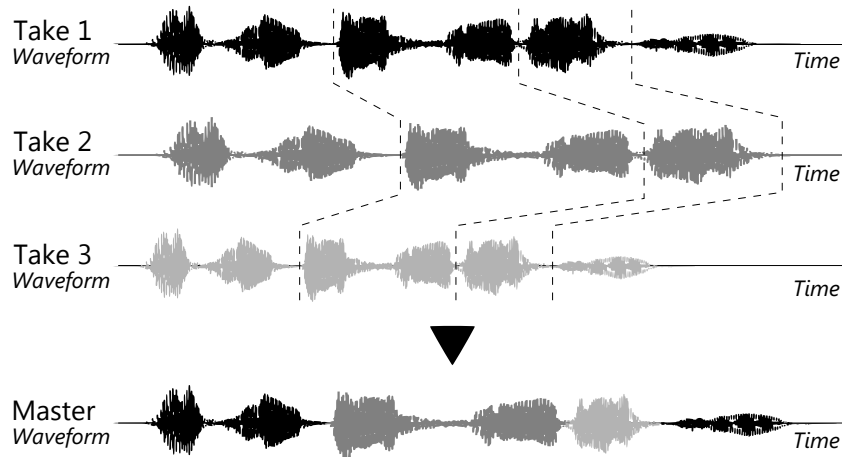


Figure 1.1: An example of recording workflow. The master recording is often a patch-work of multiple takes. Dotted lines indicate the same places in the underlying musical piece.

1.2 Uses of alignment techniques

When we want to use a computer software to browse or manipulate multiple recordings to a same piece of music, we would expect to interact with the computer software in a human-like manner, as discussed earlier. Therefore, a software program that presents different music media to a same piece of music must implement the capability to align different music media based on the common location in the underlying musical composition. This kind of functionality is important for many practical applications, some uses of which are outlined below.

1.2.1 Recording and editing

Consider recording a piece of music. When recording a piece of music for production, it is common to record a same piece multiple times (called *takes*), since the first take may not be the best one – performers may make a mistake, or they may play with unintended nuance. Thus, the final recording is often times a patch-work of the best-sounding takes, as shown in Fig. 1.1.

Since performers may play each take with a different tempo curve, this kind of patch-working requires the recording engineer to extract points over different recordings that correspond to the same point in the underlying musical piece. In other words, dotted lines in Fig. 1.1 are extracted manually, for every point at which the take is switched back and forth. To alleviate this burden, the alignment technique may be used to seamlessly switch from one take to another take [90].

1.2.2 Informed source separation

Source separation, the technique of extracting or removing a specified component from a polyphonic mixture, has gained considerable interest in recent years. It is a useful technique for removing unwanted sound, or extracting sound components of interest.

An issue in source separation lies in *specifying* which component to separate. To specify which component to separate, one could imagine providing to the separation system some additional media data, such as different audio signal or symbolic music score data. This kind of approach is called *informed source separation*, and considerable interest has been vested recently. In the context of music source separation, there are two popular approaches. First, the user may provide the symbolic music score of the piece played by the audio signal (*score-informed source separation*). Then, the user could separate the audio by specifying a particular note or a part [29,33,60,64]. Second, the user may provide a reference audio signal to separate (*separation by humming*). For example, given a polyphonic music audio signal, the user may hum the melody that he wants to separate out [55,119]. In either case, it is important to align the music score or the reference audio to the audio signal that the user wants to separate. This is required for the separation method to identify and separate the signal components in the reference that are played in the polyphonic audio signal.

1.2.3 Performance logging / comparison

Imagine that the user, in this case a music performer, is practicing a piece for an upcoming recital. The user may be interested in recording his practicing, so that he could listen back to it later on and diagnose his playing. For example, he might visualize his performance to see the tendencies of his playing [26]. He might record multiple takes, each with a different style of playing, and compare which style of playing he prefers. In this kind of use, it is valuable to align the practice recording to each other. If the user uses a software to display the music score, it is useful to align the recording to the music score [42]. These uses call for alignment between different music media.

This kind of performance comparison is also valuable for musicological purposes. For example, a musicologist might be interested in comparing how playing style has evolved in the last century (since the phonograph was first invented in 1877). Alignment methods may be used to partially automate the process of analysis to uncover trends of playing [114]. By comparing different performers, the distinguishing features of a legendary performer may be unveiled [134]. It is also valuable to listen back and forth between different recordings of a same piece of music [40]. Synchronization methods may be used for visualizing or retrieving interpretation that the user likes [81], as shown in Fig. 1.2.

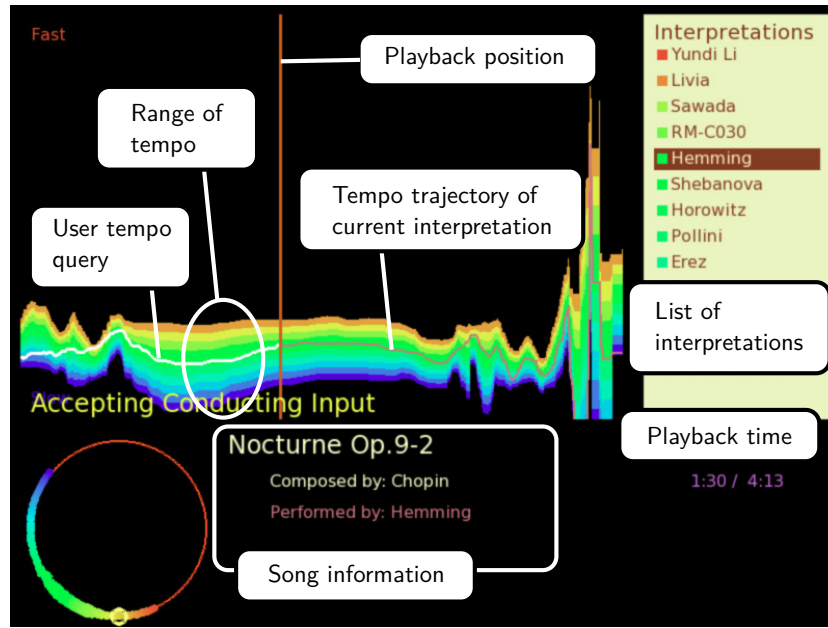


Figure 1.2: An interpretation retrieval system using alignment technique. The user “conducts” the tempo trajectory that he wants to listen. The system visualizes the tempo of various interpretations while retrieving the interpretation based on user-input tempo data, plotting the range of tempo trajectories against time.

1.3 Problem

Music alignment is the task of aligning multiple signals that represent a same piece of music. This thesis focuses on two kinds of music alignment. First, the work focuses on *audio-to-audio* alignment, which is music alignment among multiple audio renditions to a same piece of music. Second, the work focuses on *audio-to-score* alignment, which is music alignment among a symbolic score and an audio signal to the same piece of music.

Furthermore, the thesis presents and provides a solution for *subset music alignment*, a kind of music alignment where each music representation presents a *subset* to a common piece of music. For example, subset music alignment aims to align between a symphonic recording of the *Ninth* to a child humming the melodic part to the *Ninth*.

These relationships are illustrated in Fig. 1.3. Throughout the work, a piece of music is assumed to be described as a progression of multiple pitched notes.

1.4 Issues in music alignment

Music alignment is a difficult problem because a same piece of music can be played very differently. This is because, as illustrated in Fig. 1.4, the underlying musical piece only notates the sequence of pitches and its relative durations, along with weakly

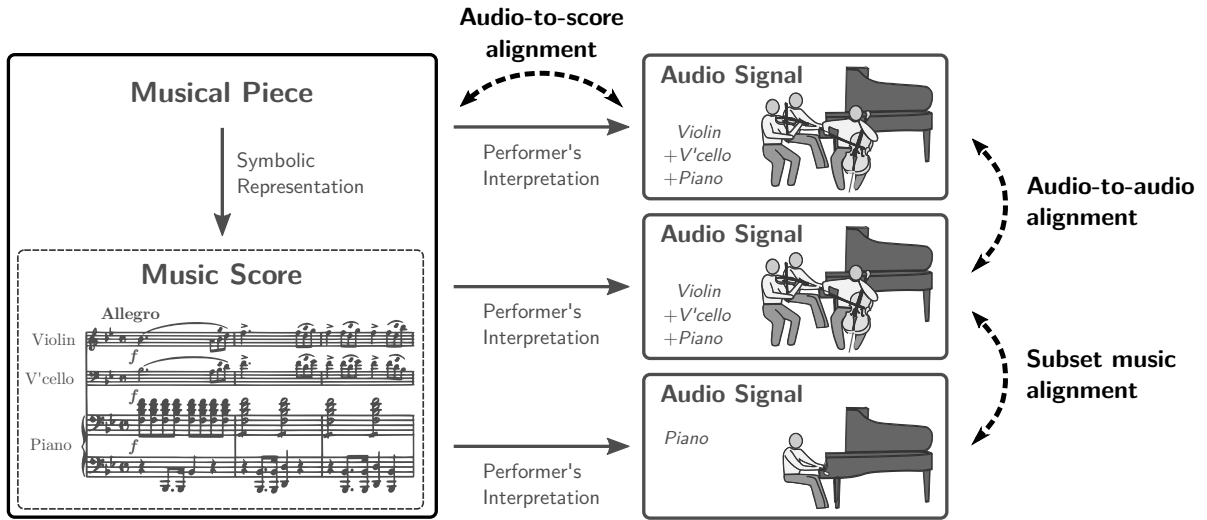


Figure 1.3: The relationship between a music piece, a music score and audio signals, and the various tasks of music alignment.

informative cues on the timbre (e.g., instrumentation markings), dynamics (e.g., dynamics markings such as “forte,” which means “strong”), or tempo (e.g., “Andante,” which means “in a walking manner”). More specifically, there are following aspects of uncertainty:

Uncertainty of the tempo trajectory – People play a same piece of music with different expression. The expression is reflected as variations in tempo, dynamics and timbre. The tempo changes non-linearly as the piece progresses; for example, structural boundaries of music are often slowed down in a non-linear, parabola-like tempo trajectory [122]. How can a machine align signals whose tempo trajectories are different and changing non-linearly?

Uncertainty of the dynamics – Different performer plays the piece with different dynamics. Different performers interpret words like “forte” differently, and execute them with different dynamics. How can a machine understand that some notes across recordings mean the same thing, even though they are played with different volume?

Uncertainty of the timbre – The timbre varies across performances. For example, different recordings are played on different instruments, so the instrumental acoustics contribute to different timbre. Moreover, different performers may employ different playing techniques to convey a same note, creating timbral differences. How can a machine understand that some notes across recordings mean the same thing, even though the timbres are different?

Uncertainty of the played parts – Imagine aligning between a child’s humming of the main melody to the *Ninth* and a polyphonic orchestral recording. The orchestral recording contains many notes that are not sung by the child, such as the bass, the accompaniment, or the counter-melody (a subsidiary melodic line that is played in conjunction with the main melody). In this case, one aligns the two audio by listen-

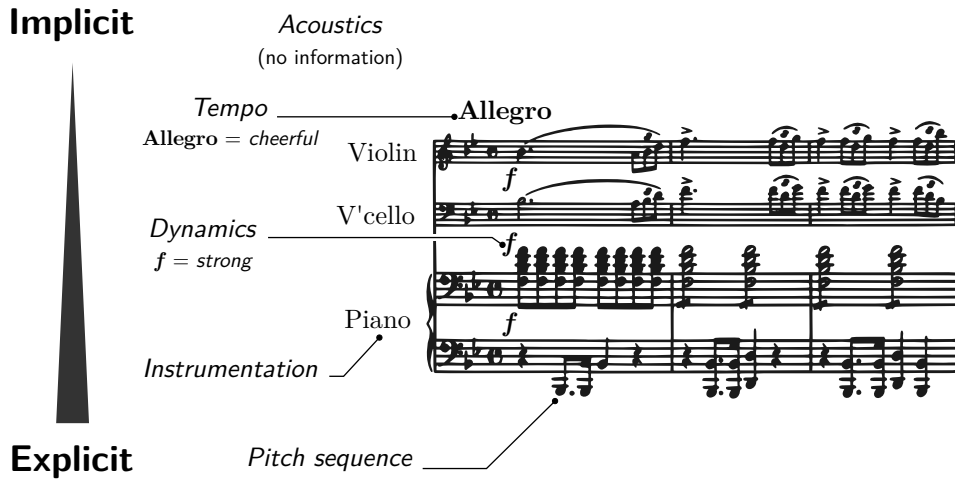


Figure 1.4: Aspects conveyed in a music score. Music score conveys different aspects of produced sound with different degree of ambiguity.

ing for components within the orchestra that is hummed by the child. How could a machine choose which audio components to listen for?

Uncertainty of the underlying music representation – We humans can align different audio signals because we *know* what the underlying musical piece is, and what it is invariant to. This means that we have an internal representation of music that is inferred from the audio signals. When aligning to a symbolic score, there is need to be concerned with such an internal representation because it is already given as a symbolic score. However, when the symbolic score is not given, one must infer the underlying music representation. This is difficult because automatic music score transcription is difficult [72]. Thus, a representation about the musical composition is necessary, such that it is informative enough to allow alignment and also inferable from the input signals.

Uncertainty of the room acoustics – A piece of music is recorded under various acoustic environments. Different acoustic environments add different durations of reverberation to the performed audio, adding pleasing acoustic effect to the audio signal generated by the performers. This kind of reverberation, however, smears the signals generated by the performers, making it difficult to attain the true goal of aligning signals that were produced by the performers. How can a machine align different performances, as if it were free of reverberation?

The above-mentioned problems are exacerbated by their mutual dependencies. For example, we humans gain certainty on the played parts if we know the tempo trajectory. This is obtained readily from alignment. However, finding the alignment requires us to know which parts are played, so that we know which parts to listen for. Likewise, we can find the tempo trajectory if we know what each instrument sounds like, *i.e.*, the timbre, but to know the timbre it is advantageous to know the tempo (so that we know which instruments are played at each time instance).

1.5 Approach

At the core of these issues lie the *partial* uncertainties about the audio signals, owing to the fact that the music score underpinning these signals is an ambiguous medium.

The fact that there are uncertain aspects calls for a probabilistic treatment to alignment. In other words, aspects that are uncertain should be encoded as random variables. Moreover, the fact that some musical aspects, such as the sequence of pitch combinations, are more “informative” than others, such as the dynamics, calls for a treatment where different musical aspects can be embodied with prior knowledge of a varying degree. For example, one may wish to convey that the sequence of pitch combination is identical, while the volume has little prior information.

To meet these criteria, this thesis tackles alignment as a Bayesian inference problem. In Bayesian inference, the likelihood of observing data X given underlying parameters Θ , $p(X|\Theta)$, and a prior distribution $p(\Theta)$ are supplied. Then, inference is done by evaluating the posterior distribution $p(\Theta|X)$. It is useful because the prior distribution $p(\Theta)$ over probabilistic model parameters Θ may be used to represent the prior constraints or knowledge about the phenomenon at hand. Then, the posterior distribution takes into account both the prior beliefs and the observed signal.

To tackle alignment in a Bayesian framework, various sources of variations in human performance are introduced as random variables Θ , whose prior distributions are expressed according to the constraints imposed on by the music score. For example, Θ may represent the volume, the timbre or the tempo curve, all of which, as discussed earlier, are conveyed only vaguely in the music score. To align signals under different reverberation, this thesis presents a dereverberation front-end for music alignment. An alignment method encodes uncertainties on human performance, while a dereverberation method encodes uncertainties on room acoustics. Therefore, in this case, Θ represents the unobservable factors pertaining to reverberation, such as room reverberation and the dry signal that drives the room. Then, given a reverberant data X , the posterior $p(\Theta|X)$ is evaluated, from which the dry signal may be recovered.

1.6 Organization of this thesis

This thesis addresses the problems discussed earlier in four respects. Fig. 1.5 provides an outline of the thesis.

To analyze audio signals with different timbre and dynamics, Chapter 3 presents an audio-to-score alignment method that jointly infers the tempo, timbre and dynamics in a parametric representation of pitched notes.

To tackle the issue of representing the underlying piece of music in audio alignment, Chapter 4 presents a generative model of audio alignment that explicitly decouples performance variations and the underlying piece. The method decouples au-

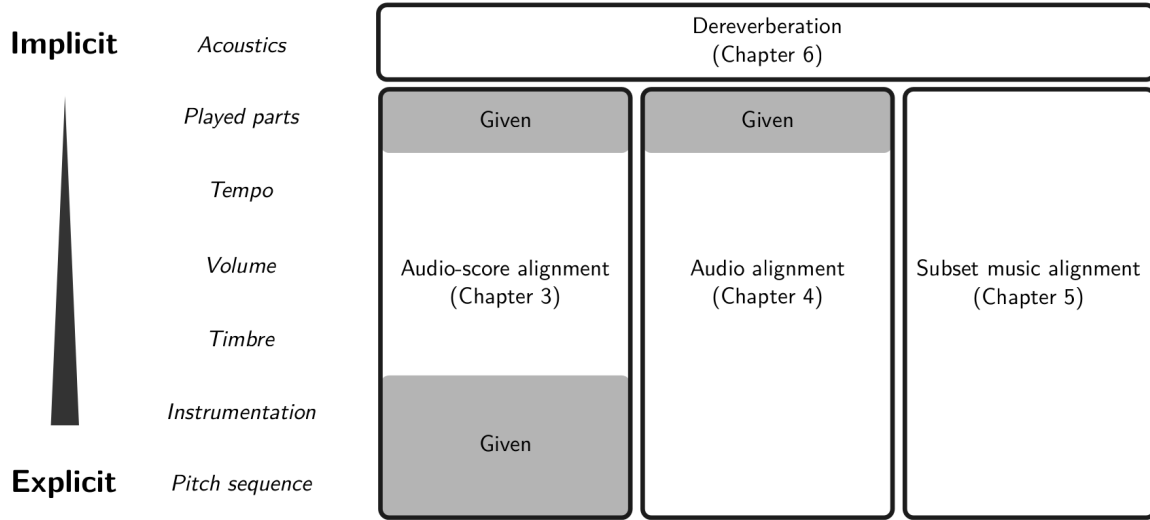


Figure 1.5: Organization of this thesis.

audio signals into (1) a common sequence of audio features, (2) a common sequence of relative note durations, and (3) sequence of audio features generated by each performance.

For analyzing audio signals that play different combination of parts, Chapter 5 presents an audio alignment method that is robust to differences in comprising parts. It does so by decomposing the input audio into components that are played in common by different recordings, and those that are unique to each recording, and simultaneously aligning the audio signals based on the common components.

Finally, to analyze audio signals with different room acoustics, Chapter 6 presents a dereverberation method that can be used as a front-end to any alignment methods. It is designed to attenuate long-term reverberation (called the *late reverberation*), which causes past musical notes to smear into the current time instance. The effectiveness of the proposed method is verified on audio-to-score alignment method, which has the widest variation of the room acoustics.

Moreover, Appendix A shows an application of alignment method for recuperating the violin fingering of an audio signal of a musical piece written for the violin.

Chapter 2

Review of Music Alignment

This chapter reviews studies that are relevant to this thesis. First, basic ideas behind music audio signals are reviewed in Section 2.1. Then, Section 2.2 reviews studies in Bayesian inference that are relevant in this thesis. Finally, Section 2.3 reviews studies in alignment methods.

2.1 Preliminaries on music and musical acoustics

Here, basic theories on tone generation are discussed, along with how it relates to Western music. Then, aspects of a music score are discussed.

2.1.1 Source-filter model

Here, the *source-filter theory* is briefly discussed, as it plays a significant role in this thesis. According to the source-filter theory, a sound is generated by convolving a source signal with a linear filter. To elaborate, suppose that an observed time domain signal x_n , where n is the sample index, is composed of I simultaneously sounding signals. Each signal is composed of a source signal, convolved by a linear filter. Let s_{in} be the time domain source signal for the i th sound, and h_{ik} is the causal impulse response of the i th sound with K coefficients. Then, the time domain signal x_n can be expressed as follows:

$$x_n = \sum_{i=1}^I \sum_{k=1}^K s_{i,n-k+1} h_{ik}. \quad (2.1)$$

It is assumed that the magnitude response of h is band-limited to below the Nyquist frequency.

For a pitched sound, the source signal s_{in} is expressed approximately as a pulse train with a frequency $f_{0,i,n}$. The pulse train creates uniform peaks at integer multiples of $f_{0,i,n}$ in the frequency domain. The frequency of the pulse-train f_0 is called the

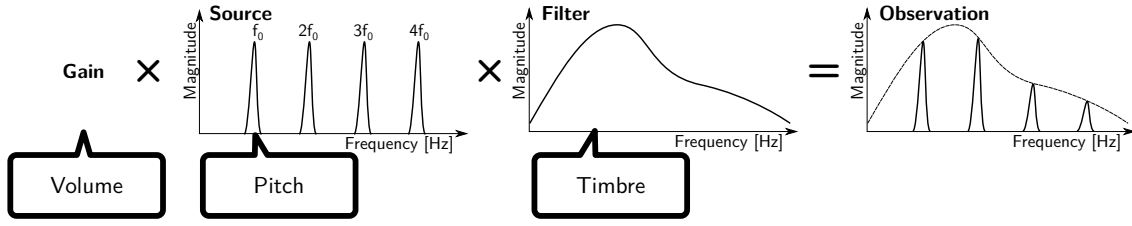


Figure 2.1: Idea of source-filter model for harmonic sounds. It is assumed that a pitched instrumental sound is generated by a pulse train convolved by a linear filter.

fundamental frequency (“ f_0 ”). Its integer multiples are called the *partials*; in particular, the signal component at frequency jf_0 is called the j th *partial*.

Signals originating from different sources overlap heavily in the time domain representation because a linear filter is being convolved. On the other hand, since convolution in time domain is multiplication in the frequency domain, the effect of each source is expected to be seen only at the integer multiples of the fundamental frequency for each source. Thus, if each instrument plays different f_0 s, the sources are expressed more discriminatively in the frequency domain representation. To express signals that slowly vary in the time domain, x_n is passed through a filter-bank. For this purpose, the short-time Fourier transform (STFT) or the constant-Q transform (CQT) are some of the popular approaches [107, 112]. In STFT, the frequency component of bin f at frame t , $\tilde{x}_{tf} \in \mathbb{C}$, is expressed as follows:

$$\tilde{x}_{tf} = \sum_{n=1}^N w_n x_{n+Ht} e^{-\frac{2\pi jfn}{N}}. \quad (2.2)$$

where w_n is a windowing function (the subsequent chapters use the Gaussian window [100]), N is the frame size and H is the hop size. When N is a power of 2, then the fast Fourier transform (FFT) may be used to efficiently compute $\tilde{x}$ [19]. The sampling frequency is assumed to be f_s Hz.

Now, suppose there is a signal generated from the source-filter model, as described in Equation 2.1, with harmonic sources, where the i th instrument plays at a fundamental frequency of $f_s \frac{f_{0,ti}}{N}$ Hz at frame t . Next, assume the sidelobe leakage caused by the windowing function w is negligible. Then, Equation 2.1 is expressed as an analytic signal representation in the STFT domain (as in Equation 2.2) as the following:

$$\tilde{x}_{tf} \approx \sum_i e^{-j\theta_{tfi}} g_{ti} \sum_j \tilde{w}_{f-jf_{0,ti}} \tilde{h}_{if}, \quad (2.3)$$

where θ_{tfi} is some constant phase offset, $\tilde{w}_f$ is the DFT coefficient of w at frequency f , $\tilde{h}_{if}$ is the DFT coefficient of h_i evaluated at frequency f , and g_{ti} is the gain of instrument i at frame t . Since h is assumed to be band-limited to the Nyquist frequency, j is effectively computed up to the Nyquist frequency. In practice, even when $f_{0,ti}$ is non-

integer values, an interpolated version of $\tilde{w}$ reasonably approximates Equation 2.3.

In the magnitude STFT domain, a harmonic source-filter model becomes more-or-less additive as well, allowing us to reasonably ignore the interference between different signal components. The additivity of the magnitude spectra can be reasonably assumed for harmonic sources because $\tilde{w}$ is expected to be distributed sparsely in the frequency domain. That is, since a single i dominates the observed power in most frequency bins, interference occurs only at a few frequency bins. Thus, letting $x = |\tilde{x}|$, $w = |\tilde{w}|$ and $h = |\tilde{h}|$, then x_{tf} is expressed as follows:

$$x_{tf} \approx \sum_i g_{ti} \sum_j w_{f-jf_{0,ti}} h_{if}, \quad (2.4)$$

Even when the additivity holds approximately, this kind of additive representation is known to be effective for separating a polyphonic mixture into comprising components [127].

Assuming that a Gaussian window is used to compute the STFT, a pulse train in the time domain becomes a periodically-spaced Gaussian in the frequency domain. This means that $w_{f-jf_{0,ti}}$ in Equation 2.4 can be approximated as a Normal distribution centered about $jf_{0,ti}$:

$$w_{f-jf_{0,ti}} \propto \mathcal{N}(f | jf_{0,ti}, \sigma^2). \quad (2.5)$$

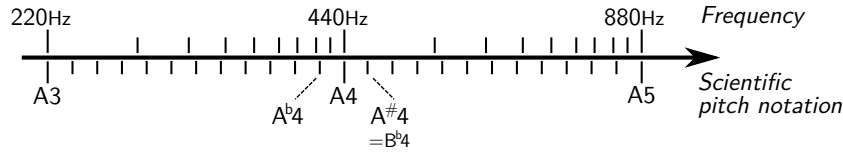
Furthermore, the relative strength of the spectral envelope evaluated at $jf_{0,ti}$ for $j \in [1 \cdots J]$ can be treated individually as $\mathbf{a}_i = [h_{i,f_{0,ti}} \cdots h_{i,Jf_{0,ti}}]$. This yields in the following representation of $\mathbf{x}$ in Equation 2.4:

$$x_{tf} = \sum_{i=1}^I g_{ti} \sum_{j=1}^J a_{tij} \mathcal{N}(f | jf_{0,ti}, \sigma^2). \quad (2.6)$$

Here a_{tij} is the relative strength of the j th partial for signal i at frame t , and $f_{0,ti}$ is the average fundamental frequency of signal i at frame t . This kind of representation is useful in fundamental frequency analysis [48, 67]. If the filter remains stationary, then a single $\mathbf{a}_i$ is associated with each source i that plays a particular fundamental frequency. This concept is illustrated in Fig. 2.1.

2.1.2 On pitch in Western music

The fundamental frequency f_0 is related closely with the human perception of pitch. Namely, we human perceive a linear increment of pitch with an exponential increase in the f_0 . We perceive sounds with f_0 that are twice apart as belonging to a same *class* of pitch. Two sounds whose f_0 are twice apart are said to be played in an *octave*. In Western music, an octave is divided into twelve *pitch-classes* labeled in an ascending order as C, C $\sharp$, D, D $\sharp$, E, F, F $\sharp$, G, G $\sharp$, A, A $\sharp$, and B. The pitch-class is cyclic,

Figure 2.2: Scientific pitch notation and f_0 .

so the note above B is a C . Furthermore, the pitch-class may also be represented by suffixing a b symbol to the alphabet (*i.e.*, those that do not end with a $\sharp$). Appending a b symbol rotates down the pitch-class by one; for example, Cb is the same as B , and Db is the same as a $C\sharp$. To indicate the pitch, the pitch-class is suffixed by a number that indicates the octave difference from a reference octave. This kind of representation, as illustrated in Fig. 2.2, is known as the *scientific pitch notation*. $A4$ is often set to 440Hz, even though this frequency varies by tens of Hertz depending on historical contexts. Two adjacent pitch labels are said to be a *semitone* apart; for example, $A4$ is two semitones below $B4$.

2.1.3 Primers on the music score in Western music

In Western music, performers play a piece of music off of a *music score*. Fig. 2.3 shows a typical music score in Western music. A music score typically notates the sequence of pitches and their relative durations, shown as circles on five adjacently-spaced horizontal rules.

Pitch is described by the vertical position on the rule. It is common to simultaneously play multiple pitches on multiple instruments. There are two important points related to pitch sequences. First, the fact that the music score notates the pitch suggests that the fundamental frequencies inside a note remains relatively stationary. Second, even though many notes may be played simultaneously, the total number of notes used is typically much fewer than the space of all possible notes. These mean that the spectral representation of Western music audio is roughly approximated as a weighted sum of a few locally stationary spectral “bases” that represent the instantaneous spectra of pitched instruments playing different notes. This can be seen in a spectrogram of a typical Western musical piece, shown in Fig. 2.4. As shown in the figure, it is composed of locally stationary and harmonic spectral time-slices. Moreover, the total number of the harmonic patterns is much fewer than the space of all possible harmonic patterns.

The duration is described by the shape of the circle; the number of thick horizontal lines attached to the vertical line placed adjacent to the circle, if any; and the dot adjacent to the circle, if any. For example, a *quarter note*, denoted as ♩ , means to play the note for one *beat*. An eighth note, denoted as ♩ , means to play the note for half a *beat*. *Beat* is a unit of time. It is described by the *tempo*, which dictates how many beats there are in one minute (beats per minute, bpm). The tempo is dictated by the mode

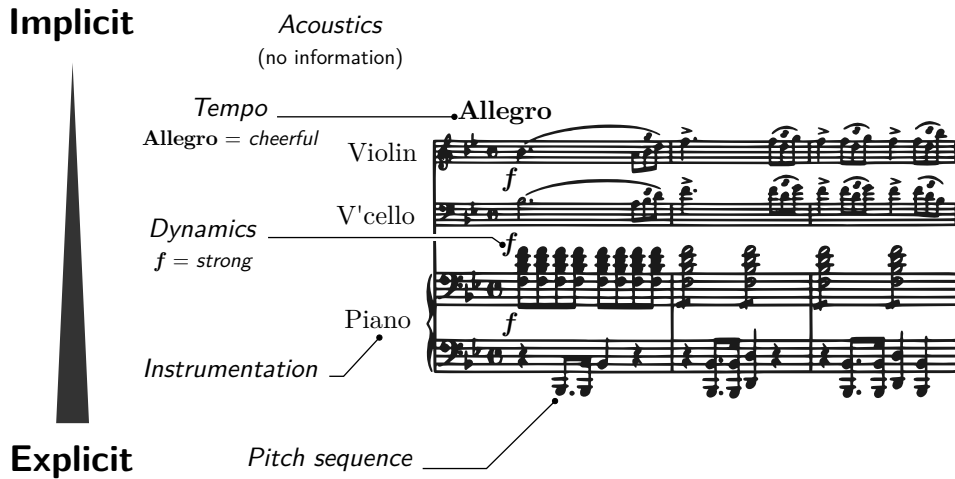


Figure 2.3: Aspects conveyed in a music score. Music score conveys different aspects of produced sound with different degree of ambiguity.

of speed to play. For example, “Allegro” on the top-left of the figure indicates the piece to be performed in a cheerful speed. The tempo tends to remain more-or-less stationary, but it is common for performers to consciously or unconsciously fluctuate the tempo. Moreover, performance instructions such as *ritardando* or *accelerando* request the players to slow down or speed up, respectively, the tempo.

In addition to the pitch sequences and their durations, the music score shows the instrumentation (e.g., the violin, the v’cello and the piano in the figure). It may also indicate the dynamics; for example, an italicized “f” denote the subsequent passage to be played strongly. Furthermore, some notations specify the mode of playing at the note-level; for example, a note with a > sign is played with an emphasis. Moreover, the music score has structural markings. For example, a portion of the music score that is enclosed with $\|$: and $: \|$ means that portion is repeated. It is common for the performer to skip some of the repeats. Finally, a music score may notate expression markings and special performance techniques. For example, a score may say *agitato*, which means to play in an agitated manner, or mark “pizzicato” on a bowed stringed instrument part, which means to pluck the string. In the piano, the score may mark places to depress or release the *sustain pedal*, a device that elongates the played notes of the piano as long as the pedal is pressed down. As an instructional aid, some music scores print *fingering* markings on instrumental parts, showing which finger should be used to play the note.

2.2 Preliminaries on Bayesian inference

Since the music score is an ambiguous representation of music, this thesis relies on statistical modeling to encode unknown aspects, *i.e.*, aspects that are unique to each performance. Moreover, since some elements are more explicitly stated than others, it

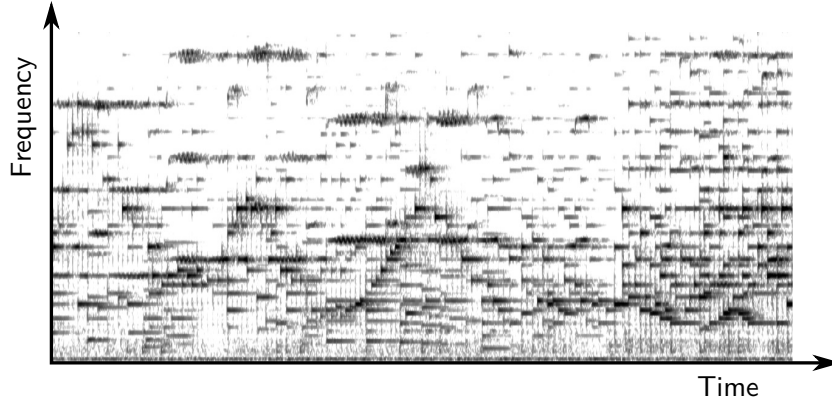


Figure 2.4: The spectrogram of a Western musical piece. Dark colors indicate strong energy in the time-frequency bin. Notice that the spectrogram consists of regularly recurring spectral patterns.

is useful to incorporate prior knowledge on the musical piece. To meet these demands, *Bayesian inference* is used extensively in this thesis, because it offers a data-driven inference that takes into account encoded prior information.

In Bayesian inference, one first designs a probabilistic model parametrized by a set of parameters Θ . The probabilistic model is described in terms of the likelihood of observing some data X , $p(X|\Theta)$ (the *likelihood*), and distribution on the parameters $p(\Theta)$ that are encoded before inference (the *prior distribution*). Then, given some data X , it infers the probability density function (pdf) of the probabilistic model parameters Θ given the data X (the *posterior distribution*) using the Bayes' rule as follows:

$$p(\Theta|X) = \frac{p(X|\Theta)p(\Theta)}{\int p(X|\Theta')p(\Theta')d\Theta'}. \quad (2.7)$$

Once the posterior distribution is found, the statistics of $p(\Theta|X)$ can be extracted to recover useful information.

Unfortunately, direct application of the Bayes' rule is intractable for many models that are practically interesting. This owes to the difficulty in evaluating the integral in Equation 2.7. To tackle this problem, there are mainly two approaches. First, the integral may be approximated with sampling [54]. Second, an easily computable upper-bound to the posterior distribution may be derived, and the bound may be tightened [4]. This thesis focuses on the second strategy, which is known as *variational Bayesian inference*.

The subsequent sections introduce the readers to variational Bayesian inference and statistical models that are used extensively in this thesis.

2.2.1 Inference with variational methods

Variational Bayesian method is an approximate method that is based on tightening an upper bound to the Kullback-Leibler (KL) divergence from the true posterior to some approximate posterior. By assuming some variables in the approximate posterior are independent of each other, an analytically tractable inference may be possible.

Assume a probabilistic model parameterized by parameters $\Theta = \{\theta_1 \cdots \theta_I\}$ and observation X . Let $q(\Theta)$ be a density function that approximates the posterior distribution. To approximate the posterior, consider minimizing the KL divergence from the variational distribution to the true posterior, where the KL divergence is given as follows:

$$D(q\|p) = - \int q(\Theta) \log \frac{p(\Theta|X)}{q(\Theta)} d\Theta. \quad (2.8)$$

Minimizing this function is analytically intractable if θ_i in Θ are mutually coupled in the posterior distribution.

To make the inference tractable, one typically considers restricting q to be in a space of distributions in which θ_i 's are mutually independent. Then, the above problem is rewritten as follows:

$$D(q\|p) = - \int \prod_{i=1}^I q(\theta_i) \log p(\Theta|X) d\theta_i + \sum_{i=1}^I \int q(\theta_i) \log q(\theta_i) d\theta_i. \quad (2.9)$$

Minimizing the KL divergence w.r.t. $q(\Theta)$ yields in the following:

$$\log q(\theta_i) = \langle \log p(X, \Theta) \rangle_{q(\neg\theta_i)} + \text{const}. \quad (2.10)$$

Here $q(\neg\theta_i)$ means all variables in Θ except for θ_i , and $\langle f(x, y) \rangle_{p(x)}$ means to take the expectation of $f(x, y)$ w.r.t. $p(x)$, i.e., $\int f(x, y) dp(x)$. Since this step decreases the KL divergence from q to the posterior for any arbitrary θ_i , iterating this step over different θ_i 's decreases the KL divergence, until it converges to a local optimum.

An important point here is that the log-expectation $\langle \log p(X, \Theta) \rangle_{q(\neg\theta_i)}$ should be analytically tractable. It is tractable if the model is conjugate-exponential, i.e., the sufficient statistics of the likelihood are the natural parameters of the prior distribution. When the model is non-conjugate, then it may become difficult to find the log-expectation. In Chapter 6, a Bayesian inference method is presented for a non-conjugate model, by finding an upper-bound to the KL divergence from the true posterior to the variational posterior, and iteratively tightening the bound.

2.2.2 Bayesian nonparametrics

This thesis is often concerned with modeling a collection, such as note collection or filter coefficients. A critical issue is in expressing the *effective number of components* used

in the collection. This is difficult because the number of components should change in a data-driven manner: new components should be introduced as the data complicates. Bayesian nonparametrics offers a flexible modeling strategy for modeling collections whose effective number of elements vary with the complexity of the data.

To allow the effective number of components in a mixture model to vary with the complexity of data, a point process called the Dirichlet Process (DP) may be used, so that the effective number of points grows with the complexity of data, and effective number of point is almost surely finite.

X is said to be a draw from a Dirichlet Process [37] with a base measure μ and a concentration parameter α , $DP(\alpha, \mu)$, if for any measurable partition on S denoted by $\{A_1, \dots, A_N\}$, $X(A_1), \dots, X(A_N)$ are drawn according to $\text{Dir}(\alpha\mu(A_1), \dots, \alpha\mu(A_N))$. Mathematical rigors aside, DP can be interpreted as generating a distribution by “sampling” points (“atoms”) from the base measure μ , such that the density of the sampled points are distributed according to μ . Each point is then associated a weight that sums to 1, where most points take on negligibly small values. The concentration parameter α controls the number of points that take on non-negligible values. Interested readers are referred to [70] for a mathematically rigorous monograph, and [36] or [133] for treatment geared towards Bayesian nonparametrics.

In order to handle the DP, it is convenient to describe samples generated from a DP in terms of the stick-breaking construction [118]. To this end, consider drawing countable number of parameters from μ , and generating the normalized weight w_i by a process that successively breaks off a stick of length 1. Specifically, the process starts with a stick of length 1. Then, at each step of the process, the process first chooses the proportion by which current piece stick is broken by drawing the proportion from $\text{Beta}(1, \alpha)$, then breaks the piece in two and assigning the length of the left side as w_i , and repeating the same process for the broken stick to the right. Such a process is defined by Griffiths-Engen-McCloskey (GEM) distribution, a distribution governed by the concentration parameter α . Specifically, w_i being drawn from a $\text{GEM}(\alpha)$ distribution is defined as follows:

$$\xi_i \sim \text{Beta}(1, \alpha) \quad (2.11)$$

$$w_i = \xi_i \prod_{i'=1}^{i-1} (1 - \xi_{i'}). \quad (2.12)$$

Then, the DP samples points from the base measure μ , one for each index i in the above equation. Thus, if a pdf G is drawn according to $DP(\alpha, H)$, then G yields in the following form of distribution:

$$G(x) = \sum_{i=1}^{\infty} w_i \delta(x - \theta_i), \quad (2.13)$$

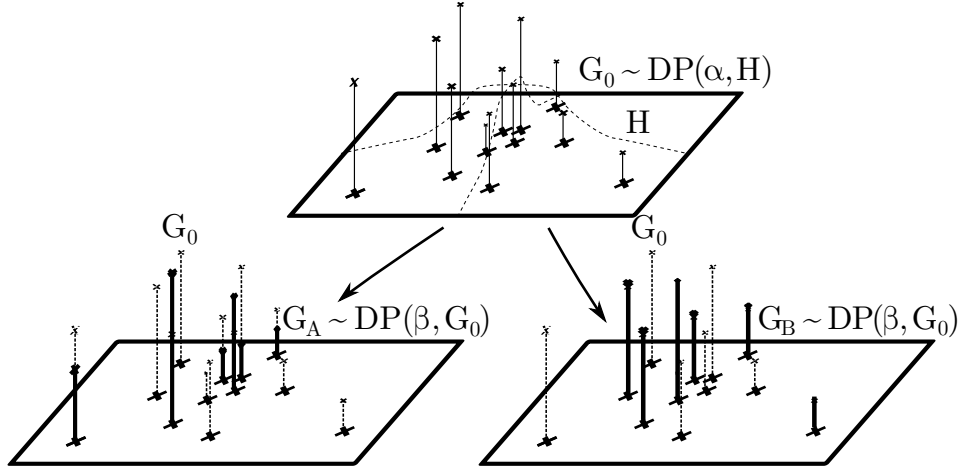


Figure 2.5: Hierarchical Dirichlet Process. A top-level DP G_0 (top; solid lines) with base measure H (top; dotted curve) becomes the base measure (bottom; dotted lines) of the bottom level DPs, G_A (bottom left; solid lines) and G_B (bottom right; solid lines). As can be seen, the atoms of G_A and G_B are subsets of G_0 , allowing us to convey the notion of subsets with an unbounded number of elements.

where w_i is drawn according to $\text{GEM}(\alpha)$ and θ_i is drawn from H .

As suggested by the stick-breaking representation, the DP tends to generate highly biased weights, where only a few elements of the countably infinite sampled points take on significant values and others negligible. This kind of “rich-gets-richer” effect makes DP valuable in clustering tasks because a parameter that has been introduced becomes increasingly likely to be reused, meaning that a large number of observations is described by much fewer, compact set of parameters.

Sometimes a subset-like relationship needs to be modeled, with a varying number of elements in each subset. This can be dealt with by expressing each subset as a DP with a shared atomic base measure, *i.e.*, the base measure is shared and discrete. If the base measure is yet another DP, both the number of elements inside each subset and the number of elements in the parent set increases with the complexity of data. This kind of DP is known as a *hierarchical DP* (HDP). In the HDP, a DP (top-level DP) first discretizes the base measure by sampling points from it. Then, another set of DPs (bottom-level DP) sample points from the atoms of the top-level DP. Thus, bottom-level DPs generate measures that are subsets of the top-level DP, as shown in Fig. 2.5.

2.2.3 Bayesian sequential model

Music score dictates the sequence of performance instructions, and music performance is rendered as an audio signal. Thus, the need to model music as a probabilistic state sequence arises very naturally. Suppose the music score is described as a state sequence of length T , $s = \{s_1 \cdots s_T\}$, whose t th state is denoted as s_t . The values that s_t ’s could take on is the *state space*. For now, let us assume that the state space is either

discrete or continuous. If the state at time t , s_t , is dependent only on the previous state s_{t-1} , then the model is called a *Markov chain*. If the likelihood of drawing the first state in a Markov chain is π and the transition likelihood between two adjacent states is parametrized by θ , then the likelihood of the chain is described as follows:

$$p(s = \{s_1 \cdots s_T\} | \pi, \theta) = p(s_1 | \pi) \prod_{t=2}^T p(s_t | s_{t-1}, \theta). \quad (2.14)$$

This thesis uses two common modeling strategies utilizing the Markov chain, each for when the state space is continuous or discrete.

First, suppose the state space is continuous. When the state space is continuous, this thesis often assumes that the current state is some linear transformation A of the previous state, plus a zero-mean Gaussian noise with precision matrix Λ :

$$s_t \sim \mathcal{N}(As_{t-1}, \Lambda^{-1}). \quad (2.15)$$

When s_t is not observable but there is an observable variable x_t , and the likelihood of x_t , $p(x_t | s_t)$, is defined as a Normal distribution, it is known as a *linear dynamical system*. If A or Λ toggles at each time between a choice of finite matrices, the model is called a *switching-state Kalman filter* [47].

In a Bayesian setting, prior distributions may be introduced on the precision matrix Λ and/or matrix A . For example, the Wishart distribution is conjugate to the precision matrix of a Normal distribution, so it is mathematically convenient to use the Wishart distribution as the prior of Λ .

On the other hand, suppose the state space is discrete, with S possible states. When the state space is discrete, it is often assumed that the current state s_t depends on a transition probability τ associated with previous state s_{t-1} :

$$s_t \sim \text{Discrete}(\{\tau_{s_{t-1}, n}\}_{n=1}^S). \quad (2.16)$$

Here, $\tau \in \mathbb{R}^{S \times S}$, and τ_{ij} is the likelihood for state i to transition to j . When s_t is not observable but it has an observation model $p(x_t | s_t)$ with an observable x_t , it is known as the *hidden Markov model* [108]. When using categorical distribution for transition probability as in Equation 2.16, it is convenient to use the Dirichlet distribution as a prior of $\tau_{i,:}$ due to its conjugacy. Namely, it is useful to set:

$$\tau_{i,:} \sim \text{Dir}(\tau_0), \quad (2.17)$$

where $\tau_0 \in \mathbb{R}^{+S}$ defines the properties of the generated $\tau_{i,:}$. For example, τ becomes sparse (*i.e.*, τ gets dominated by a few entries that have significant values) if the element of τ_0 becomes less than 1. On the other hand, if it is greater than 1, then the mode of τ gets concentrated about τ_0 . If $\tau_0 = 1$, then the prior becomes *non-informative*, *i.e.*,

it puts no preference towards a particular value of τ . It is also possible to use GEM distribution and extend the state-space of the HMM to a countably infinite states; it favors transition to lower-indexed state [25]. It is possible to use a more expressive distribution as well. For example, audio-to-score alignment formulated as HMMs often uses more elaborate transition pdfs that account for tempo constraints [18]. Chapters 3 and 4 introduces these kinds of elaborate or hierarchical transition pdfs.

2.2.4 Bayesian factor decomposition

This section introduces probabilistic latent semantic indexing (pLSI) [57], familiar in the signal processing community also as a special case of probabilistic latent component analysis (PLCA) or as a special case of non-negative matrix factorization (NMF). First, the likelihood of the model is introduced to illustrate the concept behind PLCA. Then, Bayesian extensions through incorporation of various prior distributions are introduced.

Likelihood model

In PLCA, a discretized version of the magnitude spectrogram x_{tf} is treated as a histogram that describes the number of times frequency f was observed at time t . Let c_{tn} be the bin f observed for the n th count for the t th frame. Furthermore, let N_t be the energy inside x_{tf} at frame t . Then, it is clear that $\sum_{n=1}^{N_t} 1(c_{tn} = f) = x_{tf}$.

Suppose that the data has been generated from a set of I bases. The goal of pLSI is to label each count c_{tn} with an index indicating which basis the count was generated from. Let $z = \{z_{t,n}\}_{t=1, n=1}^{t=T, n=N_t}$ be the label associated with each c_{tn} . In pLSI, each basis represents a unique spectral time-slice, such as those representing, say, a trumpet playing $E3$, or a piano playing $A\flat4$, and so on. The spectral time-slice associated with basis i is represented by w_{if} , the likelihood of observing bin f . Assume that the bases are used with different proportions at each frame, with proportion h_{ti} for the i th basis at frame t . Then, the joint likelihood of c and z is given as follows:

$$p(c, z | w, h) = \prod_{t=1}^T \prod_{n=1}^{N_t} \prod_{i=1}^I h_{ti}^{z_{tni}} w_{if}^{z_{tni} c_{tn}}. \quad (2.18)$$

The concept is illustrated in Fig. 2.6.

This kind of framework has been extended for multidimensional data, more familiar to source separation community as the probabilistic latent component analysis (PLCA) [121]. Non-negative matrix factorization (NMF) is a generalization of this kind of decomposition model, which deals with richer observation likelihoods and prior distributions [126].

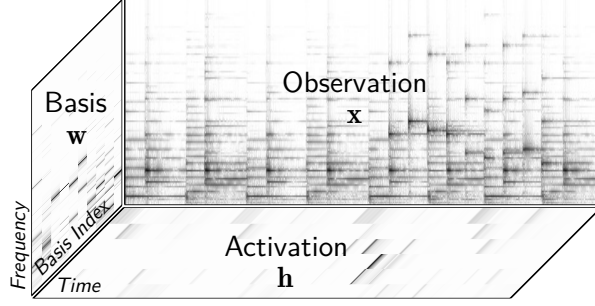


Figure 2.6: Illustration of PLCA. The input magnitude spectrogram x is decomposed into a low-rank representation, described in terms of a weighted linear combination of a few bases w and their activations h .

Bayesian extension

A Bayesian extension is possible by introducing a prior distribution [11] over the spectral basis and activations. Namely, if $h_t \sim \text{Dir}(h_0)$ and $w_i \sim \text{Dir}(w_0)$, then the values of h_0 and w_0 affect the behavior of h_t and w_i , respectively. This kind of model is called latent Dirichlet allocation (LDA).

This can be developed further; if the Dirichlet process is used to model the distribution over basis w_i , then a countably infinite number of bases can be expressed. Namely, one can imagine drawing an infinite number of w_i with weights $\phi = \{\phi_i\}$ from $\text{DP}(\alpha H)$, then drawing h_t so that $h_t \sim \text{DP}(\beta \phi)$. This kind of formulation is known as HDP-LDA [123].

Harmonic bases

In music audio, each basis should correspond to a single pitched instrument. Recall that Equation 2.6 has shown that amplitude spectrum of a pitched instrument can be approximated as a Gaussian mixture. This kind of constraint can be integrated in a LDA-like framework by constraining the basis vector, called the latent harmonic allocation (LHA) [138]:

$$p(c, z | w, h) = \prod_{t=1}^T \prod_{n=1}^{N_t} \prod_{i=1}^I \prod_{j=1}^J h_{ti}^{z_{tnij}} (a_{ij} \mathcal{N}(f | f_{0,ti}, \sigma^2))^{z_{tnij} c_{tn}}. \quad (2.19)$$

Here, $z_{tnij} = 1$ means that the n th count data at frame t was created from the j th partial of the i th instrumental sound. Moreover, h and a are normalized to 1, i.e., $\sum_i h_{ti} = 1$ for all t and $\sum_j a_{ij} = 1$ for all i . Then, prior distribution can be imposed on $f_{0,ti}$, h_{ti} and a_{ij} . For example, a prior of $f_{0,ti}$ may limit the range of the f_0 , or a prior of a_{ij} may express the harmonic structure of a particular instrument.

Recall that the Dirichlet distribution is a mathematically convenient form of prior for multinomial variables like a_i or h_t . Moreover, when the parameter to the Dirich-

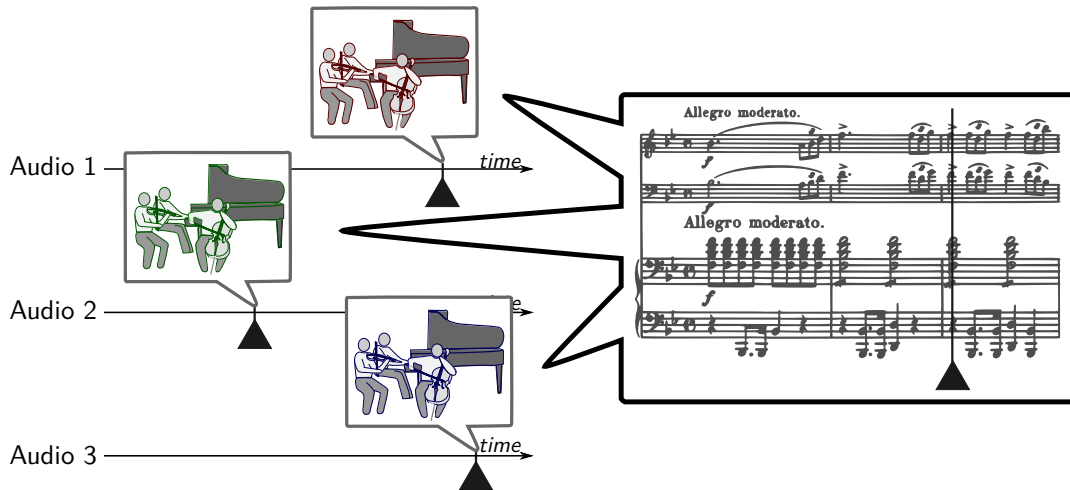


Figure 2.7: Problem statement of music alignment. The goal is to identify time instances in multiple audio signals that all play the same place in the underlying piece of music. When the music score is given, the task is called *audio-to-score alignment*.

let distribution is 1, the prior becomes *non-informative*, meaning that it has no bias towards a particular shape of the multinomial variable. This will become an important concept in achieving robustness to variations in timbre, as will be seen in Chapter 3.

2.3 Music alignment

Music alignment is the task that finds a temporal mapping among multiple media, each of which represents a common piece of music. There are two main kinds of music alignment. First is *audio alignment*, which is the task that finds a temporal mapping between different audio renditions to a same piece of music, such that the temporal mapping determines time instance in each of the audio signals that refer to a same place in the underlying musical piece. Second is the task of aligning an audio signal to a symbolic music score data to the underlying musical piece, called *audio-to-score alignment*. These problem statements are illustrated in Fig. 2.7.

Alignment has gathered considerable interest in the music information retrieval (MIR) community, mainly for three applications. First, it allows a user to compare how different musicians play a same piece of music [114, 134]. This is important especially for analyzing classical music performance: a same piece of music is recorded by many people¹ and tempo fluctuation is integral to the individuality of musical expression [89]. The alignment information can be used for performance visualization [26], switching back and forth between different performances while playing back a piece of music [40], or retrieving interpretations that a user likes [81].

¹There are about 1500 hits when searching for Beethoven's *Ninth* in an online music store as of 2014. The number is monotonically increasing.

Second, aligning music audio and music score is useful for intelligent musical accompaniment [21,111]. To elaborate, the user first prepares a music score of the piece that he wants to play. Then, the user aligns, in real-time, his playing with the part in the music score that he is playing. Finally, an accompaniment that aligns to the user's playing may be generated by synthesizing the remaining parts.

Third, it serves as a valuable cue in source separation. By aligning an audio signal and its music score, it becomes easier to separate particular musical parts from the music audio [29,33,60,64,87]. In query-by-humming applications [55,119], audio alignment becomes valuable for aligning the polyphonic mixture and a user's humming sound.

Though not elaborated in this thesis, there are other contexts in which different kinds of alignment are useful. For example, there are methods for aligning music score graphics to an audio signal [75]. This kind of alignment is useful for, say, viewing a music score while playing back an audio file of the same piece. There are also alignments that are based on matching the lyrics contained in the score [43,78]. In some music genre such as jazz, the music score is a highly simplified representation, showing only harmonic progressions and a rough guide of melody to play on top of the harmony (called a *lead-sheet*). Methods have been proposed to align in this kind of case as well [28].

2.3.1 Approaches to audio alignment

Aligning a music audio signal is difficult because a same piece can be played with different timbre, volume balance, tempo trajectory and under different acoustic conditions. To tackle alignment, it is important to find a representation that decouples factors that are invariant to performance differences and those that depend on each performer. As discussed in Section 2.1.3, the pitch sequence and the instrumentation are typically the two most invariant musical aspects. Thus, the first goal is to decouple an audio signal into intermediate representations that represent pitch and/or instrumentation. Then, alignment may be found using such intermediate representations.

2.3.2 Intermediate representation

As discussed in Equation 2.6, the spectral representation of harmonic sounds contains multiple harmonic partials. The f_0 's at each time instance correspond to the pitches that are played off of the music score at that time. Since the gain of each source and its harmonic structure are unknown, a representation that is robust to variations in relative volume and harmonic structure is useful.

To this end, it is possible to design another feature that accumulates the power contained in a particular pitch-class, to arrive at a cyclic vector that represents the power contained in each pitch-class (e.g., power in C, C $\sharp$, and so on). This feature is called the *chroma* vector [44], even though the definition is conceptual and there are many ways

to realize it (*e.g.*, one may either sum over pitch-classes, take the maximum, sum over the logarithm, and so on). A bulk of research has focused on how to remove timbral dependency in the chroma vector [34, 93, 101]. The performance is also dependent on how one preprocesses the chroma [16]. In addition to the chroma, it is possible to capture note onsets by extracting the half-wave rectified adjacent difference [34]. Chroma-based approach has been used extensively [58, 62, 80] because the resulting feature is less sensitive to timbral variations than a spectral representation.

Another approach involves extracting which instrument is playing which pitch at each time instance. This is often done by first generating “bases” of different instrumental sounds playing different pitches, and then finding which bases are active [17]. Source decomposition methods like those introduced in Section 2.2.4 are useful for finding the active bases. To attain both robustness to timbre and volume, it is possible to formulate a probabilistic model that jointly infers the dynamics and timbre, given the music score [61, 84, 103]. To achieve robustness to noise, source separation methods have been used as a front-end to segregate components that are relevant to alignment and those that are irrelevant [69]. Ways to decompose an audio signal into different components are discussed in Section 2.2.4.

Attaining robustness to reverberation

The spectrum from which the features are extracted contains reverberation. This is undesirable because reverberation adds unwanted noise components that degrades the extracted feature. To elaborate, it adds extra sounding components through *late reverberation*, smearing the spectrum in time. Therefore, suppressing the late reverberation through *dereverberation* methods is useful.

There are many dereverberation methods, each tailored for a specific application. For example, for speech recognition purposes, it is possible to dereverberate the sequence of speech recognition features [117]. For post-production purposes, the goal is to recover the dereverberated audio signal. Dereverberation, for this use, may operate either in non-invertible domain, *i.e.*, the transformed representation and the time domain signal is not bijective, such as the power spectrum, or in invertible domain such as the time domain or the complex spectral domain. Typical dereverberation methods focus on suppressing the late reverberation.

Dereverberation methods on a time domain signal or complex spectrum seek to find the transfer function of a room, and recover the dry signal using inverse filtering [88, 92]. When reverberation is assumed to be a moving-average process, and the reverberation filter satisfies a few reasonable assumptions, the problem of source signal estimation becomes that of recovering the noise that drives an auto-regressive system, using a formulation known as the multi-step linear prediction [46, 71, 96]. By changing the noise model, the reverberation method can be tuned to dereverberate a particular kind of signal. For example, white Gaussian noise [95], auto-regressive model for speech [5, 32, 97, 140, 141], or mixture of harmonic sounds [136] have been

proposed. If one over-specifies the model of the source signal, the method becomes selective. Inversely, if one under-specifies the model, the method becomes more versatile, at the expense of reduced accuracy.

On the other hand, many studies focus on suppressing reverberant signal in the power spectrum domain [31, 53, 66, 74, 76]. For example reverberation has been formulated as a moving-average process [66], or as a moving-average process with an exponential decay, *i.e.*, as an AR(1) process [2, 76, 139]. In this family of algorithms, the source model can be customized to suit the characteristics of the source signal. To name a few, additive white Gaussian noise [74] or generalized Normal distribution [66] are some of the possibilities.

Dereverberation methods based on the power domain typically assume additivity of the power spectrum, which may degrade the performance. Methods based on the complex spectrum, on the other hand, require no such assumptions, but are sensitive to slight displacement of the source signal. To see this, note that the phase response of reverberation changes drastically with a slight displacement of the source signal. However, the magnitude response changes very slightly. Therefore, methods based on the magnitude or power spectrum representation is more robust to source signal displacement than methods based on complex spectrum.

2.3.3 Alignment estimation

Once an intermediate representation has been generated for each signal, either explicitly in the form of feature sequence or implicitly as a probabilistic model, the alignment method temporally maps the multiple intermediate representations. For the remainder of this section, it is assumed that an N dimensional feature sequence $\mathbf{x}_{dt}$ for the d th music audio signal is given. When using only one audio signal, the feature sequence is written as $\mathbf{x}_t$.

To find the alignment a path optimization technique known as the dynamic time-warping (DTW) is typically used for explicitly-determined feature sequences [27, 58, 99]. For a probabilistic model, a popular choice is to formulate alignment as decoding of an HMM and its variants [18, 61, 112], or as a continuous state-space model [30, 103].

Path optimization approach

The path optimization approach considers aligning two audio signals. For audio-to-score alignment, the music score is synthesized into an audio signal using a synthesizer.

First, a distance matrix R_{ij} is computed, each entry of which contains the distance between the feature of the i th frame in the first audio and j th frame in the second audio. Euclidean distance $R_{ij} = \|\mathbf{x}_{1i} - \mathbf{x}_{2j}\|_2$, or the cosine distance $R_{ij} = 1 - \frac{\mathbf{x}_{1i}^T \mathbf{x}_{2j}}{\|\mathbf{x}_{1i}\|_2 \|\mathbf{x}_{2j}\|_2}$ are often used [34]. Note that if the two audio signals play the same piece and if

the intermediate representation represents the pitch content well enough, one will observe an almost diagonal path with a consistently low distance.

Next, a path over the similarity matrix indices that minimizes the similarity is computed, subject to various path constraints. A popular constraint is to require that the index (i, j) may only traverse to $(i+1, j)$, $(i, j+1)$ or $(i+1, j+1)$ [20]. Then, the path may be computed using dynamic programming, by first computing the following matrix:

$$C_{ij} = R_{ij} + \min\{C_{i-1,j}, C_{i-1,j-1}, C_{i,j-1}\}. \quad (2.20)$$

Here, for any $(i, j) \notin [1, \dots, I] \times [1, \dots, J]$, $C_{ij} = 0$. Then, alignment is determined by backtracking C_{ij} , starting at C_{IJ} , where I and J are the size of the feature of the first and second signals, respectively. Namely, the alignment is a list $[(i_0, j_0), \dots, (i_N, j_N)]$ where

$$(i_n, j_n) = \arg \min_{(i,j) \in \left\{ \begin{array}{l} (i_{n+1}-1, j_{n+1}) \\ (i_{n+1}, j_{n+1}-1) \\ (i_{n+1}-1, j_{n+1}-1) \end{array} \right\}} C(i, j), \quad (2.21)$$

and $(i_0, j_0) = (1, 1)$, $(i_N, j_N) = (I, J)$. N is determined given these terminal conditions. Other DP-based methods with different constraints have been also used recently [41, 51], to cope with repeats and cuts in music.

In DTW-based approach, it is infeasible to jointly align more than two renditions at a time. This owes to the computational complexity that grows exponentially with the number of signals D to align. To perform multiple audio alignment, one may compute pairwise alignments for all possible combination of signals, ignoring any inconsistencies between the pairwise alignments. It is also possible to incorporate a measure of inconsistency among pairwise alignments, and minimizing it to refine alignment [81]. To provide a consistent alignment, it is possible to create a kind of “reference” signal to which each signal aligns, refining the reference based on the observations [132].

Statistical estimation approach

For audio-to-score alignment, a statistical model of the generated audio features given the music score is first designed. For this, the HMM is often used [108]. In the HMM-based approach, each position in the score is modeled as a distinct state. Then, at frame t , the audio signal is assumed to play some position s_t of the score. Typically, a state can transition only to itself or to the next adjacent state, conveying the idea that the state sequence is monotonically increasing. This kind of constraint is called *left-to-right*; this kind of HMM is called a left-to-right HMM (LRHMM). The joint likelihood of an HMM is then given as follows:

$$p(x, s | \pi, \theta, \phi) = p(s_1 | \pi) \prod_t p(x_t | \phi_{s_t}) p(s_t | s_{t-1}, \theta). \quad (2.22)$$

Here $p(x_t|\phi_s)$ is the likelihood of observing the feature at frame t , given parameters associated with state s of $\mathbf{s}$. This is analogous to the similarity matrix used in Section 2.3.3. For example, the Euclidean distance may be conveyed by letting $p(x|\phi)$ be a Normal distribution with a fixed covariance $x|\phi \sim \mathcal{N}(x|\mu, \sigma^2)$ with $\phi = \{\mu, \sigma\}$. and the cosine distance may be conveyed by the von Mises-Fisher distribution (vMF distribution), by letting $x|\phi \sim \text{vMF}(\mu, \kappa)$, with $\phi = \{\kappa, \mu\}$.

A much richer set of distributions as $p(x|\phi)$ can be encoded. For example, the idea that a particular position in the score may emit distinctly different kinds of sound may be conveyed by modeling the observation pdf as a multimodal distribution. It is also possible to look explicitly for what kind of sounds are being simultaneously played by using a source segregation model, like those discussed in Section 2.2.4.

Since a music score conveys the relative duration (e.g., a quarter note versus an eighth note), it is relevant to incorporate duration constraints. Many studies focus on the idea that the tempo transitions smoothly [112]. Namely, the tempo is often modeled as a random walk – the tempo at the current note v_s is that at the previous note v_{s-1} plus a Gaussian noise ϵ_s :

$$v_s = v_{s-1} + \epsilon_s. \quad (2.23)$$

Then, this information, together with some small factors for correcting the phase of the beat positions, can be incorporated as the probability that a signal resides in state s [18]. This kind of formulation, in which the state duration probability is explicitly given, is called a hidden semi-Markov model (HSMM). It is also possible to use a continuous state-space s_t instead of discrete state-space of an HMM. This way, the score position s_t is expressed as a difference equation, such as $s_t = s_{t-1} + v_t$, with some additive noise [30, 91, 103].

When performing audio-to-audio alignment, there are two main approaches. First, one audio signal may be designated as the “reference” audio signal, and subject the remaining signal to be aligned to the reference [91]. In this case, the feature sequence extracted from the reference serves as the parameters to the statistical model. It is also possible to infer the model parameter by aligning to a left-to-right HMM [86]. This may be conceptually thought of as doing a partial music transcription first, then aligning to the transcribed score.

2.3.4 On asynchrony between notes

Sometimes when fine details of playing need to be analyzed, one needs to take into account slight asynchrony in the note onset timings from the music score, as shown in Fig. 2.8. These approaches are incapable of treating such an asynchrony, and have been dealt with by post-processing the alignment result [23, 87]. This thesis defines the asynchrony between different parts as *inter-part asynchrony*.

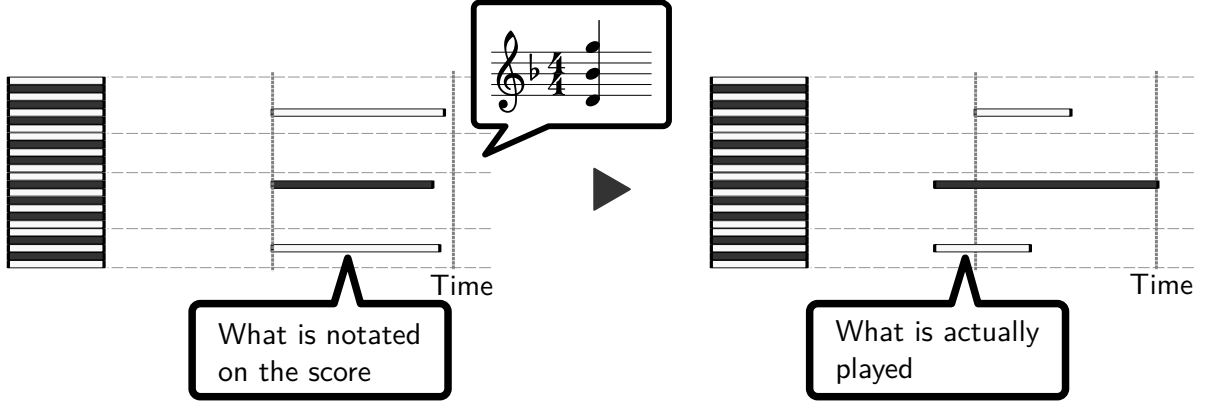


Figure 2.8: Example of inter-part asynchrony. Even though the music score specifies different parts (shown in one stave) to play three notes in sync, the performer(s) may not start the sound simultaneously.

2.4 Formulation of Bayesian music alignment

Based on the previous discussions, let us formulate music alignment from a Bayesian perspective. In music alignment, the underlying piece of music, z , generates (1) parameters to instantaneous sounds at each point in the piece of music, and (2) a sequence that represents how such instantaneous sounds are organized in time. Suppose that the underlying piece of music z is described using N states. Moreover, suppose that parameters governing the instantaneous sound, $\theta = \{\theta_{dn}\}_{d=1, n=1}^{D, N}$, are generated from a distribution $p(\theta|z)$. Furthermore, suppose that parameters governing the sequence of sound, $s = \{s_{dt}\}_{d=1, t=1}^{D, T_d}$, are generated from a distribution $p(s|z)$. To summarize, music alignment is formulated as follows:

$$p(x, s, z, \theta) = p(z) \prod_{d, t} p(x_{dt} | \theta_{d, s_{dt}}) p(s|z) p(\theta|z). \quad (2.24)$$

This concept is summarized pictorially in Fig. 2.9. In a Bayesian setting, then, the goal is to compute $p(s, z, \theta|x)$.

Different music alignment tasks may be formulated as specializations of this equation. These are summarized as follows:

Audio-to-score alignment z is *given* in terms of a music score. θ governs the kind of sound that may be generated at each point in the score, and s governs how the music score gets expanded in time.

Audio-to-audio alignment z is unknown. To allow inference, it is assumed that each point in the underlying piece of music z emits a *similar* kind of sound for each signal x_d . Thus, a typical formulation assumes that $\theta_{dn} = \theta_{d'n}$ for any d and d' .

Subset music alignment z unknown. Moreover, the kind of sound generated at each

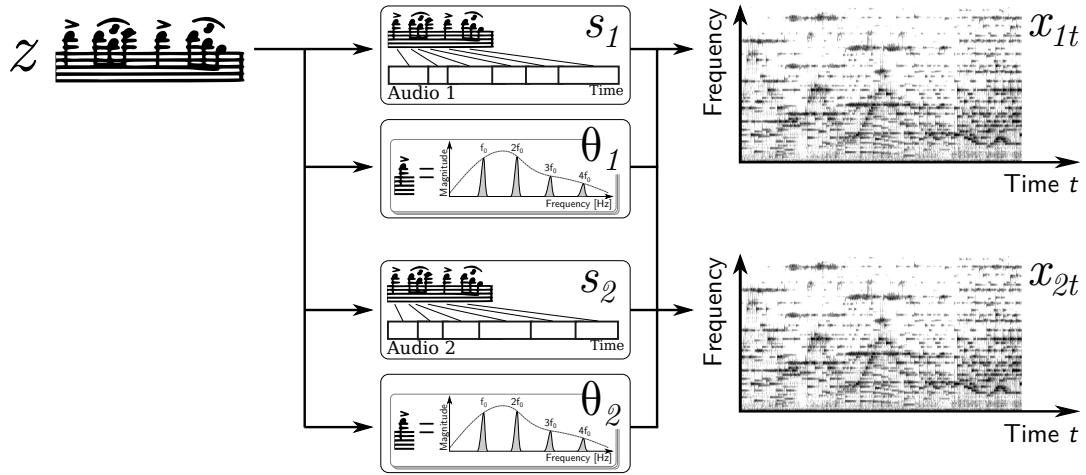


Figure 2.9: Unified formulation of music alignment.

position in the underlying piece of music z may *vary* significantly across different signals. Thus, θ no longer specifies “shared” instantaneous sounds, but a variation of instantaneous sounds across different signals.

Chapter 3

Bayesian Audio-to-Score Alignment

This chapter presents an offline method for aligning an audio signal and individual music instrument parts comprising a music score. The proposed method tackles the problem of variations in dynamics, timbre, tempo and note onset timings. It does so by formulating a generative model of magnitude spectrogram and inferring the posterior distribution.

As reviewed in Chapter 2, a key issue in audio-to-score alignment is dealing with robustness to variations in timbre, dynamics and tempo. Inter-part asynchrony, *i.e.*, notes that are notated to be played synchronously being played out-of-sync, is an important problem in music alignment. To tackle these issues, in most alignment approaches, a timbre-robust representation of the input audio is first extracted. Then, the timbre-robust representation is matched to the music score, using a distance metric that is amenable to slight variations in the volume of the played parts. Finally, the inter-part asynchrony is estimated with a post-processing method. This kind of multi-stage approach fundamentally assumes that the timbre, the volume, the alignment and the inter-part asynchrony are independent of each other.

A fundamental problem of a multi-stage approach to audio-to-score alignment is that the timbre, the volume, the alignment and the inter-part asynchrony are *mutually dependent*. That is, if the alignment is given, one gains a better understanding of the timbre that different instruments play at different pitches. This updated knowledge can then be used to create a more timbre-robust representation for the alignment stage to use. Similar argument goes for volume and inter-part asynchrony as well. The mutual dependencies among the timbre, the volume, the alignment and the inter-part asynchrony calls for joint inference.

This chapter presents an audio-to-score alignment method that jointly infers the timbre, the volume, the alignment and the inter-part asynchrony in a Bayesian manner. Bayesian inference allows us to encode prior knowledge on the musical piece, and its certainty. This is important because the alignment should be estimated in a data-driven manner, while taking into account weak assumptions that may be effectively overridden with ample observation.

For expressing the timbre, the volume, the alignment and the inter-part asynchrony in a unified model that permits joint inference, the proposed method expresses three ideas: (1) different parts agree on a common tempo trajectory, (2) each part executes the music score independently, given the tempo trajectory, (3) each position in the score is rendered as a weighted sum of harmonically spaced sinusoids. Technically, the method first expresses the shared tempo trajectory as an LDS that expresses the continuity of the tempo. Then, the method expresses the observed signal as a nested HSMM, where each HSMM corresponds to different “parts.” Asynchrony is expected to occur among different parts. For each HSMM, the duration pdf is expressed using the shared tempo trajectory, and the transition pdf describes the structure of the musical piece. Then, the method represents each audio using a Bayesian spectral model that expresses a polyphonic mixture of sinusoids spaced integer multiples apart [138].

The probabilistic model is formulated in Section 3.1, and an inference method is derived in Section 3.2. Finally, Section 3.3 evaluates the proposed method.

3.1 Formulation

The goal of the proposed method is to estimate the tempo, timing fluctuation between parts, timbre, dynamics, pitch, and sequence of states. Here “part” means an arbitrary combination of voices for which instrumentation is different and asynchrony is expected. For example, in a piece comprising of a violin and a piano, single part may be assigned to the violin and another to the piano part. If one is concerned in how the pianist coordinates the left and the right hands, different parts may be assigned to the notes played by the left hand and those played by the right hand.

Fig. 3.1 shows the idea behind the proposed method. The observed spectrogram x_{tf} is expressed in terms of an underlying generative model, which describes a top-down generative process of the spectrogram. Specifically, the method first assumes that for a given piece, the comprising parts (the violin, the cello and the piano in this case) agree on a common tempo trajectory (upper left-hand side of the figure). The tempo trajectory expresses how fast the piece is played at a given time instance. Then, once the parts agree on a common tempo trajectory, each part plays independently of others. Then, given the executed note sequences, each part renders a sequence of normalized spectral time-slices. Rendering process takes into account the notated pitches, the harmonic structure and the relative dynamics. Finally, the model outputs the observed spectrogram as a weighted mixture of the normalized spectrograms generated by these parts.

To express this kind of behavior, the proposed method represents the shared tempo trajectory as an LDS, which expresses smoothness of the tempo trajectory. Then, the proposed method expresses the note sequence of each part as an HSMM. It is assumed that the audio starts at the beginning of the score, and ends at the end of the

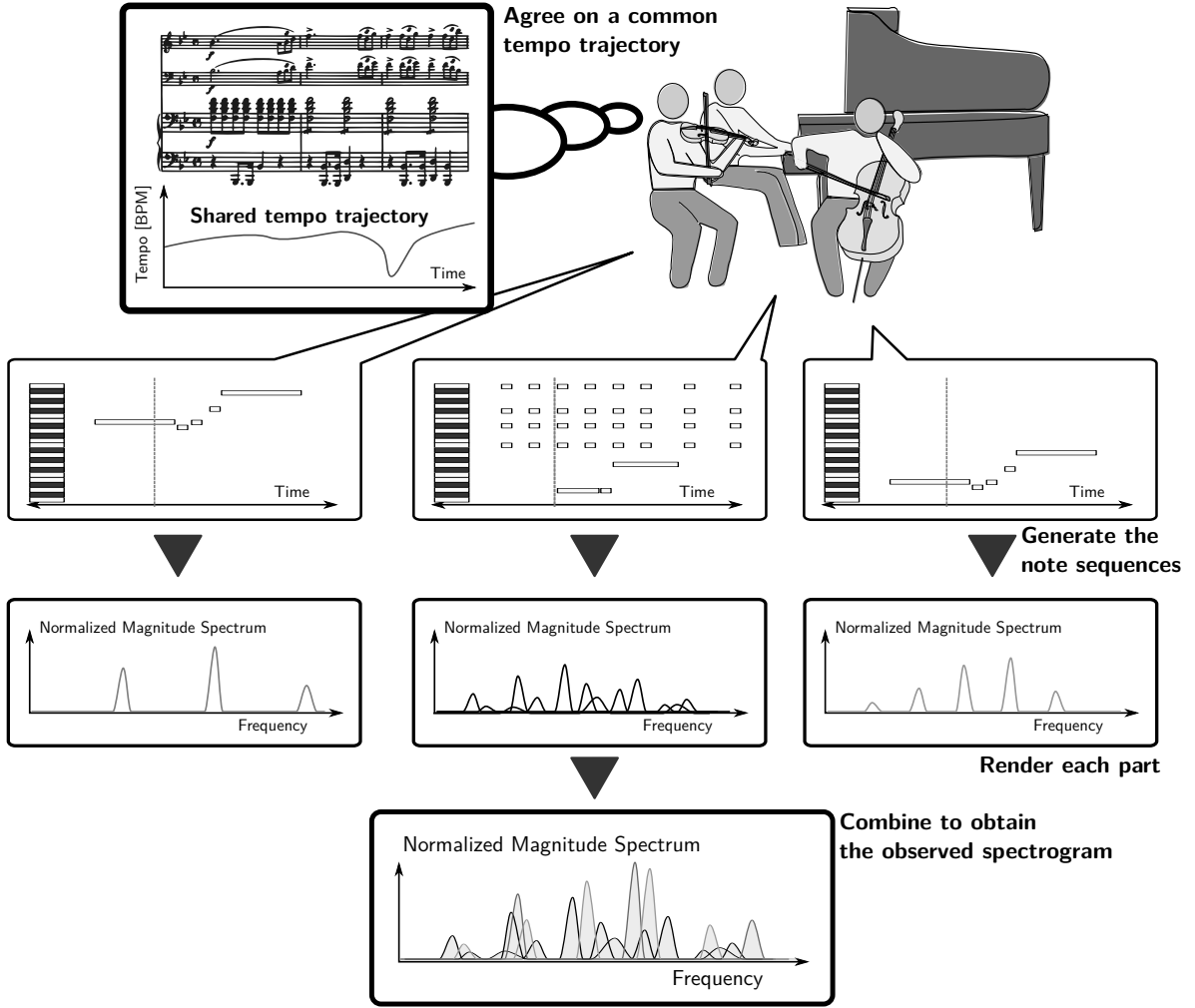


Figure 3.1: Conceptual idea of the proposed method.

score. Here, each state describes a distinct note combination and has a note duration associated with it, as shown in Fig. 3.2. How long each part plays each state depends on the shared tempo trajectory, and slight asynchrony in timing caused by musical expression and/or physical limitations.

Given the sequence of notes, the method represents a normalized spectrogram of each part. The normalized spectrogram is constrained such that significant components exist at integer multiples of the notated fundamental frequencies. The pair of pitch/instrument combination is defined as an **instrument-pitch pair (IPP)**. It is assumed that each of I IPPs is harmonic; thus, each IPP is assumed to emit a histogram count of the acoustic frequencies that are approximately integer multiples of the notated fundamental frequency. The relative strengths of these components reflect the relative volume balance between different notes within the state, and the relative strength of each harmonic partial. This concept is expressed using LHA because (1) it is a model of harmonic mixtures and thus mixes well with the proposed method, and

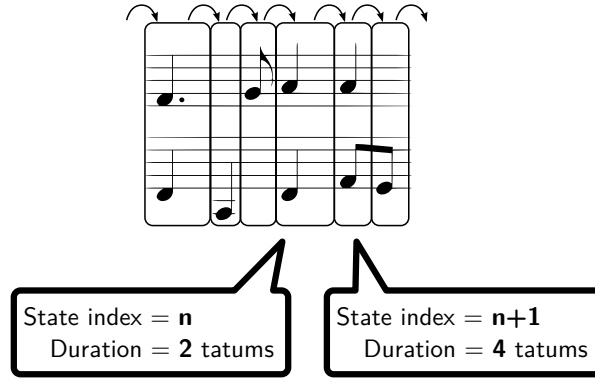


Figure 3.2: Model of the music score as a state sequence.

(2) its interpretation of the power spectrum as count data dovetails with the proposed method’s interpretation of x .

Finally, the spectrogram x_{tf} is obtained by accumulating quanta of energy over the time-frequency axis. Specifically, the quantized version of the constant-Q transform (CQT) spectrogram x_{tf} is interpreted as the number of energy quanta observed at time-frequency bin (t, f) . This kind of interpretation of audio signal as a histogram count has been successfully applied in many tasks such as f_0 estimation [138], or sound source separation [13]. Each quantum is then assigned the originating part, pitch, instrument and the score position (state). To assign which score position, part and the IPP generated a particular energy quantum, the normalized spectrograms of different parts obtained earlier is used as the likelihood of observing a quantum of energy at a particular time-frequency bin.

Now, each aspect of the proposed method is elaborated.

3.1.1 Modeling the music score state sequence

The model of state sequence is based on first generating a smooth global tempo trajectory using an LDS, and for each part, creating an HSMM such that the duration of each state is governed by the LDS.

Modeling the global tempo trajectory

The global tempo trajectory is designed as a smooth process, such that adjacent tempi tend to remain close to each other. This concept is modeled using an LDS or, equivalently, an auto-regressive model of order 1. LDS is useful because it describes the current observation in a sequence in terms of the deviation from previous observation; if the current observation is allowed to deviate only slightly from previous observation, then it naturally expresses the smoothness of tempo trajectory.

Let $\mathcal{T}_n$ be the logarithm of the number of audio frames per tatum at state n of N ; this is related to log-BPM. A tatum is the greatest common divisor of the notated

lengths of all the notes in the score. Similar to [112], the proposed model assumes that $\mathcal{T}_n$ deviates by a small amount from $\mathcal{T}_{n-1}$, following a Normal distribution centered about $\mathcal{T}_{n-1}$:

$$\mathcal{T}_n \sim \mathcal{N}\left(\mathcal{T}_{n-1}, \mathcal{L}_{n-1} \lambda_n^{(T)-1}\right) \quad (3.1)$$

$\mathcal{L}_n$ is the number of tatums of state n (for example, a state whose length is a quarter note, with tatum defined as a sixteenth note has $\mathcal{L}_n = 4$). $\lambda_n^{(T)}$ is the precision (*i.e.*, the inverse variance) of the tempo difference – the greater the value, the less $\mathcal{T}_n$ deviates from $\mathcal{T}_{n-1}$. The model assumes that this parameter is generated from the Gamma distribution:

$$\lambda_n^{(T)} \sim \text{Gam}\left(l_n^{(T)}, v_n^{(T)}\right) \quad (3.2)$$

$l_n^{(T)}$ and $v_n^{(T)}$ govern the distribution of the precision between $\mathcal{T}_{n-1}$ and $\mathcal{T}_n$. Specifically, the ratio $l_n^{(T)}/v_n^{(T)}$ is the mean of $\lambda_n^{(T)}$, and the ratio $l_n^{(T)}/v_n^{(T)2}$ is its variance. Thus, they may be set as to convey the default value of $\lambda_n^{(T)}$ along with its uncertainty. For example, steady tempo can be set by setting $l_n^{(T)}$ much greater than $v_n^{(T)}$, and making $v_n^{(T)}$ large, so that the expected tempo difference variance is small and the variance of $\lambda_n^{(T)}$ is small, conveying the idea that one is confident that the tempo change is small. On the other hand, upon encountering a tempo marking, these values can be set in such a way that $\lambda_n^{(T)}$ is distributed with a large variance, *i.e.*, the tempo can jump by a large amount.

Modeling a score part as an HSMM

The music score, as mentioned previously, is represented as a sequence of states in a manner depicted by Fig. 3.2, and a state sequence is defined for each of the H parts. The model of the state sequence must be consistent with both the allowed transition between states, and the global tempo trajectory. The first requirement means that the transition between states must abide by the music score, *i.e.*, state n can transition to state $n + 1$, plus repeats, cuts, or other structural markings, if any. The second requirement means that duration in each state must be explicitly modeled. To meet these requirements, the state sequence of the score part is modeled as an HSMM.

The proposed model is completely specified by four variables: state sequence, initial state pdf, state transition pdf, the state duration pdf. Typically, an HSMM needs to specify the cumulative density function of the state duration pdf in order to rigorously treat the terminal condition; this information is not necessary in the proposed model because we assume that the audio ends at the end of the score, *i.e.*, the terminal state is constrained to be the end of the last state.

Initial state pdf π , a categorical distribution, specifies the distribution of states at $t = 1$. It is drawn from a Dirichlet distribution:

$$\pi \sim \text{Dir}(\pi_0). \quad (3.3)$$

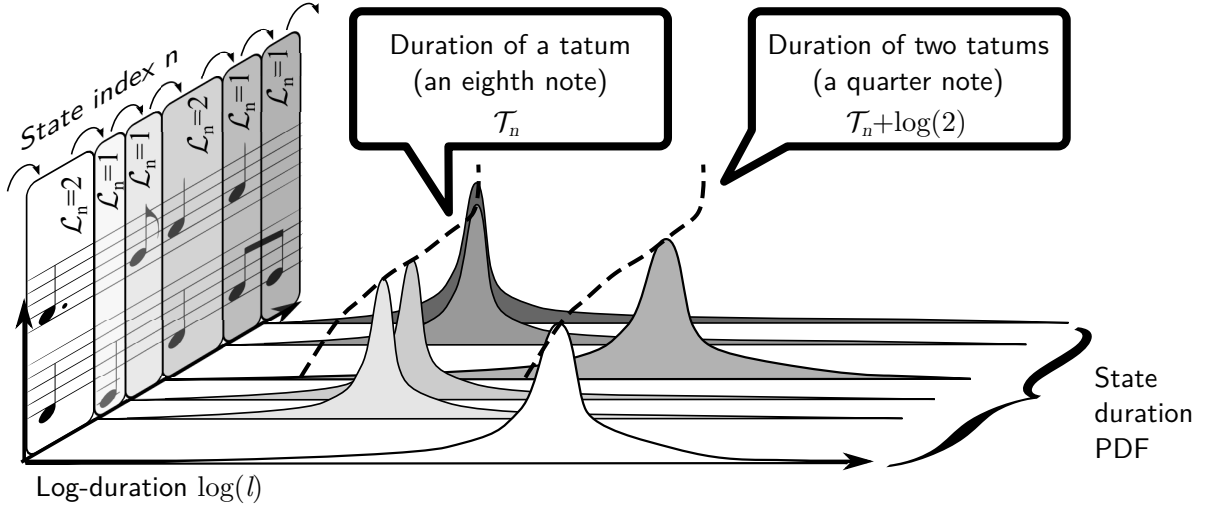


Figure 3.3: Model of the state duration pdf.

The music audio is assumed to start at the beginning of the score. To express this knowledge, π_0 is set such that the first element $\pi_{0,1}$ is set to 1, and the remaining value to a tiny positive value $\epsilon \ll 1$ (we set $\epsilon = 10^{-50}$). Since $\langle \pi \rangle \propto \pi_0$, this kind of setting encourages the initial state to be in the very beginning of the score. Other knowledge, such as optional cuts in the beginning (*e.g.*, abridging the introductory passage of a concerto), can be encoded in the initial state pdf as well.

State transition pdf $\tau_{n'n}$, a categorical distribution, is the probability of transitioning from state n' to n . It is also drawn from a Dirichlet distribution:

$$\tau_{n,\{1 \dots N\}} \sim \text{Dir}(\tau_{0n,\{1 \dots N\}}). \quad (3.4)$$

The prior τ_0 describes the possible transitions of the music score: it should be mostly left-to-right (*i.e.*, the only allowed transition is from state n to $n+1$), except for few occasions where music structure mandates otherwise, *e.g.*, repeats. The method parses the music score to determine the appropriate τ_0 : hyperparameters for allowed transitions are set to 1 (non-informative), and everything else is set to ϵ . Note that transition to the same state is not allowed, *i.e.*, $\tau_{0n,n'} = \epsilon$.

State duration pdf governs the expected number of audio frames per tatum given the notated duration of each state. It should be consistent with the global tempo trajectory $\mathcal{T}_n$, in that the expected duration is centered about $\exp \mathcal{T}_n$, as illustrated in Fig. 3.3. To this end, the log-duration at state n , $\log l$, is modeled as a Normal distribution centered about the expected duration according to the tempo model. Recall that $\mathcal{L}_n$ is the number of tatums of state n and $\mathcal{T}_n$ is the log-duration of tatum at state n generated by the LDS. The duration PDF is then modeled as a Normal distribution centered about $\mathcal{T}_n + \log \mathcal{L}_n$, with variance of σ_T^2 , $\log l \sim \mathcal{N}(\mathcal{T}_n + \log \mathcal{L}_n, \sigma_T^2)$. This kind of model allows each part to fluctuate from the global tempo trajectory, where a small value of σ_T strengthens the effect of the global tempo.

The state sequence is described as a sequence, over time (audio frame index) t , of two variables: the state n and countdown timer value l . First, at $t = 1$, the HSMM chooses the initial state according to the initial state PDF, and chooses the initial countdown timer value, according to the state duration PDF at the initial state. Then, for each time $t = [1 \cdots T]$, if the current countdown timer value is not 1, the timer is decremented by 1. If the timer value is 1, the HSMM chooses the next state according to the state transition PDF associated with the current state, and chooses the countdown timer value by drawing from the duration PDF of the next state. Let $s_{ht} = \{s_{htnl}\}_{n=1, l=1}^{N, L}$ be a 1-of- LN binary variable that indicates that the state of part h at time t is state n , with l frames remaining in n . “1-of- K binary variable” is a K -dimensional unit vector, where the dimension corresponding to the active state has a binary value of 1, and the other dimensions are set to 0. It is crucial to note that while 1-of- K vector is binary, its expectation is a continuous variable. Note that this assumes that the sequence stays at a state for at most L frames. Then, s is described as follows:

$$\begin{aligned}
 p(s_{h,\{1 \cdots T\}} | \mathcal{T}, \pi, \tau) = & \prod_{n,l}^{N,L} \pi_n^{s_{h1nl}} \underbrace{\prod_{t>1, n=1, l=1}^{T,N,L} 1^{s_{h,t-1,n,l+1} s_{htnl}} \prod_{t>1, n,l,l' \neq l+1}^{T,N,L,L} 0^{s_{h,t-1,n,l'} s_{htnl}}}_{(\dagger)} \\
 & \times \underbrace{\prod_{t>1, l,n,n' \neq n}^{T,L,N,N} \left(\tau_{n',n} \mathcal{N}(\log l | \mathcal{T}_n + \log \mathcal{L}_n, \sigma_T^2) \right)^{s_{h,t-1,n',1} s_{htnl}}}_{(\ddagger)}. \quad (3.5)
 \end{aligned}$$

Here, the convention $0^0 = 0$ is used. The terms braced by $(\dagger)$ expresses the count-down timer: the left term indicates that count-down timer must deterministically decrement until 1, and the right term makes any other transition illegal. The terms surrounded by $(\ddagger)$ expresses both the state transition and state duration. The left term in the parenthesis governs the state transition, and the right term governs the count-down timer value at the next state. It is activated only when the count-down timer value is 1, meaning that when the count-down timer terminates, it jumps to the next state, draws a duration value, and assigns it to the count-down timer.

3.1.2 Generating a histogram count

Once the state sequence is generated for all H parts, the CQT spectrogram x can be generated by (1) choosing, for each count of x , the score position n , the part h , the IPP i , and the harmonics index j that generates the count and (2) emitting a quantum at some time-frequency bin (t, f) based on the choice made in (1).

At time t at frequency f , there are x_{tf} counts. Each of these x_{tf} counts is associated with one of H parts that generated it. The part associated with the c th count is denoted as $\mathcal{M}_{tfc}$. The likelihood that a count is associated with part h follows a

categorical distribution parametrized by $\mathbf{m}_{tf} = \{m_{tfh}\}_{h=1}^H$. This parameter describes the relative gain of each part at (t, f) . It depends on the time and frequency because the likelihood of choosing the h th part clearly depends on t through the score position and f through the IPPs comprising the score of the h th part, which governs which frequency f is likely to be observed. At the same time, the independence of m from the actual count index means that once (t, f) is given, there is no bias towards choosing any particular quanta created inside (t, f) . Because which frequency bin is likely to appear is unknown *a priori*, a non-informative prior is imposed on m_{tfh} ; this is realized by posing a Dirichlet prior with unit hyperparameter $m_{0tfh} = 1$:

$$\mathbf{m}_{tf} \sim \text{Dir}(\mathbf{m}_{0tf}). \quad (3.6)$$

Then, $\mathcal{M}_{tfc} = \{\mathcal{M}_{tfch}\}_{h=1}^H$ is assigned one of H latent variables, drawn according to m_{tfh} :

$$\mathcal{M}_{tfc} \sim \text{Discrete}(\mathbf{m}_{tf}). \quad (3.7)$$

Next, the method chooses which IPP is generated. The likelihood of observing an IPP is dependent on the notated notes at each state, and the relative volume of notated notes. This is expressed as a categorical likelihood of observing IPP i at state n of part h , $\mathbf{e}_{hn} = \{e_{hni}\}_{i=1}^I$. Note that $\mathbf{e}_{hn}$ is independent of time; in other words, the relative volume of IPPs within a state is assumed to be stationary. Furthermore, $\mathbf{e}_{hn}$ is drawn from a Dirichlet distribution:

$$\mathbf{e}_{hn} \sim \text{Dir}(\mathbf{e}_{0hn}). \quad (3.8)$$

The prior information $\mathbf{e}_{0hn}$ should convey two aspects. First, the relative volume of notated IPPs should be non-informative, unless the dynamics is known through notation such as *forte* or *piano*. Second, the relative volume of IPPs that are not notated must be very close to zero. To this end, e_{0hni} is set to 1 (non-informative) for notated IPP indices i , and $e_{0hni'}$ is set to some positive constant much smaller than 1 for un-notated IPP indices i' . Finally, the IPP index i drawn from $\mathbf{e}$ is assigned to $\mathcal{I}_{hnf} = \{\mathcal{I}_{hnfi}\}_{i=1}^I$, a 1-of- I binary variable:

$$\mathcal{I}_{hnf} \sim \text{Discrete}(\mathbf{e}_{hn}). \quad (3.9)$$

Once the IPP index i is drawn, the index of the harmonic partial is drawn. Each IPP is associated with a unique harmonic structure with J partials, from the fundamental up to the J th partial. The harmonic structure is described as a single categorical distribution, $\mathbf{a}_{hfi} = \{a_{hfij}\}_{j=1}^J$, the harmonic structure of the i th IPP of part h . $\mathbf{a}$ is drawn from Dirichlet distribution:

$$\mathbf{a}_{hi} \sim \text{Dir}(\mathbf{a}_{0hi}). \quad (3.10)$$

Then, the index of harmonic partial is assigned to $\mathcal{H}_{hfi} = \{\mathcal{H}_{hfij}\}_{j=1}^J$, a 1-of- J binary variable:

$$\mathcal{H}_{hfi} \sim \text{Discrete}(\mathbf{a}_{hi}). \quad (3.11)$$

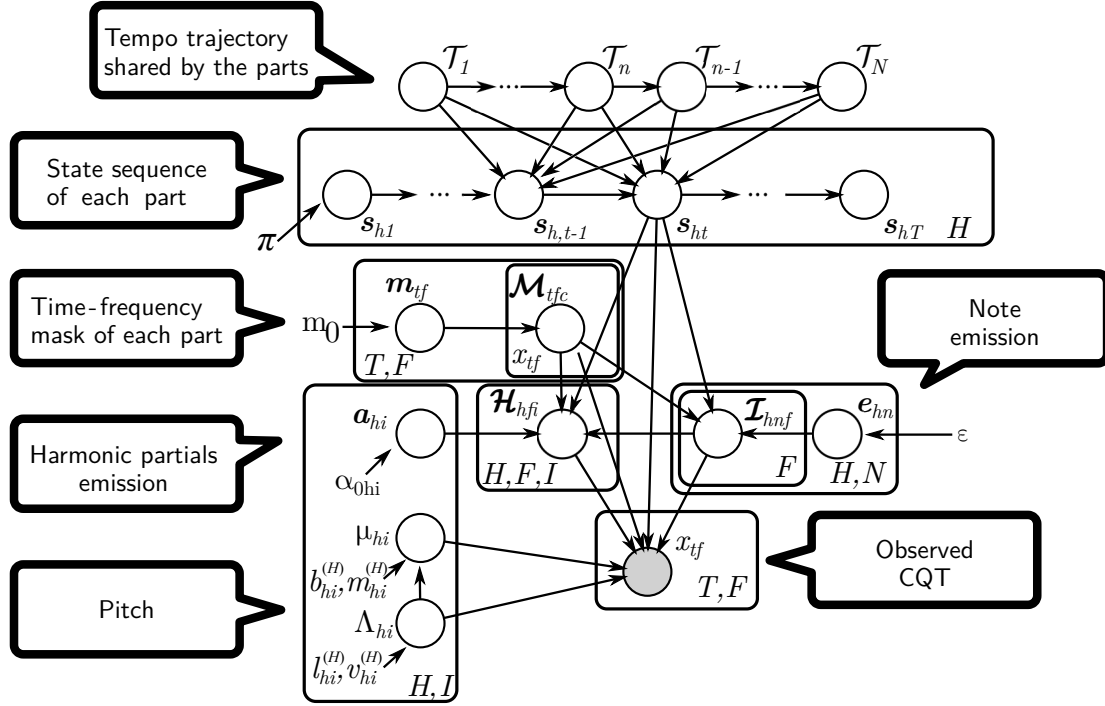


Figure 3.4: Graphical model of the proposed audio-to-score alignment method. Circles represent random variables and arrows indicate conditional dependencies.

a_{0hi} determines the prior knowledge on the harmonic structure. Specifically, the relative strength of j th partial is proportional to the j th element of a_0 , and the variance of the relative strength is governed by the norm of a_0 . Therefore, it is possible to determine a_0 in advance by training it with an instrument sound corpora, or set it to non-informative prior and determine the posterior distribution given the observed data. Note that by assuming independence from t means that the relative strength of the partials remains constant within an IPP.

Next, the fundamental frequency is generated. The method assumes that the fundamental frequency of the i th IPP, μ_{hi} , is distributed according to Normal distribution centered about the notated fundamental frequency, with precision λ_{hi} . A prior distribution on μ_{hi} is imposed, such that its distribution is concentrated about the fundamental frequency of IPP i of part h . Moreover, λ has a prior such that its expected value is large. Specifically, μ_{hi} and λ_{hi} are drawn from a Normal-Gamma distribution:

$$\mu_{hi}, \lambda_{hi} \sim \mathcal{NG}(m_{hi}^{(H)}, b_{hi}^{(H)}, l_{hi}^{(H)}, v_{hi}^{(H)}). \quad (3.12)$$

Note that $\langle \mu_{hi} \rangle = m_{hi}^{(H)}$, and $\langle \lambda_{hi} \rangle = l_{hi}^{(H)} / v_{hi}^{(H)}$. This way, the range of pitch for the i th IPP of part h to that notated in the music score may be restricted, by setting l much larger relative to v , so that the variance of the f_0 is small.

Finally, the quanta in time-frequency bin (t, f) are drawn. First, given a time t , the set of IPPs to generate the count are chosen, by referring to the state sequence and the

score. Then, the part h , IPP i and harmonics j are chosen according to the previous discussions. Then, a frequency bin inside f can be drawn from a Normal distribution centered about the frequency of the j th harmonic of the i th IPP:

$$\log f \sim \prod_{h,i,j,t,c,n,l}^{H,I,J,T,x_{tf},N,L} \mathcal{N}(\mu_{hi} + \log j, \lambda_{hi}^{-1})^{I_{hni} \mathcal{H}_{hij} S_{htnl} \mathcal{M}_{tch}} \quad (3.13)$$

This kind of model has the capability to jointly and uniquely identify the part, the volume (the relative strengths of the IPP) and the timbre (relative strengths of the harmonic partials of each IPP).

Thus, the generative process of x can be summarized as follows:

$$p(x, \mathcal{I}, \mathcal{H}, \mathcal{M} | s, m, e, a, \mu, \lambda) = \prod_{t=1, f=1, c=1, n=1, l=1, h=1, i=1, j=1}^{T, F, x_{tf}, N, L, H, I, J} (m_{tch} e_{hni} a_{hij} \mathcal{N}(\log f | \mu_{hi} + \log j, \lambda_{hi}^{-1}))^{I_{hni} \mathcal{H}_{hij} \mathcal{M}_{tch} S_{htnl}} \quad (3.14)$$

In other words, this is a latent variable model that associates with each (t, f, c) a latent variable of tuple (h, i, j, n, l) .

The overall generative process is illustrated in the graphical model of Fig. 3.4. Reflecting back on the formulation in Equation 2.24, the proposed method can be seen as fixing z , expressing θ using an LHA model, and expressing s using a set of HSMs defined for each part, tied together by a common tempo trajectory.

3.2 Inference

Having defined the model, the goal is to determine the posterior distribution $p(s, \mathcal{H}, \mathcal{M}, \mathcal{I}, \mathcal{T}, e, a, \mu, \lambda, m | x)$. The posterior distribution can then be used to find the desired statistics. For example, extracting the maximum a posteriori (MAP) estimate of the state sequence s is used as the estimate of the audio-to-score alignment.

This chapter infers an approximate posterior distribution using variational Bayes. Specifically, $p(s, \mathcal{H}, \mathcal{M}, \mathcal{I}, \mathcal{T}, e, a, \mu, \lambda, m | x)$ is approximated as a factored form $q_s(s) q_{\mathcal{H}}(\mathcal{H}) q_{\mathcal{M}}(\mathcal{M}) q_{\mathcal{I}}(\mathcal{I}) q_{\mathcal{T}}(\mathcal{T}) q_e(e) q_a(a) q_{\mu}(\mu) q_{\lambda}(\lambda) q_m(m)$. This kind of factorized approximation is called the *variational posterior*. By factorizing the posterior in this manner, posterior inference can be performed by iteratively minimizing the KL divergence from the variational posterior to the true posterior w.r.t. each factor of the variational posterior.

For simplicity, let us define the following variables:

$$\begin{aligned} l\mathcal{N}_{hijf} &= \langle \log \mathcal{N}(\log(f/j) | \mu_{hi}, \lambda_{hi}^{-1}) \rangle \\ &= -\frac{1}{2} \left(\frac{\bar{l}_{hi}}{\bar{v}_{hi}} \left(\log \frac{f}{j} - \bar{m}_{hi} \right)^2 + \frac{1}{\bar{b}_{hi}} \right) - \log 2\pi \bar{v}_{hi} + \psi(\bar{l}_{hi}) \end{aligned} \quad (3.15)$$

$$\iota_{nl} = \langle \log p(\log l | \mathcal{T}_n) \rangle \quad (3.16)$$

$$\eta_{htn} = \langle s_{htn} \rangle \quad (3.17)$$

$$lA_{hij} = \langle \log a_{hij} \rangle = \psi(\bar{\alpha}_{hij}) - \psi(\bar{\alpha}_{hi}) \quad (3.18)$$

$$lE_{hni} = \langle \log e_{hni} \rangle = \psi(\bar{e}_{hni}) - \psi(\bar{e}_{hn}) \quad (3.19)$$

$$lM_{tfh} = \langle \log m_{tfh} \rangle = \psi(\bar{m}_{0tfh}) - \psi(\bar{m}_{0tf}) \quad (3.20)$$

The method iteratively updates the variational posterior of each factor until the divergence has converged. First, the time-frequency mask $\mathcal{M}$ is updated as follows:

$$\mathcal{M}_{tfc} \sim \text{Discrete}(\gamma_{tf}^{(m)}), \quad (3.21)$$

where $\gamma_{tf}^{(m)} = \{\gamma_{tfh}^{(m)}\}$ is given as follows, up to a constant normalization coefficient:

$$\log \gamma_{tfh}^{(m)} = lM_{tfh} + \sum_{n=1, i=1}^{N, I} \eta_{htn} \left[lE_{hni} + \sum_{j=1}^J \xi_{hfij} (l\mathcal{N}_{hijf} + lA_{hij}) \right]. \quad (3.22)$$

Next, the probability the i th pitch-instrument pair occupies frequency f at state n of the HSMM, $\mathcal{I}$, is updated as follows:

$$\mathcal{I}_{hnf} \sim \text{Discrete}(\gamma_{hnf}^{(i)}), \quad (3.23)$$

where $\gamma_{hnf}^{(i)} = \{\gamma_{hnfi}^{(i)}\}$ is given as follows, up to a constant normalization coefficient:

$$\log \gamma_{hnfi}^{(i)} = \left(\sum_{t=1}^T x_{tf} \gamma_{tfh}^{(m)} \eta_{htn} \right) \left[lE_{hni} + \sum_{j=1}^J \xi_{hfij} (l\mathcal{N}_{hijf} + lA_{hij}) \right]. \quad (3.24)$$

In Equation 3.24, the first term in the right-hand side correspond to the time-averaged spectrum weighed by the probability of being in state n . The second term is the weight of the i th pitch-instrument pair based on the expected note emission probability, distributed using the harmonic partials emission probability. In other words, this distributes the expected observed spectrum at state n to different IPP.

Similarly, the probability that the j th partial occupies frequency f for the i th pitch-

instrument pair, $\mathcal{H}$, is updated as follows:

$$\mathcal{H}_{hfi} \sim \text{Discrete}\left(\gamma_{hfi}^{(h)}\right), \quad (3.25)$$

where $\gamma_{hfi}^{(h)} = \{\gamma_{hfi}^{(h)}\}$ is given as follows, up to a constant normalization coefficient:

$$\log \gamma_{hfi}^{(h)} = \left(\sum_{t=1, n=1}^{T, N} x_{tf} \gamma_{tfh}^{(m)} \eta_{htn} \gamma_{hni}^{(i)} \right) [l \mathcal{N}_{hijf} + l A_{hij}]. \quad (3.26)$$

Equation 3.26, like Equation 3.24, can be seen as filtering of the spectrum, weighted by the likelihood of observing the i th IPP, into different partials.

Next, consider updating the relative volume of each note within a state. By the independence of IPP, it is clear that the posterior of IPPs are independent across part h and across states n within each part. The variational posterior of e_{hn} is given as $e_{hn} \sim \text{Dir}(\bar{e}_{hn})$, where the hyperparameter $\bar{e}_{hn} = \{\bar{e}_{hni}\}_{i=1}^I$ is given as follows:

$$\bar{e}_{hni} = e_{0hni} + \sum_{t=1, f=1}^{T, F} x_{tf} \gamma_{tfh}^{(m)} \eta_{htn} \gamma_{hni}^{(i)} \quad (3.27)$$

Similarly with harmonic partials emission probability, the variational posterior of a_{hi} is found as $a_{hi} \sim \text{Dir}(\bar{a}_{hi})$, where $\bar{a}_{hi} = \{\bar{a}_{hij}\}_{j=1}^J$ is given as follows:

$$\bar{a}_{hij} = a_{0hij} + \sum_{t=1, f=1, n=1}^{T, F, N} x_{tf} \gamma_{tfh}^{(m)} \eta_{htn} \gamma_{hni}^{(i)} \xi_{hfi} \quad (3.28)$$

The fundamental frequency of IPP i inside part h , μ_{hi} , and its precision, λ_{hi} , is estimated as $\mu_{hi}, \lambda_{hi} \sim \mathcal{NG}(\bar{m}_{hi}^{(H)}, \bar{b}_{hi}^{(H)}, \bar{l}_{hi}^{(H)}, \bar{v}_{hi}^{(H)})$, where the hyperparameters are given as follows:

$$\bar{m}_{hi}^{(H)} = (b_{hi} + N_{hi})^{-1} (m_{hi} b_{hi} + N_{hi} \langle \log(f/j) \rangle_{\theta_{hi}}) \quad (3.29)$$

$$\bar{b}_{hi}^{(H)} = b_{hi} + N_{hi} \quad (3.30)$$

$$\bar{l}_{hi}^{(H)} = l_{hi} + N_{hi} \quad (3.31)$$

$$\begin{aligned} \bar{v}_{hi}^{(H)} = v_{hi} + \frac{1}{2} \frac{b_{hi} N_{hi}}{b_{hi} + N_{hi}} & \left(\left(\langle \log(f/j) \rangle_{\theta_{hi}} - \bar{m}_{hi}^{(H)} \right)^2 \right. \\ & \left. + N_{hi}^2 \langle \log(f/j) - \langle \log(f/j) \rangle_{\theta_{hi}} \rangle_{\theta_{hi}}^2 \right). \end{aligned} \quad (3.32)$$

Here, $\theta_{hi} = \{\theta_{hifj}\}_{f=1,j=1}^{F,J}$ is the following density:

$$\theta_{hifj} = \sum_{n=1,t=1}^{N,T} x_{tf} \eta_{htn} \gamma_{tfh}^{(m)} \gamma_{hnfi}^{(i)} \xi_{hfig} / N_{hi}, \quad (3.33)$$

$$N_{hi} = \sum_{n=1,t=1,f=1,j=1}^{N,T,F,J} x_{tf} \eta_{htn} \gamma_{tfh}^{(m)} \gamma_{hnfi}^{(i)} \xi_{hfig}. \quad (3.34)$$

The variational posterior of the LDS, $q_{\mathcal{T}}(\mathcal{T})$, is determined in a manner similar to the Kalman smoothing algorithm. Minimizing the KL divergence from the variational posterior to the true posterior w.r.t. $q_{\mathcal{T}}(\mathcal{T})$ leads to the following expression that resembles an LDS, where $\theta_{nl}^{(s)} = \sum_{h,t>1,n' \neq n} \langle s_{h,t-1,n',1} s_{htnl} \rangle$:

$$\log q_{\mathcal{T}}(\mathcal{T}) \stackrel{c}{=} \sum_{n=1}^N \left[\sum_{l=1}^L \theta_{nl}^{(s)} \log \mathcal{N}\left(\log \frac{l}{\mathcal{L}_n} | \mathcal{T}_n, \sigma_T^2\right) + \left\langle \log \mathcal{N}\left(\mathcal{T}_n | \mathcal{T}_{n-1}, \frac{\mathcal{L}_{n-1}}{\lambda_n^{(T)}}\right) \right\rangle \right]. \quad (3.35)$$

The equation is highly similar to an LDS in that the second term expresses continuity of $\mathcal{T}_n$. However, the first term describes the cross-entropy between the expected duration pdf $\theta_{nl}^{(s)}$ and the duration pdf expected from the LDS; this is different from an LDS, which typically simply describes the log-likelihood of observing a particular value given the state variables.

Similar to the Kalman smoother, $q(\mathcal{T})$ is updated using the forward-backward algorithm. The forward algorithm is described as the following forward recursion:

$$\begin{aligned} \alpha_n^{(\mathcal{T})}(\mathcal{T}_n) &\triangleq p(\mathcal{T}_n | \theta_{1,\dots,n}^{(s)}) = \mathcal{N}(\mathcal{T}_n | u_n, s_n) \\ &\propto \int \alpha_{n-1}^{(\mathcal{T})}(\mathcal{T}_{n-1}) \mathcal{N}\left(\mathcal{T}_n | \mathcal{T}_{n-1}, \mathcal{L}_{n-1} \lambda_n^{(L)-1}\right) \prod_{l=1}^L \mathcal{N}\left(\log(l/\mathcal{L}_n) | \mathcal{T}_n, \sigma_T^2\right)^{\theta_{nl}^{(s)}} d\mathcal{T}_{n-1}. \end{aligned} \quad (3.36)$$

Integrating $\mathcal{T}_{n-1}$ out and completing the square with respect to $\mathcal{T}_n$ gives the following for u_n and s_n :

$$m_n^{-1} = \frac{1}{s_{n-1}} + \frac{\langle \lambda_n^{(T)} \rangle}{\mathcal{L}_{n-1}}, \quad (3.37)$$

$$s_n^{-1} = \frac{\theta_{n:}^{(s)}}{\sigma_T^2} + \frac{\langle \lambda_n^{(T)} \rangle}{\mathcal{L}_{n-1}} - m_n \left(\frac{\langle \lambda_n^{(T)} \rangle}{\mathcal{L}_{n-1}} \right)^2, \quad (3.38)$$

$$u_n = s_n \left(m_n \frac{\langle \lambda_n^{(T)} \rangle}{\mathcal{L}_{n-1}} \frac{u_{n-1}}{s_{n-1}} + \sum_{l=1}^L \frac{\theta_{nl}^{(s)}}{\sigma_T^2} \log \frac{l}{\mathcal{L}_n} \right). \quad (3.39)$$

The backward algorithm is described as the following backward recursion:

$$\begin{aligned}\beta_{n-1}^{(\mathcal{T})}(\mathcal{T}_{n-1}) &\triangleq p(\theta_{n \dots T}^{(s)} | \mathcal{T}_{n-1}) = \mathcal{N}(\mathcal{T}_{n-1} | v_{n-1}, q_{n-1}) \\ &= \int \beta_n^{(\mathcal{T})}(\mathcal{T}_n) \mathcal{N}(\mathcal{T}_n | \mathcal{T}_{n-1}, \mathcal{L}_{n-1} \lambda_n^{(T)-1}) \prod_{l=1}^L \mathcal{N}(\log(l/\mathcal{L}_n) | \mathcal{T}_n, \sigma_T^2)^{\theta_{nl}^{(s)}} d\mathcal{T}_n.\end{aligned}\quad (3.40)$$

By completing the square, we obtain the following:

$$n_n^{-1} = \frac{1}{q_{n+1}} + \frac{\langle \lambda_{n+1}^{(T)} \rangle}{\mathcal{L}_n} + \frac{\theta_{n+1,:}^{(s)}}{\sigma_T^2}, \quad (3.41)$$

$$q_n^{-1} = \frac{\langle \lambda_{n+1}^{(T)} \rangle}{\mathcal{L}_n} - n_n \left(\frac{\langle \lambda_{n+1}^{(T)} \rangle}{\mathcal{L}_n} \right)^2, \quad (3.42)$$

$$v_n = n_n q_n \frac{\langle \lambda_{n+1}^{(T)} \rangle}{\mathcal{L}_n} \left(\sum_{l=1}^L \frac{\theta_{n+1,l}^{(s)}}{\sigma_T^2} \log \frac{l}{\mathcal{L}_{n+1}} + \frac{v_{n+1}}{q_{n+1}} \right). \quad (3.43)$$

Using these, the variational posterior is obtained as follows:

$$q(\mathcal{T}_n | l_{1:T}) = \alpha_n^{(\mathcal{T})}(\mathcal{T}_n) \beta_n^{(\mathcal{T})}(\mathcal{T}_n) = \mathcal{N}\left(\mathcal{T}_n | \frac{1}{q_n^{-1} + s_n^{-1}} \left(\frac{v_n}{q_n} + \frac{u_n}{s_n} \right), \frac{1}{q_n^{-1} + s_n^{-1}}\right). \quad (3.44)$$

This distribution can be used to generate the duration pdf of the HSMM, namely:

$$\begin{aligned}\iota_{nl} &= \left\langle -\frac{1}{2\sigma_T^2} \left(\log \frac{l}{\mathcal{L}_n} - \mathcal{T}_n \right)^2 - \log(2\pi\sigma_T^2) \right\rangle \\ &= -\frac{1}{2\sigma_T^2} \left(\log \frac{l}{\mathcal{L}_n} - \frac{1}{q_n^{-1} + s_n^{-1}} \left(\frac{v_n}{q_n} + \frac{u_n}{s_n} \right) \right)^2 - \frac{1}{2\sigma_T^2 (q_n^{-1} + s_n^{-1})} - \log(2\pi\sigma_T^2).\end{aligned}\quad (3.45)$$

Notice that the effect of the tempo model on the actual note duration is dependent on σ_T ; it is, roughly speaking, a coupling parameter between the LDS and the HSMM. Too small a σ_T makes the proposed method “myopic,” in that each state cannot “see” past itself as constrained by the tempo model. Too large a σ_T makes the proposed method ignorant of the tempo model, as the actual note duration is penalized less for deviating from the tempo model.

Minimizing the KL divergence from the variational posterior to the true posterior

w.r.t. $q_s(s)_{h,1\dots T}$ leads to the following:

$$\log q_s(s) = \sum_{l=1, n=1}^{L, N} \left[s_{h1nl} \langle \log \pi_{nl} \rangle + \sum_{t=2, n' \neq n}^{T, N} s_{h, t-1, n', 1} s_{htnl} \left(\langle \log \tau_{n'n} \rangle + \iota_{nl} \right) \right. \\ \left. + \sum_{t=1, f=1}^{T, F} s_{htnl} x_{tf} \gamma_{tfh}^{(m)} \log \kappa_{hnf} \right], \quad (3.46)$$

where

$$\log \kappa_{hnf} = \sum_{i=1, j=1}^{I, J} \gamma_{hnfi}^{(i)} \xi_{hfij} \left[l E_{hni} + l A_{hij} + l N_{hijf} \right]. \quad (3.47)$$

This has a same functional form as an HSMM, with κ_{hnf} being an unnormalized histogram that shows the spectrum state n is likely to emit, and $x_{tf} \gamma_{tfh}^{(m)}$ is the expected CQT spectrogram of part h . Thus, like an HMM, Baum-Welch algorithm can be used to infer the state expectation. Let us define $\alpha_{htnl}^{(S)} \triangleq p(s_{htnl} = 1 | x_{1\dots t, f})$ and $\beta_{htnl}^{(S)} \triangleq p(x_{t+1\dots T, f} | s_{htnl} = 1)$ as the forward and backward variables of the HSMM, respectively. Then, the following recurrences are obtained:

$$\alpha_{htnl}^{(S)} \propto O_{htn} (\alpha_{h, t-1, n, l+1}^{(S)} + \sum_{n'=1}^N e^{\langle \log \tau_{n'n} \rangle + \iota_{nl}} \alpha_{h, t-1, 1, n'}^{(S)}), \quad (3.48)$$

$$\beta_{htnl}^{(S)} = \begin{cases} O_{h, t+1, n} \beta_{h, t+1, n, l-1}^{(S)} & l > 1 \\ \sum_{n'=1}^N O_{h, t+1, n'} e^{\langle \log \tau_{n'n} \rangle} \sum_{l'=1}^L \beta_{h, t+1, l', n'}^{(S)} e^{\iota_{n'l'}} & l = 1. \end{cases} \quad (3.49)$$

Here O_{htn} is the pseudo-emission probability of state n at time t for part h , given as follows:

$$\log O_{htn} = \sum_{f=1}^F x_{tf} \gamma_{tfh}^{(m)} \log \kappa_{hnf}. \quad (3.50)$$

Because the state sequence is assumed to end at the downbeat of the terminal state, $\beta_{hTnl}^{(S)}$ is set to 1 for $n = N$ and $l = 1$, and 0 everywhere else. Based on these variables, the following expectations are given:

$$\langle s_{htnl} \rangle \propto \alpha_{htnl}^{(H)} \beta_{htnl}^{(H)} \quad (3.51)$$

$$\langle s_{h, t-1, n', 1} s_{htnl} \rangle \propto \alpha_{h, t-1, 1, n'}^{(H)} O_{htn} e^{\langle \log \tau_{n, n'} \rangle + \iota_n} \beta_{htnl}^{(H)}. \quad (3.52)$$

This information is then used to compute the expectation in Equation 3.35, and to find an estimated trajectory of part h at time t , $\arg \max_{nl} \langle s_{htnl} \rangle$.

Using this information, the variational posterior of τ is found as follows:

$$q_{\tau} = \prod_{n=1}^N \text{Dir}(\tau_{n,\{1\cdots N\}} | \bar{\tau}_{n,\{1\cdots N\}}), \quad (3.53)$$

where

$$\bar{\tau}_{nn'} = \tau_{0nn'} + \zeta_{:::nn'}. \quad (3.54)$$

Moreover, π is updated as $q_{\pi} = \text{Dir}(\pi_0 + \eta_{h,1,\{1\cdots N\}})$.

3.3 Evaluation

To evaluate the proposed method, this section assesses the capability to (1) align the music score to audio, assuming that players play synchronously, (2) detect pre-notated repeated segments, and (3) detect un-notated timing discrepancies among different players.

In the subsequent evaluations, the CQT was computed at every eighth of a semi-tone from C1 to E8 (where A4 = 440 Hz), at a rate of 20 frames per second. Unless otherwise mentioned, $a_0(h, i)$ are non-informative for all h and i , and evaluated up to $J = 5$ th partial. Moreover, $m_{hi}^{(H)}$ is set to the notated fundamental frequency, $l_{hi}^{(H)} = 10^{20}$, and $v_{hi}^{(H)} = b_{hi}^{(H)} = (1/2)^2 10^{20}$; this sets the prior distribution of μ_{hi} concentrated about the notated fundamental frequency with standard deviation of a quarter tone. Also, $\langle \lambda_d^T \rangle$ is set to 10^{-9} for all d except for the very beginning and the end, which is set to 10. Thus, tempo LDS is allowed to vary more in the very beginning and the end. Moreover, σ_T is swept from 10 to 0.06 over the course of VB iterations. This effectively weakens the influence of the tempo LDS model in the initial iterations of the inference. These values were set based manually based on a real-world musical audio that is unused in the evaluation due to the lack of reliable ground truth data. Moreover, the model was first inferred with all parts collapsed into a single part (*i.e.*, a score with $H=1$), and used the variational posterior of this model to initialize the proposed model with each HSMM for each part. The maximum duration to stay in a state was set to 1 second, or $L = 20$.

3.3.1 Evaluation of the global alignment

This experiment compares the effectiveness of the proposed method for audio-score alignment with three methods.

Experimental setup

First, the proposed method was compared with an alignment method based on DTW ("CHR"). Method CHR minimized the net cosine distance between chroma vector of

Table 3.1: Percentile of the absolute error in milliseconds.

		25%	50%	75%	90%	95%
Piano solo	CHR	90	304	1363	6422	11736
	IT	17	48	224	891	2040
	FS	1485	4520	10468	19415	26728
	Proposed	9	21	50	126	269
Instrument + Piano Accomp.	CHR	68	182	619	2714	9848
	IT	14	32	86	255	473
	FS	863	2549	6437	9373	11219
	Proposed	8	21	45	93	163
Small ensemble	CHR	90	259	891	2804	4710
	IT	16	46	131	393	816
	FS	1927	4296	8827	16260	25178
	Proposed	10	22	45	88	133
Orchestral	CHR	123	394	1384	6688	36550
	IT	38	104	574	4793	16768
	FS	3111	10463	21788	34275	44847
	Proposed	23	51	119	805	2996
Bach10 dataset (Real performance)	CHR	66	205	340	1297	1988
	IT	11	30	60	166	350
	FS	159	350	667	978	1213
	Proposed	11	29	57	165	366

the audio and the score that has been warped in time, using path constraints in [58].

Second, the proposed method was compared with its variant, where the note durations are assumed to be mutually independent. In other words, the tempo LDS was ignored, and assumed the tempo is distributed identically at $p(\mathcal{T}_d) = \delta(\mathcal{T}_d - 10)$ for all d (“IT,” or Independent Tempo). This is similar to [107] in that the note duration are mutually independent given some fixed tempo.

Finally, the proposed method was compared with its variant using a fixed timbre/volume/part balance (“FS,” or Fixed Spectrum). To elaborate, the expected emission spectrum in the HSMM was fixed; that is, κ_{hnf} in Equation 3.46 was switched into a fixed normalized spectrum. This is similar in spirit to [112].

The proposed method was evaluated on both synthetic and real datasets. A synthetic dataset is desirable because it has an accurate ground truth. Moreover, alignment performance obtained from a synthetic dataset is indicative of the performance in real-world audio signals generated by human players, as suggested by [93]. Therefore, 60 pieces from the music scores (in Standard MIDI File Format, SMF) from the RWC Classical Music Database [49] were synthesized using Freepats patch [129]. The SMF contains expressive data including tempo changes and note strength (velocity) variations. Then, the expected beat positions and the estimated beat positions obtained from score alignment were computed. Furthermore, to test with real audio,

ten instrumental performances of Bach chorales were used as another test data [29].

Experimental results

From the obtained alignment, the percentile of the absolute error of the beat positions extracted was evaluated for the proposed method and the three baselines described above. The result is presented in Table 3.1. Notice that method IT performs better than CHR. This suggests that incorporating a Bayesian model of spectrum is effective in alignment. The importance of a Bayesian treatment of spectrum is also prominent by comparing the proposed method and method FS – by fixing the spectrum (*i.e.*, fixing the observation likelihood), the performance deteriorates significantly. Moreover LDS tempo model is effective, since the proposed method performs much better than LH.

The proposed method performs the worst in orchestral pieces. This is because orchestra is a complex polyphonic mixture, making it difficult to infer the spectral model. Surprisingly, the proposed method performs somewhat poorly on piano solo pieces as well. This is because a pianist may elongate the note duration much longer than is notated, using the “sustain pedal.” This is problematic because the proposed model assumes that no signal component exists at pitches that are not notated inside each state. To elaborate, recall that the proposed method expresses normalized histograms and spectrum is expressed as a count data over the histogram. Therefore, the proposed method penalizes little for having a few counts in prominent peaks of the normalized histogram, *i.e.*, the proposed method is fine not observing a strong signal component at notated frequencies and their partials. the proposed method, however, penalizes a lot for having many counts in small-valued components of the normalized histogram, *i.e.*, it penalizes a lot for observing a strong component *outside* notated frequencies and their partials.

3.3.2 Evaluation of music structure inference

This experiment evaluates the capability of MAHLER to detect pre-notated musical structures such as repeats and cuts. Because cuts and repeats are mathematically dealt in the same way (as non-adjacent state transitions of the HSMM), it suffices to evaluate the capability to detect repeats.

Experimental setup

For each polyphony p from 2 to 6, 100 random scores were generated with polyphony p and a duration of 256 beats, including a repeated segment in the middle of length 2, 4, 8, 16, and 32 bars long. These files were then synthesized with 100BPM using a piano patch. Finally, these audio files were aligned against the music score. Here, the estimation is defined to be “correct” if the alignment result has the same number of backwards skips as there are number of repeats.

Table 3.2: True positive rate (TPR) of repeated segment detection. True negative rate was 100%.

(a) As a function of the repeated segment length.

Beat Length	2	4	8	16	32
TPR	96%	87%	83%	89%	94%

(b) As a function of the number of polyphony.

No. Notes	2	3	4	5	6
TPR	100%	100%	99%	88%	63%

Experimental results

Table 3.2a shows the result as a function of repeat bar duration, averaged over the number of polyphony. The table shows that as the polyphony increases, the detection capability decreases. This behavior mirrors the result from Table 3.1, where genres with increased instrumentation performs relatively poorly.

Table 3.2b shows the result as a function of polyphony, averaged over bar duration. This result shows that the accuracy decreases for a repeat of about 8 to 12 beats. In practice, such repeats are rare, so it should not hinder the merit of incorporating pre-notated non-adjacent state transitions.

3.3.3 Evaluation of inter-part onset timing differences inference

This experiment evaluates the capability of the proposed method to find small un-notated onset timing fluctuations among different parts.

Experimental setup

For each number of polyphony p , 1000 SMF files were synthesized. Each file consecutively plays three chords with p notes, each with duration of 1 second and played by a piano patch. Notes are assigned in one of p parts, such that all p parts are monophonic. Moreover, the onset times are dilated by a uniformly-distributed noise between -250ms and 250ms. Each of the $p \times 1000$ audio files was aligned to the score using (1) single HSMM used for entire part, and (2) p HSMMs, each HSMM defined for each part. The cumulative percentage of the absolute error of the estimated onsets was evaluated.

Experimental results

The result, presented in Table 3.3, shows the percentage of frames whose error lie within an error margin, for different polyphony. It shows that the proposed method almost always has a higher percentage of frames whose error lies within a given

Table 3.3: Cumulative percentage of the absolute error of the estimated onsets w.r.t. polyphony p .

Method	p	10ms	20ms	50ms	100ms	200ms	500ms	1000ms
Single HSMM	2	20%	46%	57%	76%	94%	99%	100%
Proposed	2	25%	58%	66%	80%	94%	99%	99%
Single HSMM	3	14%	37%	48%	70%	93%	100%	100%
Proposed	3	18%	46%	56%	74%	92%	99%	100%
Single HSMM	4	10%	29%	41%	66%	92%	99%	100%
Proposed	4	16%	39%	50%	70%	91%	99%	100%
Single HSMM	5	9%	27%	39%	66%	92%	99%	100%
Proposed	5	14%	37%	48%	69%	90%	99%	100%
Single HSMM	6	9%	25%	37%	64%	92%	99%	99%
Proposed	6	12%	34%	45%	68%	91%	99%	100%

bound, meaning that the proposed method is more capable of detecting asynchrony between parts than when assuming that every part is completely synchronized.

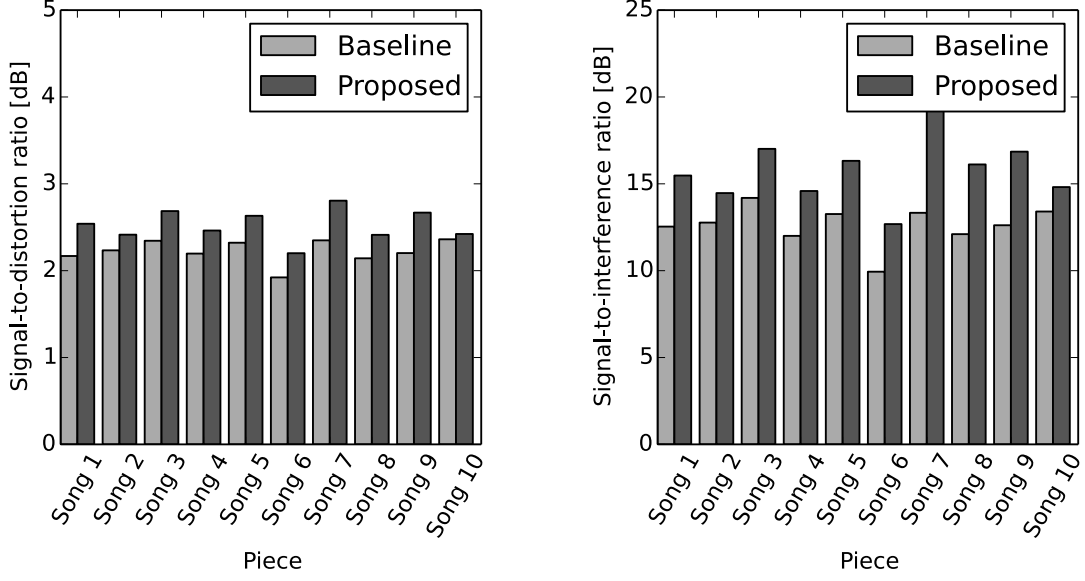
3.3.4 Evaluation of informed source separation

Here, the proposed method is evaluated when used directly in an informed source separation setting. This is important for two reasons. First, since informed source separation is a useful method on its own, it is interesting to see the applicability of the proposed method for source separation. Second, it is a good test of how well our generative model expresses the audio signals. To elaborate, note that separation of a polyphonic mixture into comprising parts depends on how well we estimate each part. Moreover, the proposed method describes a generative model of how each part is mixed in a polyphonic mixture. Therefore, the better the proposed method separates audio, it is a better generative model of each part.

Experimental setup

The proposed method was compared with its variant that does not update the timbre or dynamics. This kind of baseline is useful for assessing whether posterior inference makes the proposed model more representative of the audio signal than was before.

The posterior distribution of the proposed method and the baseline were inferred using ten quartet from BACH10 database [29]. Then, the signal corresponding to each part was extracted by Wiener filtering. To elaborate, audio for part h was extracted by first evaluating a STFT $\bar{x}_{t\bar{f}}$ of the input signal with the same analysis framerate as that used for computing the CQT. Then, $\gamma_{t\bar{f}h}^{(m)}$ was interpolated to the linear frequency axis $\bar{f}$. Finally, the inverse-STFT of $\bar{x}_{t\bar{f}}\gamma_{t\bar{f}h}^{(m)}$ was computed. Then, for each separated signal, the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) were evaluated using BSS_EVAL [125]. SDR is indicative of how well the method



(a) SDR of the separated parts.

(b) SIR of the separated parts.

Figure 3.5: SDR and SIR of the separated parts, using score-informed separation based on the proposed method.

kept wanted components, and SIR is indicative of how well the method suppressed unwanted components.

Experimental results

Fig. 3.5 shows the result of source separation. The average SDR improved from 2.2 dB to 2.5 dB, and SIR from 12.6 dB to 15.8 dB. It shows that the inference of timbre and dynamics is effective for informed separation, showing that the proposed method better represents the underlying audio signal. This suggests that the proposed model is capable of representing musical audio signals.

Using a similar dataset, an existing informed separation method reports a SDR of about 5.5 dB and SIR of 12.9 dB [29]. The proposed method generates slightly lower SDR and a higher SIR. This means that the proposed method is capable of removing unwanted sound, but distorts the signal in the due process. This is because the proposed method assumes that the dynamics is fixed within a note and the timbre for each IPP is fixed. This approximation is good enough for identifying where the noisy signals are, but it fails to capture minute spectral variations within a note. This may be alleviated by incorporation of a more elaborate spectral model that captures the fine spectral detail at each time instance.

3.4 Summary

This chapter presented a Bayesian audio-score alignment method. The key issues that (1) inter-part asynchrony and alignment is mutually dependent, and (2) timbre, dynamics and alignment are mutually dependent, were handled by jointly inferring timbre, dynamics, tempo, pitch, score alignment, and inter-part asynchrony in a Bayesian manner.

Experimental evaluation demonstrated that the proposed method achieves a median accuracy of about 30 ms on real chamber music and 50 ms on synthetic orchestral recording, making it a viable choice for detailed performance analysis. Furthermore, the proposed method was shown to be capable of estimating inter-part asynchrony; this is important for informed source separation applications, which require high-quality note onset timing information; and application to musicology, *e.g.*, for analyzing the interaction between different performers within an ensemble.

Chapter 4

Bayesian Audio-to-Audio Alignment

This chapter presents a method for aligning different audio signals that play a same piece of music. In order to align multiple music audio signals that represent a same piece of music, it is necessary to express both what kind of piece the audio signals are playing, and how the piece is played by each audio signal. In order to unify different musical constraints in a single framework, generative audio alignment is attractive. Furthermore, since how a piece is played by each performance and what kind of piece is played in common are mutually dependent, joint inference through statistical modeling is desirable.

Audio-to-audio alignment is difficult because, unlike audio-to-score alignment, the symbolic music score is not given. This means that the sequence of note combinations and their relative durations, which are common to all signals, are unknown. Thus, generative audio-to-audio alignment needs to specify a model of the note combination sequences and their relative durations (*i.e.*, musical *composition*), and the alignment between each audio signal and the musical composition (*i.e.*, music *performance*).

To represent the generative process of a musical composition, the proposed method focuses on the general fact that small fragments consisting of multiple musical notes form the basic building blocks of music and are organized into a larger work. For example, the sonata form is based on developing two contrasting fragments known as the “subject groups,” and a song form essentially repeats the same melody. Hence, a composition model should express the observation that basic melodic patterns are reused to form the sonata or the song. Yet, such a property has been neglected [86,91], assuming that different part of the underlying musical composition is distributed independently.

To represent the generative process of each performance, the proposed method focuses on temporal fluctuations from a common music composition. Since each performance plays the same musical composition, the small fragments should appear in the same order. On the other hand, each performance can be played by a different set of musical instruments with a unique tempo trajectory.

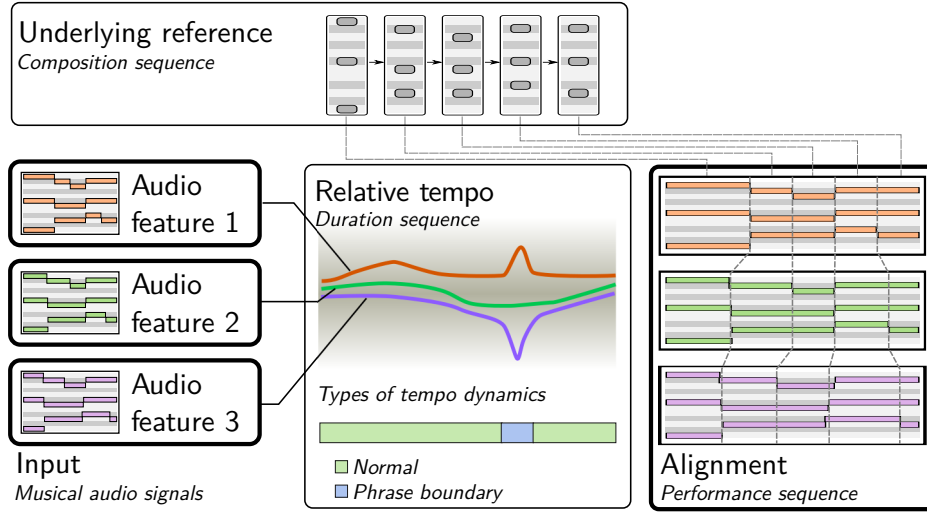


Figure 4.1: An overview of generative audio alignment.

Unlike audio-to-score alignment, the tempo trajectory can be expressed only implicitly because the note durations are unknown. To tackle this issue, the proposed method assumes that the *relative tempo*, the speed at which each audio signal plays a piece relative to a reference, changes slowly over time. Tempo is a special case of the relative tempo, where the reference is a fixed metronome click. As seen in Chapter 2, the tempo is often expressed as a random walk [18, 30, 102, 112, 113]. Likewise, in probabilistic audio alignment [91], the relative tempo is assumed to change slowly over time, expressed using a random walk.

A random walk, in two ways, poorly approximates the actual dynamics of the relative tempo. First, it is incapable of expressing the appropriate range of the relative tempo. This is problematic since the relative tempo evolves within a narrow confine defined by the physical playability and musical appropriateness. Second, it is incapable of handling sporadic changes in the underlying dynamics. This is problematic because the relative tempo changes dramatically at structural boundaries of music, such as a phrase ending [105]: the tempo *does* change slowly over time, but the rate of change varies by the structural context of the music. The rate of change is fixed in a random walk, so it cannot handle this kind of sporadic change.

To deal with these issues, the proposed method employs a hierarchical Bayesian model of music composition and performance. Note duration of each performance is then modeled as a process whose smoothness may vary sporadically. The proposed method first addresses the need to express the kind of structure that appear in music composition. It does so by first decoupling the generative process of music audio signals into those originating from the underlying music score and those from the unique performance. Then, inspired by music structure analysis methods [1, 106], the proposed method represents a music composition as a sequence generated from a compact, ergodic Markov model (“latent composition”). This allows the method to capture frequently-occurring note patterns. Each music performance is represented

as a left-to-right Markov chain that traverses the latent composition with the state durations unique to each performance.

Second, the proposed method addresses the problems associated with expressing the dynamics of the relative tempo as a random walk. It does so using two approaches. First, it limits the range of the relative tempo, instead of allowing it to diverge without bound. This is achieved by expressing the dynamics of the relative tempo as a mean-reverting process (or Ornstein-Uhlenbeck process), instead of a random walk. A mean-reverting process is like a random walk, except it has a restoring force to a predefined value. The restoring force is proportional to how far the relative tempo deviates from a predefined relative tempo value. Second, the method allows sporadic changes in the relative tempo dynamics, which an LDS cannot express. To do so, it, inspired by tempo trajectory models [52], expresses the relative tempo as a switching-state LDS (SSLDS). An SSLDS is capable of toggling the smoothness of the relative tempo, allowing the smoothness of the relative tempo change sporadically at structural boundaries. To incorporate such a model into an audio alignment method in a unified manner, the proposed method integrates the model of the relative tempo as a duration probability at each position in the underlying reference performance.

A conceptual overview is provided in Section 4.1, and then the concept is mathematically formalized in Section 4.2. Section 4.3 provides an inference scheme. Finally, the proposed method is evaluated in Section 4.4.

4.1 Conceptual overview

The proposed method formulates a generative model of alignment that aligns D performances. It first extracts short-time audio features from each of D performances. Denote the feature sequence for the d th performance at frame $t \in [1, T_d]$ as $\mathbf{x}_{dt}$, where T_d is the total number of frames for the d th audio signal. Here, the kind of feature is arbitrary, and depends on the generative model of the short-time audio. Then, $\mathbf{x}_{dt}$ is modeled as a set of D state sequences. Each state is associated with a unique generative process of short-time audio feature. In other words, each state represents a distinct audio feature, *e.g.*, distinct chord, f_0 and so on, depending on how the generative model of the feature is designed.

For audio alignment, the state sequence must abide by two rules. First, the order in which each state appears is the same for all D feature sequences. In other words, every performance is described by one sequence of distinct audio features, *i.e.*, the musical piece that the performances play in common. Such a sequence is called the *latent composition* in this chapter. Second, the duration that each performance resides in a given state in the latent composition can be unique to the performance. In other words, each performance traverses the latent composition with a unique “tempo curve.” The sequence that each performance traverses over the latent composition sequence is called the *performance sequence* in this chapter. This concept is illustrated in Fig. 4.2.

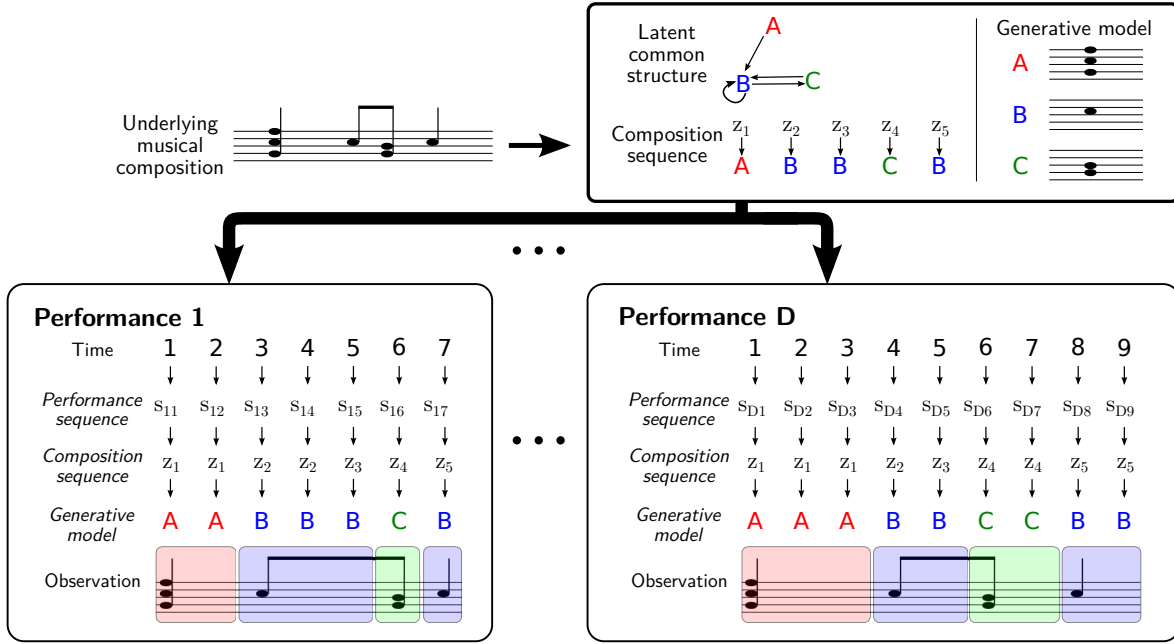


Figure 4.2: The idea of how a piece is generated. Music composition is modeled as a sequence (composition sequence) from an ergodic Markov model. Each performance plays the composition sequence by traversing the composition sequence in the order it appears. Each performance stays in each state with a unique duration, as governed by the duration sequence (omitted for the sake of simplicity).

The latent composition is a sequence of length N drawn from an ergodic Markov model, which is called the *latent common structure*. The latent composition is described as a sequence of length N and S states, z_n , where each state describes a distinct audio feature. In other words, the musical piece is described by at most N distinct audio events, using at most S distinct sounds. The latent common structure encodes the structure inherent to the music. The transition probabilities of each state shed light on a “typical” performance, *e.g.*, melody line or harmonic progression. Therefore, the latent common structure provides a generative model of music composition.

The performance sequence provides a generative model of performance. Each audio signal is modeled as an emission from a N -state left-to-right Markov model, where the n th state refers to the generative model associated with the n th position in the latent composition. Specifically, write the performance sequence for audio d as s_{dt} , which is a state sequence of length T_d and N states, such that state n refers to the n th element of the latent composition. Each performance sequence is constrained such that (1) it begins in state 1 and ends at state N , and (2) state n may traverse only to itself or state $n + 1$. In other words, each performance sequence is constrained to traverse the latent composition in the same order but with a unique duration. Such a model conveys the idea that each performance can independently play a piece in any tempo trajectory.

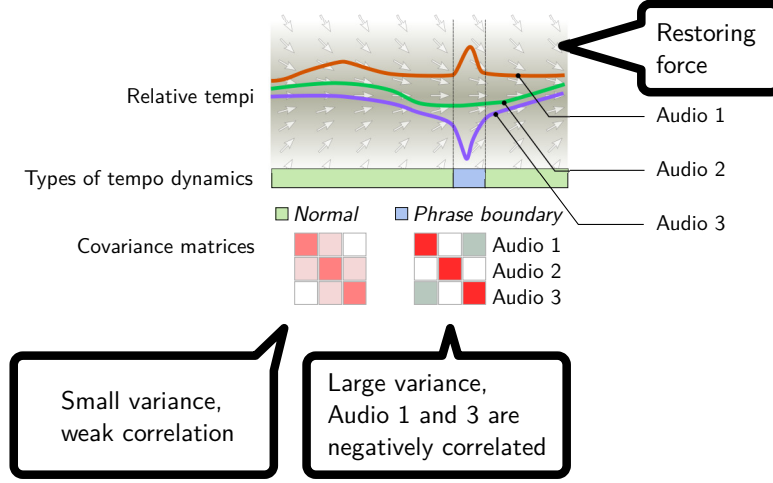


Figure 4.3: The idea of the duration modeling.

To express the tempo trajectory, the proposed method considers a “template” of tempo dynamics, and switching back and forth between the templates in the course of the piece. It is illustrated in Fig. 4.3. Each template encodes the covariance matrix of tempo change among different performances. For example, one template might correspond to every recording maintaining a smooth tempo, and another template might correspond to, say, tempo of signal 1 and 4 changing rapidly in a negatively correlated manner. By sporadically changing the template assignment, it is possible to express a tempo trajectory whose dynamics changes sporadically. In the proposed method, one template is assigned to each position in the composition sequence. This way, the tempo of every signal transitions with the same covariance structure at any given position in the underlying piece. From this kind of model, the covariance templates are learned and the sequence of covariance templates used is inferred.

An example of performance and composition sequence

Fig. 4.2 clarifies the relationship between the performance and composition sequence model. In the example, $S = 3$ and $N = 5$, where state “A” corresponds to a combination of notes G, C and F, “B” corresponds to the note C, and so on; moreover, z_n encodes the state sequence “ABBCB,” as to reflect the underlying common music composition that the performances play. Note that a single note may be expressed using more than one state in the latent composition, *e.g.*, both z_2 and z_3 describe the note C. Next, each performance aligns to the latent composition, through the performance sequence. Each state of the performance sequence is associated to a position in the latent composition. For example, $s_{1,3}$ is associated to position 2 of z , z_2 . Then, at each time, the observation is generated by emitting from the state in latent common structure referred by the current frame of the current audio. This is determined hierarchically by looking up the state n of the performance sequence of audio d at time t , and referring

to the state s of the n th element of the latent composition. In the example, $s_{1,3}$ refers to state $n = 2$, so the generative model corresponding to $z_{n=2}$, or “B,” is referred.

4.2 Formulation

The proposed method probabilistically expresses D feature sequences, each being an audio performance to a common piece of music.

This kind of multiple feature sequences is assumed to be generated by four steps. First, a reference performance that every sequence plays is generated. Second, each signal determines how long each position in the reference should be played. Third, each signal expands the reference in time, such that each position of the reference is played with a duration determined from the second step. Finally, based on the expanded sequence, audio features that correspond to each position in the reference are created. To express such a generative process, a probabilistic model is designed, which is composed of (1) *composition sequence*, a state sequence that describes the underlying reference performance common to the D signals, (2) *duration sequence*, a state sequence that describes how long each signal resides in each position in the composition sequence, (3) *performance sequence*, a state sequence that describes which position in the composition sequence is being played at each time instance, and (4) a model of audio feature that is likely to be observed at each state of the composition sequence.

4.2.1 Composition sequence

The composition sequence describes the underlying musical piece, *i.e.*, the ordering of features that are emitted in common to all signals. It is described as an S -state sequence of length N , $z = \{z_n\}_{n=1}^N$, where each state s is associated with a parameter θ_s , which describes what kind of acoustic features are likely to be observed at that state. Here, z_n is an S -dimensional binary variable, whose value z_{ns} is 1 if the state of z_n is s , and 0 otherwise. Specifically, we model z as an ergodic Markov chain whose initial state pdf is π and the transition pdf is τ :

$$p(z|\pi, \tau) = \prod_{s=1}^S \pi_s^{z_{1,s}} \prod_{n=2}^N \prod_{s=1}^S \prod_{s'=1}^S \tau_{s,s'}^{z_{n-1,s'} z_{n,s}}. \quad (4.1)$$

τ_s is generated from a conjugate prior, a Dirichlet distribution with hyperparameter $\tau_{0,s}$, *i.e.*, $\tau_s \sim \text{Dir}(\tau_{0,s})$. The same is done for π as well, *i.e.*, $\pi \sim \text{Dir}(\pi_0)$. The hyperparameters τ_0 and π_0 are set to a positive value less than 1, which induces a sparse transition structure. Sparse transition structure decreases the number of redundant states, which helps the inference of θ_s .

The latent composition and structure implicitly convey the information about how the music is structured and what its building blocks are. Fig. 4.4 shows a similarity

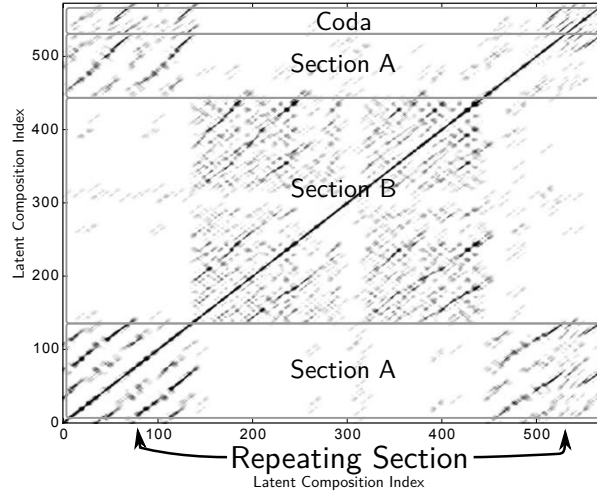


Figure 4.4: Structural annotation of Chopin Op. 41-2 and the similarity matrix computed from its latent composition.

matrix derived from the estimated latent composition of Op. 41-2 by F. Chopin¹ having the ternary form (*a.k.a.* ABA form). The first “A” section repeats a theme of form “DEDF” repeated twice. The second section is in a modulated key. Finally, the last section repeats the first theme, and ends with a short coda, borrowing from “F” motive from the first theme. Noting that the diagonal lines of a similarity matrix represent strong similarity, such a trend may be unveiled by analyzing the matrix. The bottom-left diagonal lines in the first section, for example, show that a theme repeats, and the top-left diagonal suggests that the first theme is repeated at the end. This suggests that the latent composition reflects the organization of music.

Notice that this kind of structure arises because the organization of music is explicitly modeled as an ergodic Markov model; simply aligning multiple performances to a single left-to-right HMM [83,86] is insufficient because it cannot revisit a previously visited state.

4.2.2 Performance sequence

Every audio signal is assumed to be generated from the sequence $\mathbf{z}$, with each signal staying in a given state inside $\mathbf{z}$ with a different duration. Thus, when signal d of D signals generates a feature sequence of length T_d , it is necessary to (1) traverse the sequence in the same order as $\mathbf{z}$, and (2) explicitly model the duration that each signal d spends in each position n inside $\mathbf{z}$. To achieve these goals, the proposed method considers a sequence of “pointer” to $\mathbf{z}$, Φ_d , defined for each signal d . The pointer is designed such that it starts at $\mathbf{z}_{n=1}$. Then, the pointer value stays the same for some

¹The similarity matrix $R_{i,j}$ was determined by removing self-transitions from $\mathbf{z}_n$ and assigning it to $\mathbf{z}'$, and setting $R_{i,j} = 1$ if $\mathbf{z}'_i = \mathbf{z}'_j$, and 0 otherwise. Next, R was convolved by a two-dimensional filter that emphasizes diagonal lines.

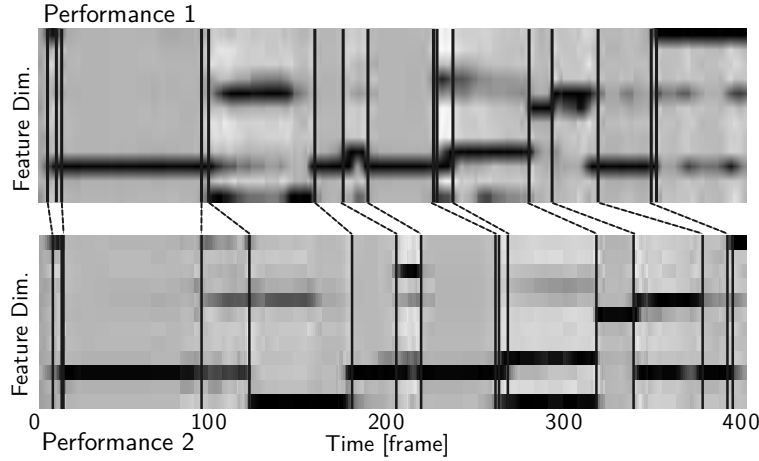


Figure 4.5: Feature sequences (chroma vector) of two performances, overlaid by points where the state of the latent composition changes.

time according to $p(l_{nd}|\Psi)$, which is the likelihood, for some parameter Ψ , that signal d stays in position n of $\mathbf{z}$ for l_{nd} frames. Then, it points to $\mathbf{z}_{n+1}$, repeating the same process until reaching $\mathbf{z}_N$.

To model such a behavior, the sequence of the position inside $\mathbf{z}$ that each signal traverses is expressed as a Markov chain of length T_d over state-space $[1 \dots N] \times [1 \dots L]$, $\Phi_d = \{s_{dt}\}_{t=1}^{T_d}$ (also known as a *semi-Markov chain*). Here, $s_{dtnl} = 1$ means signal d at time t points to $\mathbf{z}_n$, with l more frames remaining in $\mathbf{z}_n$; L is the maximum allowed duration at each state. The state dynamics is set such that for some n , l keeps decrementing until reaching 1, where n gets incremented by 1 and l is set according to $p(l_{n+1,d}|\Psi)$. Furthermore, s_{dt} is constrained to start at $\mathbf{z}_1$ and end at $\mathbf{z}_N$. With $p(l_{nd}|\Psi)$, the duration that s resides in n can be explicitly modeled, and a left-to-right dynamics ensures that the signal starts at $\mathbf{z}_1$ and ends at position $\mathbf{z}_N$. Then, the likelihood is given as follows:

$$p(\Phi|\Psi) = \prod_{d=1}^D \left[\prod_{n=1}^N \left(\delta(n, 1)^{s_{d,1,n,1}} \delta(n, N)^{s_{d,T_d,n,1}} \right) \times \prod_{t=2}^{T_d} \prod_{n=2}^N \left[\prod_{l=1}^L p(l_{nd}|\Psi)^{s_{d,t-1,n-1,1} s_{dtnl}} \prod_{l=2}^L 1^{s_{d,t-1,nl} s_{dtn,l-1}} \right] \right]. \quad (4.2)$$

Fig. 4.5 shows excerpts of the feature sequences obtained from two performances, overlaid by blue lines that indicate time instances where the state of the composition sequence has changed. The figure suggests that the state changes with a notable change in the feature, such as when new notes are played. Since, by the definition of a left-to-right Markov model, the number of vertical lines is identical for all performances, audio signals can be aligned by mapping the occurrences of the s th vertical line for all performances, for each s .

4.2.3 Duration sequence

The duration sequence models the state duration pdf $p(l_{nd}|\Psi)$ introduced in Equation 4.2, as to encourage the relative tempo of each signal to be similar and changing smoothly, with a sporadically varying degree of smoothness. To provide an overview, the idea is to decouple the state duration l_{nd} into (1) the average state duration at each n , $\mathcal{L}_n$, and (2) the relative tempo $\mathcal{T}_n = [\mathcal{T}_{n1} \cdots \mathcal{T}_{nd}]$, *i.e.*, the multiplicative offset from the average $\mathcal{L}_n$ for each signal d at each position n in $\mathbf{z}$. Then, $\mathcal{T}_{nd}$ is modeled as an switching-state LDS (SSLDS) [47] with a *mean-reverting* dynamics with mean 1. SSLDS allows the smoothness of the relative tempo to change sporadically, and mean-reverting dynamics encourages the relative tempo to “return” to a default value, more so when the relative tempo is farther away from the default value.

The state duration is first decoupled into the average state duration $\mathcal{L}_n$ and the relative tempo $\mathcal{T}_n$. Then, l_{nd} , given $\Psi = \{\mathcal{L}_n, \mathcal{T}_n\}_n \cup \{\lambda_0\}$, is assumed to be distributed Normally about $\mathcal{L}_n \mathcal{T}_n$ with precision λ_0 :

$$p(l_n|\mathcal{L}_n, \mathcal{T}_n, \lambda_0) = \mathcal{N}(l_n|\mathcal{L}_n \mathcal{T}_n, \lambda_0^{-1}). \quad (4.3)$$

Since there is little prior information on the value of $\mathcal{L}_n$, and since $\mathcal{L}_n$ is likely to be independent of each other, $\mathcal{L}_n$ is assumed to be distributed mutually independently, with a mean of zero and a small precision ι (*i.e.*, large variance):

$$p(\mathbf{a}|\iota) = \prod_{n=1}^N \mathcal{N}(\mathcal{L}_n|0, \iota^{-1}). \quad (4.4)$$

By setting ι to a small value, the prior assumption on $\mathcal{L}_n$ weakens, and the posterior distribution, with $\mathcal{T}_n = 1$, resembles a Normal distribution centered about the average state duration.

To model $\mathcal{T}$, the proposed method considers the two requirements for the relative tempo. First, the proposed method should encourage the signals to play with a similar relative tempo trajectory, *i.e.*, the $\mathcal{T}$ should “favor” a value of 1. Second, it should express the smoothness of the relative tempo, with a sporadically varying degree of smoothness. To achieve these, $\mathcal{T}$ is modeled as a mean-reverting process with mean $\mathbf{m} = 1$, and an innovation noise with a switching covariance:

$$\mathcal{T}_n - \mathcal{T}_{n-1} = -\alpha(\mathcal{T}_n - \mathbf{m}) + \epsilon_n. \quad (4.5)$$

This kind of process can be seen as a discrete-time analogue of the Ornstein-Uhlenbeck process. The adjacent difference (left-hand side of Equation 4.5) contains (1) a restoring force proportional to $\alpha \in [0, 1]$ and difference from $\mathbf{m}$ (first term in the r.h.s. of Equation 4.5), plus (2) some innovation ϵ_n (second term in the r.h.s. of Equation 4.5). Therefore, the farther $\mathcal{T}_n$ deviates from $\mathbf{m}$, the greater it tends to revert back to $\mathbf{m}$.

To allow the dynamics of $\mathcal{T}$ to change sporadically, ϵ_n is assumed to be a Gaus-

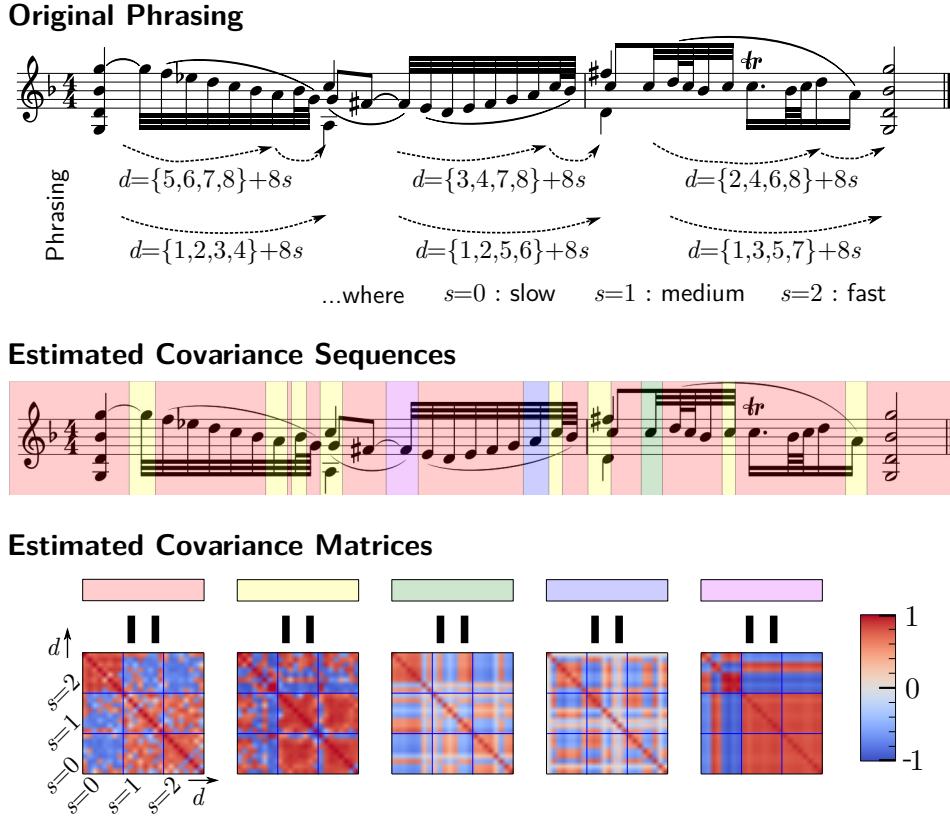


Figure 4.6: A short phrase played with 24 different interpretations, and the estimated $\mathbf{u}$ and Λ . Dotted arrows indicate the choice of executed phrasing.

sian noise with zero mean and precision matrix Λ_n , where Λ_n may vary sporadically. Here, Λ expresses different relative tempo dynamics, *e.g.*, “there is little change in the relative tempo,” “the relative tempo of signals 1 and 3 vary significantly in a correlated manner,” and so on. To model a time-varying Λ_n while constraining it to vary only spuriously, Λ_n is modeled as an M -state Markov chain, $\mathbf{u}$, each state of which describes a unique covariance structure for Λ . By setting M to a reasonable value, the tempo dynamics is expressed as a sequence of a few covariances. The initial state pdf is defined as $\mathbf{v}$ and the state transition pdf is defined as ξ_m . Then, the likelihood of $\mathbf{u}$ is given as follows:

$$p(\mathbf{u}|\xi, \mathbf{v}) = \prod_{m=1}^M v_m^{u_{1,m}} \prod_{n=1}^N \prod_{m=1}^M \prod_{m'=1}^M \xi_{m,m'}^{u_{n-1,m} u_{n,m'}}. \quad (4.6)$$

Then, $\mathcal{T}$ is described as an SSLDS with a time-varying covariance:

$$p(\mathcal{T}_n|\mathcal{T}_{n-1}, \Lambda, \mathbf{u}_n) = \mathcal{N}\left(\mathcal{T}_n \middle| \frac{\mathcal{T}_{n-1} + \alpha \mathbf{m}}{1 + \alpha}, \Lambda_m^{-1}\right)^{u_{nm}}. \quad (4.7)$$

Here, ξ_m and $\mathbf{v}$ are generated from a conjugate Dirichlet distribution, *i.e.*, $\xi_m \sim \text{Dir}(\xi_0)$

and $\mathbf{v} \sim \text{Dir}(\mathbf{v}_0)$; furthermore, Λ_m is generated from a conjugate Wishart distribution $\Lambda_m \sim \mathcal{W}(n_m, \mathbf{W}_m)$.

The proposed model naturally extends duration models based on the smoothness of tempo [91, 112], to deal with a sporadically varying degree of smoothness, and to encourage the relative tempo trajectory to be similar among audio signals. Note that how the proposed method treats the SSLDS for relative tempo modeling is different from that used in beat tracking [15] or score following [104]. Namely, while these studies assume a switching-state dynamics of the *observation noise* and a fixed dynamics of the underlying tempo trajectory, the proposed method assumes a switching-state dynamics of the *underlying tempo trajectory*. It allows the underlying tempo to exhibit different dynamics, making it similar in spirit to expressive tempo modeling [52].

Preliminary analysis suggests that the state of $\mathbf{u}$ changes at structural boundaries of musical interpretation. To illustrate, the *maximum a posteriori* (MAP) estimate of $\mathbf{u}$ and Λ for a short violin phrase is presented in Fig. 4.6, played with 24 different tempo trajectories. Specifically, the d th recording is generated by permuting over three binary decisions over the phrasing (dotted lines) and choosing an overall tempo (variable s in the figure). The MAP estimate of $\mathbf{u}$ is shown as color labels over the music score, and Λ is shown in the bottom of the figure. It shows that performances with similar overall tempo are correlated, as shown by the block-like covariance structure. The red-colored label is a “default” covariance, and the remaining indicate spurious changes in the degree of smoothness. Note that the spurious changes (*i.e.*, non-red labels) often occur at phrase boundaries (*i.e.*, ends of arrows in Fig. 4.6), perhaps since the relative tempo changes significantly at musical structure boundaries [22].

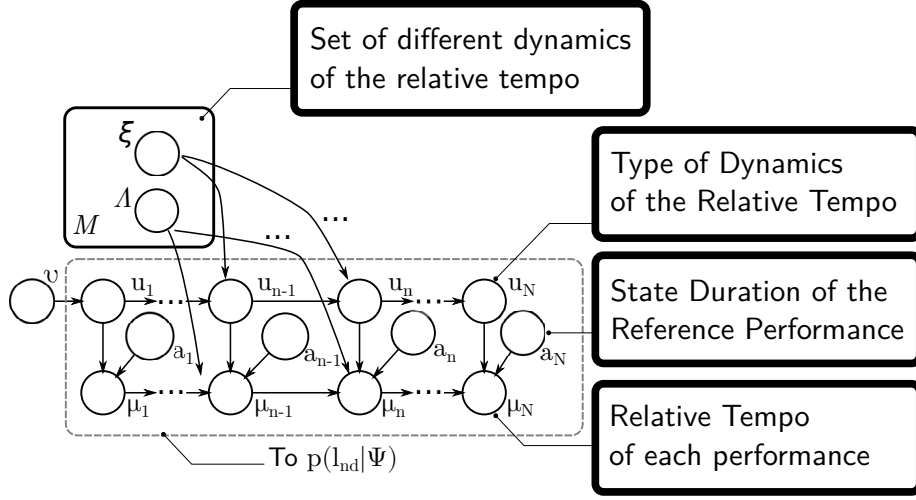
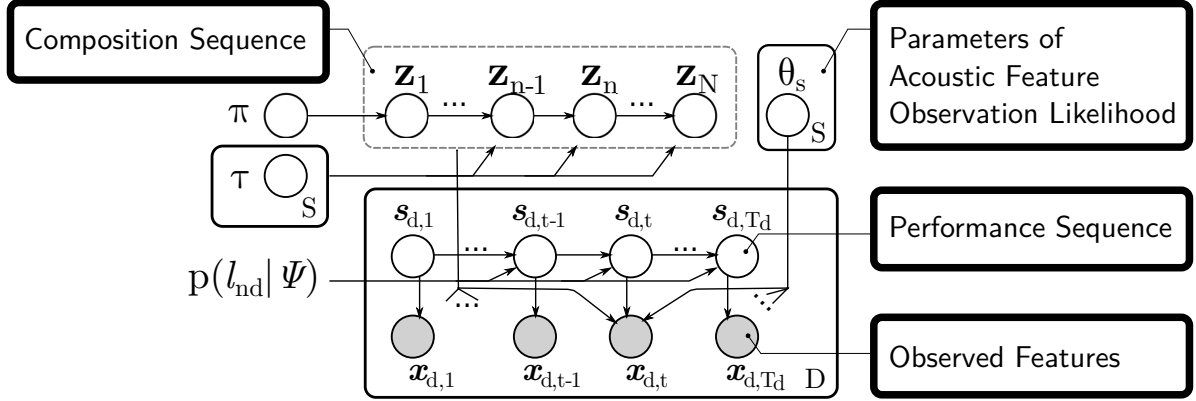
To summarize, $p(l_{nd}|\Psi)$ in Equation 4.2 is expressed as follows:

$$\begin{aligned}
 p(\{\mathbf{l}_n\}_{n=1}^N|\Psi) = & \prod_{n=1}^N \left[\mathcal{N}(\mathbf{l}_n|\mathcal{L}_n\mathcal{T}_n, \lambda_0^{-1}) \mathcal{N}(\mathcal{L}_n|0, \iota^{-1}) \right. \\
 & \times \prod_{m=1}^M \left[\mathcal{N}\left(\mathcal{T}_n \middle| \frac{\mathcal{T}_{n-1} + \alpha}{1 + \alpha}, \Lambda_m^{-1}\right)^{u_{nm}} v_m^{u_{1m}} \prod_{m'=1}^M \xi_{mm'}^{u_{n-1,m}u_{nm'}} \right] \\
 & \times \text{Dir}(\mathbf{v}; \mathbf{v}_0) \prod_{m=1}^M \left[\text{Dir}(\xi_m; \xi_0) \mathcal{W}(\Lambda_m; n_m, \mathbf{W}_m) \right]. \quad (4.8)
 \end{aligned}$$

Fig. 4.7 shows the graphical model of the duration model.

4.2.4 Feature sequence

Finally, the proposed method generates the audio feature $\mathbf{x}_{dt}$ by emitting data from a distribution with an appropriate θ_s associated with frame t of signal d . To do this, s_{dt} is referred to find which state n of $\mathbf{z}$ is referred at the current frame. Then the


 Figure 4.7: The model of state duration pdf ($p(l_{nd})$ in Fig. 4.8).

 Figure 4.8: The model of the performance and composition sequence ($p(l_{nd})$ comes from Fig. 4.8).

appropriate θ_s is retrieved by finding the state s to which z_n refers. In other words:

$$p(\mathbf{x}_{dt} | \mathbf{z}, \Phi, \theta) = \prod_{s=1}^S \prod_{n=1}^N \prod_{l=1}^L p(\mathbf{x}_{dt} | \theta_s)^{z_{ns} s_{dtnl}}. \quad (4.9)$$

Here, $p(\mathbf{x} | \theta_s)$ is the likelihood of observing audio feature $\mathbf{x}$ at state s of $\mathbf{z}$, and θ_s is generated from some prior distribution $p(\theta_s; \theta_0)$. In this chapter, the proposed method uses the chroma vector and its half-wave rectified increment as the feature, assuming that $p(\mathbf{x}_{dt} | \theta_s)$ is a $\dim(\mathbf{x}_{dt})$ -dimensional Normal distribution, whose parameter θ_s is generated from its conjugate prior, the Normal-gamma prior. In other words, the proposed method assumes that $\theta_s = \{\tilde{\mu}_s, \tilde{\lambda}_s\}$, $\theta_0 = \{\mathbf{m}_0, \nu_0, u_0, \mathbf{k}_0\}$, and assumes $\mathbf{x}_{dt} | \tilde{\mu}_s, \tilde{\lambda}_s \sim \mathcal{N}(\tilde{\mu}_s, \tilde{\lambda}_s^{-1})$ and $\tilde{\mu}_{si}, \tilde{\lambda}_{si} \sim \mathcal{NG}(m_{0,i}, \nu_{0,i}, u_{0,i}, k_{0,i})$. The graphical model of the performance and composition sequence is shown in Fig. 4.8.

Reflecting back on the formulation in Equation 2.24, the proposed method can be seen as generating z from an underlying ergodic HMM, expressing θ using a single distribution of the chroma vector at each state, and expressing s using a set of HSMs with a duration model that encourage every signal to play in a similar relative tempo.

4.3 Inference

The posterior distribution to the model described above is derived in this section. Since direct application of Bayes' rule to arrive at the posterior is difficult, the variational Bayes method [7] is employed to find an approximate posterior of form $q(s, z, \theta, \eta, \pi, \tau) = \prod_d q(s_{d,\cdot}) q(z) q(\pi) \prod_{d,n} q(\eta_{d,n}) \prod_s q(\theta_s) q(\tau_s)$ that minimizes the Kullback-Leibler (KL) divergence to the true posterior distribution.

$q(s)$ and $q(z)$ can be updated in a manner analogous to an HMM. $q(z)$ may be updated using the forward-backward algorithm, with the state emission probability g_n at position n of the latent composition and the transition probability v_s from state s given as follows:

$$\log g_{n,s} = \sum_{d,t} \langle s_{d,t,n,\cdot} \rangle \langle \log p(x_{d,t} | \theta_s) \rangle, \quad (4.10)$$

$$\log v_{s,s'} = \langle \log \tau_{s,s'} \rangle. \quad (4.11)$$

Specifically, the forward message is computed as follows:

$$\alpha_{ns} = g_{ns} \sum_{s'} \alpha_{n-1,s'} v_{s's}, \quad (4.12)$$

starting at $\log \alpha_{1,s} = \langle \log \pi_s \rangle$. The backward message is computed as follows:

$$\beta_{ns} = \sum_{s'} v_{ss'} g_{n+1,s'} \beta_{n+1,s'}, \quad (4.13)$$

starting at $\beta_{N,s} = 1$. Then, π can be updated as follows:

$$q(\pi) = \text{Dir}(\pi_0 + \langle z_1 \rangle), \quad (4.14)$$

and τ as follows:

$$q(\tau_s) = \text{Dir}\left(\tau_{0,s} + \sum_{n>1}^N \langle z_{n-1,s} z_n \rangle\right). \quad (4.15)$$

Likewise, for $q(s_{d,t})$, forward-backward algorithm is used to update the variational posterior, with the state emission probability $h_{d,n}$ and transition probability

$w_{d,s}$ given as follows:

$$\log h_{dtn} = \sum \langle z_{ns} \rangle \langle \log p(\mathbf{x}_{dt} | \boldsymbol{\theta}_s) \rangle \quad (4.16)$$

$$\log w_{dnl} = \left\langle \log \mathcal{N}(l | a_n \mathcal{T}_{nd}, \lambda_0^{-1}) \right\rangle. \quad (4.17)$$

Then, the forward message is computed as follows:

$$\alpha_{dtnl} = h_{dtn} (\alpha_{d,t-1,n,l+1} + \alpha_{d,t-1,n-1,1} w_{dnl}), \quad (4.18)$$

starting at $\log \alpha_{d,1,n,l} = \delta_{n,1} \frac{1}{L}$. Furthermore α_{dtnl} for all $l > L$ is set to 0, and α_{dtnl} for all $n < 1$ is set to 0. The backward message is computed as follows:

$$\beta_{dtnl} = \begin{cases} \sum_{l'} h_{d,t+1,n+1,l'} \beta_{d,t+1,n+1,l'} w_{d,n+1,l'} & l = 1 \\ h_{d,t+1,n,l-1} \beta_{d,t+1,n,l-1} & 1 < l \leq L, \end{cases} \quad (4.19)$$

starting at $\beta_{dT_dnl} = 1$ for all d, n and l .

Based on these parameters, the generative model of audio features can be updated. Some commonly-used statistics for state s include the count $\bar{N}_s$, the mean $\bar{\boldsymbol{\mu}}_s$ and the variance $\bar{\boldsymbol{\Sigma}}_s$, which are given as follows:

$$\bar{N}_s = \sum_n \langle z_{ns} \rangle \langle s_{::n} \rangle \quad (4.20)$$

$$\bar{\boldsymbol{\mu}}_s = \frac{1}{\bar{N}_s} \sum_{d,t,n} \langle z_{ns} \rangle \langle s_{dtn} \rangle \mathbf{x}_{dt} \quad (4.21)$$

$$\bar{\boldsymbol{\Sigma}}_s = \frac{1}{\bar{N}_s} \sum_{d,t,n} \langle z_{ns} \rangle \langle s_{dtn} \rangle (\mathbf{x}_{dt} - \bar{\boldsymbol{\mu}}_s)^2 \quad (4.22)$$

With these, the observation model can be updated as follows:

$$q(\boldsymbol{\mu}_s, \boldsymbol{\lambda}_s) = \mathcal{NG} \left(v_0 + \bar{N}_s, \frac{v_0 \mathbf{m}_0 + \bar{N}_s \bar{\boldsymbol{\mu}}_s}{v_0 + \bar{N}_s}, \right. \\ \left. u_0 + \frac{\bar{N}_s}{2}, \mathbf{k}_0 + \frac{1}{2} \left(\bar{N}_s \bar{\boldsymbol{\Sigma}}_s + \frac{v_0 \bar{N}_s}{v_0 + \bar{N}_s} (\bar{\boldsymbol{\mu}}_s - \mathbf{m}_0)^2 \right) \right) \quad (4.23)$$

Hyperparameters may be set manually, or optimized by minimizing the KL divergence from q to the posterior.

To update the duration sequence, the following statistics are introduced:

$$\mathcal{X}_{ndl} = \sum_{t=1}^{T_d} \langle s_{d,t-1,n-1,1} s_{dtnl} \rangle \quad (4.24)$$

$$\mathcal{C}_{nd} = \mathcal{X}_{nd} \quad (4.25)$$

$$\mathcal{M}_{nd} = \frac{1}{\mathcal{C}_{nd}} \sum_{l=1}^L l \mathcal{X}_{ndl} \quad (4.26)$$

To update the assigned covariance, the forward-backward algorithm is used, with emission probability O_{nm} and transition probability $T_{m,m'}$ given as follows:

$$\log O_{nm} = -\frac{1}{2} \text{tr}(\Gamma_n \Lambda_m) + \frac{1}{2} \langle \log \det \Lambda_m \rangle \quad (4.27)$$

$$\log T_{m,m'} = \langle \log \xi_{m,m'} \rangle \quad (4.28)$$

Then the forward message is computed as follows:

$$\alpha_{nm} = \sum_{m'} \alpha_{n-1,m'} O_{nm} T_{m',m} \quad (4.29)$$

and the backward message is computed as follows:

$$\beta_{nm} = \sum_{m'} \beta_{n+1,m'} O_{n+1,m'} T_{m,m'} \quad (4.30)$$

From these, $\langle \mathbf{u}_n \rangle \propto \alpha_n \beta_n$ may be estimated. With the estimated expectations, $q(\mathbf{v})$ can be updated as follows:

$$q(\mathbf{v}) = \text{Dir}(\mathbf{v}_0 + \langle \mathbf{u}_1 \rangle), \quad (4.31)$$

and $q(\xi)$ as follows:

$$q(\xi_{m,:}) = \text{Dir}\left(\xi_0 + \sum_{n>1}^N \langle u_{n-1,m} \rangle \langle z_{n,:} \rangle\right) \quad (4.32)$$

Next, consider updating $q(\mathcal{T})$. Since the functional form is like an LDS, the expectations may be computed using the Kalman smoother. The forward message $p(\mathcal{T}_n | \mathcal{X}_{1\dots n}, \dots)$, whose functional form is $\mathcal{N}(\mathcal{T}_n | \mathbf{g}_n, \mathbf{V}_n)$, is computed. To this end, let us introduce the following variables:

$$\Gamma_n = \langle (\boldsymbol{\mu}_n - \beta \boldsymbol{\mu}_{n-1} - \boldsymbol{\gamma})(\boldsymbol{\mu}_n - \beta \boldsymbol{\mu}_{n-1} - \boldsymbol{\gamma})^T \rangle \quad (4.33)$$

$$\mathbf{S}_n = \sum_{m=1}^M \langle u_{nm} \rangle \langle \Lambda_m \rangle \quad (4.34)$$

$$\mathbf{A}_n = \text{diag} \langle \mathcal{L}_n \rangle \quad (4.35)$$

Then, the forward message is computed follows:

$$\mathbf{P}_n^{-1} = \mathbf{V}_{n-1} + (1 + \alpha)^{-2} \mathbf{S}_n \quad (4.36)$$

$$\mathbf{V}_n = \mathbf{S}_n + \mathbf{A}_n \Lambda_n \mathbf{A}_n - (1 + \alpha)^{-2} \mathbf{S}_n \mathbf{P}_n \mathbf{S}_n \quad (4.37)$$

$$\mathbf{g}_n = \mathbf{V}_n^{-1} \left[\frac{\mathbf{S}_n \mathbf{P}_n}{1 + \alpha} \left(\mathbf{V}_{n-1} \mathbf{g}_{n-1} - \frac{\alpha \mathbf{S}_n \mathbf{m}}{(1 + \alpha)^2} \right) + \frac{\alpha \mathbf{S}_n \mathbf{m}}{1 + \alpha} + \mathbf{A}_n \Lambda_n \mathcal{M}_n \right]. \quad (4.38)$$

Similarly, the backward message $p(\mathcal{X}_{n+1 \dots N} | \mathcal{T}_n, \dots)$ can be computed to yield a functional form $\mathcal{N}(\mathcal{T}_n | \mathbf{h}_n, \mathbf{W}_n)$, whose parameters are given as follows:

$$\mathbf{Q}_n^{-1} = \mathbf{W}_n + \mathbf{S}_n + \mathbf{A}_n \Lambda_n \mathbf{A}_n \quad (4.39)$$

$$\mathbf{W}_{n-1} = (1 + \alpha)^{-2} \mathbf{S}_n (\mathbf{I} - \mathbf{Q}_n \mathbf{S}_n) \quad (4.40)$$

$$\mathbf{h}_{n-1} = \frac{\mathbf{W}_n^{-1} \mathbf{S}_n}{1 + \alpha} \left[\mathbf{Q}_n \left(\mathbf{W}_n \mathbf{h}_n + \frac{\alpha \mathbf{S}_n \mathbf{m}}{1 + \alpha} + \mathbf{A}_n \Lambda_n \mathcal{M}_n \right) - \frac{\alpha \mathbf{m}}{1 + \alpha} \right]. \quad (4.41)$$

Combining these, the variational posterior for $\mathcal{T}_n$ is obtained as follows:

$$q(\mathcal{T}_n) = \mathcal{N}(\mathbf{U}_n (\mathbf{V}_n \mathbf{g}_n + \mathbf{W}_n \mathbf{h}_n), \mathbf{U}_n), \quad (4.42)$$

where

$$\mathbf{U}_n = (\mathbf{V}_n + \mathbf{W}_n)^{-1}. \quad (4.43)$$

Using these, the expectations $\langle \mathcal{T}_n \mathcal{T}_n \rangle$ and $\langle \mathcal{T}_{n-1} \mathcal{T}_n \rangle$ may be updated as follows:

$$\langle \mathcal{T}_n \mathcal{T}_n^T \rangle = \mathbf{U}_n \quad (4.44)$$

$$\langle \mathcal{T}_{n-1} \mathcal{T}_n^T \rangle = \frac{\mathbf{P}_n^{-1} \mathbf{S}_n}{1 + \alpha} \left(\mathbf{Q}_n^{-1} + \frac{\mathbf{S}_n^T \mathbf{P}_n^{-1} \mathbf{S}_n}{(1 + \alpha)^2} \right)^{-1}. \quad (4.45)$$

Next, the update for the relative note duration $q(a)$ is obtained:

$$\mathcal{L}_n \sim \mathcal{N} \left(\frac{\iota \kappa + \lambda_0 \sum_{d=1}^D C_{nd} \langle \mathcal{T}_{nd} \rangle \mathcal{M}_{nd}}{\iota + \lambda_0 \sum_{d=1}^D C_{nd} \langle \mathcal{T}_{nd}^2 \rangle}, \left(\iota + \lambda_0 \sum_{d=1}^D C_{nd} \langle \mathcal{T}_{nd}^2 \rangle \right)^{-1} \right). \quad (4.46)$$

Finally, the tempo covariance for each template Λ_m is updated:

$$\Lambda_m \sim \mathcal{W} \left(n_m + \langle u_{:m} \rangle, \left[\mathbf{W}_m^{-1} + \sum_{n=1}^N \langle u_{nm} \rangle \langle \Gamma_n \rangle \right]^{-1} \right) \quad (4.47)$$

4.4 Evaluation

Here, three experiments are conducted to assess the proposed method. First experiment assesses the effectiveness of the proposed method against existing methods that ignore the organization of music [83, 86]. Second experiment investigates how well similar sounding fragments are grouped into one state; to evaluate this, the onset detection capability of the proposed method is tested, assuming that a change of state corresponds to an onset. Last test evaluates the robustness of the proposed method to the length of the latent composition, a parameter that needs to be fixed in advance.

4.4.1 Experimental conditions

The proposed method is evaluated by computing the alignment error percentile using nine pieces from Chopin’s *Mazurka*, each with 2 to 5 recordings. For each song, the estimated alignment was compared with a human-annotated data [114]. The feature was computed using a sampling frequency of 44.1 kHz at 25 frames per second. Then, the proposed method was compared against six different methods:

DTW DTW that minimizes the net cosine distance using path constraint in [58]. Since it practically cannot deal with alignment of more than two signals at a time, the error percentile is evaluated over all possible pairwise alignments.

HMM Our model, with an HMM to model ϕ , ignoring any duration constraints [82]. This serves as the baseline for probabilistic audio alignment.

Independent Treating ϕ as an HSMM, assuming independent state distribution both across signals d and positions in z, n .

Mean Assuming, for each position n , a single Normal distribution as the duration pdf, which ignores continuity of relative tempo but encourages signals to play with similar relative tempo trajectories.

Dynamics (Single-state) Setting $\alpha = 0$ and setting $M = 1$ in Equation 4.8, making the proposed model a LDS (no switching-state) with a random-walk dynamics. It ignores the deviation of relative tempo trajectory from each other and assumes a single underlying dynamics, thus making it similar in spirit to existing duration models [91, 112].

Dynamics (Switching-state) Same as above, except a switching-state dynamics is used by setting $M = 5$.

Proposed The proposed method, *i.e.*, method **Dynamics (Switching-state)** with a mean-reverting dynamics.

Table 4.1: Onset detection rate

Method	Prec.	Rec.	F-meas.
HMM	43%	26%	31%
Dynamics (Single-state)	72%	70%	70%
Proposed	75%	74%	74%

The hyperparameters to the composition sequence were set to $\pi_0 = 0.1$ and $\tau_{0,s,s'} = 0.9 + 10\delta(s, s')$, which encourages a sparse transition and self-transitions. The hyperparameters related to the mean-reverting dynamics were set to $\alpha = 0.1$, which moderately encourages the relative tempo to be near $m = 1$. The prior precision of a_n, ι , was set to 0.01, allowing for standard deviation of about three frames. The prior precision of l_{nd} is set to $\lambda_0 = 30$. For the SSLDS, the hyperparameters were set such that $\xi_0 = 0.1$, $\nu_0 = 0.1$, and $M = 5$, which encourages a sparse transition structure. The prior of Λ_m was set to $n_m = D$ and $\mathbf{W}_m = 100\mathbf{I}_d$. For the observation models, the hyperparameters were set such that $u_0 = \mathbf{k}_0 = 1$, $\nu_0 = 0.1$ and $\mathbf{m}_0 = 0$, which encourages a sparse variance and a dispersed mean. Finally S was set to 100 and N to $0.3 \min_d T_d$.

4.4.2 Experimental results

The result is shown in Fig. 4.9. The proposed method performs comparably to DTW, a commonly-used pair-wise audio alignment method, attaining a median alignment error of about 60 ms. Moreover, it was confirmed that it is effective to (1) encourage the similarity of the tempo trajectory among the input signals and (2) encourage the smoothness of relative tempo with sporadically switching degree of smoothness. This suggests that existing alignment methods, that use LDS-based tempo dynamics, including audio-to-score alignment, may benefit from tying the mean duration using mean-reverting dynamics, and incorporating switching-state dynamics. This kind of modeling strategy is impractical for DTW-based alignment, whose memory and computational complexity increases exponentially with number of signals D to simultaneously align, and incorporating explicit duration constraint is difficult.

The result suggests a room for improvement in modeling the composition sequence, performance sequence and the acoustic observation model. That is, when duration constraints are ignored, DTW (red line) significantly outperforms HMM (orange line). This result is in accord with results from audio-score alignment literature, where the DTW typically outperforms HMM when only simple feature such as the chroma is used inside a simple model of observed feature [63]. Improvement is expected by incorporating better features and sophisticated models of notes, as in [63].

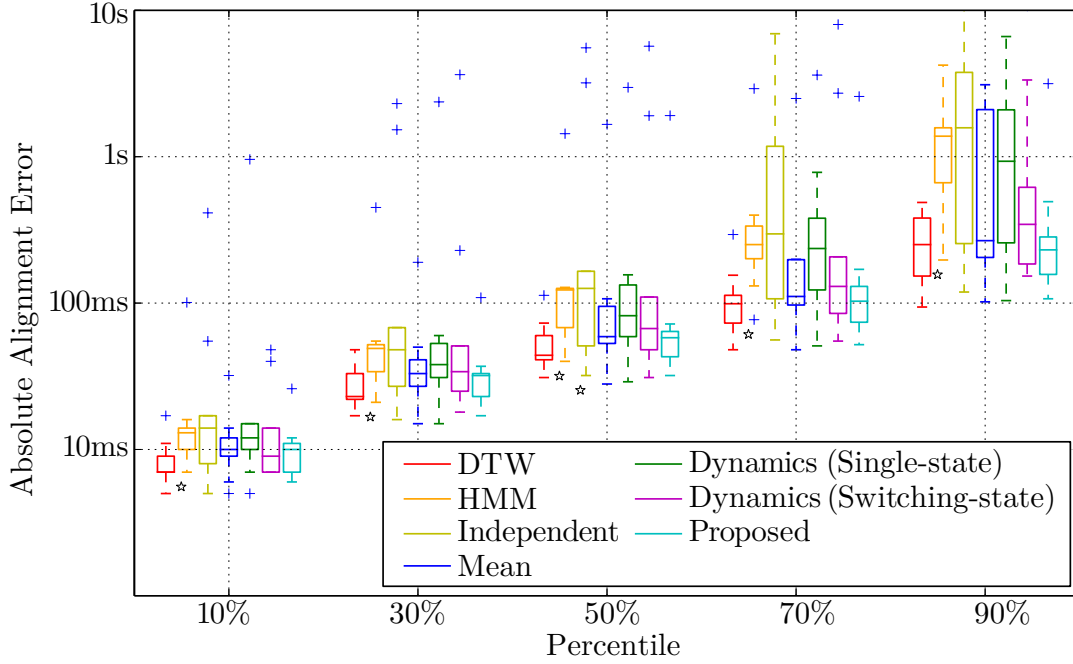


Figure 4.9: Percentile of absolute alignment error. Asterisks indicate statistically significant difference over DTW ($p=0.05$) and circles indicate statistically significant difference over LRHMM ($p=0.05$), using Kruskal-Wallis H-test.

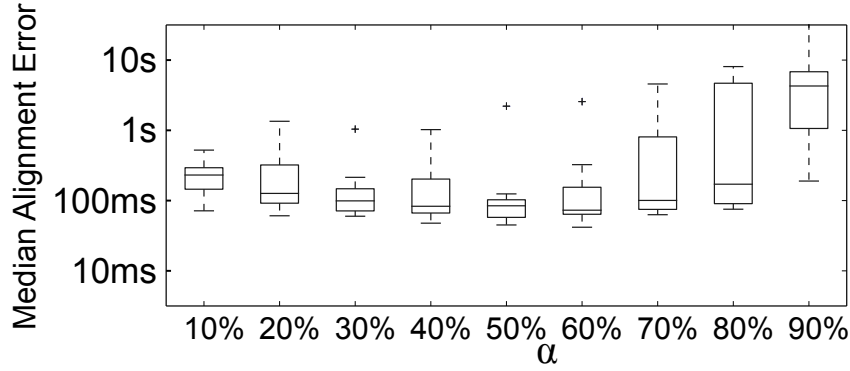
4.4.3 Onset detection

This experiment verifies that the places where the state of latent composition changes correspond to changes in tonal characteristics. To this end, the onset detection accuracy of the proposed method was computed. The onset is defined the positions where the state of the latent composition have changed. Then the estimated onset position was compared to the onset data from the Mazurka Project [114]. The onset position was defined to be correct if it lied within 0.1 s of an annotated onset position.

Table 4.1 shows the precision, recall and the F-measure. This shows that the proposed method perform significantly better than the left-to-right HMM method, with the semi-Markov performance sequence performing better than the Markovian performance sequence.

Robustness to the length of the latent composition

Since the proposed method requires the user to set the length of latent composition N , this experiment evaluated the quality of alignment as N is varied. To evaluate the performance of the proposed method with different values of N , the alignment of the proposed method was evaluated when N is set to $N = \alpha |T_{d=1}|$, with α ranging from $\alpha = 0.1$ to $\alpha = 0.9$ with an increment of 0.1. Fig. 4.10 shows the median alignment error. When α is too small, the error increases when there is an insufficient number of states to describe a composition. The error also increases when α is too large, since

Figure 4.10: Median alignment error against α .

the maximum total allowed deviation decreases (*i.e.*, to about $(1 - \alpha)T_{d=1}$). However, outside such extremities, the performance is relatively stable for moderate values of α around 0.5. This suggests that the proposed method is relatively insensitive to a reasonable choice of N .

4.5 Summary

This chapter presented an audio alignment method based on a probabilistic generative model. Based on the insight that a generative model of musical audio alignment should represent both the underlying musical composition and how it is performed by each audio signal, this chapter presented a unified generative model of musical composition and performance. The proposed generative model contributed to a significantly better alignment performance than existing methods, attaining a median alignment error of 60 ms. The result brings generative alignment on par with DTW-based alignment, opening door to alignment problem settings that require integration of various sources of uncertainties. For example, alignment under a noisy mixture is a problem where integration of denoising techniques and alignment seems preferable. In such a case, one must approach the problem in a generative formulation.

Chapter 5

Bayesian Subset Music Alignment

This chapter presents a method for aligning musical media that *contains* some shared note sequences. Only audio signals will be discussed in this chapter, but a symbolic music score may be treated in the same manner by rendering it with a synthesizer.

In music alignment, it is often useful to align different music media that *contain* an identical sequence of notes plus some other elements. In performance analysis, for example, an amateur singer might want to compare his solo singing of an opera with an orchestral recording of the same piece. This calls for an alignment method that aligns solo vocal audio to full opera audio. In informed source separation, the user may want to provide the music score for only the parts that he is interested in separating out, instead of all parts. This requires the alignment between an audio signal and a musical score data that represents a subset of the notes played by the audio signal. When a user wants to hum a melodic line that he wants to separate out from a polyphonic mixture (known as separation by humming [119]), alignment between user's humming and polyphonic mixture allows the separation to be done on humming that is out of sync with the mixture. Yet, such a problem has been neglected in conventional studies, which have focused on how to align audio signals and/or music score data that represent an *identical* sequence of notes.

This chapter proposes *subset music alignment*, a task for aligning different musical media that *contain* parts that are played in common. Furthermore an offline subset music alignment method is presented. Subset music alignment must concurrently address two tasks: identifying the common set of notes played by two or more media, and aligning the media based on the common components. The two problems are mutually dependent, so they must be determined jointly. To elaborate, common set of notes can be identified if the times at which D signals play the same location in the music score are known. But the components are played in common must be known for identifying the location. To merge the two problems in a unifying framework, joint estimation is tackled by inferring a probabilistic generative model that expresses these two subproblems. Constraints pertaining to these subproblems are expressed probabilistically, through the use of prior distributions.

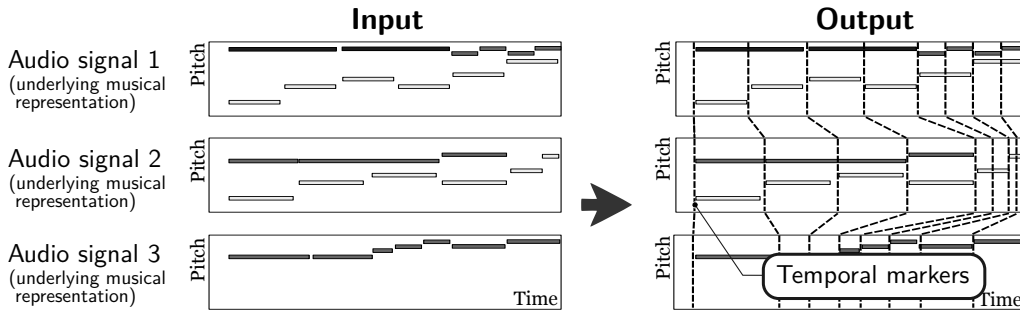


Figure 5.1: The idea behind subset music alignment. Assuming audio 1 and 2 play subsets of a same score, the method aligns the audio signals by focusing on the notes played by both signals (shaded noteheads).

Section 5.1 defines subset music alignment and positions it with respect to existing studies. Section 5.2 presents a generative model to subset music alignment. A posterior inference scheme is presented in Section 5.3, and the proposed method is evaluated in Section 5.4.

5.1 Subset music alignment

Here, subset music alignment is defined and compared with existing audio alignment methods.

5.1.1 Problem Specification

Subset music alignment is a task for aligning different musical media. The problem statement is illustrated in Fig. 5.1 and summarized below:

Input D music media (audio signals or music score). Each medium represents a non-disjoint subset to a common piece of music.

Output Temporal alignment among the media, *i.e.*, a collection of sets of D temporal “markers,” each set of D markers of which contains the temporal locations of the D media that play the same location in the underlying common piece of music.

Assumption The common piece of music is describable as a sequence of a set of notes.

Subset music alignment is nontrivial because it needs to find out not only the alignment but also the components within the input signals that are relevant for alignment.

5.1.2 Related Studies

Aligning the time axis of multiple musical representations has gathered considerable interest in the music information retrieval (MIR) community. However, as noted in Chapter 2, most studies have focused on how to align different musical media that

represents an *identical* musical score. On the contrary, subset music alignment requires alignment among audio signals that contain some shared note sequences that are not necessarily identical.

Subset music alignment methods should be able to (1) discern different *notes* within the signals, and (2) align the signals so that same notes played are “shared” among different representations. On the other hand, these requirements are not necessarily handled in existing alignment methods.

Many existing alignment methods are designed to discern different pitch *combinations*, instead of the notes themselves. This is sensible when aligning audio signals that play an identical score, since it is reasonable to expect to observe the same sequence of pitch combinations. This makes it sensible to use, say, the chroma vector to extract the combination of pitch-classes played, and use it as the basis for aligning audio signals. Some studies discern different notes by segregating the input audio into different pitched sounds [17, 61, 84, 98, 103]. These studies, however, use the segregated information to ultimately find the note combination that is played. Subset music alignment requires differentiation between different notes, suggesting that segregation approach is more sensible.

Existing methods align the signals so that the temporally aligned signals look as *identical* to each other as possible, instead of making the signals share as many components as possible. For example, a path minimization approach, typically using dynamic time-warping (DTW) [27, 58, 99], seeks to minimize the difference between two signals that are temporally warped. For a probabilistic approach, often formulated as a hidden Markov model (HMM) and its variants [18, 61, 84, 112] or as a continuous state-space model [30, 103], observation likelihood is modeled as to “favor” signals that play identically to a predefined pattern. In subset audio alignment, however, the alignment should maximize the degree of “sharedness,” instead of sameness: aligned signals may look very different from each other, as long as there are many shared components. This is similar in philosophy to ignoring components that are irrelevant for alignment [69].

5.2 Formulation

This chapter presents a subset audio alignment method, which is capable of aligning multiple audio renditions to a same piece of music, each rendition of which *contains* a common piece of music. The proposed method renders a digital music score using a high-quality synthesizer, allowing symbolic music score data to be treated like an audio data. Therefore, the following exposition is based on aligning signals, and formulates a generative model that describes a collection of musical audio signals.

5.2.1 Concept

Broadly speaking, the proposed method assumes that each signal plays a subset of a “master” score. Here, the master score is a sequence of note combinations common to all input signals. Then, the model expresses the note sequence associated with each signal by temporally warping the master score, and selecting which notes from the master score to play. Finally, the model converts the note sequence to audio signal by expressing the signal as a weighted mixture of the spectrum associated with each note that is being played. Given this kind of generative model, the alignment is estimated by working *backward*: the method starts from the given input audio signals, and infers the model parameters that best describe the input signals.

Expressing the note sequence

First, Fig. 5.2 illustrates the concept behind the generative process of the note sequence. This figure shows three audio data (right half). Shaded boxes indicate the notes that are played at a given pitch at a given time, and different parts are indicated by different shades of gray. Each signal represents a common piece of music composed of three parts, indicated by the “master score” shown in the left half of the figure, with some notes missing (indicated by dotted notes).

As shown in the figure, each signal is interpreted as a subset of the master score that has been warped in time (indicated by dotted vertical line). The master score is described as a state sequence with a known ordering. Moreover, each state of the master score emits a combination of “bases.” Each basis describes a distinct note. Then, the note sequence is generated by first warping the master score in time, and then choosing which notes in the master score are emitted at each position.

An issue with the proposed method is that the number of notes comprising the piece is unknown, since it depends on the degree of polyphony of each signal at each time instance. That is, the number of bases, the number of notes in each state of the master score, and the number of notes that each signal chooses to play within each state are unknown. Thus, a model that could infer the effective number in a data-driven manner is necessary.

Expressing the observed signals

The note combination is represented as a discretized constant-Q transform (CQT) [12]. In other words, the proposed method infers the model described in Section 5.2.1 from the CQT spectrogram of the input signals.

Given the sequence of emitted notes, the CQT spectrogram is generated by accumulating quanta of energy that has been generated from different bases. To elaborate, the discretized CQT spectrogram is interpreted as a histogram of energy quanta, where the power contained in each time-frequency bin corresponds to the number of quanta observed at the given time-frequency bin. This kind of interpretation has

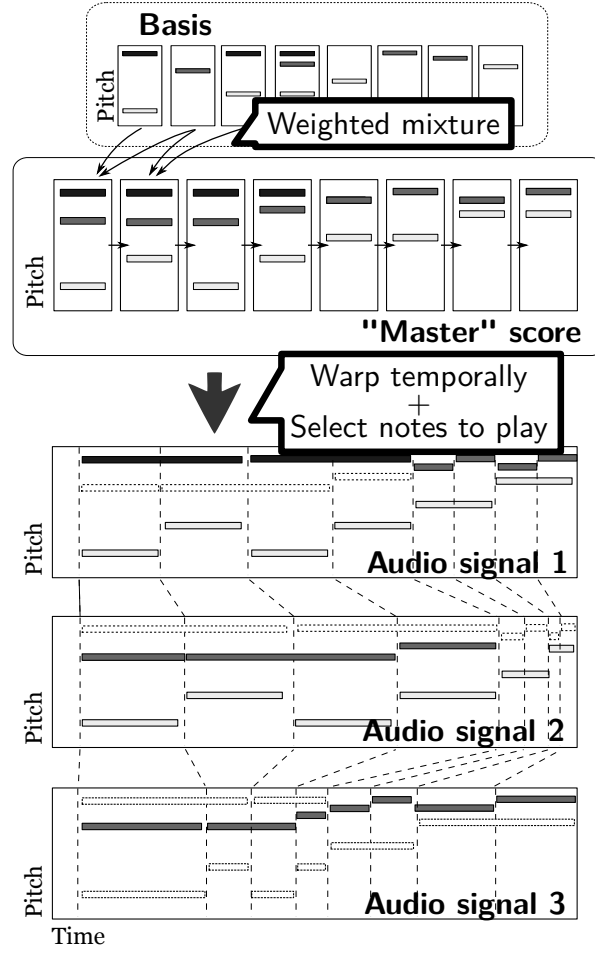


Figure 5.2: The proposed method considers a latent “master” music score, subsets of which generates the sequence of notes played for each music media, temporally warped by the signal-dependent note durations.

been shown to be effective for decomposing spectrogram into spectral time-slice that repeatedly appear throughout the data [13, 119, 120, 138]. As illustrated in Fig. 5.3, that each count is assumed to be generated from a different basis (expressed as different shades of gray). Which basis is likely to appear at each time is governed by which notes are likely to appear at that time, as given by the model of note activation described earlier.

Now the model is described in more detail. For the moment, let K be the number of spectral basis and F be the number of frequency bins. The method expresses some spectral time-slice $\mathbf{x} = [x_1 \cdots x_F]$. The k th spectral basis is expressed as a normalized histogram over the frequency axis, $\boldsymbol{\theta}_k = [\theta_{k1} \cdots \theta_{kF}]$. For example, $\boldsymbol{\theta}_1$ may represent the normalized spectral time-slice of a piano playing A4, $\boldsymbol{\theta}_2$ may represent that of a violin playing F#4, and so on. Moreover, let the likelihood of observing each basis be $\boldsymbol{\pi} = [\pi_1 \cdots \pi_K]$, where π_k is the likelihood of observing basis k . Then, $\mathbf{x}$ is expressed in two steps. First, $N = \sum_f x_f$ indicator variables $\mathbf{z} = [z_1 \cdots z_N]$ are drawn. z_c ex-

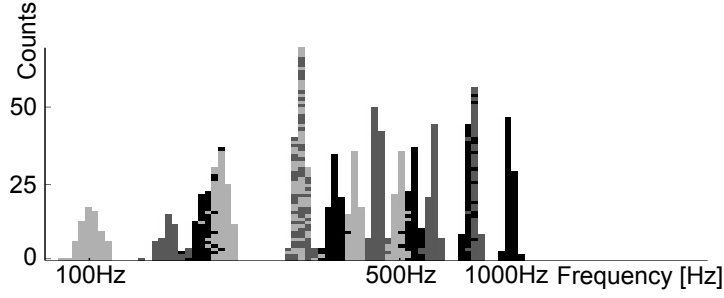


Figure 5.3: A sample drawn from the model of CQT timeslice. The proposed method treats the timeslice as a histogram count. Different shades of gray indicate different spectral bases, each associated with a note in Fig. 5.2 (inside the box labeled “Basis”).

presses the spectral basis that is associated with the c th count of the histogram. Since the likelihood of observing each basis is governed by π , z is drawn independently according to π :

$$z_c \sim \text{Categorical}(\pi). \quad (5.1)$$

Then, the frequency bin to assign to the c th count is chosen. Let c_c be the assigned frequency bin. Since the likelihood of assigning the frequency bin is dependent on the shape of the spectral basis θ , c_c is drawn according to it:

$$c_c \sim \text{Categorical}(\theta_{z_c}). \quad (5.2)$$

Note that if z is integrated out, the following mixture distribution is obtained:

$$p(c_c = f | \pi, \theta) = \sum_{k=1}^K \pi_k \theta_{kf} \quad (5.3)$$

The form of distribution is independent of c , and describes a weighted mixture of the spectral bases. Here, K should express the number of notes that are contained in $\mathbf{x}$.

In the proposed method, θ represents the notes that exists for a given signal at a given time, as shown in Fig. 5.2. Moreover π indicates the relative loudness of the notes for a given signal at a given time, and K reflects the number of notes for a given signal at a given time. In other words, θ , π and K varies at each state for each signal. This means that, as discussed in Section 5.2.1, the number of bases K should “vary” with the nature of the observed spectra. Moreover, the bases θ should maintain a hierarchical relationship among the set of all bases, the master score, and the notes played for each signal.

5.2.2 Model

Now, the concepts introduced in Section 5.2.1 and illustrated in Fig. 5.2 shall be mathematically formulated. The key idea is to exploit the Dirichlet Process (DP) to express

spectral bases (different boxes inside “Bases” illustrated in Fig. 5.2). DP has a property that the effective number varies with the complexity of the signals. It then uses hierarchical DP (HDP) to express the hierarchical relationship depicted in Fig. 5.2 – namely, those among the set of all notes used, the notes inside the master music score, and the notes emitted by each signal. To model how the master sequence develops over time, the proposed method uses a left-to-right hidden semi-Markov model (LRHSM) with fixed terminal states. This way, each signal traverses the master music score in the same order, while the actual duration that each signal stays in a given state in the master music is unique to each signal.

In the following exposition, D audio signals are given, each of which is an audio rendition of a subset of a given piece of music. Let x_{dtf} be the CQT spectrogram of the d th audio signal, evaluated at time t of duration T_d and f th of F filter banks comprising the CQT.

Modeling the note sequence

As shown in Fig. 5.2, the proposed method expresses the note emitted by each signal as a subset of the master score, which in turn is a sequence of subsets of all possible notes. This is realized by expressing the hierarchy of notes using an HDP, and expressing how the master score evolves over time using a LRHSM. The proposed method is pictorially shown in Fig. 5.4.

To elaborate, a three-level hierarchical DP is used to express the hierarchy of notes. First, the set of basis used (the box labeled “Basis” in Fig. 5.2 and Fig. 5.4) is represented using a DP. The DP represents countably infinite number of spectral bases, each basis of which is associated a weight. Thus, it is an infinite-dimensional extension of the spectral model described in Equation 5.1, Equation 5.2 and Equation 5.3. However, the DP prefers a compact representation, so only a finite number of the infinite number of bases gets used. Second, the notes played inside a master score is expressed as a sequence of HDP whose base measure is the set of bases used (the box labeled “Master score” in Fig. 5.2 and Fig. 5.4). Finally, the actual spectral basis emitted by each signal is an HDP whose base measure is based on the master score (the boxes labeled “Audio Signal d ” in Fig. 5.2 and Fig. 5.4). The key point here is that the numbers of elements in these sets depend on how many notes are sounding for each signal at each time instance. HDP expresses this kind of concept naturally.

Now the proposed model is mathematically specified. First, assume that the space of possible spectral basis is defined by a measure H . The specific form of H in will be developed in Section 5.2.2. Given H , the distribution of spectral basis G_0 is defined as a DP with base measure H and concentration parameter α , where α controls the growth of the effective number of spectral bases in G_0 :

$$G_0 \sim \text{DP}(\alpha, H). \quad (5.4)$$

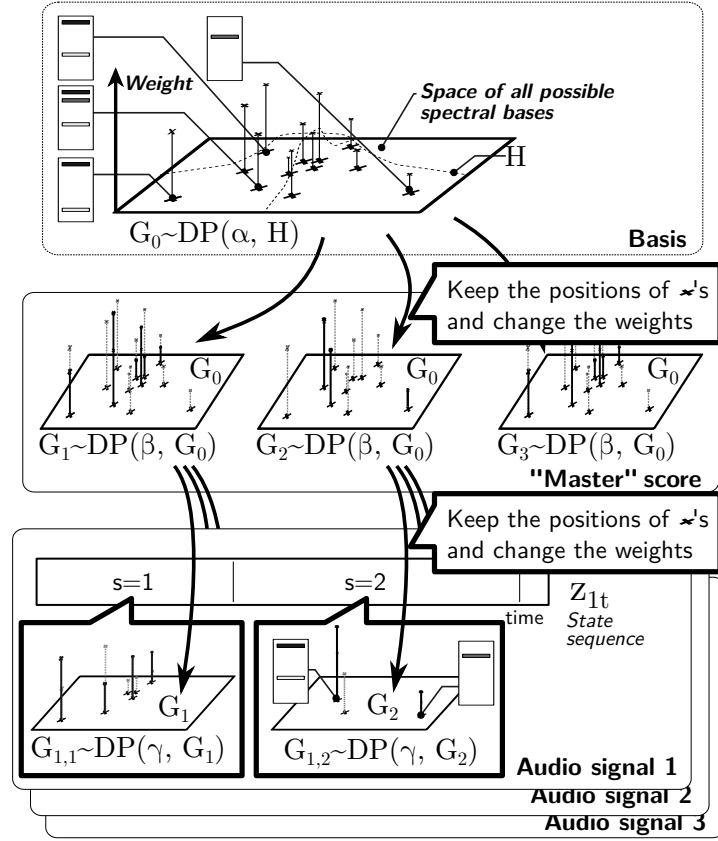


Figure 5.4: Illustration of the proposed three-layer HDP-HSMM model. Each 3D plot represents a weighted mixture of spectral basis. The lower-level DP's inherit the atoms from the upper-level DP's, while making the mixture more concentrated.

Thus, a countable number of histograms over the frequency axis is sampled, and each histogram is assigned a weight, sparseness of which is controlled by α . The weights exhibit a property that the more a basis is used, the more likely it is for the basis to be used again in another observation. This induces G_0 to have a few atoms with non-negligible associated weight, providing a compact representation of the observed θ 's whose number of atoms with non-negligible grows with the complexity of x .

Fig. 5.5 illustrates an example of the maximum a posteriori (MAP) estimate of the samples drawn from G_0 and the weights, sorted in descending order. Note that most bases have dark bands at $\log_2 n$ ($n \geq 1$) octaves plus some offset, meaning that they represent monophonic pitched sounds. The most frequently occurring basis (basis 0) describes noisy components that could not be described the harmonic sounds. Note that only a few basis (about 15), are used and remaining basis are rarely used, if at all. This suggests that G_0 represents a compact set of pitched sounds that describes the notes that appear in a piece.

Next, the master music score is modeled as a sequence of N states, each state emitting a subset of the atoms in G_0 . Each state n of the master music score is associated with a subset of G_0 , denoted G_n . Each signal in state n is allowed to emit a subset of

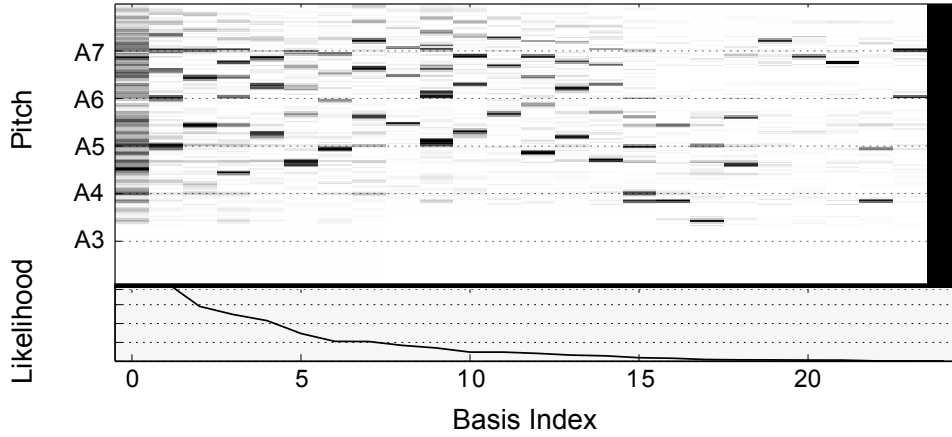


Figure 5.5: MAP estimate of the basis $\mathbf{g}$ (top; color in logarithmic scale, larger values are darker) and how frequently each basis is used (bottom).

the atoms (*i.e.*, spikes in Fig. 5.4) in G_n . G_n is assumed to be drawn from a DP with concentration parameter β and base measure G_0 :

$$G_n \sim \text{DP}(\beta, G_0). \quad (5.5)$$

Intuitively, G_n correspond to the notated notes in the master music score, which tends to emit a subset of the atoms in G_0 .

To ensure that every signal plays with the same ordering of states in the master score, the sequence that signal d traverses over the master music score $\mathbf{s}_{d,1 \dots T_d}$ is expressed as a N -state LRHSMM:

$$\mathbf{s}_{d,1 \dots T_d} \sim \text{LRHSMM}(\mathbf{l}). \quad (5.6)$$

Here, $\text{LRHSMM}(\mathbf{l})$ denotes a left-to-right HSMM with state durations $\mathbf{l} = \{l_1 \dots l_N\}$. Specifically, the probability that $\mathbf{l}$ remains in state n for l frames is assumed to be drawn from $p_n(l_n) = \mathcal{N}(l_n | \mu_n, c\mu_n^2)$. The state sequence is furthermore constrained such that $\mathbf{s}_{d11} = 1$ and $\mathbf{s}_{dT_d1} = N$ (*i.e.*, $\mathbf{s}_{d111} = 1/L$ and $\mathbf{s}_{dT_dN1} = 1$ in 1-of- N notation) for all d , meaning that every signal is constrained to start at the same state and end at the same state. By definition, $\mathbf{s}$ is constrained such that state n may only switch to $n+1$ after staying in it for some duration l_n . Note that a left-to-right structure ignores structural discrepancies among signals, such as repeats and cuts.

To perform inference over an HSMM, the semi-Markov model is replaced with an equivalent HMM. Namely, $\mathbf{s}$ is considered as an HMM state sequence that traverses over an augmented state space $(n, l) \in [1 \dots N] \times [1 \dots L]$, where n is the state as described in the HSMM, l is the number of frames remaining in state n , and L is the maximum allowed duration at any state. Then, the $\mathbf{s}$ transitions in a way such that (1) for $l > 0$, state (n, l) always transitions to $(n, l-1)$, (2) for $l = 0$, state $(n, 0)$ transitions to $(n+1, l')$ with probability $p_{n+1}(l')$. A set of other illegal transitions from state (n, l)

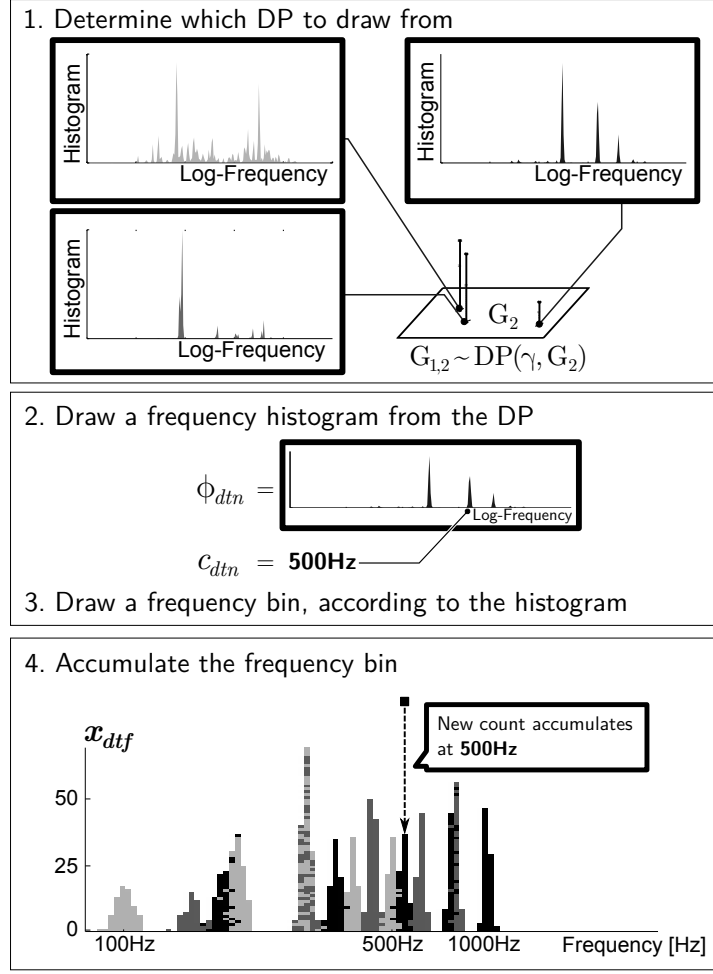


Figure 5.6: How the CQT spectrum is created in the proposed model.

to (n', l') is denoted as $\mathcal{S}_{\text{illeg.}} = \{(n, l, n', l')\}$.

Finally, the model considers how each audio signal renders the music score. At each state n in the score, each audio signal chooses a subset of G_n ; for example, an audio signal of a violin solo that plays a duet consisting of the violin and the piano emits only the atoms corresponding to the violin part. Therefore, the actual atoms emitted by signal d at state n , $G_{d,n}$ is expressed as a draw from a DP with a concentration parameter γ and a base measure G_n :

$$G_{d,n} \sim \text{DP}(\gamma, G_n). \quad (5.7)$$

The proposed model is different from infinite HMM [8], which also uses the HDP. First, the proposed model applies for a finite number of states. Since forced alignment requires known terminal states and ordering of states, the proposal method necessarily uses a finite number of states. Second, the proposed model is hierarchical with respect to signals as well as states. Therefore, the model can share not only basis

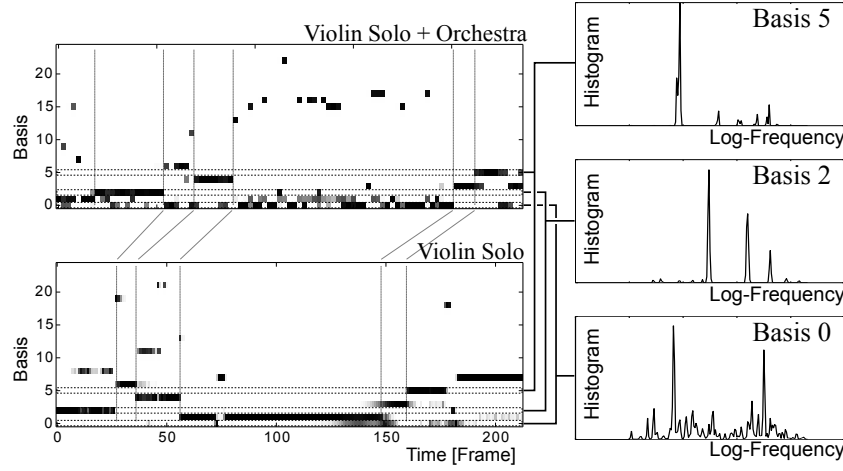


Figure 5.7: An excerpt of the expectation of the spectral basis activation, and the histogram associated with some of the basis (logarithmic scale). The alignment is manually annotated (dashed) for the sake of clarity.

among states, but also basis in a chord at a given state among different signals.

Modeling the observed signals

The proposed method interprets x_{dtf} as a count data – *i.e.*, the number of times a time-frequency bin (t, f) was observed for the d th signal – and identifies the spectral basis that generated each count of the histogram. To represent x as a count data, let us introduce a frequency indicator variable c_{dte} that takes on discrete values between 1 and F , where c is defined for $c \in [1, 2, \dots, \sum_f x_{dtf}]$, and such that $\sum_c \delta(c_{dte}, f) = x_{dtf}$. Each element in c_{dte} denotes a unique count in the histogram at signal d at time t , and indicates to which frequency bin the particular count was accumulated. Fig. 5.6 illustrates the generative process of the spectrum, which shall be elaborated.

Each count c_{dte} is assumed to originate from one of countable spectral basis, each of which describes the likelihood of emitting a count data from a given frequency:

$$c_{dte} \sim \text{Categorical}(\theta_{dte}). \quad (5.8)$$

Here, $\text{Categorical}(\theta)$ is a categorical distribution with the likelihood of observing frequency f given as θ_{dte} ; thus, θ resembles a normalized CQT spectrum of sounds comprising x_{dtf} . Recall that the kind of spectral basis θ is defined by the three-level HDP-HSMM model described earlier. Thus, at time t in signal d , the state sequence θ is referenced to find the position in the master score θ_{dt} . Then, the spectral atom θ_{dte} is generated from the corresponding DP inside the HDPs:

$$\theta_{dte} \sim G_{d,s_{dt}}. \quad (5.9)$$

Since the atoms of $G_{d,s_{dt}}$ should be a categorical distribution and since the domain

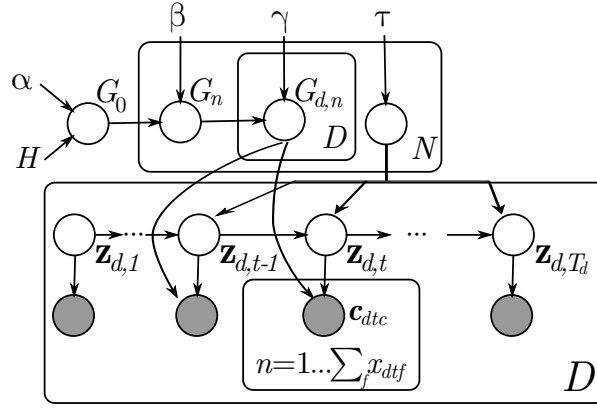


Figure 5.8: Graphical model of the proposed method.

of G_{dn} is the same as G_0 , the base measure of G_0 , or H , should be defined over all possible categorical distributions. To this end, H is chosen to be a Dirichlet distribution of dimension F with hyperparameter g_0 , $\text{Dir}(g_0)$. The Dirichlet distribution is used because it is conjugate to the categorical distribution, which simplifies inference. Further extensions might be possible; for example, one may use a distribution over harmonically-constrained categorical distributions [138].

An excerpt of the MAP estimate of the proposed method is illustrated in Fig. 5.7, when aligning a professional violin concerto recording to a recording of the solo violin part of an end-user. Here, the proposed method inferred one basis for a “summary” of orchestral parts (Basis 0), and other basis corresponding to a single pitched note, presumably that of the solo violin (Basis 2 and so on). Thanks to the DP, only a few components are used, yielding in a compact representation of music audio signals.

With this kind of model, one could align different musical audio by finding the state sequence of the master music score for each signal, as to maximize the posterior estimate of s_{dtc} . Since HDP seeks a parsimonious representation, the estimated master score tends to maximize the degree to which basis are shared among each signal (as mandated by G_n). This is possible because the input spectrogram is decomposed into spectral bases, such that the total number of spectral bases used is also parsimonious (as mandated by G_0). The proposed method is summarized as a graphical model in Fig. 5.8.

Reflecting back on the formulation in Equation 2.24, the proposed method can be seen as generating z as a sequence of iid variables, expressing θ using the three-level hierarchical DP, and expressing s using a set of left-to-right HSMMs.

5.2.3 Re-writing as a conjugate model

To compute the alignment, the posterior distribution of the previous model needs to be inferred, but it is difficult because the current HDP representation is nonconjugate. To make the proposed model amenable to closed-form Bayesian inference, it is rewrit-

ten using Sethuraman's stick-breaking construction [130], making the model fully conjugate. The basic idea is to use the stick-breaking representation (Equation 2.11 and Equation 2.12 in the appendix) to represent all three DPs. Hereon, the latent variables are represented as 1-of- K variables when the discrete latent variables are not shown in boldface, *e.g.*, $\mathbf{s}_{dt} = n'$ is denoted as $s_{dtn} = 1$ for $n = n'$ and 0 otherwise.

First, consider rewriting G_0 . To this end, let us draw $\tilde{I} \rightarrow \infty$ atoms from the base measure H using the stick-breaking process (SBP). Specifically, $\mathbf{g}_i$ is drawn from the following distribution, with $\mathbf{h}^{(1)} \sim \text{GEM}(\alpha)$:

$$\mathbf{g}_i \sim \text{Dir}(\mathbf{g}_0), \quad (5.10)$$

$$\mathbf{h}^{(1)} \sim \text{GEM}(\alpha), \quad (5.11)$$

where $\text{GEM}(\alpha)$ translates to the following:

$$\xi_i^{(1)} \sim \text{Beta}(1, \alpha); \mathbf{h}_i^{(1)} = \xi_i^{(1)} \prod_{i'=1}^{i-1} (1 - \xi_{i'}^{(1)}). \quad (5.12)$$

In other words, G_0 is described in terms of the atoms (spectral basis) $\mathbf{g}_i$ and its associated weight $\mathbf{h}^{(1)}$.

Next, G_n is rewritten by an equivalent stick-breaking representation as done above. To this end, let us draw $\tilde{J} \rightarrow \infty$ indicator variables $\mathbf{u}_{nj}$, j th of which indicates to which global atom the j th atom within a chord in state s refers. Specifically:

$$\mathbf{u}_{nj} \sim \text{Categorical}(\mathbf{h}^{(1)}), \quad (5.13)$$

$$\mathbf{h}_n^{(2)} \sim \text{GEM}(\beta), \quad (5.14)$$

where $\text{GEM}(\beta)$ translates to the following:

$$\xi_{nj}^{(2)} \sim \text{Beta}(1, \beta); \mathbf{h}_{nj}^{(2)} = \xi_{nj}^{(2)} \prod_{j'=1}^{j-1} (1 - \xi_{nj'}^{(2)}). \quad (5.15)$$

In other words, G_n is realized by drawing $\tilde{J}$ indices ranging from 1 to $\tilde{I}$ according to $w_i^{(g)}$, the length of the stick broken for the top-level DP. Then, each draw is stored by setting $\mathbf{u}_{nj} := i$. Finally, for each j , a categorical weight is associated with it.

Next, G_{dn} is rewritten similarly. Namely, $\tilde{K} \rightarrow \infty$ indicator variables $\mathbf{v}_{dnk}$ are drawn, k th of which indicates which atom of G_n the k th atom of G_{dn} refers to:

$$\mathbf{v}_{dnk} \sim \text{Categorical}(\mathbf{h}_n^{(2)}), \quad (5.16)$$

$$\mathbf{h}_{dn}^{(3)} \sim \text{GEM}(\gamma), \quad (5.17)$$

where $\text{GEM}(\gamma)$ translates to the following:

$$\xi_{dnk}^{(3)} \sim \text{Beta}(1, \gamma); h_{dnk}^{(3)} = \xi_{dnk}^{(3)} \prod_{k'=1}^{k-1} (1 - \xi_{dnk'}^{(3)}). \quad (5.18)$$

In other words, $\tilde{K}$ indices are drawn according to $h_n^{(2)}$, and each draw is associated with a categorical likelihood generated from the SBP.

Next, each count c_{dtc} is modeled as having originated from one of k atoms of G_{dn} , given s_{dt} . To this end, a latent variable w is introduced, which indicates the atoms of G_{dn} that generated c_{dtc} , given n :

$$w_{dct}|s_{dt} \sim \text{Categorical}(h_{d,s_{dt}}^{(3)}). \quad (5.19)$$

Then, the observation likelihood can be modeled as follows:

$$c_{dtc} \sim \text{Categorical}[g_{u_{s_{dt}}, v_{d,n}, w_{dctn}}]. \quad (5.20)$$

Comparing with Equation 5.8, note that the variable θ_{dtc} has been replaced with $g_{u_{s_{dt}}, v_{d,n}, w_{dctn}}$. This means that θ is constrained to be one of draws from G_0 , and the hierarchy of DP's used to arrive at s , explored in Section 5.2.2 is realized by traversing the index associated with an element of each layer's DP.

Then, the complete log-joint likelihood is given as follows, up to a constant normalization factor:

$$\begin{aligned} & \sum_{i,j,k,n,l,f,d,c,t} c_{dtcf} u_{nji} v_{dnkj} w_{dctk} s_{dtnl} \log g_{if} + \sum_i \log \text{Dir}(g_i | g_0) \\ & + \sum_{t=1, n=2, l=1, d=1}^{T-1, N, L, D} s_{d,t,n-1,1} s_{dtnl} \log \mathcal{N}(l | \mu_n, c \mu_n^2) + \sum_{t,d,(n,l,n',l') \in \mathcal{S}_{\text{illeg.}}} s_{d,t-1,n,1} s_{dtn'l'} \log 0 \\ & + \sum_{n,i,j} u_{nji} \log h_i^{(1)} + \log \text{GEM}(h^{(1)} | \alpha) + \sum_{d,n,j,k} v_{dnkj} \log h_{nj}^{(2)} + \sum_n \log \text{GEM}(h_n^{(2)} | \beta) \\ & + \sum_{t,c,f,k,d,n} w_{dctk} s_{dtn} \log h_{dnk}^{(3)} + \sum_{d,n} \log \text{GEM}(h_{dn}^{(3)} | \gamma). \end{aligned} \quad (5.21)$$

Now this is fully conjugate, so standard variational Bayesian method [7] and truncation approximation to Sethuraman's stick-breaking representation [130] can be used to arrive at an approximate posterior distribution.

5.3 Inference

Having defined the model, the goal is to determine the posterior distribution of the model. To this end, this section derives an inference scheme using variational Bayesian

[7] (VB) method. The conjugacy of the stick-breaking construction used in the proposed method allows for an analytically tractable and simple inference. The inference method assumes that the posterior is decomposed in the following form:

$$q(\mathbf{s}, \mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{g}, h^{(1,2,3)}) = \prod_d q(\mathbf{s}_{d,1\dots T_d}) \prod_i q(\mathbf{g}_i) \\ \times \prod_{n,j} q(\mathbf{u}_{nj}) \prod_{d,n,k} q(\mathbf{v}_{dnk}) \prod_{d,c,f,n} q(\mathbf{w}_{dct} | \mathbf{s}_{dt}) q(h^{(1)}) \prod_n q(h_n^{(2)}) \prod_{d,n} q(h_{dn}^{(3)}). \quad (5.22)$$

Moreover, for some I , J , and K , it is assumed that $q(\xi_I^{(1)}) = 1$, $q(\xi_J^{(2)}(n)) = 1$ and $q(\xi_K^{(3)}(d, n)) = 1$. This means that coefficients greater than I , J and K , respectively, will take on values 0, and can be ignored in the inference. Then, update each factor is updated iteratively until convergence.

First, let us update the latent variables $q(\mathbf{u})$, $q(\mathbf{v})$ and $q(\mathbf{w})$. Standard VB technique yields in the following:

$$\log q(\mathbf{u}_{nji}) \propto \sum_{f,k,d,c,t} x_{dtf} \omega_{dnkj} \chi_{dtfnk} v_{dtn} \langle \log g_{if} \rangle + \langle \log h_i^{(1)} \rangle, \quad (5.23)$$

$$\log q(\mathbf{v}_{dnkj}) \propto \sum_{i,f,c,t} x_{dtf} v_{nji} \chi_{dtfnk} v_{dtn} \langle \log g_{if} \rangle + \langle \log h_{nj}^{(2)} \rangle, \quad (5.24)$$

$$\log q(\mathbf{w}_{dctk} | \mathbf{s}_{dt} = n) \propto \sum_{i,j,k,t} v_{nji} \omega_{dnkj} v_{dtn} \langle \log g_{i,c_{dct}} \rangle + \sum_t v_{dtn} \langle \log h_{dnk}^{(3)} \rangle, \quad (5.25)$$

where $\chi_{dtfnk} = \langle w_{dct} = k | \cup_l \mathbf{s}_{dt} = (n, l) \rangle$ for every c such that $c_{dct} = f$, $v_{nji} = \langle u_{nji} \rangle$, $\omega_{dnkj} = \langle v_{dnkj} \rangle$, and $v_{dtn} = \langle s_{dtn} \rangle$. The equations intuitively show that the log-expectations of the latent variables are generated by applying a soft mask to x_{dtf} as to assign appropriate atoms in each level of the hierarchy.

To update $\mathbf{s}$, the expectations may be computed efficiently using the forward-backward algorithm [108]. Specifically, the following unnormalized histogram is treated as the log emission likelihood of an HMM:

$$\log o_{dtn} = \sum_{f,k} x_{dtf} \chi_{dtfnk} \langle \log h_{dnk}^{(3)} \rangle + \sum_{f,i,j,k} v_{nji} \omega_{dnkj} \chi_{dtfnk} x_{dtf} \langle \log g_{if} \rangle. \quad (5.26)$$

Then, the forward variable is computed as follows:

$$\alpha_{dtnl} \propto o_{dtn} (\alpha_{d,t-1,n,l+1} + \alpha_{d,t-1,n-1,1} e^{\langle \log p_n(l | \mu_n) \rangle}), \quad (5.27)$$

where α_{dtnl} for $n < 0$ is set to 0, $\alpha_{d,t,n,L+1}$ is set to 0, and α_{d11l} is set to $o_{d1n} p_n(l | \mu)$.

Then, the backward variables is computed as follows:

$$\beta_{dtnl} = \begin{cases} \beta_{d,t+1,n,l-1} o_{d,t+1,n} & l \leq \mathbb{L} \\ \beta_{d,t+1,n+1} \sum_{l'} o_{d,t+1,n+1} e^{\langle \log p_n(l'|\mu_{n+1}) \rangle} & l = 1. \end{cases} \quad (5.28)$$

Finally, the state expectation can be computed as $\nu_{dtn} \propto \sum_l \alpha_{dtnl} \beta_{dtnl}$. Moreover, $\iota_{dtnl} = \langle s_{d,t-1,n-1,1} s_{dtnl} \rangle$, the expectation of state transition and its duration can be computed as follows:

$$\iota_{dtnl} \propto \alpha_{d,t-1,n-1,1} e^{\langle \log p_n(l|\mu_n) \rangle} o_{d,t-1,n-1} \beta_{dtnl}, \quad (5.29)$$

where the normalization constant accounts for both the transition from augmented state $(n, l+1)$ to (n, l) , and $(n, 1)$ to $(n+1, l)$.

Next, the parameters are optimized. Optimizing w.r.t. $q(g(i))$ yields the following, with $\zeta_{dnfk} = \sum_t \nu_{dtn} x_{dtf} \chi_{dtfnk}$:

$$g_i \sim \text{Dir}(g_0 + \bar{g}_i), \quad (5.30)$$

$$\bar{g}_{if} = \sum_{j,n,k,d,c} v_{nji} \omega_{dnkj} \zeta_{dfnk}, \quad (5.31)$$

$$\xi_i^{(1)} \sim \text{Beta}\left(1 + v_{::i}, \alpha + \sum_{n,j,i'>i} v_{::i'}\right), \quad (5.32)$$

$$\xi_{nj}^{(2)} \sim \text{Beta}\left(1 + \omega_{:n:j}, \beta + \sum_{j'>j} \omega_{:n:j'}\right), \quad (5.33)$$

$$\xi_{dnk}^{(3)} \sim \text{Beta}\left(1 + \zeta_{d:nk}, \gamma + \sum_{k'>k} \zeta_{d:nk'}\right), \quad (5.34)$$

$$\mu_n = \sum_l l_{t::nl}. \quad (5.35)$$

After the algorithm has converged, audio d can be aligned to d' by finding the MAP state of d , $\hat{n}$, and finding the time at signal d' such that its MAP state is $\hat{n}$.

5.4 Evaluation

This section evaluates the capability of the proposed method to align musical audio signals that contains common parts, and compare it with existing alignment methods that implicitly assume that two audio signals play identical note sequences.

5.4.1 Experimental Setup

The BACH10 dataset [29] was used to evaluate the proposed method. BACH10 dataset contains ten instrumental recordings of four-voice Bach chorales played using a bas-

soon, saxophone, clarinet and a violin. The four parts are called “Soprano,” “Alto,” “Tenor” and “Bass,” respectively, in the decreasing order of the average pitch. Each part is recorded in a synchronized and isolated manner, allowing the Bach chorales to be synthesized with an arbitrary combination of parts. It also contains standard MIDI file (SMF) to the ten chorales, additional set of audio with different part combinations to be generated, each with different timbre, dynamics, and tempo trajectory from the instrumental recording. Moreover, it contains a ground-truth SMF-to-audio alignment, allowing the absolute alignment error at note onsets to be evaluated.

For the SMFs of the ten chorales in the BACH10 dataset, all possible combination of parts were synthesized, resulting in 15 recordings for each chorale. The same was done for real audio part recordings of BACH10 audio data. Next, pairs of audio with parts played in common were aligned. This resulted in 175 pairs to be aligned for each chorale, totaling in 1750 data, or about 20 hours of unique pairs of audio signals. Then, for each pair, the alignment extracted was compared using three methods: (1) the proposed method, (2) MATCH toolchest [27], and (3) DTW based on minimizing the net cosine distance between two CQT spectrograms, warped in time. MATCH is an effective alignment method with a publicly available implementation, and is convenient as a reproducible reference implementation.

For the proposed method, the parameters for the HDP were set to $\alpha = 1.0$, $\beta = 0.1$, and $\gamma = 0.01$. This encourages the lower-level DP’s to be much more parsimonious than the upper-level. The prior for the spectral basis was set to $g_0 = 0.01$, which encourages the basis to be sparse. The CQT was evaluated at semitone interval, using sampling frequency of 44.1 kHz and analysis framerate of 50 frames per second. The number of states N was set to $0.1\min(T_d)$, and the maximum number of frames occupied by any one state, L , was set to 50.

DTW based on minimization of CQT cosine distance (“CQT-DTW”) was chosen as a baseline that represents a popular approach to offline audio alignment. Since existing formulations typically incorporate not only the feature itself but the half-wave rectified first-order difference as well [34], the DTW was evaluated over both CQT spectrogram and its half-wave rectified first-order difference. In the DTW method, the path constraint is set such that position i in the first audio and j in the second audio (denoted (i, j)) may traverse to $(i + 1, j + 1)$, $(i + 1, j)$ or $(i, j + 1)$, a popular choice in DTW-based alignment [58]. Local constraints were not used.

In addition to DTW based on the feature used by the proposed method, DTW based on the chroma vector was evaluated as well. The result is not presented because it had consistently underperformed CQT-DTW. This owes to the invariance of the chroma to pitch transposition by octaves: the octave information – *e.g.*, whether one phrase is played in the register of a bass or that of a soprano – is a valuable cue when performing subset music alignment.

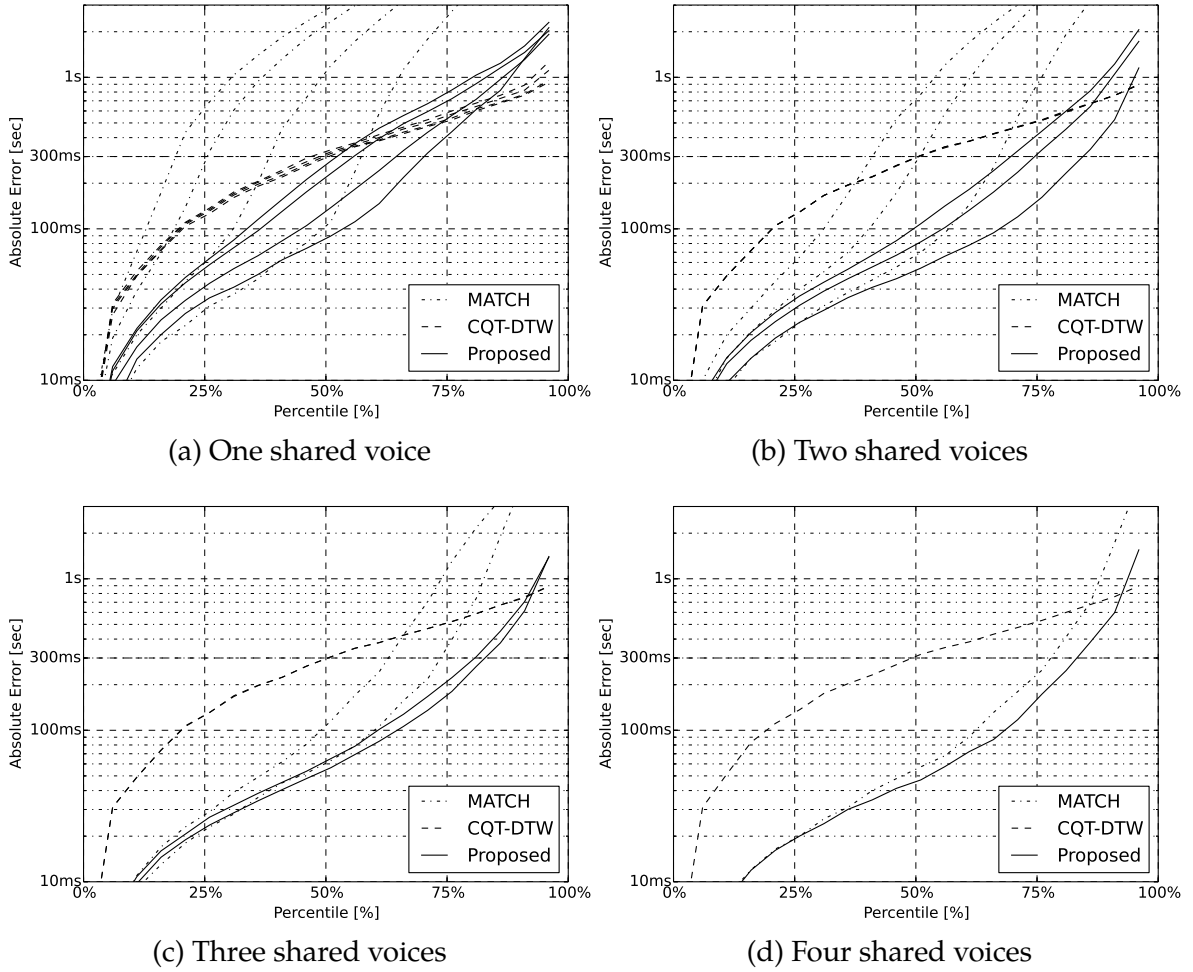


Figure 5.9: Absolute alignment error percentile as a function of number of simultaneously shared notes of a four-part ensemble. Different shades of gray indicate total number of simultaneously sounding distinct notes, D , where $D = 0$ (no note unique to a recording) is black, and the shade becomes lighter as D increases.

5.4.2 Experimental Result

Fig. 5.9 shows the percentile of alignment absolute error, evaluated with different number of overlapping and disjoint parts. Comparing MATCH and the proposed method, the proposed method is much more robust to differences in the number of disjoint voices. For example, in the case of one part played in common (Fig. 5.9a), the median (50 percentile) error of MATCH varies by order of seconds with different number of disjoint parts, whereas the proposed method varies only by about 300 ms. It also performs comparably or better than MATCH in the context of conventional audio-to-audio alignment ($D = 0$). When comparing CQT-DTW and the proposed method, the proposed method performs much better than CQT-DTW at most percentiles. CQT-DTW seems to perform better at higher percentile, suggesting that the proposed method has outliers that behave poorly. This presumably owes to the local

minima introduced by the local search of the variational Bayesian algorithm, exacerbated by the introduction of latent variables for Sethuraman’s stick-breaking construction. In this evaluation, the path generated by DTW tended to miss onsets, despite incorporating delta features. This suggests that semitone-feature is good enough for roughly aligning part mixtures, but the delta-feature cannot capture the note onsets. The median alignment error was about 200 ms, making it useful in rough music performance analysis tasks.

5.5 Summary

This chapter presented *subset music alignment*, a method for aligning music scores and/or audio signals that *contain* a common piece of music. This is unlike existing audio alignment or score alignment formulations that assume two signals represent identical music.

Subset music alignment opens door to alignment problems that links recording of a single part and the full parts, allowing a user to compare professional recordings or a concerto to a solo recording of an end-user, or separating a complex mixture of sound with off-sync humming of the end-user.

To attack the problem of jointly finding parts that are played in common and aligning the signals based on the common parts, a three-layer HDP-LRHMM model that emits a CQT was formulated. Evaluation demonstrated the robustness of the proposed method against parts that are played mutually exclusively among different music audio, attaining median alignment error of about 200 ms.

Chapter 6

Bayesian Dereverberation for Music Alignment

The key theme in music synchronization is to arrive at representations that are invariant to sources of variations in playing. An often-overlooked but critical aspect in music synchronization is that of room acoustics. This is important because an audio signal, in the real world, is a mixture of various kinds of source signals, marred by different kinds of reverberation. Reverberation is typically characterized by the *early reflection* component and the *late reverberation* component: the former refers to components of impulse responses that contribute to coloration of the spectrum, and the latter refers to components that add decaying sound. Many studies have tackled the problem of suppressing the reverberation through dereverberation methods, methods for recovering the source signal (“dry” signal) that has driven the reverberant acoustic environment.

In context of alignment methods, dereverberation is important because alignment is based on matching local spectral time-slices of different audio renditions. Thus, it is vitally important that each frame of the audio renditions represent the sound produced by the performer. In many musical audio signals, however, the hall acoustics smears the musician’s performance. Therefore, achieving robustness to different acoustics is necessary. Dereverberation may be helpful in this kind of situation, by making the input audio more representative of the music score. Other MIR tasks may benefit from dereverberation as well.

This chapter presents a dereverberation method that suppresses the late reverberation of a variety of audio signals, including music and speech, recorded under a variety of room acoustics. Such a goal mandates the following criteria for the design of dereverberation algorithm:

1. The proposed method should operate using a single-channel input. Many audio signals are recorded as monaural tracks, *e.g.*, digital remasters of early LP records or an audio signal recorded on an end-user’s smartphone.

2. The proposed method should be robust to minor variations in the spatial placement of sources that drive the room acoustics, for musical audio has many sources that are close to each other, meaning that each source is convolved by impulse responses with similar magnitude response but different phase response.
3. The proposed method should be robust to various source signals, since musical audio is a mixture of a wide variety of musical instruments, ranging from a whistle to a snare drum.
4. The user should not have to specify the complexity of the reverberation in advance. This means that the method should be robust to various room acoustics, since the room acoustics of many musical audio cannot be known in advance: the acoustics can range from dead-dry (*e.g.*, close-microphone), to highly reverberant (*e.g.*, a cathedral).

In order to achieve (1) and (2) simultaneously, the proposed method models a reverberant signal on the power spectrogram domain, since a method that operates in the power spectrum domain is relatively insensitive to speaker movements [66].

In order to tackle (3), the constraints on the source signal should be minimal. To this end, the sparsity of the source signal is expressed, similar to [66,139]. Here, “sparsity” means that the mode of the probability density function of each time-frequency bin of the power spectrogram is zero.

Finally, (4) is realized by estimating the reverberation complexity in a data-driven manner. To this end, the Dirichlet process (DP) is used. Specifically, the DP is used to express a reverberant power spectrogram as an infinite mixture of count data, with the mixture component derived from a non-negative auto-regressive (AR) process defined for each frequency bin. Modeling the reverberation as a mixture of count data allows us to use the DP to model the AR coefficients, such that the AR model complexity grows with the complexity of the observed data. By using the DP, the model would introduce additional AR model prediction coefficient when it is sensible to do so. Thus, it infers AR model coefficients that are compact yet expressive enough for the given audio signal.

Section 2.3.2 introduces the readers to existing studies. The proposed method is introduced in Section 6.1, and an inference algorithm in Section 6.2. The proposed method is evaluated in Section 6.3.

6.1 Formulation

This chapter aims to design a dereverberation method that works with various kinds of music audio signals. It models the power spectrogram as a histogram count data, and incorporate prior information that guides the reverberation model towards a

good posterior distribution. Namely, it incorporates a sparse prior on the source signal, and uses the DP to model the AR coefficients, allowing the model order of AR to increase without bound as the data mandates.

The proposed method develops on the exponential decaying model (AR(1) model) of reverberation in the power spectrogram domain [76], and incorporate ideas from methods formulated in the complex spectrum representation [141] to arrive at a higher order AR model – AR(∞), in fact – over the power spectrogram. The infinite limit of the AR order is captured using the DP, and allows the “effective” number of active components in the AR model to be tuned to the input audio signal.

With these in mind, the main contribution of this chapter is the extension of the AR(1) model of reverberation in the power spectrum domain [76] to infinite order. First, the higher order AR model is justified, and the foundation of the statistical model is laid. Next, a novel model of reverberant audio that uses the DP to express the reverberation coefficients is introduced. Finally, since the proposed method cannot be solved using standard techniques [7], a novel posterior inference algorithm based on minorization maximization is presented.

6.1.1 Signal model

In order to model the power spectrogram as a mixture model, it is necessary to first define what kind of components would comprise the mixture. To this end, the multi-channel linear prediction model (MCLP) [141] is extended to model the observed power spectrogram as a mixture of a source signal and previous observations.

Let $\bar{y}_{ft} \in \mathbb{C}^{F \times T}$ be the short-time Fourier transform (STFT) of the observation, and $\bar{s}_{ft} \in \mathbb{C}^{F \times T}$ be that of the source signal. Here, F and T denote the number of frequency bins and the number of frames, respectively. Then, under reasonable assumptions [141], reverberation can be modeled as follows, for some time-invariant filter of order I , $\bar{h}_{ft} \in \mathbb{C}^{F \times I}$:

$$\bar{y}_{ft} = \bar{s}_{ft} + \sum_{i=1}^I \bar{h}_{fi} \bar{y}_{f,t-i}. \quad (6.1)$$

This kind of formulation is known as MCLP, which is originally formulated as a multichannel problem, but is known to work with single-channel audio as well [141]. This formulation ignores early reflection, and takes into account of the late reverberation, whose impulse response is greater than the frame length of the STFT.

The proposed method extends MCLP to the power domain. Note that for all (f, i) , $\bar{h}_{fi}$ is zero-mean because its expectation should be invariant to rotations in the complex plane. In other words, where one defines the reverberation to begin should not affect the distribution inherent to the reverberation. Now, let us ignore the short-term correlation typically present in h and assume that for each f , $\{\bar{h}_{f1}, \dots, \bar{h}_{fI}\}$ are mutually independent. Moreover, assume $\bar{s}_{ft}$ and $\bar{h}_{ft}$ are independent for all $t \in [1, T]$. Then, in terms of expectation, the following model of the power spectrogram may be

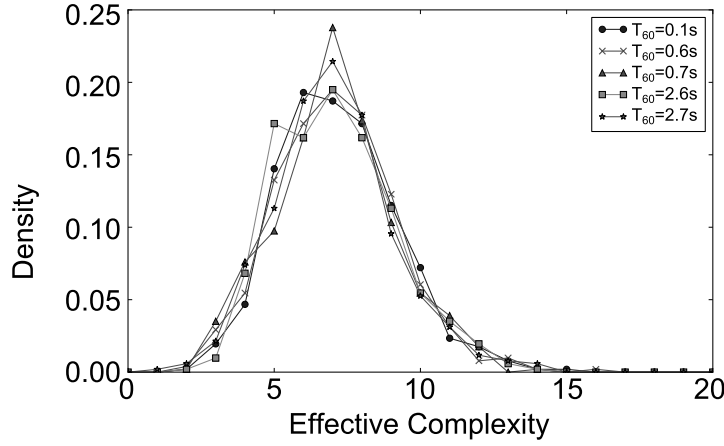


Figure 6.1: Distribution of the effective complexity.

formulated, where $|\cdot|$ denotes the l_2 -norm:

$$|\bar{y}_{ft}|^2 = |\bar{s}_{ft}|^2 + \sum_{i=1}^I |\bar{h}_{ft}|^2 |\bar{y}_{f,t-i}|^2. \quad (6.2)$$

Hereon, $|\bar{y}_{ft}|^2$ is denoted y_{ft} , $|\bar{s}_{ft}|^2$ is denoted s_{ft} , and $|\bar{h}_{fi}|^2$ is denoted h_{fi} .

This model can be thought of as a generalization of the exponentially decaying model of reverberation used in [76], where $I = 1$ reduces Equation 6.2 to an exponential decay model. Note that increasing I makes the reverberation model more expressive but complicates the inference due to the increased number of free parameters. Hence, it is important to choose an I that is compact yet expressive enough for the observed signal.

An investigation suggests that I is best modeled by some finite number that is dependent on the nature of reverberation. Specifically, the AR coefficients of the impulse responses used in Section 6.3 was determined. A frame length of 1024 samples at 16kHz sampling rate with no overlap was used to compute the power spectrogram of each impulse response. Then, for each impulse response, each frequency bin of the power spectrogram was modeled as a non-negative AR(200) process. The non-negative AR coefficients were estimated using non-negative least squares (NNLS). For each frequency bin, the coefficients was sorted in descending order, evaluated the cumulative sum, and determined the smallest number for which the cumulative sum exceeds 99% of the total sum. This is called the “effective complexity” at a given frequency bin, in that a significant portion of reverberation is expressed using this number of components.

Fig. 6.1 shows the density of the effective complexity for different reverberation time (T_{60} , the time it takes for the impulse response of the reverberation to attenuate by 60dB) associated with each impulse response. This figure suggests that the effective complexity is a random variable that is relatively invariant to the kind of reverberation

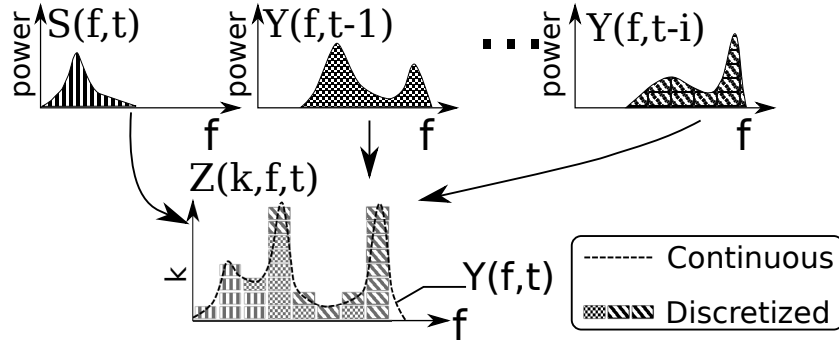


Figure 6.2: Conceptual diagram of the proposed statistical model. Power spectrum is treated as a histogram, and each count is associated with a latent component that created it.

(the reverberation time of the impulse responses ranges from 0 to 2.7 seconds). From this, the figure also suggests that an exponential decay (non-negative AR(1)) model such as [76] may merit from a higher order non-negative AR process.

With this signal model, the fraction of y_{ft} generated by the source s_{ft} , r_{ft0} , is given as follows:

$$r_{ft0} = \frac{s_{ft}}{s_{ft} + \sum_{i=1}^I h_{fi} y_{f,t-i}}. \quad (6.3)$$

Moreover, the fraction of y_{ft} generated by the contribution from frame $t - i$, r_{fti} , is given as follows:

$$r_{fti} = \frac{h_{fi} y_{f,t-i}}{s_{ft} + \sum_{i=1}^I h_{fi} y_{f,t-i}}. \quad (6.4)$$

These form the basis of the statistical model, which will be discussed next.

6.1.2 Statistical model

The proposed method treats a discretized version of y_{ft} as a histogram count data and formulates a latent variable model that assigns a latent component to each count of the histogram. Such an interpretation are often used either explicitly [13] or implicitly [24] in source separation literatures. It is useful because treating the spectrum as a histogram allows one to apply techniques used in mixture modeling tasks, including the DP.

In order to formulate a latent variable model, the method discretizes y_{ft} and sets $y_{ft} := \text{round}(\nu y_{ft})$ for some $\nu > 0$. y_{ft} is then treated as the number of times a time-frequency bin (f, t) was observed. According to Equation 6.2, each of y_{ft} counts in time-frequency bin (f, t) is generated by either the source s_{ft} with a likelihood of Equation 6.3 or previous observation $y_{f,t-i}$ with a likelihood of Equation 6.4. The independence of Equation 6.3 and Equation 6.4 on the actual count index shows that the counts are independently and identically distributed given (f, t) . This concept is

illustrated in Fig. 6.2.

To record the originating component for the k th count in time-frequency bin (f, t) , let us introduce z_{kfti} , a one-of- $I+1$ binary variable¹. “1-of- $I+1$ ” means that z is set such that given (k, f, t) , exactly one element of $I + 1$ choices is 1 and the rest is 0. $z_{kft0} = 1$ indicates that the k th count in time-frequency bin (f, t) originated from the source s_{ft} . For $i \in [1, I]$, $z_{kfti} = 1$ indicates that k th count originated from observation from i frames ago, $y_{f,t-i}$. Note that for a given f and t , z_{kft} is defined up to the number of observation at (f, t) , that is, $k \in [1, y_{ft}]$.

Equation 6.3 and Equation 6.4 may be rewritten as follows, by introducing, for each f , $\beta_f \in [0, 1]$ and an I -simplex w_f :

$$\frac{1 - \beta_f}{\beta_f} = \sum_{i=1}^I h_{fi}, \quad (6.5)$$

$$w_{fi} = \frac{h_{fi}}{\sum_{i'=1}^I h_{fi'}}. \quad (6.6)$$

Together, they give the following:

$$r_{ft0} = \frac{\beta_f s_{ft}}{n_{ft}}, \quad (6.7)$$

$$r_{fti} = \frac{(1 - \beta_f) w_{fi} y_{f,t-i}}{n_{ft}}, \quad (6.8)$$

where

$$n_{ft} = \beta_f s_{ft} + \sum_{i=1}^I (1 - \beta_f) w_{fi} y_{f,t-i}. \quad (6.9)$$

With this kind of reparametrization, β_f may be thought of as the source signal-to-reverberation ratio for frequency f : the source s_{ft} dominates the observation when $\beta_f = 1$, and the late reverberation dominates the observation when $\beta_f = 0$. w_{fi} may be thought of as the relative contribution of the observation from i frames before, for frequency f .

Equation 6.7 and Equation 6.8 respectively indicate the likelihoods that, for a given k, f and t , $z_{kft0} = 1$ and $z_{kfti} = 1$ for $i \in [1, I]$. With these in mind, the conditional

¹In this chapter, an index that is normalized to 1 is denoted using subscripts, e.g., $\sum_i z_{kfti} = 1$ for all (k, f, t) . An exception to this convention is the hyperparameters of the prior distribution, which by convention is denoted using a subscript zero. The exceptions to the hyperparameter notation are the variables z and ϕ , which are zero-indexed variables but are not hyperparameters.

posterior of z given y is described as follows:

$$p(z|y, s, \beta, w) = \prod_{f=1, t=1, k=1}^{F, T, y_{ft}} \left[\left(\frac{\beta_f s_{ft}}{n_{ft}} \right)^{z_{kft0}} \prod_{i=1}^I \left(\frac{(1 - \beta_f) w_{fi} y_{f, t-i}}{n_{ft}} \right)^{z_{kfti}} \right]. \quad (6.10)$$

Next, prior information on the parameters are incorporated to the proposed model. Prior information is essential because the maximum likelihood estimation of Equation 6.10, *i.e.*, maximizing Equation 6.10 w.r.t. s , z , β , and w , is susceptible to degenerate solutions. To see why, observe that the maximum value of Equation 6.10 is 1, for the terms to be exponentiated is at most 1 and the exponent is a binary variable. The maximum likelihood of 1 is attained under many degenerate conditions, such as (1) setting $\beta_f = 1$, (2) setting $\beta_f = 0$ and $w_{fi} = \delta(i - i')$ for some $(i, i') \in [1, I] \times [1, I]$ where $\delta(i)$ is the Kronecker delta, or (3) setting $\beta_f > 0$ and letting $s_{ft} \rightarrow \infty$. Therefore, it is critical to incorporate prior on both the source and the reverberation parameters.

Prior of w as a Dirichlet process

Recall that in Section 6.1.1, it was observed that I was a random quantity that is dependent on the frequency bin f . Therefore, the proposed model should convey the idea that the effective complexity grows without bound with the complexity of the observation. To this end, the DP is used, allowing a data-driven approach to determine the effective complexity.

With these in mind, let us apply the DP to model the complexity of reverberation, or the number of effective components in w . Specifically, the model assumes $I \rightarrow \infty$, and assumes that w_{fi} is an infinite-dimensional simplex drawn from the GEM distribution. That is, for $i \in [1, \infty]$, w_{fi} has a prior of the following form:

$$w_{fi} = \theta_{fi} \prod_{j=1}^{i-1} (1 - \theta_{fj}), \theta_{fi} \sim \text{Beta}(1, \alpha). \quad (6.11)$$

Therefore, reverberation is assumed to be a non-negative AR(∞) process, but only a finite number of coefficients contribute to the observation.

Prior of S

In order to prevent the degeneracy of Equation 6.10 by increasing s arbitrarily, a sparse prior on s is set, such that the mode of the distribution is 0. Specifically, the proposed model lets $s_{ft} \sim \text{Gam}(u_{ft}, v_{ft})$. By setting $u = 1$, the prior of S has a mode at zero, with an exponentially decaying likelihood proportional to $\exp(-s_{ft}/v_{ft})$. Moreover, the Gamma distribution is also the simplest distribution that allows for an analytically tractable inference scheme – namely, the sufficient statistics include $-s$ and $\log s$.

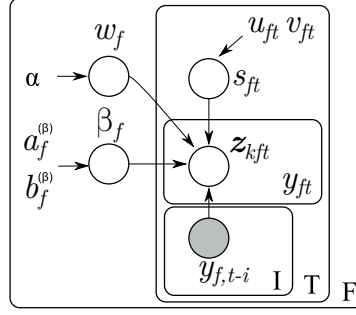


Figure 6.3: Graphical model of the proposed method.

The Gamma distribution is chosen because not only is it mathematically convenient but it is also an acceptable posterior distribution for both musical audio and speech signals. Power spectrograms of musical audio signals [29] and speech signals [45] were prepared, using conditions used in Table 6.1. Then, for each frequency bin, parameters to the Gamma distribution was estimated using maximum-likelihood estimation using 90% of the frames. Finally, Kolmogorov-Smirnov test was performed on the remaining 10% of the dataset to evaluate the goodness-of-fit. The null hypothesis that the underlying distribution of the power spectrogram is distributed as a Gamma was accepted 44% of the time for speech signals, and 67% of the time for musical audio. This shows that the Gamma distribution is an acceptable model of the posterior distribution.

Prior of β

The model lets $\beta_f \sim \text{Beta}(a_f^{(\beta)}, b_f^{(\beta)})$. By setting $a_f^{(\beta)} = b_f^{(\beta)} = 1$ for all f , β becomes non-informative, in that no assumption is made on the signal-to-reverberation ratio, and it is inferred solely from the data. Even though β may cause degenerate solutions as argued previously, the sparse prior of s and the prior of w are anticipated to “compete” with each other to yield in a meaningful posterior distribution.

With these in mind, the joint posterior pdf is given as follows, up to a constant:

$$\begin{aligned} & \log p(z, s, \beta, w | y, \alpha, a^{(\beta)}, b^{(\beta)}, u, v) \\ &= \sum_{f=1, t=1, k=1}^{F, T, y_{ft}} z_{kft0} \log(\beta_f s_{ft}) + \sum_{i=1}^{\infty} z_{kfti} \log((1 - \beta_f) w_{fi} y_{f,t-i}) - y_{ft} \log n_{ft} \\ & \quad + \log p(s_{ft} | u_{ft}, v_{ft}) + \log p(w_f | \alpha) + \log p(\beta_f | a_f^{(\beta)}, b_f^{(\beta)}). \end{aligned} \quad (6.12)$$

The graphical model is shown in Fig. 6.3. Once the posterior distribution is found, the dry signal may be recovered by multiplying the following time-frequency mask m_{ft}

to the power spectrogram:

$$m_{ft} = \frac{\sum_{k=1}^{y_{ft}} \langle z_{kft0} \rangle}{y_{ft}} = \langle z_{\hat{k}ft0} \rangle, \quad (6.13)$$

for any $\hat{k} \in [1, y_{ft}]$. Unless otherwise mentioned, the expectations are taken under an approximate posterior distribution introduced in Section 6.2.

6.2 Inference

In order to compute Equation 6.13, it is necessary to marginalize the posterior over all variables but z . In other words, it is necessary to find the expectation of z , weighed by all possible combinations of reverberation parameters and source signals. Unfortunately, direct integration $p(z|\mathbf{y}) = \iiint p(z, s, \beta, w|\mathbf{y}) ds d\beta dw$ is analytically intractable. Therefore, this integral is approximated in a manner similar to VB. In this study, a bound to a constant ($= 0$) is tightened, using an approximate posterior distribution $q(z, s, \beta, w)$, which is assumed to be of a factored form (“mean-field” approximation, MFA), as done in VB. Similar approach based on MFA has been employed for discriminative models [128] in the context of conditional random fields. The proposed method, however, uses MFA to approximate the posterior distribution to a Bayesian latent variable model. To this end, the posterior $q(z, s, \beta, w)$ is approximated as $\prod_{f,t} q(z_{ft})q(s_{ft})q(w_f)q(\beta_f)$. Moreover, $q(w)$ considers only up to a finite dimension $\bar{I}$. This is known as the truncation approximation [10]. Note that in this case, the marginal posterior w.r.t. z_{ft} simply becomes $q(z_{ft})$. To find an approximate posterior, the following lower bound is optimized, obtained using Jensen’s inequality, i.e., $\sum_{i=1}^{I'} \phi_i f(x_i) \leq f(\sum_{i=1}^{I'} \phi_i x_i)$ for a convex function f and an I' -simplex:

$$\begin{aligned} \log \int p(s, z, \beta, w|\mathbf{y}) ds dz d\beta dw &\geq \int q(s, z, \beta, w) \log \frac{p(s, z, \beta, w|\mathbf{y})}{q(s, z, \beta, w)} ds dz d\beta dw \\ &= \langle \log p(z|\mathbf{y}, s, \beta, w) \rangle_{q(s, z, \beta, w)} - \langle \log q(z) \rangle_{q(z)} + \langle \log p(s|u, v) \rangle_{q(s)} - \langle \log q(s) \rangle_{q(s)} \\ &\quad + \langle \log p(\beta|a^{(\beta)}, b^{(\beta)}) \rangle_{q(\beta)} - \langle \log q(\beta) \rangle_{q(\beta)} + \langle \log p(w|\alpha) \rangle_{q(w)} - \langle \log q(w) \rangle_{q(w)}. \end{aligned} \quad (6.14)$$

The bound is tightened when q is closest, in Kullback-Leibler divergence sense, to the posterior distribution. Note that when no restriction on the functional form of q is assumed, the bound is tightened when $q(s, z, \beta, w) = p(s, z, \beta, w|\mathbf{y})$.

Optimization of Equation 6.14 w.r.t. $q(z)$ yields the following:

$$q(z) = \prod_{f=1, t=1, k=1, i=0}^{E, T, y_{ft}, \bar{I}} \phi_{fti}^{z_{kfti}}, \quad (6.15)$$

where $\phi_{fti} = \bar{\phi}_{fti} / \sum_{j=0}^{\bar{I}} \bar{\phi}_{ftj}$, and

$$\log \bar{\phi}_{fti} = \begin{cases} \langle \log \beta_f \rangle + \langle \log s_{ft} \rangle, & i = 0 \\ \langle \log(1 - \beta_f) \rangle + \langle \log w_{fi} \rangle + \log y_{f,t-i}, & i > 0. \end{cases} \quad (6.16)$$

ϕ_{ft} , or $\langle z_{kft} \rangle$, is independent of k because given f and t , z_{kft} 's are distributed i.i.d². From this follows that, since $\sum_{k=1}^{y_{ft}} \langle z_{kfti} \rangle = y_{ft} \phi_{fti}$, Equation 6.13 simplifies to $m_{ft} = \phi_{ft0}$.

Updating with respect other variables is complicated since the term $\langle -\log n_{ft} \rangle$ in Equation 6.12 makes our model non-conjugate, meaning that the expectations in Equation 6.14 cannot be computed analytically. To attack this problem, a lower-bound to the objective function Equation 6.14 is optimized, using minorization-maximization (MM) technique.

6.2.1 Design of the minorization function

In order to find an analytical inference rule for the model, the MM algorithm [59], also known as auxiliary function method in the context of signal decomposition [66,77], is used. The MM algorithm is a convergent algorithm for maximizing an objective function by designing a lower bound to the objective by introducing additional variables (auxiliary variables), and iteratively tightening the bound with respect to the auxiliary variables and the parameter. The key is to design a lower bound such that it is easy to update w.r.t. both the auxiliary variables and the parameter.

Inspecting Equation 6.10 and Equation 6.12 reveals that inference is complicated by the term $\langle -\log n_{ft} \rangle$ that appears in the log posterior distribution. Computation of the expectation is complicated because the term is cannot be expressed as a weighted sum of the natural parameters of the posterior. Analytical computation of the expectation is possible if it could be expressed as a weighted sum of the expectations of the natural parameters, $\langle \log \beta_f \rangle$, $\langle \log(1 - \beta_f) \rangle$, $\langle \log w_{fi} \rangle$, $\langle \log s_{ft} \rangle$, and $\langle -s_{ft} \rangle$. Hence, a lower bound to $\langle -\log n_{ft} \rangle$ that satisfies these properties should be designed.

Let us introduce a set of auxiliary variables $\mathcal{A}$ to design a lower bound $J(z, s, w, \beta, \mathcal{A})$, such that maximizing J w.r.t. z, s, w, β and $\mathcal{A}$ is possible. The lower bound J shall satisfy the following criteria:

1. $\langle \log p(y, z, s, w, \beta) \rangle_q = \max_{\mathcal{A}} J(z, s, w, \beta, \mathcal{A})$
2. $\langle \log p(y, z, s, w, \beta) \rangle_q \geq J(z, s, w, \beta, \mathcal{A})$
3. $J(z, s, w, \beta, \mathcal{A})$ is a weighted sum of the expectations of the natural parameters of the posterior, plus $\mathcal{A}$ and a constant.

²This is not to say that, for a given (f, t) , z 's are identical for all k , i.e., they are constrained to take on the same value: they are separate random variables that have the same expectations.

To simplify the notation, the time-frequency index (f, t) is abbreviated when it is obvious, and $y_{f,t-i}$ is notated as $y^{(i)}$.

First, $\langle -\log n_{ft} \rangle$ is lower-bounded by a first-order Taylor expansion about $\mathcal{R}_{ft}$:

$$\langle -\log n_{ft} \rangle \geq -\log \mathcal{R}_{ft} - \frac{\langle n_{ft} \rangle - \mathcal{R}_{ft}}{\mathcal{R}_{ft}}. \quad (6.17)$$

This bound is expressed in terms of a weighted sum of $\log s_{ft}$ and $-s_{ft}$, so it is possible to optimize the objective w.r.t. $q(s_{ft})$. Optimizing w.r.t. $q(s_{ft})$ yields:

$$s_{ft} \sim \text{Gam}(\bar{u}_{ft}, \bar{v}_{ft}), \quad (6.18)$$

where

$$\bar{u}_{ft} = u_{ft} + y_{ft} \phi_{ft0}, \quad (6.19)$$

$$\bar{v}_{ft} = v_{ft} + y_{ft} \langle \beta_f \rangle / \mathcal{R}_{ft}. \quad (6.20)$$

The bound is tightened by setting $\mathcal{R}_{ft} = \langle n_{ft} \rangle$.

Equation 6.14 may be optimized w.r.t. $q(w)$ and $q(\beta)$ by lower-bounding $-\langle n_{ft} \rangle$ in the right-hand side of Equation 6.17. Namely, $-\langle n_{ft} \rangle$ is lower-bounded by the following, for a set of auxiliary variables $\mathcal{V}_{ft} > 0$ and some constant $C_{ft} > \mathcal{V}_{ft}$:

$$-\langle n_{ft} \rangle \geq \mathcal{V}_{ft} - C_{ft} - \mathcal{V}_{ft} \log \mathcal{V}_{ft} + \mathcal{V}_{ft} \left\langle \log \left(C_{ft} - \langle n_{ft} \rangle_{q(s_{ft})} \right) \right\rangle_{q(w)q(\beta)}. \quad (6.21)$$

The variable C can be chosen arbitrarily, provided that:

$$C_{ft} > C_{\min, ft} = \sqrt{\langle s_{ft} \rangle_{q(s_{ft})}^2 + \sum_{i=1}^{\bar{I}} y_{f,t-i}^2}. \quad (6.22)$$

Section 6.2.2 presents the detailed proof for this bound and the derivation of $C_{\min}$.

The lower bound of Equation 6.21 is still not a linear combination of the natural parameters, as it now contains the logarithm of a linear combination of the exponents of natural parameters. Observe, however, that $\log(C_{ft} - \langle n_{ft} \rangle_{q(s_{ft})})$ is expressed as a logarithm of a weighted combination of the exponent of natural parameters. Therefore, extract the summation can be extracted outside the logarithm by applying Jensen's

inequality and obtain the following bound, where $\mathbf{Q}_{ft}$ is a $(\bar{I} + 1)$ -simplex:

$$\begin{aligned}
 & \left\langle \log \left(C_{ft} - \langle n_{ft} \rangle_{q(\mathbf{s}_{ft})} \right) \right\rangle_{q(\boldsymbol{\beta})q(\mathbf{w})} \\
 &= \left\langle \log \left[\beta_f \left(C_{ft} - \langle \mathbf{s}_{ft} \rangle_{q(\mathbf{s}_{ft})} \right) + \sum_{i=1}^{\bar{I}} (1 - \beta_f) w_{fi} (C_{ft} - y_{f,t-i}) \right] \right\rangle_{q(\boldsymbol{\beta})q(\mathbf{w})} \\
 & \quad [-1mm] \geq \mathbf{Q}_{ft0} \left\langle \log \left(\beta_f \left(C_{ft} - \langle \mathbf{s}_{ft} \rangle_{q(\mathbf{s}_{ft})} \right) \right) \right\rangle_{q(\boldsymbol{\beta})q(\mathbf{w})} \\
 & \quad + \sum_{i=1}^{\bar{I}} \mathbf{Q}_{fti} \left\langle \log \left[(1 - \beta_f) w_{fi} (C_{ft} - y_{f,t-i}) \right] \right\rangle_{q(\boldsymbol{\beta})q(\mathbf{w})} - \sum_{i=0}^{\bar{I}} \mathbf{Q}_{fti} \log \mathbf{Q}_{fti}. \quad (6.23)
 \end{aligned}$$

The right-hand side is a weighted sum of the expectations of the natural parameters of $\boldsymbol{\beta}$ and $\mathbf{w}$ used by the joint posterior. Therefore, this bound can be used to analytically optimize the objective function in Equation 6.14 with respect to $q(\boldsymbol{\beta})$ and $q(\mathbf{w})$.

Optimizing the function w.r.t. $q(\mathbf{w})$ yields the following update:

$$\theta_{fi} \sim \text{Beta}(\bar{a}_{fi}^{(\theta)}, \bar{b}_{fi}^{(\theta)}), \quad (6.24)$$

where

$$\bar{a}_{fi}^{(\theta)} = 1 + \xi_{fi}, \quad (6.25)$$

$$\bar{b}_{fi}^{(\theta)} = \alpha + \sum_{j=i+1}^{\bar{I}} \xi_{fj}, \quad (6.26)$$

$$\xi_{fi} = \sum_{t=1}^T y_{ft} \left(\phi_{fti} + \mathbf{Q}_{fti} \frac{\mathcal{V}_{ft}}{\mathcal{R}_{ft}} \right). \quad (6.27)$$

Likewise, optimizing w.r.t. $q(\boldsymbol{\beta})$ leads to the following:

$$\beta_f \sim \text{Beta}(\bar{a}_f^{(\beta)}, \bar{b}_f^{(\beta)}), \quad (6.28)$$

where $a^{(\beta)}$ and $b^{(\beta)}$ are the following:

$$\bar{a}_f^{(\beta)} = a_f^{(\beta)} + \xi_{f0}, \quad (6.29)$$

$$\bar{b}_f^{(\beta)} = b_f^{(\beta)} + \sum_{i=1}^{\bar{I}} \xi_{fi}. \quad (6.30)$$

The bound is tightened by updating Q as follows, normalizing it such that $Q_{ft} = 1$:

$$Q_{ft0} \propto (C_{ft} - \langle s_{ft} \rangle) \exp \langle \log \beta_f \rangle, \quad (6.31)$$

$$Q_{f,t,i>0} \propto (C_{ft} - y_{f,t-i}) \exp(\langle \log(1 - \beta_f) \rangle + \langle \log w_{fi} \rangle). \quad (6.32)$$

6.2.2 Proof of Equation 6.21

Theorem 1. For a differentiable and strictly monotonic function $f(x)$, a positive density function q over the domain of x and some C such that $0 < f(x) < C$ for all x in its domain, the following holds for all $\mathcal{V} \in (0, C)$:

$$\langle -f(x) \rangle_{q(x)} \geq \mathcal{V} \langle \log(C - f(x)) \rangle_{q(x)} - \mathcal{V} \log \mathcal{V} - C + \mathcal{V}. \quad (6.33)$$

The bound is tight at $\mathcal{V} = \exp \langle \log(C - f(x)) \rangle_{q(x)}$.

Proof 1. Consider a function $\bar{u} \log(C - f(x)) + \bar{v}$ for some $\bar{u} > 0$ and $\bar{v}$. Match the tangent of this function and that of $-f(x)$ about $x = x_0$ such that $f(x_0) \in (0, C)$, and solve for $\bar{u}$ and $\bar{v}$. Set $\mathcal{V} = C - f(x_0)$ and take the expectation of both sides to obtain Equation 6.33. Since the r.h.s. is strictly concave, it meets the l.h.s. at exactly one point.

To tighten the bound, optimize the r.h.s. with respect to x_0 to obtain $f(x_0) = M - \exp \langle \log(C - f(x)) \rangle_{q(x)}$. Substitute this to the definition of $\mathcal{V}$ to tighten the bound w.r.t. $\mathcal{V}$ by setting it to $\exp \langle \log(C - f(x)) \rangle_{q(x)}$.

Apply this theorem, where $f(\beta, w) = \langle N \rangle_{q(s)}$ and $q(x) = q(\beta, w)$ to obtain the desired inequality in Equation 6.21.

Next, $C_{\min}$ shall be derived. First note that $\log(C_{ft} - \langle n_{ft} \rangle_{q(s)q(z)})$ is defined for the domain of $q(w)q(\beta)$; therefore, for all w' in $\bar{I}$ -simplex and $\beta' \in [0, 1]$, $C_{ft} - \langle n_{ft} \rangle_{q(s)q(z)}|_{w=w', \beta=\beta'}$ should be greater than zero. Note that $\langle n_{ft} \rangle_{q(s)}$ can be expressed an inner product of two vectors $\mathbf{v}(f, t)$ and $\mathbf{w}(f, t)$, where

$$\mathbf{v}(f, t) = \begin{pmatrix} \beta_f \\ (1 - \beta_f w_{f1}) \\ (1 - \beta_f w_{f2}) \\ \dots \\ (1 - \beta_f w_{f\bar{I}}) \end{pmatrix} \text{ and } \mathbf{w}(f, t) = \begin{pmatrix} \langle s_{ft} \rangle_{q(s_{ft})} \\ y_{ft-1} \\ y_{ft-2} \\ \dots \\ y_{ft-\bar{I}} \end{pmatrix},$$

with $\sum_i v_{fti} = 1$. Since $|\mathbf{v}_{ft}| \leq 1$, it is possible to bound $|\mathbf{v}_{ft} \cdot \mathbf{w}_{ft}| \leq |\mathbf{w}_{ft}|$. Therefore, the bound is valid if $C_{ft} > |\mathbf{w}_{ft}|$.

6.2.3 Expectation of the sufficient statistics

The sufficient statistics are given as follows:

$$\langle \log \beta_f \rangle = \psi(\bar{a}_f^{(\beta)}) - \psi(\bar{a}_f^{(\beta)} + \bar{b}_f^{(\beta)}), \quad (6.34)$$

$$\langle \log(1 - \beta_f) \rangle = \psi(\bar{b}_f^{(\beta)}) - \psi(\bar{a}_f^{(\beta)} + \bar{b}_f^{(\beta)}), \quad (6.35)$$

$$\langle \beta_f \rangle = \frac{\bar{a}_f^{(\beta)}}{\bar{a}_f^{(\beta)} + \bar{b}_f^{(\beta)}}, \quad (6.36)$$

$$\langle s_{ft} \rangle = \bar{u}_{ft} \bar{v}_{ft}, \quad (6.37)$$

$$\langle \log s_{ft} \rangle = \log(\bar{u}_{ft}) + \psi(\bar{v}_{ft}), \quad (6.38)$$

$$\langle \log w_{fi} \rangle = \psi(\bar{a}_{fi}^{(\theta)}) + \sum_{j=1}^{i-1} \psi(\bar{b}_{fj}^{(\theta)}) - \sum_{j=1}^i \psi(\bar{a}_{fj}^{(\theta)} + \bar{b}_{fj}^{(\theta)}), \quad (6.39)$$

$$\langle w_{fi} \rangle = \frac{\bar{a}_{fi}^{(\theta)} \prod_{j=1}^{i-1} \bar{b}_{fj}^{(\theta)}}{\prod_{j=1}^i (\bar{a}_{fj}^{(\theta)} + \bar{b}_{fj}^{(\theta)})}. \quad (6.40)$$

6.3 Evaluation

Three experiments are conducted to assess the proposed method. First, the performance of the proposed method is compared against two existing methods in the power-spectrum domain, one which relies on source sparsity [66] and another that relies on an AR(1) model of reverberation [76]. Moreover, the behavior of the algorithm is explored when a few key parameters are changed. Second, the proposed method is evaluated through a listening test. Finally, the effectiveness of dereverberation as an audio-to-score alignment front-end is assessed.

In the following experiments, unless otherwise stated, the parameters shown in Table 6.1 are used. The dereverberated signal is created by first multiplying the STFT of the observation $\tilde{y}_{ft}$ and the square-root of the time-frequency mask m_{ft} defined in Equation 6.13. Then, the time domain signal is recovered using inverse STFT. This way, the recovered power spectrogram becomes $m_{ft} y_{ft}$, as specified in Equation 6.13.

6.3.1 Dereverberation performance

Data

For the source signals, utterances from the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus's core test dataset of 24 speakers [45] were used. For musical

Table 6.1: List of the parameters used.

Parameter	Description	Value
N/A	Waveform Format	mono, 16 bit unsigned
N/A	Sampling Frequency	16 kHz
F	Frame Length	1024
N/A	Hop Size	25% overlap
N/A	Window function	Modified Bartlett-Hann Window
$a^{(\beta)}, b^{(\beta)}$	Signal-to-Reverb Ratio	1.0, 1.0
u, v	Source Sparseness	1.0, 10^{-3}
α	Reverberation Concentration	50.0
$\bar{I}$	Truncation order	50
N/A	Number of iterations of the MM algorithm	20

sources, the BACH10 dataset [29], a database of Bach chorales recorded in an anechoic chamber, was used. Each signal has a duration of 20 to 30 seconds.

For the impulse response, real and simulated responses were prepared. Five real-world impulse responses were prepared, three of which are from the “Impulse response and speech data with microphone array” data from the RWCP Sound Scene Database in Real Acoustic Environment [94], which consist of impulse responses of reverberations whose reverberation time is up to 780 ms (33 frames). They include an impulse response of an echo room (denoted “E2A,” reverb length of 120 ms), a tatami-floored room (“JR1,” 600 ms), and a conference room (“OFC,” 780 ms). Here, the length of reverberation is defined as the time that the power of reverberation decays by -60 dB compared to the start. The remaining two are from the Promenadikeskus Concert Hall impulse response [85], which contains long reverberation of up to about 3 s (130 frames). One impulse response is recorded with a source on stage and mic on stage (“s1_r1_o,” 2.7s); another is recorded with a source on stage and mic on the audience seat (“s1_p1_o,” 2.6s). Therefore, even though they have similar reverberation times, dereverberating the latter is more difficult: the direct-path component is weaker relative to the reverberation, so the source signal is less clear. In other words, the posterior distribution of the variable β is likely to be different, where s1_r1_o is expected to have a smaller $\langle\beta\rangle$ compared to s1_p1_o.

For the synthesized response, the image method [3] was used to synthesize artificial impulse responses. Eighteen impulse responses were generated, with uniform lengths of 20 m, heights of 8 m, absorption coefficients of 0.19, and widths of 5 m to 95 m in 5 m increment.

Table 6.2: Comparison of different methods, averaged over impulse response.

(a) Itakura-Saito Distance.						
Method	Wet	KAM09	LEB01	Proposed (finite)	Proposed (infinite)	
E2A	3.73	6.24	13.01	3.10	3.09	
JR1	3.73	6.55	12.94	3.29	3.28	
OFC	4.14	6.66	12.30	3.34	3.33	
s1_p1_o	5.35	7.53	10.19	3.98	3.93	
s1_r1_o	5.12	7.37	9.66	3.96	3.91	
Ave.	4.41	6.87	11.62	3.53	3.51	
(b) Signal-to-Reverberation Modulation Energy Ratio.						
Method	Dry	Wet	KAM09	LEB01	Proposed (finite)	Proposed (infinite)
E2A	5.01	7.03	9.94	8.80	8.52	8.61
JR1	5.01	4.47	6.52	7.60	5.72	5.83
OFC	5.01	4.60	6.75	7.57	5.88	6.01
s1_p1_o	5.01	3.49	5.40	6.38	4.72	4.88
s1_r1_o	5.01	3.13	4.76	7.55	4.12	4.23
Ave.	5.01	4.54	6.67	7.58	5.79	5.91

Metrics

Two metrics were used to evaluate the proposed method: Signal-to-Reverberation Modulation Energy Ratio (SRMR) [35] and the Itakura-Saito Distance (ISD)³.

SRMR is a non-intrusive measure that increases for signals with shorter late reverberation. Therefore, a dereverberation method should improve the SRMR of the estimated signal from that of the wet signal. The SRMR, however, should not be the only figure of merit because it ignores signal distortion caused by dereverberation: a method may attenuate the late reverberation drastically but sound extremely poorly.

In order to evaluate how close the dereverberated signal sounds to the true dry signal, the ISD was also evaluated. ISD is an asymmetric measure of spectral dissimilarity defined as follows:

$$\text{ISD}_{x,\hat{x}} = \frac{1}{FT} \sum_{f,t} \left(\frac{x_{ft}}{\hat{x}_{ft}} - \log \frac{x_{ft}}{\hat{x}_{ft}} - 1 \right) \quad (6.41)$$

where x_{ft} is the power spectrogram of the ground-truth dry signal and $\hat{x}_{ft}$ is that of the dereverberated signal. Since the degree to which spectral *shapes* are distorted in

³Note that there is yet no single validated measure for dereverberation. Therefore, the numerical results presented here should be considered an indication, and not conclusion, on the effectiveness of the proposed method.

the process of dereverberation should be evaluated, the evaluation metric should be invariant to the output gain. Therefore, the estimated signal $\hat{x}$ is scaled by a constant that minimizes the ISD from x to $\hat{x}$. Moreover, a small constant $10^{-4} \times \bar{x}$ is added to both x and $\hat{x}$; because ISD is sensitive to signal power ratios, adding a small floor prevents minute differences in weak signal components from adversely affecting the ISD. Since ISD is a good measure of speech perception [9], ISD is expected to be indicative of the perceptual similarity between the estimated dry signal and the ground-truth dry signal.

These metrics were computed from the last 10 seconds of the dereverberated audio, since all the methods require some time (about 2 s) to settle-in.

Comparison with existing methods

The proposed method was compared with two methods that perform dereverberation in the power spectrum domain. The first method, denoted “KAM09,” uses a frequency bin-wise convolutive model of reverberation with a sparse source prior [66]. It is similar to the proposed method in that it relies on the sparseness of the source, but different in that it uses a convolutive reverberation model that only incorporates prior on the source signal. The second method, denoted “LEB01,” uses a frequency bin-wise AR(1) model of reverberation [76]. It is similar to the proposed method in that it uses an auto-regressive model of reverberation, but different in that it assumes very little on the source signal, incorporates no prior information on the reverberation coefficients, and is restricted to AR(1).

Moreover, the proposed method was compared with a finite-mixture version of the proposed method, which is realized by replacing $\text{GEM}(\alpha)$ (with truncation approximation $\bar{I}$ th component) with a finite $\bar{I}$ -dimensional $\text{Dir}(\alpha/50)$. This is an approximation of the DP at $\bar{I} = 50$, and the but the expected number of effective component increases with $\bar{I}$ whereas it is fixed in $\text{GEM}(\alpha)$. The inference algorithm is quite similar, and so is omitted; the difference is in the update of $q(w)$, given by Equation 6.24, and the computation of the sufficient statistics in Equation 6.39 and Equation 6.40.

Table 6.2 shows the evaluation measures averaged over the source signals. Table 6.2a shows that the proposed method is capable of decreasing the ISD compared to the wet signal. This is unlike the existing methods, where the ISD *increases* by attenuating the late reverberation. Comparing the infinite version and the finite version of the proposed method suggests that the difference between the two is negligible, in terms of the ISD. Table 6.2b shows that all the dereverberation methods under consideration are capable of attenuating the late reverberation, since the SRMR increases relative to the wet signal. The data shows that the proposed method is better capable of suppressing the late reverberation compared to the finite version. Even though the existing methods consistently attain higher SRMR than the proposed method, it comes at the price of vastly increased ISD. In this respect, the proposed method is advantageous because its SRMR is comparable to that of the true dry signal, while its

Table 6.3: Evaluation measures averaged over source signals.

(a) Itakura-Saito Distance.

	Wet	KAM09	LEB01	Proposed (finite)	Proposed (infinite)
Voice	4.31	6.51	11.41	3.43	3.41
Music	5.35	10.02	13.47	4.46	4.37

(b) Signal-to-Reverberation Modulation Energy Ratio.

	Dry	Wet	KAM09	LEB01	Proposed (finite)	Proposed (infinite)
Voice	5.11	4.63	6.97	7.78	5.91	6.04
Music	2.06	3.02	3.73	5.86	3.37	3.39

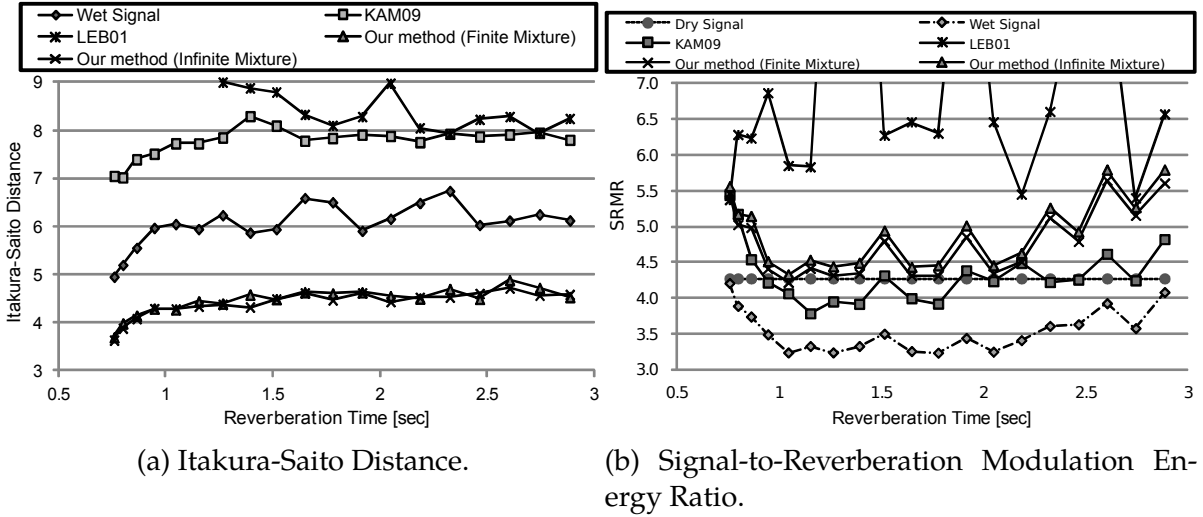


Figure 6.4: Evaluation measures of synthetic impulse response as the reverberation time is varied.

ISD is smaller than the wet signal.

Next, Table 6.3 shows the evaluation measures averaged over source signals. Table 6.3a shows that the proposed method, again, is capable of making the ISD closer to the true dry signal, and Table 6.3b shows that the resulting signal is at least as dry as the true dry signal. The table also shows that the SRMR for music is substantially lower than that of speech, even if they are averaged over the same impulse responses. This presumably owes to the fact that pitched musical instrument sounds are stationary compared to speech, which induces a low-frequency envelope modulation, which in turn decreases the SRMR; in other words, sustained pitched notes sound like reverb.

Next, the evaluation measures of the synthetic impulse data is presented in Fig. 6.4. Fig. 6.4a shows that the ISD increases with the reverberation time, reflecting the intuition that recuperating source signal that has been convolved by a long reverberation

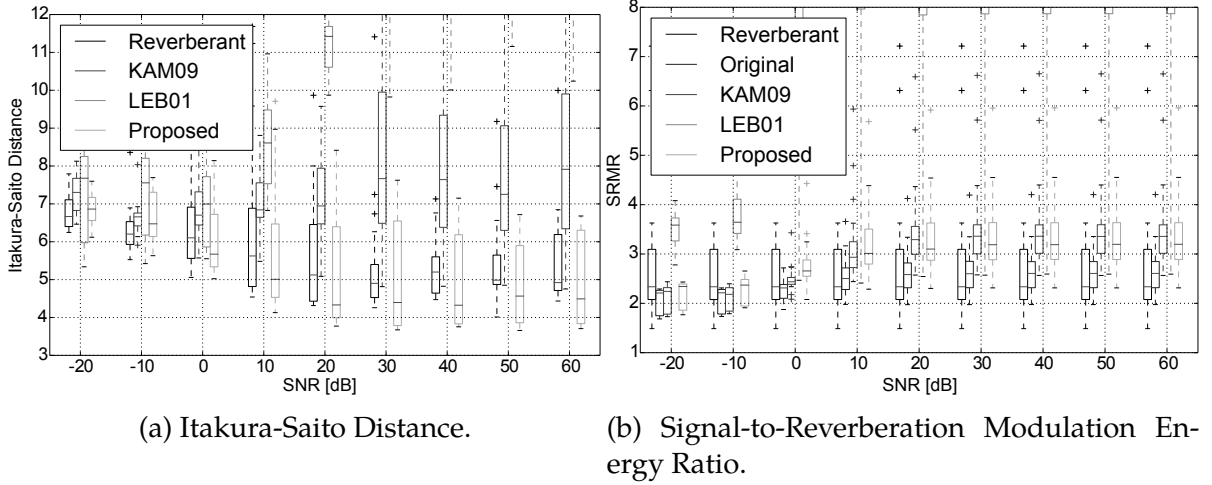


Figure 6.5: Evaluation measures with different SNR.

is difficult. Fig. 6.4b shows that the SRMR initially decreases, then increases. In the decreasing region (reverb time < 1s), early reflection is dominant because the simulated room width is less than the wavelength of the framesize; dereverberation fails in this case because the proposed method only suppresses the late reverberation. In the increasing region, the proposed method starts to suppress late reverberation since late reverberation starts to dominate over early reflection with increased room size.

Robustness to noise

The robustness of the proposed method against noise was tested. This is important because the proposed method assumes that the observation is an output from a purely autoregressive system – it ignores the presence of additive noise, such as the LP noise, which is not driven by the AR system that the proposed method tries to estimate.

To this end, the reverberant audio was corrupted by a pink noise, the power spectrum of which rolls off at -10dB per decade. The signal-to-noise ratio (SNR) ranged from -20 dB to 60 dB, in 10 dB increment.

Fig. 6.5 shows the result. Fig. 6.5a shows that the proposed method fails to dereverberate the signal for a low SNR of less than 0 dB. As the SNR increases, the ISD improvement increases, and saturates after about SNR > 50 dB. The proposed method suffers at a low SNR because the noise masks the wet signal, in particular the weak late reverberation component; the weak late reverberation component provides a valuable cue on higher order AR model parameters, so loss of late reverberation leads to decreased performance. Fig. 6.5b shows that the capability to attenuate the late reverberation saturates after SNR of 10 dB or better. These results suggest that the proposed method is admits a weak noise with SNR of about 10 dB. This is acceptable for post-production and preprocessing for MIR tasks: in a typical music audio, very little noise is introduced after the recording phase.

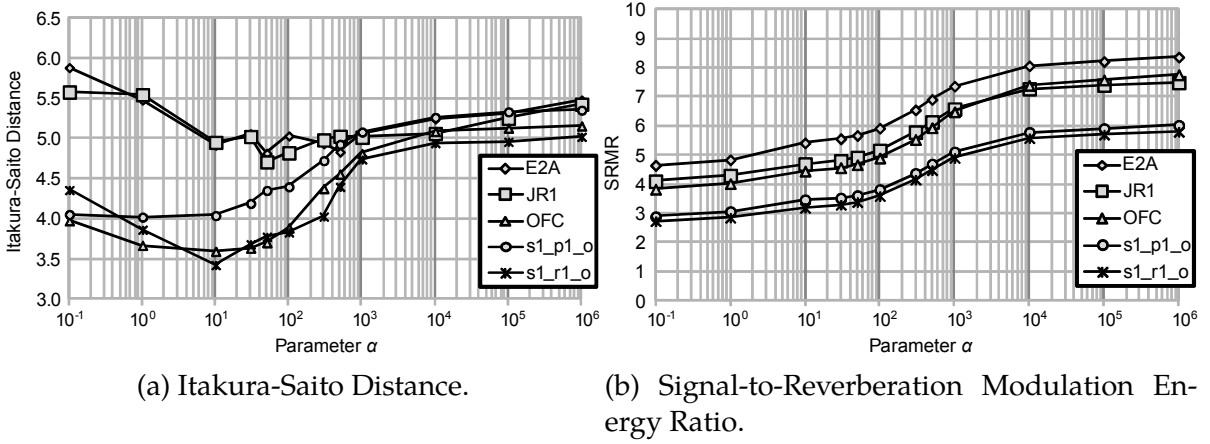

 Figure 6.6: Evaluation measures as α is varied.

 Table 6.4: Standard deviation, evaluated over I , of ISD and SRMR for finite and infinite mixtures, averaged over impulse responses.

	Finite (ISD)	Proposed (ISD)	Finite (SRMR)	Proposed (SRMR)
Voice	0.66	0.42	0.68	0.17
Music	0.12	0.07	0.28	0.14

Analysis of key parameters

The effect of manually-set room acoustics parameter $\bar{I}$ is evaluated. It was shown in Table 6.2 and Table 6.3 that the infinite model performs better than the finite model, in that the infinite version exhibits roughly the same ISD as the finite version while having higher mean SRMR. This experiment compares the robustness of finite and infinite model by changing $\bar{I}$, which governs the underlying model complexity for the finite case. Recall that the finite version assumes reverberation of $\text{AR}(\bar{I})$, and the infinite version assumes $\text{AR}(\infty)$ but evaluates an approximation that computes up to $\bar{I}$. Therefore, the infinite version is expected to be more robust against choice of $\bar{I}$, that is, it is less prone to misspecification of $\bar{I}$. Table 6.4 shows the standard deviation of ISD and SRMR taken for each $\bar{I} \in [32, 48, 64, 96, 128, 192, 256]$. Note that the standard deviation of the infinite model is smaller than the finite model except for SRMR of speech, where the difference is small. This suggests that the infinite model is less sensitive to the manually predetermined complexity of room acoustics, $\bar{I}$, than the finite model. Moreover, in the finite version, both the SRMR and the ISD tends to increase with increased $\bar{I}$ after a certain point; this suggests that the finite version tends to over-attenuate the reverberation by increasing $\bar{I}$ beyond some optimal point.

Next, the effect of α , the parameter that governs the tendency for new filter component to be activated, is evaluated. In this experiment, only the TIMIT data was used. The ISD and the SRMR were evaluated for $\alpha \in [10^{-1}, 10^6]$. The ISD and SRMR

Table 6.5: Average rating (on a 5 point scale) of excerpts averaged over the musical genre. Number following $\pm$ is the standard deviation.

	KAM09	Proposed
Orchestra	1.72 $\pm$ 0.74	3.40 $\pm$ 0.81
Chamber Music	1.64 $\pm$ 0.67	3.74 $\pm$ 0.76
Popular Song	1.70 $\pm$ 0.74	3.17 $\pm$ 0.98
Speech	2.00 $\pm$ 1.13	3.17 $\pm$ 1.04

are shown in Fig. 6.6a and Fig. 6.6b, respectively. These figures show two tendencies. First, SRMR increases and saturates by increasing α . This matches the intuition that as more components of w are activated, the capability to suppress reverberation tails increase. Second, as α increases, the ISD decreases first and then increases. This behavior is likely to be caused by two factors at play. First, as α is increased, the reverberation tail is attenuated, which causes the dereverberated signal to look more like the dry signal. However, after a certain point, the decaying component is attenuated too excessively that the ISD starts to increase. Fortunately, the tendencies of ISD and SRMR as a function of α does not vary too significantly, so very little manual intervention is required for practical purposes.

6.3.2 Listening tests

A subjective evaluation of the proposed method was performed. First, eight 15 second excerpts were prepared. Of eight, three are orchestral excerpts, three are chamber music excerpts (piano solo, piano solo + tenor singer, piano trio), one is a popular music excerpt and one is a male speech. For each excerpt, each subject was first presented, using a headphone, the original reverberant signal. Then, the dereverberated signal using KAM09 and the proposed method were presented in random order⁴. Finally, the subjects were asked to rate the sound quality on a 1–5 scale. LEB01 was not used because Table 6.2 suggests that KAM09 is preferable over LEB01 both in terms of ISD and SRMR. The test subjects (17 subjects) were all male between the age range of 20 to 40, and deal with sound signal processing technology on a daily basis.

Table 6.5 shows the average rating and the standard deviation. The data seems to be in favor of the proposed algorithm. To investigate this, two-way ANOVA was performed on the collected ratings data. With a significance level of 0.01, interaction between the excerpt and deverboration method was not seen (p-value of 0.02), nor did the excerpt itself (p-value of 0.29), but the effect of dereverberation method was significant (p-value of 1.1^{-16}). This suggests that the proposed method produces a more natural sounding dereverberated signal than the existing method.

⁴LEB01 was omitted because KAM09 consistently outperformed LEB01 in terms of ISD, suggesting that KAM09 produces dereverberated signals that perceptually sounds closer to the dry signal.

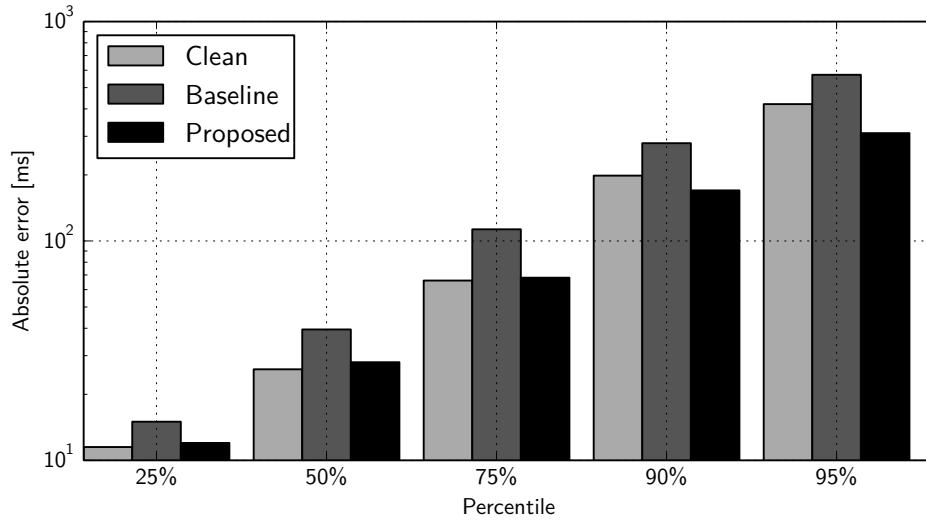


Figure 6.7: Percentile of Audio-to-Score Alignment Error.

6.3.3 Application to audio-to-score alignment

To evaluate the effect of dereverberation on music alignment, audio-to-score method introduced in Chapter 3 was used to align audio signals that have been reverberated and dereverberated.

To this end, three kinds of audio signals were synthesized. First, ten SMF files from the RWC Classical Music Database [50] are used to synthesize a musical audio signal with a synthesizer, along with a mapping of ground truth score position to audio position. The instrumentations of the chosen SMFs are duets of a single instrument and a piano. Hence, the SMFs are a polyphonic mixture of both sustained and decaying instrumental sound. Next, another set of audio signal is created by convolving the synthesized audio with a reverberation [85] (s1_r1_o). Finally, another set of audio signal is created by dereverberating the reverberant audio.

The music score was aligned to three kinds of audio signals, and the alignment error percentile was computed. The result is shown in Fig. 6.7. The result shows that that audio reverberation multiplicatively corrupts the alignment. At the same time, dereverberation brings the performance of score alignment to the same level as the ideal case, where no reverberation is added. The dereverberated signal sometimes performs better than the ideal case presumably because the proposed method suppressed not only the reverberation but also sustain pedals of the piano parts; sustain pedal causes the piano string to vibrate freely, elongating duration of the played notes than that notated in the music score.

Dereverberation is clearly beneficial for improving score alignment accuracy. This is particularly beneficial for score-aided source separation, where the fundamental assumption is that the music score and the audio signal is aligned perfectly.

6.4 Summary

This chapter presented a dereverberation method based on the power spectrogram representation of an audio signal. The proposed method is robust to variations in reverberation length because the complexity of reverberation is determined in a data-driven manner. Moreover, the proposed method works with many kinds of signal like music or speech, since the assumption on the dry signal is minimal. Namely, it assumes that the dry signal is weakly sparse.

To infer the model, an iterative, convergent method to approximate the marginal posterior distribution given the joint posterior distribution was presented. The iterative algorithm was inspired by variational Bayesian method and minorization maximization techniques.

Most importantly, the proposed method was shown to be effective in audio-to-score alignment. Namely, the quality of alignment on dereverberated signal was as good as that on ideal, reverberation-free signal. Moreover, the proposed method attained a significantly better dereverberation performance, especially in terms of the perceptual closeness to the original signal. This was demonstrated through both numerical evaluation and listening tests. Finally, the proposed method was capable of attenuating the reverberation under noisy condition with SNR of greater than 10 dB. Thus, the proposed method is applicable on noisy audio signal, such as a recording on an old disc record.

Chapter 7

Conclusion

This thesis presented a Bayesian approach to alignment of music signals. The goal of the thesis was to temporally align multiple audio signals that represent a common piece of music. This is important for applications such as music performance analysis, music production, or informed source separation. This chapter first reviews the contributions of this thesis in this light, and then presents directions for future research.

7.1 Contribution of this thesis

An audio-to-score alignment was presented in Chapter 3. The method expressed an audio signal as a sum of harmonic sounds, where the fundamental frequencies are specified in terms of the notated pitches, but the shapes of the harmonic structure and their balances are non-informative. The method also addressed note onset timing asynchrony in terms of a “shared” global tempo trajectory and note onset timings that are defined independently for each part. The proposed method (1) achieved a low alignment error of about 30 ms on real-world ensemble pieces and 50 ms on synthetic orchestral recordings, (2) was capable of estimating inter-part asynchrony, and (3) estimated timbre and volume such that the segregated signal had a mean SIR and SDR of 2.5 dB and 15.8 dB, respectively. This kind of alignment accuracy allows, in recording applications, a recording engineer to seamlessly switch back and forth between multiple takes, given the symbolic music score that the recordings play. Furthermore, the capability to estimate inter-part asynchrony allows musicological analysis of the kind of asynchrony that occurs in a piece of music. Hence, in musicological applications, one could now compare not only the tempo and inter-part asynchrony, but also the dynamics and the timbre as well. The capability to infer timbre and volume is useful in informed source separation applications, allowing the user to separate out a particular instrument from a polyphonic mixture. Moreover, this kind of capability allows a musicologist to, for example, compare the kinds of melodic lines that are emphasized in different recordings.

An audio-to-audio alignment was addressed in Chapter 4. The proposed method

decoupled the generative process of the musical piece and that of music performance by different performers. Moreover, the proposed method incorporated a duration model of music performance that has a potential to uncover phrase boundaries and similarities in tempo dynamics. Experimental evaluation on real-world piano pieces showed a median alignment error of 60 ms, outperforming existing generative methods and approaching the performance of a widely-used pairwise alignment method based on local feature matching. This is significant because the proposed method allows one to easily extend the model to uncover different aspects of the mechanisms involved in music composition or performance. This kind of extension is non-trivial in the local feature matching approach. Thus, the proposed method provides a groundwork, which performs comparably to the current audio alignment methods, for expressive models that uncovers various aspects of music composition or performance.

A subset audio alignment method was presented in Chapter 5. The proposed method aligned audio signals by listening for parts that are played in common. Experimental evaluation showed that the proposed method, even when there were many parts that are joint, achieved a median alignment error of 200 ms. The proposed method allows one to compare audio recording of a full part versus that of a single part. This is crucial for comparing professional recordings (often recorded with full orchestration) with an amateur recording (say, a violinist playing the solo violin part to a violin concerto).

Finally, a method to suppress late reverberation was presented in Chapter 6. This was important because in the real world, a music audio signal is recorded under various room acoustics. By modeling the reverberation process as an autoregressive process over the power spectral domain, and by using Bayesian nonparametrics to express the effective order of the AR process, the proposed method attained robustness to room acoustic difference. Most importantly, evaluation showed that alignment done on dereverberated signals performs comparably to that done on dry signals. This shows that many real-world signal that are contaminated with reverberation may be aligned with a high accuracy, as if the recording were reverberation-free.

7.2 Directions for future research

The author concludes by presenting some directions for future research.

In the Bayesian audio-to-score alignment method proposed in this thesis, each part was assumed to play conditionally independently of others, given a global tempo trajectory. This kind of conditional independence was the key ingredient in estimating the inter-part asynchrony. The inter-part asynchrony, however, is the result of interaction between different performers. Therefore, it may be interesting to consider the process of interaction that produces the asynchrony. For example, one might envision Markov chains defined for each performer, whose transition probabilities are coupled. Through analyzing the degree of asynchrony, it may be possible to find where

performers play “freely” and where players pay attention to “lock-in” to each other.

Another interesting direction of future research is in finer modeling of timbre and volume. To elaborate, in the proposed audio-to-score alignment method, the timbre within a note within each instrument was assumed to remain stationary, and the relative dynamics within a chord was assumed to be constant. This kind of assumption is limiting if one needs to use the proposed method for detailed analysis of playing techniques, or high-quality informed source separation. Therefore, one might consider incorporating a model that is capable of conveying subtle fluctuations in the volume, the timbre, or the pitch.

A direction of future research in Bayesian audio alignment includes generative model inference across different musical pieces. Since the proposed method decouples the process of music performance and music composition, it might be interesting to find, for a given performer, common patterns of how the relative tempo fluctuates. By comparing these patterns, one might be able to uncover global trends in how a particular performer plays, uncover the influential performer of the time, or discover “factions” of musicians. Moreover, since the model of music composition is decoupled, one might model the music composition hierarchically, uncovering commonly used phrases across a set of musical pieces. Moreover, applying the proposed modeling approach to audio alignment problems that merit from generative processes is an interesting direction for future research.

For subset music alignment, it may be interesting to incorporate richer classes of source signal models, such as those that incorporate harmonic structure. This way, the proposed method could align based on subset-like relationship in different musical aspects. To elaborate, one might allow the harmonic structure to vary across different audio signals, only mandating that the played fundamental frequencies are organized hierarchically. Moreover, application of the proposed method to informed separation is an interesting direction for further research.

In Bayesian dereverberation, a method that is free of human intervention is an important future work. In particular, a model that does require the user to set the hyperparameters regarding the source signal u , v and the concentration parameter of the reverberation process, α , is expected to make the system more robust, as these parameters tend to be counter-intuitive for the end-users.

Integrating the proposed methods into a single framework is another interesting direction of research. To elaborate, the presented chapters focused on different aspects of the “master” generative model of music alignment, given in Equation 2.24. Namely, Chapter 4 focused on $p(s|z)$ and $p(z)$, expressing $p(z)$ using a generative process of musical piece, and $p(s|z)$ using a duration model that focuses on the idea that different performances more-or-less play in the same tempo. In this chapter, however, the generative model of instantaneous sound $p(\theta|z)$ was kept simple. On the other hand, Chapter 5 focused on $p(\theta|z)$, allowing each performance to play different set of notes, provided that some notes are played in common. The generative process of the piece $p(z)$ was kept simple, in order to assess the performance of the proposed

model. Therefore, integrating the results of the presented chapters to improve the result may be an interesting direction for research. Moreover, incorporating a model that incorporate explicit harmonic structure, like those used in Chapter 3, is an interesting direction of research for audio-to-audio alignment.

Application of the generative model approach to signals other than audio is another interesting area for future research. For example, one might express the generative process of time-stamped text annotations, videos or motion data. Then, one may compare different video recordings or motion data to analyze what kinds of differences in motion contribute to the differences in sound.

As for the inference scheme, all of these methods and deverbation may merit from online inference. There are many online Bayesian inference strategies being proposed recently [56,65,115,131], so it may be interesting to apply them to the proposed methods. Moreover, for Bayesian audio-to-audio alignment and subset music alignment, the inference over hierarchical discrete models is an important problem. The author noted that the problem of local optima in variational Bayesian method using mean-field approximation is severe in hierarchical discrete models. Improved inference scheme for a discrete hierarchical model is thus an important future work.

Appendix A

Application: Violin Fingering Inference

In music instrument performance, deciding the sequence of finger placements needed to produce a given sequence of pitches, known as the *fingering*, is an important and sometimes difficult problem musicians need to solve. Fingering decision can be difficult because fingering must be both musical and ergonomic, two often-conflicting ideals. For example, the “easiest” fingering on a violin often involves unmusically abrupt changes of timbre. This is because most pitches can be found in more than one location on the instrument, namely, on more than one string; each string, however, has a different timbre (for reasons to be explained later), and often the easiest fingering involves changes between strings. Experienced musician would choose a balanced fingering that satisfies both ergonomic finger placements and expresses the artistic value of the musician.

The essence of violin fingering resides in finding an ergonomic sequence of finger placements, and choosing a musically appropriate *fingerboard location sequence*, the sequence of positions on the fingerboard on which the finger presses the string. Each string is tuned differently and has a distinct timbre. Therefore, it is essential for a violinist to choose a fingerboard location sequence that sounds musically well-motivated. For example, in *Air on the G String*, August Wilhelmj’s arrangement of the second movement of J.S. Bach’s *Orchestral Suite No. 3*, the arranger specifies the entire solo violin part to be played on one string (the G string), most likely to keep a consistent, warm timbre that is characteristic of the G string.

This study aims to develop a method for analyzing an audio recording of a violin and estimating the fingering required to recreate the “sound” of a particular artist. Such a method would allow a beginner, for example, to analyze the recording of past masters and gain insights on how to imitate them. It could also help violin students to compare different fingerings played by different players to appreciate different musical values. Finding similarities of fingering among musicians would allow one to find a stylistically suitable way to play a particular piece, a difficult task for a violinist

studying alone without a teacher’s help.

This kind of goal requires analysis of an audio signal to estimate the fingerboard location sequence, and decision of an ergonomic finger placement sequence that satisfies the sequence of notated notes and the estimated fingerboard location. Yet, existing methods in fingering estimation do not focus on timbral differences caused by different fingerboard locations, and existing methods for estimating the fingerboard location sequence needs robustness in a polyphonic mixture.

Most studies in fingering estimation research intend not to retrieve a fingering from a particular recording, but rather to determine an ergonomic fingering. In these studies, the general framework involves designing a cost function for moving the hand from one shape to another, and finding a sequence of hand movements that minimizes the accrued cost. In the seminal work by Sayegh [116], the cost uses the distance that the hand traverses horizontally and vertically across the instrument. Other works involves similar kind of cost based on the actual distance that the hand needs to traverse, based on different constraints posed by different instruments, such as the guitar [109], or the piano [68, 137]. The costs may be given *a priori* [68] or trained using training data [110]. The formulation of fingering estimation as a cost minimization problem has been given a probabilistic interpretation using Hidden Markov Models (HMM) [137]. These studies all seem to have the perspective that musicality resides in the sequence of notated notes, which overlooks the expression added by the performer. Such an orientation implies that given a symbolic representation of music, a fingering may be determined on a one-to-one basis. Yet each person, given a piece of music, might play it using a different fingering, based on his/her musical values and physiological constraints. Clearly, performance analysis based on audio signal is essential for realizing the goal of reconstructing fingerings from recorded performances.

Studies in detailed performance analysis fails either in a polyphonic mixture of audio, or when the sound of the instrument that needs to be analyzed sounds significantly different from the training data. Yet, in practical situations, analysis of a musical phrase in polyphonic mixture is essential. Methods for analyzing the control input to the violin from audio spectra [6, 73], or the guitar [124] do not work well to instruments with acoustic characteristics than those used in the training, or makes highly restrictive assumptions on how the instrument is played. Greater accuracy is reported using audio-visual fusion [79, 142]. Some studies use only visual information to analyze the fingering of a guitar [14]. In either case, the fact that videos are needed severely limits the kind of musical recordings that one could analyze.

This chapter presents a method to recuperate the fingering to an audio recording of a violin repertoire in a polyphonic mixture by analyzing the audio recording and finding an ergonomic fingering that captures the timbre unique to the recording as expressed through unique fingerboard location sequence. The proposed method is a two-step procedure. First, it analyzes the audio signal to estimate the fingerboard location sequence. Second, it determines an ergonomic fingering that satisfies the fin-

gerboard location sequence. When analyzing the fingerboard location sequence, the method uses features that are robust in a polyphonic mixture. Moreover, the method uses feature adaptation mechanism to improve the robustness when the instrument sounds different from one used for training. For ergonomic fingering decision, the method incorporates a cost function that reflects practices of violin playing (that might not be necessarily applicable to other stringed instruments such as the guitar, or the cello). The performance of the fingerboard location estimator and the fingering decision method are evaluated.

A.1 Violin fingering: a primer

Here, the essences of violin fingering and its terminologies are briefly reviewed, as they play important roles in understanding the proposed method. The left hand defines the pitch and the string on which a note is played. While there are two defining factors in fingering (finger and string), violinists often talk about fingering also in terms of the left hand position (“position”), the general placement of the left hand required to play a given note in a certain string using certain finger. As shown in Fig. A.1, this chapter labels the string on which a note is played with the variable s , where $s = 1$ is the lowest-tuned string and $s = 4$ is the highest. These strings are conventionally tuned seven semitones apart, where $s = 3$ is typically tuned to A4=440Hz. The strings are typically notated in a music score using a Roman numeral, where the lowest string ($s = 1$) is associated with IV, and the highest ($s = 4$) with I. Placed fingers are labeled by numbers, where 1 refers to the index finger, 2 the middle, 3 the ring, and 4 the little finger. The number 0 refers to an open string, *i.e.*, no finger is pressed and the string vibrates at the tuned pitch; only four pitches on a violin can be played as an open string. In this chapter, the position is defined by the number of semitones that the first finger must traverse from the *nut* (the ridge over which the string passes on the end of the fingerboard near the tuning pegs) in order to play a particular fingering.

Finger placement is determined by considering technical ease and musical effect. Using a certain finger (*e.g.*, index finger instead of the little finger) facilitates the execution of some musical effects related to pitch, such as a smooth transition between two notes (*glissando*), or a low-frequency modulation of a note (*vibrato*). At the same time, some sequences of finger placement are easier to execute than others. For example, rapid movement of the little finger is considerably more tiresome than that of the index finger. The choice of string is determined by considering the consistency of sonority. For example, since each string has a distinct sonority, violinists often play on one string to prevent abrupt changes of the sound quality. The difference in sonority given a pitch played on two different strings results from (1) the differences in the physical attributes of each string itself, such as the diameter, the tension and the material, (2) differences in the position along the string at which the finger must be placed

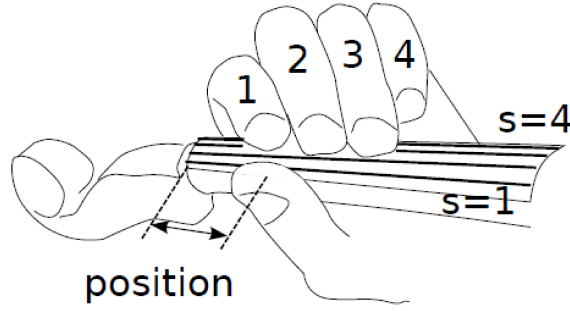


Figure A.1: The essences of violin fingering.

(since each string is tuned differently), and (3) in cases when the pitch is available as an open string, the difference between the rigid termination provided by the nut and the soft termination provided by the finger.

The choice of string is the one factor in fingering that is most likely to produce audible differences. By contrast, the position is sometimes chosen for visual effect and to show off the violinist’s technical skills. For example, violinist may present a “flashy” playing style through a wide change of position.

The consistency of sonority offered by playing on one string and the consistency of position that facilitates playing are often at a trade-off. In a fast piece, some abrupt changes of sonority caused by a certain fingering may be overlooked if it simplifies playing. On the other hand, in a slow piece, a violinist may choose a difficult fingering that produces certain sonority. A violinist may, for example, value consistency of sonority in a slow “singing” (*cantilena*) passage by playing on one string.

A.2 Formulation

The proposed method is based on determining the fingerboard location sequence from a polyphonic audio mixture that contains a musical phrase for solo violin, and finding a sensible fingering that satisfies the estimated fingerboard location sequence. The method assumes that the violin plays a monophonic melody using normal bowing technique (*i.e.*, this chapter ignores extended playing technique such as plucking (*pizzicato*), striking the string with the wooden side of the bow (*col legno*) or playing very close to the bridge to produce a shimmering sound (*sul ponticello*)).

The proposed method needs to address three issues:

1. Designing a feature that performs well in a polyphonic mixture.
2. Designing a scheme such that the performance does not degrade from an acoustical mismatch with the training data, whether because of characteristics of the violins, the rooms, or the recording process.

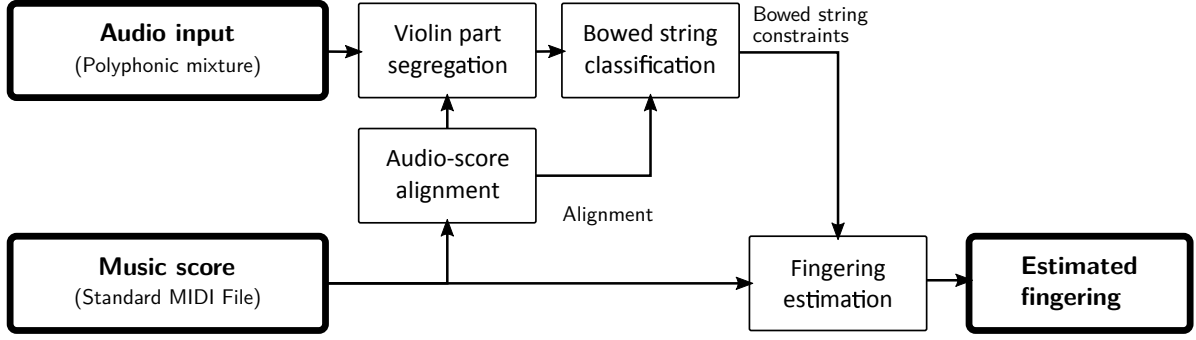


Figure A.2: Block diagram of the proposed method.

3. Designing a fingering model that reflects the practices of violin performance.

The first problem is tackled by fitting a sum of sinusoids plus noise model to the observed spectrum, and using the estimated harmonic parameters as the feature. The second problem is attacked by normalizing the average features of the training set and doing the same for the recording whose fingerboard location sequence needs to be classified. The third problem is tackled by introducing a new model of violin fingering, the *violin pedagogical model*, which incorporates features that are inspired by practices of violin playing. The block diagram of the proposed method is shown in Fig. A.2.

A.2.1 Fingerboard location sequence estimation through Viterbi alignment

In order to estimate a sensible fingering from the audio, the fingerboard location sequence must be estimated from a polyphonic mixture of sound that contains a violin melody. Since the string on which a note is played (the *bowed string*) is the primary factor that influences the timbre for different fingerboard locations, the first goal is to estimate the sequence of bowed strings. The estimation method is based on a bowed string classifier using features that are robust to polyphonic accompaniments, and classifier that takes into account the playability of a particular fingerboard location sequence.

A.2.2 Feature extraction by harmonic model fitting

The method first extracts, from a polyphonic audio mixture that contains a violin melody, the relative strength of each overtone that constitutes the harmonics of the violin (up to the N -th partial, where $N = 10$). The parameter is dependent on the material property of the string, the body resonance characteristics, and on how the instrument is played, *e.g.*, bow force, bow velocity, and the contact point of the bow and the string [39]. The feature extraction involves fitting a sum of harmonically spaced

sinusoids onto the observed short-time Fourier transform (STFT) representation of the input audio signal, y_{tf} , where f is the frequency bin, taking on values between 1 and F , and t is the frame index, taking on values between 1 and T . Because the input contains not only harmonic sound of the violin but also transients of the violin and accompaniments, harmonic sound of the violin part must be segregated. This is achieved by generating a time-frequency mask that passes the harmonic component of the violin part, based on both the observed spectrogram, the music score given as a standard MIDI file (SMF), and the alignment between the SMF and the audio signal, obtained using the audio-to-score method proposed in Chapter 3. As a signal processing front-end, the DC component is suppressed and the high-frequency component is emphasized by applying a two-tap high-pass filter with filter coefficients $[1, -1]$. Then, the proposed method iterates the following steps for a fixed number of iterations:

1. Update the time-frequency mask to segregate the harmonic components of the violin part, using the audio-score alignment and the estimated fundamental frequency as the cue.
2. Estimate the fundamental frequency of the violin melody, taking into account the pitch notated on the score, the observed spectrogram, and the audio-score alignment.

In the first step, the method generate a time-frequency mask, which approaches 1 for time-frequency bins that contain the violin sound and 0 otherwise. The method incorporates the information of the SMF and the audio-score alignment to find which bins are expected to contain the sound of the violin. Audio-score alignment gives the time-sequence of the notated pitch of the violin part, $\hat{f}_{0,t}$, for all time t . Given this, the method simultaneously segregate the violin part and estimate the fundamental frequency of the violin part, $f_{0,t}$. The method employs an approach similar to [67]. Namely, the method first segregates the violin part, by applying a mask that resembles a comb-filter to the spectrum centered about $\hat{f}_{0,t}$. Using the segregated signal, the method then re-estimates the fundamental frequency. Then, the comb-like mask is applied again but this time centered about the refined fundamental frequency. Repeating these steps refines the estimate of $f_{0,t}$ until it converges to a local optimum.

The method first segregates the violin part by separating the signal into two sub-signals – the violin part and the residual (*i.e.*, the rest of the signal). Assume the likelihood of observing the violin part at frequency f is $h_{v,f,f_{0,t}}$, and is of the following form:

$$h_{v,f,f_{0,t}} = \sum_{n=1}^N \pi_n \mathcal{N}(f | n f_{0,t}, \sigma_f^2). \quad (\text{A.1})$$

Here, the model expresses up to the N th partial. Moreover, assume that the likelihood of observing the rest is $h_{R,f}$ and is of form $\frac{1}{F}$. $\pi_t = \{\pi_{tn}\}$ is a multinomial variable that

indicates the relative strength of the n th partial. The likelihood of observing the violin part at frame t is assumed to be α_t , and the residual is $1 - \alpha_t$. $\{\alpha_t\}_{t=1}^T$ is denoted α . For each frequency bin, a latent variable $\mathbf{u} = \{u_{tf}\}$ is associated, which indicates whether the violin or the rest contributed to the observed power. $u_{tf} = 1$ indicates that at time t , the power contained in frequency f originated from the violin part, and $v_{f tn} = 1$ indicates that of frames generated from the violin, it originated from the n th partial of the violin part. The joint likelihood of the observed signal and the latent variable is then given as follows:

$$p(\mathbf{y}, \mathbf{z} | \mathbf{f}_0, \alpha, \pi) = \prod_{f=1, t=1}^{F, T} \left[\left((1 - \alpha_t) h_{R, f} \right)^{y_{tf}(1-u_{tf})} \times \prod_{n=1}^N \left(\alpha_t \pi_n \mathcal{N}(f | n f_{0, t}, \sigma_f^2) \right)^{y_{tf} u_{tf} v_{f tn}} \right]. \quad (\text{A.2})$$

This model is estimated using the Expectation-Maximization (EM) algorithm, in a manner similar to an existing work . In the E-step, the parameter from the previous step is used to find the distribution of $\mathbf{u}$ and $\mathbf{v}$ given $\mathbf{y}$ and the parameters. Namely, the algorithm assigns the following:

$$\bar{u}_{tf} = p(u_{tf} = 1 | \mathbf{y}, \mathbf{f}_0, \alpha, \pi) = \frac{\alpha_t \sum_{n=1}^N \pi_{tn} \mathcal{N}(f | n f_{0, t}, \sigma_f^2)}{(1 - \alpha_t) h_{R, f} + \alpha_t \sum_{n=1}^N \pi_{tn} \mathcal{N}(f | n f_{0, t}, \sigma_f^2)}, \quad (\text{A.3})$$

$$\bar{v}_{f tn} = p(v_{f tn} = 1 | u_{tf} = 1, \mathbf{y}, \mathbf{f}_0, \alpha, \pi) = \frac{\pi_{tn} \mathcal{N}(f | n f_{0, t}, \sigma_f^2)}{\sum_{\tilde{n}=1}^N \pi_{t\tilde{n}} \mathcal{N}(f | \tilde{n} f_{0, t}, \sigma_f^2)}. \quad (\text{A.4})$$

In the M-step, the fundamental frequency and other parameters are updated. To do this, the knowledge that the fundamental frequency does not deviate too much from the notated pitch is incorporated, by incorporating a prior distribution of the fundamental frequency as $p(f_{0, t} | \hat{f}_{0, t}) = N(f_{0, t} | \hat{f}_{0, t}, \sigma_{\hat{f}}^2)$. Then, the M-step yields in

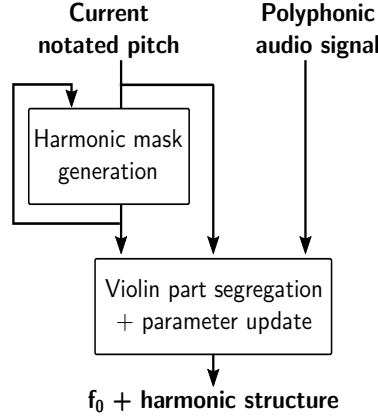


Figure A.3: Block diagram of feature extraction step.

the following updates:

$$f_{0,t} = \frac{\sigma_f^2 \hat{f}_0 + \sigma_{\hat{f}}^2 \sum_{f=1}^F \sum_{n=1} y_{tf} \bar{u}_{tf} \bar{v}_{f tn} n f}{\sigma_f^2 + \sigma_{\hat{f}}^2 \sum_{f=1}^F \sum_{n=1} y_{tf} \bar{u}_{tf} \bar{v}_{f tn} n^2}, \quad (\text{A.5})$$

$$\pi_{tn} = \frac{\sum_{f=1}^F y_{tf} \bar{u}_{tf} \bar{v}_{f tn}}{\sum_{f=1}^F y_{tf} \bar{u}_{tf}}, \quad (\text{A.6})$$

$$\alpha_t = \frac{\sum_{f=1}^F y_{tf} \bar{u}_{tf}}{\sum_{f=1}^F y_{tf}}. \quad (\text{A.7})$$

A block diagram of this method is shown in Fig. A.3 .

Finally, the logarithm of π_t is taken to enable feature adaptation, as will be discussed. Furthermore, the log-relative strength is decorrelated using the Discrete Cosine Transform (DCT). That is, $\pi_t = \text{DCT}(\log \pi_t)$.

On inharmonicity

Stringed instruments such as the violin do not produce overtones that are exact integer multiples of the fundamental frequency. Such deviation from integer harmonics is caused by the *inharmonicity* of the violin string, which in turn is created due to torsion of the string. Because inharmonicity is dependent on the material of the string, the proposed model initially incorporated inharmonicity, using Beta distribution with small shape parameter as its prior distribution. It was found that such model produced lower accuracy than that without inharmonicity. Because of the nature of the EM algorithm, it was found that the model tends to “explain” partials of the accompaniment as arising from the violin sound with extremely high inharmonicity.

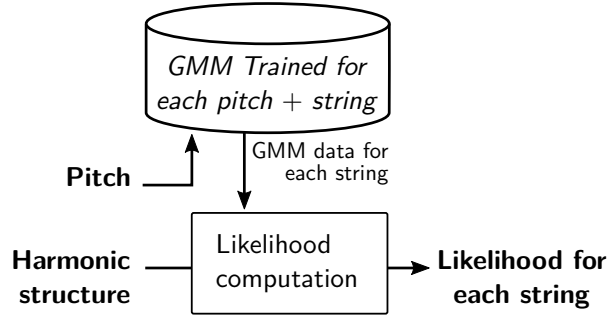


Figure A.4: Block diagram of the GMM.

A.2.3 Bowed string identification

Once the parameters $\Theta_t = \{f_{0,t}, \pi_t\}$ are extracted, the likelihood of the observed data is evaluated for each bowed string, *i.e.*, the probability at each point in time that the violinist is bowing a given string.

The likelihood of bowed string s_i is modeled using a Gaussian mixture model (GMM). The GMM models the density of π for each pitch $p_t = \text{pi}(f_{0,t})$, where $\text{pi}(f)$ converts frequency f into the closest MIDI note number. Let φ_{ipj} , μ_{ipj} , and Σ_{ipj} indicate respectively, the weight, the mean, and the covariance, of the j th component of the GMM for the i th bowed string at pitch p . Then, the likelihood of observing bowed string i given the feature Θ_t and the GMM parameters $\Theta_{GMM} = \{\varphi_{ipj}, \mu_{ipj}, \Sigma_{ipj} | \forall i, j, p\}$ is given as follows:

$$p(s = i | \pi_t, \Theta_{GMM}(p_t)) = \sum_j \varphi_{ip_t j} \mathcal{N}(\pi_t | \mu_{ip_t j}, \Sigma_{ip_t j}). \quad (\text{A.8})$$

Θ_{GMM} is trained using violin audio examples with various playing styles, each of which has a known sequence of pitch and fingerboard position. Using examples with a variety of playing styles has the effect of averaging out statistical discrepancies of the feature arising from playing the example in a particular playing style. Hence, different parameters in GMM corresponds to timbral difference arising from variety of fingerboard position and pitch, and not playing style, *e.g.*, bow pressure or bow velocity. Fig. A.4 shows the block diagram of the GMM.

The log-likelihood of the bowed string for the k th note of a piece, then, is a sum of bowed string likelihoods for all audio frames that play the k th note. Suppose the k th note on the score is played between frame St_k to Ed_k . Then, the bowed string likelihood of string s at note k , ρ_{ks} , is given as follows:

$$\rho_{ks} = \sum_{t=St_k}^{Ed_k} \log p(s | \pi_t, \theta_{GMM}(p_t)). \quad (\text{A.9})$$

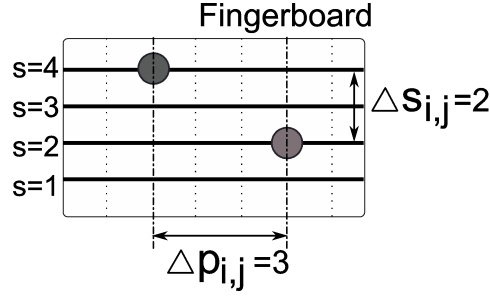


Figure A.5: Simplified horizontal-vertical model used in the Viterbi algorithm.

Sequence estimation using the Viterbi algorithm

The fingerboard location sequence is estimated by finding the most likely bowed string sequence, taking into account the observed acoustic signal and the difficulty of traversing from one fingerboard location to another. To incorporate both the likelihood of a given bowed string sequence and the likelihood of observing the features given a particular fingerboard position that a note is played, the bowed string sequence is modeled as an HMM.

The string the note is played on is treated as the hidden state that needs to be estimated, given the observed features π_t , fundamental frequency $f_{0,t}$, and the inherent difficulty from traversing from one fingerboard position to another. The observation likelihood is given readily as in Equation A.9. On the other hand, the transition probability between two strings $p(s_i|s_j; v)$ is dependent on both which strings are being crossed, and where the fingers must be placed on the fingerboard. Thus, the method expresses the transition pdf as a parametrized form of (1) the differences in the string number, and (2) the differences in the fingerboard position, in spirit of a simplified violin fingering model. The transition pdf, as shown in Fig. A.5, is then given as the following form:

$$p(s_i|s_j; v) \propto \exp\left(-v\left(\left(\frac{\Delta p_{i,j}}{7}\right)^2 + \Delta s_{i,j}^2\right)\right). \quad (\text{A.10})$$

Here, $\Delta p_{i,j}$ is the amount of change between fingerboard positions and $\Delta s_{i,j}$ is the amount of change between string numbers. The constant 7 is an empirically determined constant that models the violinists' tendencies to finger an interval of up to 7semitones by either playing on the same string or crossing a string but tend to finger an interval greater by crossing a string.

Then, the most likely bowed string is determined as follows:

$$\hat{S} = \arg \max_{S=\{s_0, \dots, s_K\}} \prod_{k=1}^K p(s_k|s_{k-1}) \exp(\rho_{k,s_k}). \quad (\text{A.11})$$

Feature adaptation

Two violins typically sound very different because each violin has a unique body resonance characteristic. Therefore, it is essential for us to adapt the proposed model to violin with different body resonance characteristics.

Let $y_{1,t,f}$ and $y_{2,t,f}$ be the observed log-magnitude spectrum of two violins. Let $B_{1,f}$ and $B_{2,f}$ be the log-magnitude response of the bodies of the two violins, and $S_{t,f}$ be the log-magnitude spectrum of the bow-string interaction. Assuming that the acoustic property of bow-string interaction does not change significantly with different instrument, the following relations hold:

$$\begin{aligned} Y_{1tf} &= B_{1f} + S_{tf} \\ Y_{2tf} &= B_{2f} + S_{tf}. \end{aligned}$$

By taking the expectation, $Y_2 = Y_1 - \langle Y_1 \rangle_{p(S)} + \langle Y_2 \rangle_{p(S)}$. The proposed method assumes that Y_1 is the observed feature, and generate Y_2 , which is the feature when Y_1 is played on instrument with body resonance characteristic B_2 . $\langle Y_1 \rangle_S$ is obtained readily by taking the average of the observed signal, and $\langle Y_2 \rangle_S$ is obtained by summing the average feature of each bowed string, weighted by how often a particular string is played at a given pitch for the given music score. Because the features are linear transformations of the logarithm of the relative powers of harmonic peaks, the same rationale holds. Initially, the probability distribution of S is set as follows:

$$p(S|pitch, \beta) \propto \begin{cases} \exp\left(-\frac{pitch - pitch_0(S)}{\beta}\right) & pitch > pitch_0(S) \\ 0 & pitch \leq pitch_0(S). \end{cases} \quad (\text{A.12})$$

Here, $pitch$ is the played pitch, $pitch_0$ is the pitch of the open string played on string S , and β is a positive parameter that assigns a greater probability to lower fingerboard positions, which are more commonly used and somewhat easier for the violinist.

A.2.4 Fingering estimation

At the core of the proposed fingering estimation is the violin fingering model, which models the difficulty of particular fingering. The fingering is determined by designing a cost function between multiple states of hand position, which reflects how difficult it is for the hand to traverse from one to the other.

Violin fingering model

This method extends on the existing *Horizontal-vertical cost model* presented in the previous section. In particular, the presented method includes fingering practices that are unique to the violin, which this thesis calls the *violin pedagogical model*. It is mainly inspired by practices of violin fingering as suggested by Yampolsky or Flesch [38, 135].

Violin, in particular, is a small, unfretted bowed stringed instrument, making it susceptible to missed intonation. Hence, violin fingerings are often set such that weak finger (the little finger) is used sporadically, and other fingers in such a way that it is kept in a natural hand position.

A *fingering* $\mathbf{n}$ is defined as a 4-tuple (n, s, f, p) , where $n \in \mathbb{N}$ is the pitch in MIDI note number, $s \in \{1, 2, 3, 4\}$ is the bowed string, $f \in \{0, 1, 2, 3, 4\}$ is the pressed finger (0=open string, *i.e.*, no finger placed), and $p \in \mathbb{N}$ is the position. Let $\mathcal{F}$ be a set of all possible fingerings. Also, let $n(\mathbf{n})$ be a function that retrieves the pitch of $\mathbf{n} \in \mathcal{F}$, $s(\mathbf{n})$ the bowed string, $f(\mathbf{n})$ the finger, and $p(\mathbf{n})$ the position.

An *unnotated score* $S_u \in \mathcal{F}^M$ is defined as a M -tuple of notes, where M is the number of notes contained in the music score, and the i th element, $S_{u,i}$, is the i th note of the music score. A *bowed string constraint*, $c_{bow} \in \{1, 2, 3, 4\}^M$ associated with an unnotated score S_u is defined as a M -tuple, where the i th element indicates the bowed string for the i th note. Finally let a *notated score*, $S_n \in \mathcal{F}^M$, contain, for the i th element, the fingering for the i th note.

Then, fingering inference is formulated as finding the optimal notated score S_{opt} that satisfies both the note sequence of the unnotated score obtained using the SMF and the bowed string constraint, obtained from the estimated bowed string sequence. Namely, it defines cost functions for traversing a sequence of two notes or three notes, and finds the fingering with smallest net cost that satisfies the constraints.

Let $C_b(\mathbf{n}_i, \mathbf{n}_{i-1}) : \mathcal{F}^2 \rightarrow \mathbb{R}$ be a cost function defined over sequence of two notes. Then, the optimal fingering is a notated score F_{opt} that satisfies the following:

$$F_{opt} = \arg \min_{\mathbf{n}_1 \dots \mathbf{n}_M \in \mathcal{F}} \sum_{i=1}^M C_b(\mathbf{n}_i, \mathbf{n}_{i-1}) \text{ s.t. } \forall i, n(\mathbf{n}_i) = S_{u,i} \text{ and } s(\mathbf{n}_i) = c_{bow,i}. \quad (\text{A.13})$$

Let $\tau_p = 3$, $\delta_{i,j}$ be Kronecker's delta, and $1(c)$ be a function that is 1 if condition c is true and 0 otherwise. Furthermore, let $p_0(s)$ denote the pitch of the open string of string s . Moreover, let us define the following quantities for convenience:

$$P_{\max} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 4 & 6 \\ 0 & 2 & 0 & 2 & 5 \\ 0 & 4 & 2 & 0 & 3 \\ 0 & 6 & 5 & 3 & 0 \end{pmatrix}; P_{\text{nat}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 3 & 5 \\ 0 & 2 & 0 & 1 & 2 \\ 0 & 3 & 1 & 0 & 2 \\ 0 & 5 & 2 & 2 & 0 \end{pmatrix}; P_{\min} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2.1 & 3.5 \\ 0 & 1 & 0 & 1.1 & 2.5 \\ 0 & 2.1 & 1.1 & 0 & 1.5 \\ 0 & 3.5 & 2.5 & 1.5 & 0 \end{pmatrix}. \quad (\text{A.14})$$

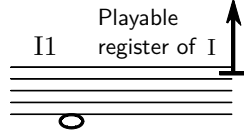


Figure A.6: Graphical description of $x_{BN,1}$. Roman numeral indicates the string (IV=lowest, I = highest), and number indicates the pressed finger (1=index finger, 4=little finger).

$$\Delta_p(i, j) = p(\mathbf{n}_i) - p(\mathbf{n}_j) \quad (\text{A.15})$$

$$\Delta_s(i, j) = s(\mathbf{n}_i) - s(\mathbf{n}_j) \quad (\text{A.16})$$

$$\Delta_f(i, j) = f(\mathbf{n}_i) - f(\mathbf{n}_j) \quad (\text{A.17})$$

$$\Delta_{pr}(i, j) = |n(\mathbf{n}_i) - p_0(s(\mathbf{n}_i))| - |n(\mathbf{n}_j) - p_0(s(\mathbf{n}_j))| \quad (\text{A.18})$$

$$R(i, j) = [P_{\min, f(\mathbf{n}_i), f(\mathbf{n}_j)}, P_{\max, f(\mathbf{n}_i), f(\mathbf{n}_j)}] \quad (\text{A.19})$$

Then, a two-note generic fingering model (denoted HVM-2) is defined as the following set of features x_{BN} :

$$b_0 = |\Delta_p(i, i-1)| \cdot \mathbf{1}(f(\mathbf{n}_i) \neq 0 \wedge f(\mathbf{n}_{i-1}) \neq 0) \quad (\text{A.20})$$

$$b_1 = \mathbf{1}(f(\mathbf{n}_i) \neq 0 \wedge \Delta_f(i, i-1) = 0 \wedge \Delta_p(i, i-1) \neq 0) \quad (\text{A.21})$$

$$b_2 = \mathbf{1}(\Delta_f(i, i-1) \cdot \Delta_{pr}(i, i-1) < 0 \wedge |\Delta_p| < \tau_p) \quad (\text{A.22})$$

$$b_3 = |\Delta_{pr}(i, i-1)| - P_{nat, f(\mathbf{n}_i), f(\mathbf{n}_{i-1})} \cdot \mathbf{1}(|\Delta_{pr}(i, i-1)| \in R(i, i-1)) \quad (\text{A.23})$$

$$x_{BN,1} = \begin{cases} 0 & n(\mathbf{n}_i) - p_0(s(\mathbf{n}_i)) - p(\mathbf{n}_i) \in [P_{\min, 1, f(\mathbf{n}_i)}, P_{\max, 1, f(\mathbf{n}_i)}] \\ \infty & \text{otherwise} \end{cases} \quad (\text{A.24})$$

$$x_{BN,(2,3)} = (b_0, b_0^2) \quad (\text{A.25})$$

$$x_{BN,4} = b_1 \quad (\text{A.26})$$

$$x_{BN,(5,6)} = (|\Delta_s(i, i-1)|, |\Delta_s(i, i-1)|^2) \cdot \mathbf{1}(f(\mathbf{n}_i) \cdot f(\mathbf{n}_{i-1}) \neq 0) \quad (\text{A.27})$$

$$x_{BN,7} = \mathbf{1}(|\Delta_{pr}(i, i-1)| - P_{nat}(f(\mathbf{n}_i), f(\mathbf{n}_{i-1})) \neq 0) \quad (\text{A.28})$$

$$x_{BN,8} = b_2 \cdot \mathbf{1}(f(\mathbf{n}_i) \neq 0 \wedge f(\mathbf{n}_{i-1}) \neq 0) \quad (\text{A.29})$$

$$x_{BN,9} = b_3 \cdot \mathbf{1}(f(\mathbf{n}_i) = 0 \vee f(\mathbf{n}_{i-1}) = 0). \quad (\text{A.30})$$

$x_{BN,1}$ determines whether a note is physically playable on a particular string, *i.e.*, this feature discards any fingering that is physically unplayable, as shown in Fig. A.6. $x_{BN,(2,3)}$ applies penalty for change of position. $x_{BN,4}$ penalizes the change of position using the same finger (*i.e.*, a glissando). $x_{BN,(5,6)}$ adds penalty for change of fingerboard location. $x_{BN,7}$ penalizes playing in an unnatural hand position. $x_{BN,8}$ penalizes playing a sequence in which the second note is placed higher on the finger-

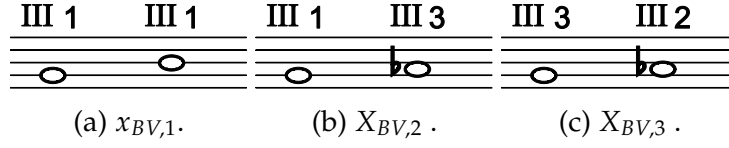


Figure A.7: Graphical descriptions of some features.

board, but the finger traverses from high finger to low (e.g., little finger to the index finger), and vice versa. Finally, $x_{BN,9}$ prevents change of position (wrist movement), when it is completely natural not to do so.

Furthermore, a two-note fingering model inspired by literatures in violin pedagogy (denoted VPM-2) is defined as the following set of features x_{BV} :

$$c = f(n_i) \cdot f(n_{i-1}) \neq 0 \wedge |\Delta_p(i, i-1)| < \tau_p \quad (\text{A.31})$$

$$x_{BV,1} = 1(c \wedge \Delta_{pr}(i, i-1) \neq 0 \wedge \Delta_f(i, i-1) = 0) \quad (\text{A.32})$$

$$x_{BV,2} = 1(c \wedge |\Delta_{pr}(i, i-1)| = 1 \wedge |\Delta_f(i, i-1)| > 1) \quad (\text{A.33})$$

$$x_{BV,3} = 1(f(n_i) \cdot f(n_{i-1}) \neq 0 \wedge \Delta_p(i, i-1) \neq 0 \wedge \Delta_{pr}(i, i-1) \cdot \Delta_f(i, i-1) < 0) \quad (\text{A.34})$$

$$x_{BV,4} = 1(f(n_i) = f(n_{i-1}) = 4 \wedge |n(n_i) - n(n_{i-1})| > 1) \quad (\text{A.35})$$

$$x_{BV,5} = p(n_i) \quad (\text{A.36})$$

$$x_{BV,(6,7,8)} = (\delta_{2,p(n_i)}, \delta_{5,p(n_i)}, \delta_{1,f(n_i)}). \quad (\text{A.37})$$

$x_{BV,1}$ penalizes using the same finger to move to a different position, as shown in Fig. A.7a. $x_{BV,2}$ penalizes chromatic change of the relative position using nonadjacent finger, as shown in Fig. A.7b; such movement involves wrist movement, which is considerably more harder to tune than placing a finger with slight stretch. $x_{BV,3}$ lessens the penalty for shifting up from higher-numbered finger to lower-numbered finger, or vice versa, when the shift is small, as shown in Fig. A.7c. $x_{BV,4}$ penalizes an adjacent use of the little finger, which is weak and prone to intonation errors. $x_{BV,5}$ adds penalty to extremely high position, which is harder to play in tune. $x_{BV,(6,7)}$ adds a preference to play in the first, second or third position, which are typically the three easiest positions to play in. Finally, $x_{BV,8}$ penalizes the half-position, the lowest possible position but an unconventional one. Thus, while it is perhaps the easiest position to play in tune from a physical perspective, it is unnatural for violin playing.

Using these features, the two-note cost function is defined as follows:

$$C_b = \text{diag}(\mathbf{W}_{BN} \mathbf{W}_{BV}) [\mathbf{x}_{BN} \mathbf{x}_{BV}]^T, \quad (\text{A.38})$$

where $\mathbf{W}_{BN}$ and $\mathbf{W}_{BV}$ define the relative weight of each feature. Given ample training data, statistical machine learning of these weights might be possible. In this chapter, the weights are adjusted manually, since the number of violin fingering corpus is too small to permit statistical machine learning.

A.3 Experiments

Three experiments are conducted to evaluate the proposed method. The first experiment assesses the performance of feature-adaptation and error-correction algorithms on bowed string estimation, the second experiment assesses the performance of the bowed string classifier in a polyphonic mixture, and the third experiment evaluates the playability of the proposed fingering model through subjective experiments. In the first two experiments, recording of three pieces of classical music were prepared, using two significantly different fingerings (207 notes for three pieces, yielding a total of 414 notes). For each fingering, the pieces were recorded in the following three conditions:

1. The same violin and strings used to record the training data (denoted as setup *SS*)
2. Same violin but with different brand of strings (setup *SD*)
3. Different violin with different brand of strings (setup *DD*)

This results in about eighteen minutes of audio files for validation. The training data, which lasted about 24 min, consisted of two-octave chromatic scales played on each string of an electric violin, using various dynamics.

Setup *SS* and *SD* were played on the same electric violin, and *DD* on an acoustic violin. There are two reasons why the training data was recorded using an electric violin and setup *DD* was recorded with an acoustic violin. First, it is easy to record a noise-free training data using an electric violin. Second, because electric and acoustic violins sound extremely different, the evaluation of feature adaptation mechanism can be regarded as the worst-case performance. Therefore, in real-life application, it is expected that the proposed method performs somewhere between setup *SD* and *DD*, as long as the training data uses an acoustic violin. In the EM algorithm, σ_f^2 was set to 0.1 and $\sigma_{\hat{f}}^2$ was set to 50 at the start of the EM algorithm, and narrow down inversely with each iteration of the algorithm. The value β used in model adaption was set to 0.1.

A.3.1 Experiment 1: evaluation of bowed string identifier

The accuracy of the bowed string estimator was evaluated with and without feature adaptation, each time evaluating the accuracy (1) of a baseline, *i.e.*, without considering any sequential information, (2) considering sequential information using a previous study, and (3) considering sequential information using the proposed method.

The value of ν used for the Viterbi algorithm used in the bowed sequence estimation was set to experimentally, by evaluating the accuracy for various values of ν . Fig. A.8 shows the result when varying ν . The following evaluations used $\nu = 30$.

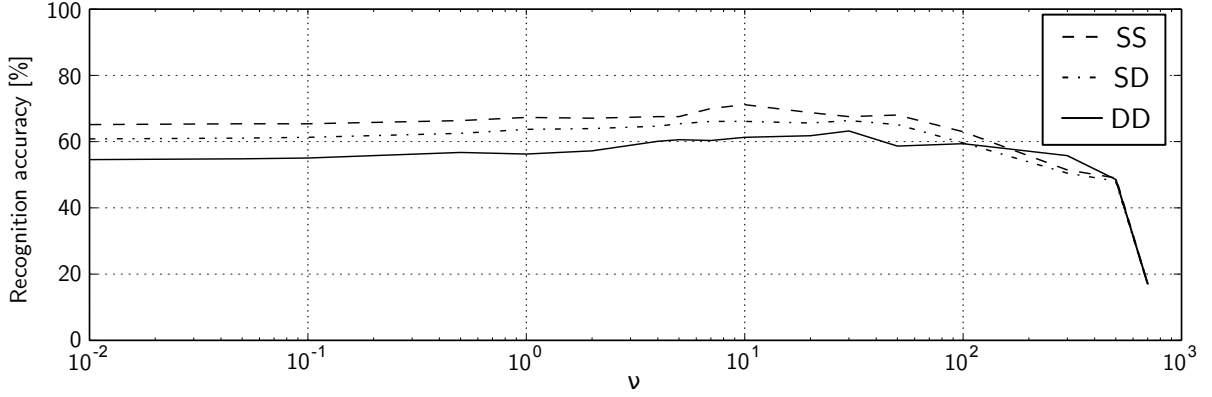


Figure A.8: Recognition accuracy for different values of ν (all values are for data with data adaptation). SS = same violin and same strings as training data; SD = same violin with different strings; DD = different violin and different strings.

Fig. A.9 compares the result obtained with the baseline methods and the proposed method. The results show that the proposed method consistently outperforms the baseline. In all cases, adaptation decreases the performance when the training data and the validation data are from the same instrument and the same brand of strings. This is because adaptation itself depends on the estimated sequence of fingerboard positions, which contains errors; the discrepancy between actual sequence of fingerboard position and the estimated one is small enough to be effective in absorbing the differences in body resonance characteristics between two different violins or strings, but significant when the same string and violin is used.

A.3.2 Experiment 2: effect of accompaniment on feature extraction

Piano accompaniments were generated for two pieces of the three used, using a synthesizer with uniform note velocity, and the amplitude was adjusted such that the peak values of the solo violin and the accompaniment were identical. Next, the recognition accuracy was evaluated for each violin/string type, by changing the level of accompaniment. The amplitude of the violin part was scaled such that the root mean-square (RMS) power of the violin part relative to that of the accompaniment is set to a given value of Signal-to-Accompaniment Ratio. Fig. A.10 shows the result. The result shows that the bowed string estimator performs similarly for different ratios, as long as the ratio is positive, *i.e.*, the signal level of the violin is greater than the “noise” level (*i.e.*, the accompaniment).

A.3.3 Experiment 3: evaluation of the estimated fingering

First, 10 excerpts from three classical pieces were prepared (approx. 200 notes total). For each excerpt, two kinds of fingerings were generated: one using HVM-2,

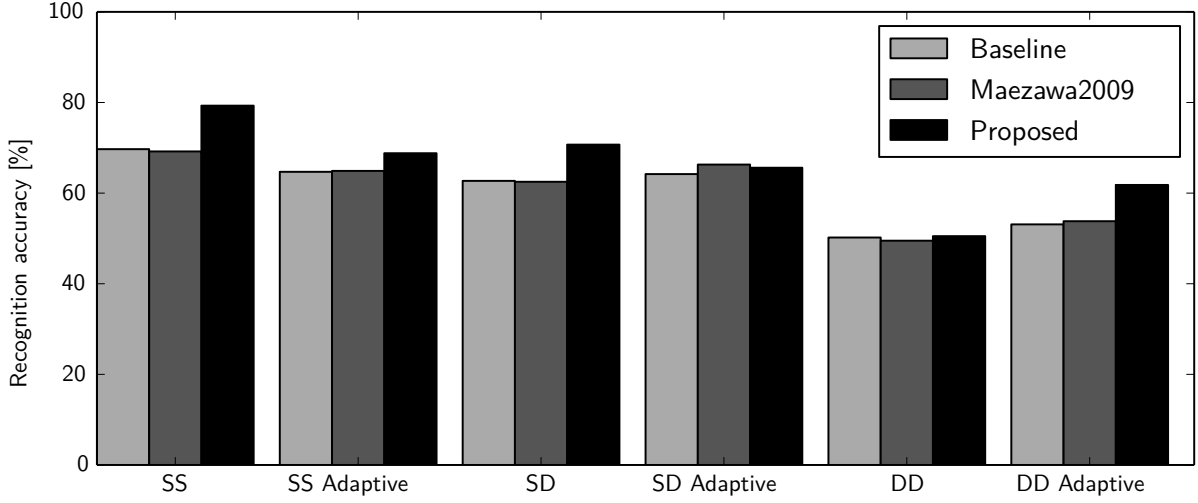


Figure A.9: Comparison of recognition accuracy for different sequence estimation methods. SS = same violin and same strings as training data; SD = same violin with different strings; DD = different violin and different strings.

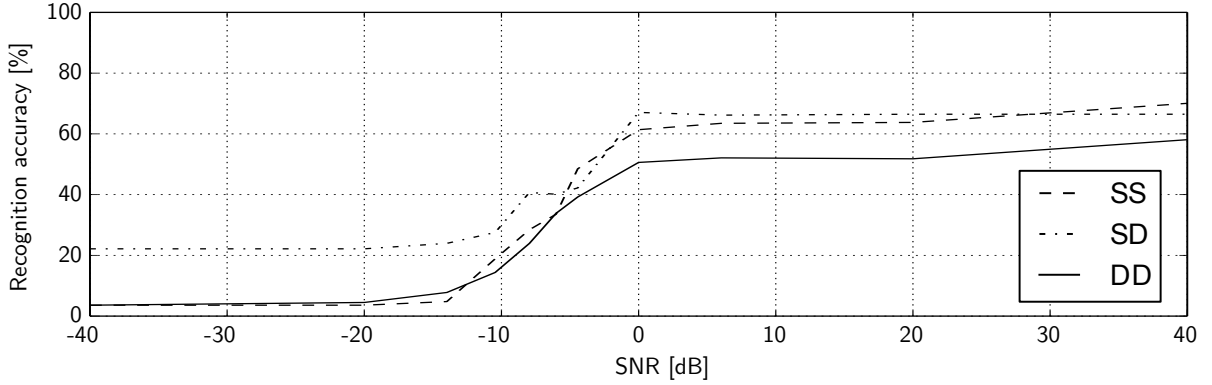


Figure A.10: Recognition accuracy with different levels of accompaniment.

and the other using HVM-2+VPM-2. Here, it is assumed that bowed string estimation is perfectly accurate, by not incorporating constraints based on the estimated bowed string. The feature weight $\mathbf{W}$ was manually tuned, by first adjusting the parameters for $\mathbf{W}_{BN}$ until it generated satisfactory fingerings for the repertoires considered. Then, parameters pertaining to $\mathbf{W}_{BV}$ were adjusted. They were set to $\mathbf{W}_{BN} = [1, 0.5, 1, 1, 5, 1, 0.5, 20, 10]$ and $\mathbf{W}_{BV} = [1, 10, -0.5, 10, 1.1, -0.3, -0.2, 0.2]$. Excerpts that generated different fingerings for each of the setups were then extracted.

Next, seven violinists of various skills (amateur and professional, 10+ years of experience) evaluated the generated fingerings. First, each violinist was presented with fingerings generated using VPM-2 and VPM-2+HVM-2, and was asked to choose the better of the two, if any. Table A.1 shows the result of the survey. Wilcoxon signed-rank test shows statistically significant difference between HVM-2 and VPM-2+HVM-2 ($p = 0.01$), suggesting that the proposed model is favored over the baseline.

Table A.1: Subjective evaluation of the proposed method.

Setup	Total count
HVM-2	5
HVM-2 + VPM-2	54
No preference	7

A.4 Discussion

The result in experiment 1 shows that in setup DD – the most realistic situation – adaptation improves the recognition accuracy whereas in SS and SD, the accuracy decreases. This occurs presumably due to the mismatch of the distribution of actual bowed string sequence to that assumed in the proposed model. Moreover, the improved performance of the proposed method demonstrates the effectiveness of considering the playability of a particular sequence of notes over a particular sequence of bowed string.

Experiment 2 suggests that the features used in the proposed method are robust in polyphonic audio mixture, as long as the signal level of the violin solo part is as loud as the accompaniment. This condition is reasonable when the violin has an important melody; in this kind of situation the choice of fingerboard location becomes even a greater musical issue. Therefore, the proposed method is presumably useful for practical applications involving works for violin and piano.

Finally, the estimated fingering using the proposed method was more natural than the baseline. The author speculates that incorporating the preference for the first and third positions (index finger placed 2 semitones and 5 semitones above the nut of the violin, respectively) is the chief reason - they are the most frequently used positions on the violin, and fingering generated on these positions are more natural for violinists to play. This work has shown that incorporating these kinds of heuristics could drastically improve the playability.

A.5 Application: de-mystifying historical recordings

The proposed method may be used to analyze a recording from the past to recover the fingering of a legendary violinist. In this example, the proposed method is used to recover the fingering of *Romanze in C* composed by Joachim, a legendary violinist of the late 19th century, using a recording on Gramophone played by Joachim himself in 1903. The recording has been pitch-shifted such that the notated A4 was set to 440 Hz. VPM-2+HVM-2 fingering model was used to estimate the fingering, an excerpt of which is shown in Fig. A.11.

The author speculates, however, that the actual fingering may be very different from that estimated by the proposed method, illuminating directions for future research. For example, there are notes that are playable using open strings (*i.e.*, no fin-

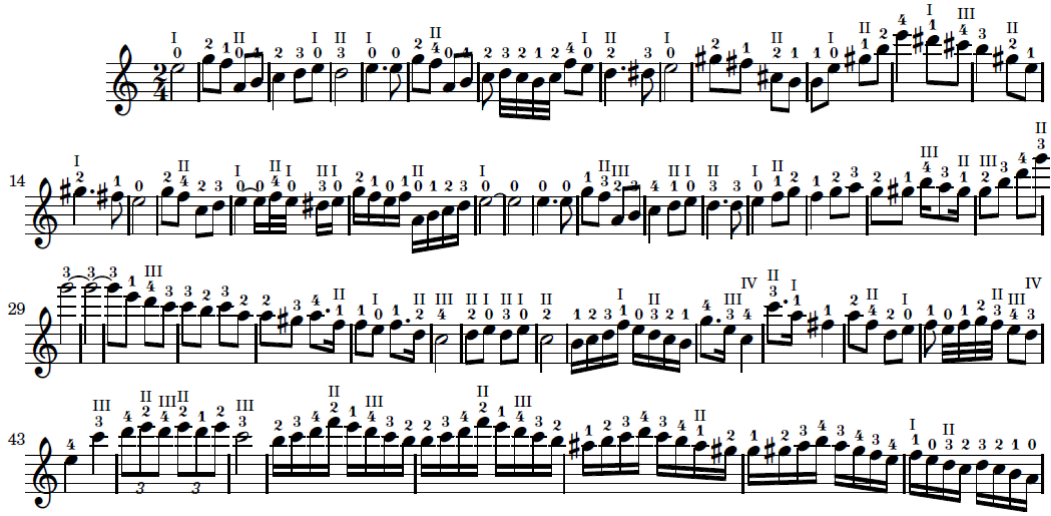


Figure A.11: Estimated fingering from first forty bars of Joachim's *Romanze in C* played by Joachim (1903).

ger pressed) whose estimated fingerings on the notated score show open strings; yet, on the recording, one clearly hears the note played with a *vibrato*, meaning that the pitch is played by stopping a string with a finger. These kinds of errors might have occurred for these three reasons:

1. In Joachim's times, the material used for the violin string was significantly different. In particular, the E-string (the highest pitched string) used gut as its core, which has a much mellower timbre than the kind of string used today. Such discrepancies in the material property of the string itself would cause the performance of the bowed string estimator to deteriorate.
2. A high quality recording of Joachim is long lost. The remastered recording we used had frequency range that extended up to only approximately 5 kHz; the rest was cut-off perhaps by applying an equalizer. Hence, there were only few partials that conveyed meaningful information.
3. Joachim's use of pitch-based playing techniques such as vibrato and glissando gives many clues to violinist who wants to infer Joachim's playing. For example, if the transition from one note to another is completely smooth, it strongly suggests that the two notes were played on the same string. Vibrato can be a give-away for distinguishing between a note that is played using an open string or not. The proposed method, however, does not incorporate the pitch trajectory for fingerboard position estimation, and hence, misses such clues.

This output suggests a future direction for research, namely, incorporating pitch-based cues for fingerboard inference, and perhaps robustness to different materials used for a same string. The latter, however, can be ameliorated with the proposed

method to a certain extent; the training data may be recorded using the string that is thought to be used in the recording whose fingering one wishes to infer.

A.6 Conclusion

This chapter presented a method for recovering the violin fingering from an input audio signal and a music score by analyzing the bowed string sequence, and using it as a constraint to determine the optimal fingering. Bowed string sequence classification was based on features that are robust in a polyphonic mixture. Moreover, a sequential model based on violin playing was introduced. It was shown that such sequential modeling drastically improved the accuracy. The fingering estimation incorporated features that are specific to violin playing practices (the pedagogical model), in addition to some of the more fundamental features applicable to other instruments as well. Incorporating such heuristics generated a drastically easier fingering.

Future direction may involve improved recognition accuracy, refining fingering model, and more applications. For example, recognition accuracy may be improved by exploiting the smoothness of pitch trajectory (glissando, vibrato, etc.), as it seems that violinists listen to the smoothness of pitch transition to infer the fingerboard location and the fingering. The fingering model may further be improved by incorporating more features that are inspired from violin pedagogy or through machine learning of feature weight by preparing a large violin fingering corpus. Another possibility is to perform joint estimation of fingerboard location and fingering – the problem of fingerboard location estimation and fingering estimation is dependent on each other, suggesting that joint estimation may improve the performance of both. From a musical perspective, artist classification, skill assessment, and analysis of historical recordings may be interesting applications of the proposed method.

Appendix B

List of Distributions

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sqrt{\frac{|\boldsymbol{\Lambda}|}{(2\pi)^D}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})\right), \mathbf{x} \in \mathbb{R}^D$$

$$\text{Gam}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$$

$$\mathcal{NG}(\boldsymbol{\mu}, \boldsymbol{\lambda}|\boldsymbol{\mu}_0, m, a, b) = \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \frac{1}{\lambda m}\right) \text{Gam}(\boldsymbol{\lambda}|a, b)$$

$$\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{\Lambda}_0, n) = \frac{|\mathbf{W}|^{\frac{n-D-1}{2}} \exp\left(\text{tr}\left(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\Lambda}\right)\right)}{2^{\frac{nD}{2}} |\boldsymbol{\Lambda}_0|^{\frac{n}{2}} \Gamma_D\left(\frac{n}{2}\right)}, \boldsymbol{\Lambda} \in \mathbb{R}^{D \times D}$$

$$\text{Discrete}(\mathbf{x}|\mathbf{w}) = \prod_{i=1}^I w_i^{x_i}, \mathbf{x} \text{ a 1-of-I binary vector}$$

$$\text{Beta}(x|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}$$

$$\text{Dir}(\mathbf{w}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_d \alpha_d)}{\prod_{d=1}^D \Gamma(\alpha_d)} \prod_{d=1}^D w_d^{\alpha_d-1}, \mathbf{w} \in \Delta^D$$

where $\Gamma_D(x) = \pi^{\frac{D(D-1)}{4}} \prod_{d=1}^D \Gamma\left(x + \frac{1-d}{2}\right)$

Bibliography

- [1] S. A. Abdallah, K. Noland, M. B. Sandler, M. A. Casey, and C. Rhodes. Theory and evaluation of a Bayesian music structure extractor. In *Proc. International Conference on Music Information Retrieval*, pages 420–425, 2005.
- [2] A. Abramson, E. A. P. Habets, S. Gannot, and I. Cohen. Dual-microphone speech dereverberation using GARCH modeling. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 4565–4568, 2008.
- [3] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustic Society of America*, 65(4):943–950, April 1979.
- [4] H. Attias. A variational Bayesian framework for graphical models. In *Proc. Advances in Neural Information Processing Systems*, pages 209–215, 2000.
- [5] H. Attias, J. C. Platt, A. Acero, and L. Deng. Speech denoising and dereverberation using probabilistic models. In *Proc. Advances in Neural Information Processing Systems*. MIT Press, 2001.
- [6] I. Barbancho. Transcription and expressiveness detection system for violin music. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 189–192, 2010.
- [7] M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University College London, 2003.
- [8] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden Markov model. In *Proc. Advances in Neural Information Processing Systems*, pages 577–584, 2001.
- [9] J. Benesty, M. M. Sondhi, and Y. Huang, editors. *Springer Handbook of Speech Processing*. Springer, Berlin, 2008.
- [10] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *BAyesian Analysis*, 1:121–144, 2005.

- [11] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [12] J. C. Brown and M. S. Puckette. An efficient algorithm for the calculation of a constant Q transform. *The Journal of the Acoustical Society of America*, 92(5):2698–2701, 1992.
- [13] N. J. Bryan and G. J. Mysore. An efficient posterior regularized latent variable model for interactive sound source separation. In *Proc. International Conference on Machine Learning*, pages 208–216, 2013.
- [14] A. M. Burns and M. M. Wanderly. Visual methods for retrieval of guitarist fingering. In *Proc. New Interfaces for Music Expression*, pages 196–199, 2006.
- [15] A. T. Cemgil, B. Kappen, P. Desain, and H. Honing. On tempo tracking: Tempogram representation and Kalman filtering. *Journal of New Music Research*, 29(4):259–273, 2000.
- [16] T. Cho, R. J. Weiss, and J. P. Bello. Exploring common variations in state of the art chord recognition systems. In *Proc. Sound and Music Computing Conference*, pages 1–8, 2010.
- [17] A. Cont. Realtime audio to score alignment for polyphonic music instruments, using sparse non-negative constraints and hierarchical HMMs. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume 5, page V, 2006.
- [18] A. Cont. A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):974–987, 2010.
- [19] J. Cooley and J. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.
- [20] R. B. Dannenberg and N. Hu. Polyphonic audio matching for score following and intelligent audio editors. In *Proc. International Computer Music Conference*, Sept. 2003.
- [21] R. B. Dannenberg and C. Raphael. Music score alignment and computer accompaniment. *CACM*, 49(8):38–43, Aug. 2006.
- [22] P. Desain and H. Honing. Does expressive timing in music performance scale proportionally with tempo? *Psychological Research*, 56(4):285–292, 1994.
- [23] J. Devaney and D. P. W. Ellis. Handling asynchrony in audio-score alignment. In *Proc. International Computer Music Conference*, pages 29–32, 2009.

- [24] O. Dikmen and C. Févotte. Maximum marginal likelihood estimation for non-negative dictionary learning in the Gamma-Poisson model. *IEEE Transactions on Signal Processing*, 60(10):5163–5175, 2012.
- [25] N. Ding and Z. Ou. Variational nonparametric Bayesian hidden Markov model. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 2098–2101, 2010.
- [26] S. Dixon. Live tracking of musical performances using on-line time warping. In *Proc. International Conference on Digital Audio Effects*, pages 92–97, 2005.
- [27] S. Dixon and G. Widmer. MATCH: A music alignment tool chest. In *Proc. International Conference on Music Information Retrieval*, London, UK, September 11-15 2005.
- [28] Z. Duan and B. Pardo. Aligning semi-improvised music audio with its lead sheet. In *Proc. International Conference on Music Information Retrieval*, pages 513–518, 2011.
- [29] Z. Duan and B. Pardo. Soundprism: An online system for score-informed source separation of music audio. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1205–1215, 2011.
- [30] Z. Duan and B. Pardo. A state space model for online polyphonic audio-score alignment. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 197–200, 2011.
- [31] J. Erkelens and R. Heusdens. Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1746–1765, Sept. 2010.
- [32] C. Evers and J. Hopgood. Multichannel online blind speech dereverberation with marginalization of static observation parameters in a Rao-Blackwellized particle filter. *Journal of Signal Process. Systems*, 63(3):315–332, 2011.
- [33] S. Ewert and M. Müller. Using score-informed constraints for NMF-based source separation. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 129–132, 2012.
- [34] S. Ewert, M. Müller, and P. Grosche. High resolution audio synchronization using chroma onset features. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 1869–1872, Taipei, Taiwan, 2009.
- [35] T. Falk, C. Zheng, and W.-Y. Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1766–1774, 2010.

- [36] S. Favaro, Y. W. Teh, et al. MCMC for normalized random measure mixture models. *Statistical Science*, 28(3):335–359, 2013.
- [37] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.
- [38] C. Flesch. *The Art of Violin Playing*, volume I. Carl Fischer, L.L.C., New York, NY, second ed. edition, 2000.
- [39] N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments*. Springer, second ed. edition, 1998.
- [40] C. Fremerey, F. Kurth, M. Müller, and M. Clausen. A demonstration of the Sync-Player system. In *Proc. International Conference on Music Information Retrieval*, pages 131–132, 2007.
- [41] C. Fremerey, M. Müller, and M. Clausen. Handling repeats and jumps in score-performance synchronization. In *Proc. International Conference on Music Information Retrieval*, pages 243–248, 2010.
- [42] C. Fremerey, M. Müller, F. Kurth, and M. Clausen. Automatic mapping of scanned sheet music to audio recordings. In *ISMIR*, pages 413–418, 2008.
- [43] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno. Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals. In *Proc. International Symposium on Multimedia*, pages 257–264, 2006.
- [44] T. Fujishima. Realtime chord recognition of musical sound: A system using Common Lisp Music. In *Proc. International Computer Music Conference*, pages 464–467, Beijing, 1999.
- [45] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93:27403, 1993.
- [46] N. D. Gaubitch, P. A. Naylor, and D. B. Ward. On the use of linear prediction for dereverberation of speech. In *Proc. of the International Workshop on Acoustics Echo and Noise Control*, pages 99–102, 2003.
- [47] Z. Ghahramani and G. E. Hinton. Variational learning for switching state-space models. *Neural computation*, 12(4):831–864, 2000.
- [48] M. Goto. A predominant-F0 estimation method for CD recordings: MAP estimation using EM algorithm for adaptive tone models. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 3365–3368, 2001.

- [49] M. Goto. Development of the RWC music database. In *Proc. International Congress on Acoustics*, volume I, pages 553–556, 2004.
- [50] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In *Proc. International Conference on Music Information Retrieval*, pages 287–288, 2002.
- [51] M. Grachten et al. Automatic alignment of music performances with structural differences. In *Proc. International Conference on Music Information Retrieval*, pages 607–612, 2013.
- [52] Y. Gu and C. Raphael. Modeling piano interpretation using switching Kalman filter. In *Proc. International Conference on Music Information Retrieval*, Porto, Portugal, October 8-12 2012.
- [53] E. A. P. Habets. *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*. PhD thesis, Technische Universiteit Eindhoven, 2007.
- [54] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, Apr. 1970.
- [55] R. Hennequin, J. J. Burred, S. Maller, and P. Leveau. Speech-guided source separation using a pitch-adaptive guide signal model. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 6672–6676, 2014.
- [56] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent Dirichlet allocation. In *Proc. Advances in Neural Information Processing Systems*, pages 856–864, 2010.
- [57] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.
- [58] N. Hu, R. B. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 185–188, 2003.
- [59] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American statistician*, pages 30–37, 2004.
- [60] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Simultaneous processing of sound source separation and musical instrument identification using Bayesian spectral modeling. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 3816 –3819, May 2011.

- [61] C. Joder, S. Essid, and G. Richard. A conditional random field viewpoint of symbolic audio-to-score matching. In *Proc. ACM Multimedia*, pages 871–874, 2010.
- [62] C. Joder, S. Essid, and G. Richard. An improved hierarchical approach for music-to-symbolic score alignment. In *Proc. International Conference on Music Information Retrieval*, pages 39–44, 2010.
- [63] C. Joder, S. Essid, and G. Richard. An improved hierarchical approach for music-to-symbolic score alignment. In *Proc. International Conference on Music Information Retrieval*, pages 39–44, Utrecht, The Netherlands, August 9-13 2010.
- [64] C. Joder and B. Schuller. Score-informed leading voice separation from monaural audio. In *Proc. International Conference on Music Information Retrieval*, pages 277–282, 2012.
- [65] M. Johnson and A. Willsky. Stochastic variational inference for Bayesian time series models. In *Proc. International Conference on Machine Learning*, pages 1854–1862, 2014.
- [66] H. Kameoka, T. Nakatani, and T. Yoshioka. Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 45–48, Apr. 2009.
- [67] H. Kameoka, T. Nishimoto, and S. Sagayama. A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):982–994, March 2007.
- [68] A. A. Kasimi, E. Nichols, and C. Raphael. A simple algorithm for automatic generation of polyphonic piano fingerings. In *Proc. International Conference on Music Information Retrieval*, pages 355–356, 2007.
- [69] B. King, P. Smaragdis, and G. J. Mysore. Noise-robust dynamic time warping using PLCA features. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 1973–1976, 2012.
- [70] J. F. C. Kingman. *Poisson processes*, volume 3. Oxford university press, 1992.
- [71] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi. Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):534–545, 2009.
- [72] A. Klapuri. *Signal processing methods for the automatic transcription of music*. Tampere University of Technology Finland, 2004.

- [73] A. Krishnaswamy and J. O. Smith. Inferring control inputs to an acoustic violin from audio spectra. In *Proc. New Interfaces for Music Expression*, pages 733–736, 2003.
- [74] K. Kumar, B. Raj, R. Singh, and R. Stern. An iterative least-squares technique for dereverberation. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 5488–5491, May 2011.
- [75] F. Kurth, M. Müller, C. Fremerey, Y.-h. Chang, and M. Clausen. Automated synchronization of scanned sheet music with audio recordings. In *Proc. International Conference on Music Information Retrieval*, pages 261–266, 2007.
- [76] K. Lebart and J. Boucher. A new method based on spectral subtraction for speech dereverberation. *ACUSTICA*, 87(3):359–366, May-Jun 2001.
- [77] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Proc. Advances in Neural Information Processing Systems*, pages 556–562, 2000.
- [78] K. Lee and M. Cremer. Segmentation-based lyrics-audio alignment using dynamic programming. In *Proc. International Conference on Music Information Retrieval*, pages 395–400, 2008.
- [79] H. Lu, B. Zhang, Y. Wang, and W. K. Leow. iDVT: an interactive digital violin tutoring system based on audio-visual fusion. In *Proc. ACM Multimedia*, page PAGES, 2008.
- [80] R. Macrae and S. Dixon. Accurate real-time windowed time warping. In *Proc. International Conference on Music Information Retrieval*, pages 423–428, 2010.
- [81] A. Maezawa, M. Goto, and H. G. Okuno. Query-by-conducting: An interface to retrieve classical-music interpretations by real-time tempo input. In *Proc. International Conference on Music Information Retrieval*, pages 477–482, 2010.
- [82] A. Maezawa, K. Itoyama, K. Yoshii, and H. G. Okuno. Bayesian audio alignment based on a unified model of music composition and performance. In *Proc. International Conference on Music Information Retrieval*, pages XX–XX, 2014.
- [83] A. Maezawa and H. G. Okuno. Audio part mixture alignment based on hierarchical nonparametric Bayesian model of musical audio sequence collection. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 5232–5236, 2014.
- [84] A. Maezawa, H. G. Okuno, T. Ogata, and M. Goto. Polyphonic audio-to-score alignment based on Bayesian latent harmonic allocation hidden Markov model. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 185–188, May 2011.

- [85] J. Merimaa, T. Peltonen, and T. Lokki. Concert hall impulse responses - Pori, Finland: Reference. Retrieved from <http://www.acoustics.hut.fi/projects/poririrs> on 1/27/2003, 2005.
- [86] R. Miotto, N. Montecchio, and N. Orio. Statistical music modeling aimed at identification and alignment. In *Proc. The International Workshop on Advances in Music Information Research*, pages 187–212, 2010.
- [87] M. Miron, J. J. Carabias-Orti, and J. Janer. Audio-to-score alignment at note level for orchestral recordings. In *Proc. International Conference on Music Information Retrieval*, pages 125–130, 2014.
- [88] M. Miyoshi and Y. Kaneda. Inverse filtering of room acoustics. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(2):145–152, Feb. 1988.
- [89] M. Molina-solana and G. Widmer. Evidence for pianist-specific rubato style in Chopin nocturnes. In *Proc. International Conference on Music Information Retrieval*, pages 225–230, 2010.
- [90] N. Montecchio and A. Cont. Accelerating the mixing phase in studio recording productions by automatic audio alignment. In *Proc. International Conference on Music Information Retrieval*, pages 627–632, 2011.
- [91] N. Montecchio and A. Cont. A unified approach to real time audio-to-score and audio-to-audio alignment using sequential Montecarlo inference techniques. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 193–196, 2011.
- [92] R. Mukai, S. Araki, and S. Makino. Separation and dereverberation performance of frequency domain blind source separation. In *Proc. Eurospeech*, pages 2599–2602, 2001.
- [93] M. Müller and S. Ewert. Towards timbre-invariant audio features for harmony-based music. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):649–662, 2010.
- [94] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition. In *Proc. International Conference on Language Resources and Evaluation*. European Language Resources Association, 2000.
- [95] T. Nakatani and M. Miyoshi. Blind dereverberation of single channel speech signal based on harmonic structure. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 92–95, 2003.

- [96] T. Nakatani, M. Miyoshi, and K. Kinoshita. One microphone blind dereverberation based on quasi-periodicity of speech signals. In *Proc. Advances in Neural Information Processing Systems*, pages 1417–1424. MIT Press, 2003.
- [97] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1717 –1731, Sept. 2010.
- [98] B. Niedermayer. Towards audio to score alignment in the symbolic domain. In *Proc. Sound and Music Computing Conference*, 2009.
- [99] B. Niedermayer and G. Widmer. A multi-pass algorithm for accurate audio-to-score alignment. In *Proc. International Conference on Music Information Retrieval*, pages 417–422, 2010.
- [100] A. V. Oppenheim and R. W. Schaffer. *Discrete-Time Signal Processing*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009.
- [101] N. Orio, S. Lemouton, and D. Schwartz. Score following : State of the art and new developments. In *Proc. New Interfaces for Music Expression*, pages 36–41, 2003.
- [102] T. Otsuka et al. Incremental Bayesian audio-to-score alignment with flexible harmonic structure models. In *Proc. International Conference on Music Information Retrieval*, pages 525–530, 2011.
- [103] T. Otsuka, K. Nakadai, T. Takahashi, T. Ogata, and H. G. Okuno. Real-time audio-to-score alignment using particle filter for co-player music robots. *EURASIP Journal of Advances in Signal Processing*, 2011(1):Article ID 384651, 13 pages, 2011.
- [104] T. Otsuka, T. Takahashi, H. G. Okuno, K. Komatani, T. Ogata, K. Murata, and K. Nakadai. Incremental polyphonic audio to score alignment using beat tracking for singer robots. In *International Conference on Intelligent Robots and Systems*, pages 2289–2296. IEEE, 2009.
- [105] C. Palmer. Music performance. *Annual review of psychology*, 48(1):115–138, 1997.
- [106] J. Paulus, M. Müller, and A. Klapuri. State of the art report: Audio-based music structure analysis. In *Proc. International Conference on Music Information Retrieval*, pages 625–636, Utrecht, The Netherlands, Aug. 2010.
- [107] P. Peeling, A. T. Cemgil, and S. Godsill. A probabilistic framework for matching music representations. In *Proc. International Conference on Music Information Retrieval*, pages 267–272, 2007.

- [108] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [109] D. P. Radicioni, L. Anselma, and V. Lombardo. A segmentation-based prototype to compute string instruments fingering. In *Proc. Conference on Interdisciplinary Musicology*, 2004.
- [110] A. Radisavljevic and P. F. Driessen. Path difference learning for guitar fingering problem. In *Proc. International Computer Music Conference*, 2004.
- [111] C. Raphael. Automatic segmentation of acoustic musical signals using hidden Markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):360–370, Apr 1999.
- [112] C. Raphael. A hybrid graphical model for aligning polyphonic audio with musical scores. In *Proc. International Conference on Music Information Retrieval*, pages 387–394, 2004.
- [113] S. Sako, R. Yamamoto, and T. Kitamura. Ryry: A real-time score-following automatic accompaniment playback system capable of real performances with errors, repeats and jumps. In *AMT*, pages 134–145, 2014.
- [114] C. S. Sapp. Comparative analysis of multiple musical performances. In *Proc. International Conference on Music Information Retrieval*, pages 2–5, 2007.
- [115] M.-A. Sato. Online model selection based on the variational Bayes. *Neural Computation*, 13(7):1649–1681, 2001.
- [116] S. I. Sayegh. Fingering for string instruments with the optimal path paradigm. *Computer Music Journal*, 13(3):76–83, 1989.
- [117] A. Sehr, R. Maas, and W. Kellermann. Frame-wise HMM adaptation using state-dependent reverberation estimates. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 5484–5487, 2011.
- [118] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1991.
- [119] P. Smaragdis and G. J. Mysore. “Separation by Humming”: User guided sound extraction from monophonic mixtures. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009.
- [120] P. Smaragdis and B. Raj. Shift-invariant probabilistic latent component analysis. *Journal of Machine Learning Research*, 2007.
- [121] P. Smaragdis, B. Raj, and M. Shashanka. A probabilistic latent variable model for acoustic modeling. *Advances in models for acoustic processing, NIPS*, 148, 2006.

- [122] D. Stowell and E. Chew. Maximum a posteriori estimation of piecewise arcs in tempo time-series. In *International Symposium on Computer Music Multidisciplinary Research*, pages 387–399, 2012.
- [123] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.
- [124] C. Traube and J. O. Smith. Estimating the plucking point on a guitar string. In *Proc. International Conference on Digital Audio Effects*, pages 153–158, 2000.
- [125] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, July 2006.
- [126] T. Virtanen, A. T. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008*, pages 1825–1828, 2008.
- [127] T. Virtanen and A. Klapuri. Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. *Proc. Advances in Neural Information Processing Systems*, 2006.
- [128] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy. Accelerated training of conditional random fields with stochastic gradient methods. In *Proc. International Conference on Machine Learning*, pages 969–976, New York, NY, USA, 2006. ACM.
- [129] E. A. Walsh. Freepats project. Retrieved from <http://freepats.zenvoid.org/> on 8/1/2013, 2013.
- [130] C. Wang, J. W. Paisley, and D. M. Blei. Online variational inference for the hierarchical Dirichlet process. *Journal of Machine Learning Research*, 15:752–760, 2011.
- [131] C. Wang, J. W. Paisley, and D. M. Blei. Online variational inference for the hierarchical Dirichlet process. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 752–760, 2011.
- [132] S. Wang, S. Ewert, and S. Dixon. Robust joint alignment of multiple versions of a piece of music. In *Proc. International Conference on Music Information Retrieval*, pages 83 – 88, 2014.
- [133] Y. Wang and L. Carin. Levy measure decompositions for the beta and gamma processes. In *Proc. International Conference on Machine Learning*, 2012.

- [134] G. Widmer, S. Dixon, W. Goebel, E. Pampalk, and A. Tobudic. In search of the Horowitz factor. *AI Magazine*, 24(3):111, 2003.
- [135] I. M. Yampolsky. *Principles of Violin Fingering*. Oxford University Press, 1967.
- [136] N. Yasuraoka, T. Yoshioka, T. Nakatani, A. Nakamura, and H. G. Okuno. Music dereverberation using harmonic structure source model and Wiener filter. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 53–56, 2010.
- [137] Y. Yonebayashi, H. Kameoka, and S. Sagayama. Automatic decision of piano fingering based on a hidden Markov model. In *Proc. International Joint Conference on Artificial Intelligence*, pages 2915–2921, 2007.
- [138] K. Yoshii and M. Goto. A nonparametric Bayesian multipitch analyzer based on infinite Latent Harmonic Allocation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):717–730, Mar. 2012.
- [139] T. Yoshioka. *Speech enhancement in reverberant environments*. PhD thesis, Kyoto University, 2010.
- [140] T. Yoshioka, H. Kameoka, T. Nakatani, and H. G. Okuno. Statistical models for speech dereverberation. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 145–148, 2009.
- [141] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno. Blind separation and dereverberation of speech mixtures by joint optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):69–84, 2011.
- [142] B. Zhang, J. Zhu, and W. Leow. Visual analysis of fingering for pedagogical violin transcription. In *Proc. ACM Multimedia*, pages 521–524, 2007.

List of Publications

Journal papers

- 1) Akira Maezawa, Katsutoshi Itoyama, Kazuyoshi Yoshii and Hiroshi G. Okuno: Nonparametric Bayesian Dereverberation of Power Spectrograms Based on Infinite-Order Autoregressive Processes, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.22, no.12, pp.1918–1930, Dec. 2014. **Chapter 6**
- 2) Akira Maezawa and Hiroshi G. Okuno: Bayesian Audio-to-Score Alignment Based on Joint Inference of Timbre, Volume, Tempo, and Note Onset Timings, *Computer Music Journal*, vol. 39, no. 1, in print, 2015. **Chapter 3**
- 3) Akira Maezawa, Katsutoshi Itoyama, Kazunori Komatani, Tetsuya Ogata and Hiroshi G. Okuno: Automated violin fingering transcription through analysis of an audio recording, *Computer Music Journal*, vol. 36, no. 3, pp.57–72, 2012. **Appendix A**

International conference papers (peer-reviewed)

- 1) Akira Maezawa, Katsutoshi Itoyama, Kazuyoshi Yoshii and Hiroshi G. Okuno: Bayesian audio alignment based on a unified generative model of music composition and performance, *Proc. International Conference on Music Information Retrieval (ISMIR)*, pp. 233–238, 2014. **Chapter 4**
- 2) Akira Maezawa and Hiroshi G. Okuno: Audio part mixture alignment based on hierarchical nonparametric Bayesian model of musical audio collection, *Proc. International Conference on Audio, Signal and Speech Processing (ICASSP)*, pp. 5212–5216, 2014. **Chapter 5**
- 3) Akira Maezawa, Hiroshi G. Okuno, Tetsuya Ogata and Masataka Goto: Polyphonic audio-to-score alignment based on Bayesian latent harmonic allocation hidden Markov model, *Proc. International Conference on Audio, Signal and Speech Processing (ICASSP)*, pp. 185–188, 2011. **Chapter 3**

- 4) Akira Maezawa, Masataka Goto, Hiroshi G. Okuno: Query-by-Conducting: An interface to retrieve classical-music interpretations by real-time tempo input. *Proc. International Conference on Music Information Retrieval (ISMIR)*, pp. 477–482, 2010.
- 5) Akira Maezawa, Katsutoshi Itoyama, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno: Violin fingering estimation based on violin pedagogical fingering model constrained by bowed sequence estimation from audio input. *Trends in Applied Intelligent Systems*, LNAI 6098, vol. 3, pp. 249–259, 2010.
- 6) Akira Maezawa, Katsutoshi Itoyama, Toru Takahashi, Tetsuya Ogata, Hiroshi G. Okuno: Bowed string sequence estimation of a violin based on adaptive audio signal classification and context-dependent error correction. *Proc. International Symposium on Multimedia (ISM)*, pp. 9–16, 2009.
- 7) Akira Maezawa, Bo Gong, Bryan Waltrip, Gerald FitzPatrick, Yicheng Wang: Using a precision bridge for magnitude and phase measurement of a multiplying DAC with an AC reference. *Proc. International Instrumentation and Measurement Technology Conference*, pp. 777–780, 2008.

Domestic conference papers (w/o review)

- 1) 前澤 陽, 糸山 克寿, 吉井 和佳, 奥乃 博, 河原 達也: 結合動的モデルに基づく音響信号アライメント, 情報処理学会 音楽情報処理研究会, 2014-MUS-104, 2014.
- 2) 前澤 陽, 糸山 克寿, 吉井 和佳, 奥乃 博: 潜在共通構造モデルに基づくオーディオ信号間のアライメント, 情報処理学会 音楽情報処理研究会, 2014-MUS-102, 2014.
- 3) 前澤 陽, 奥乃 博: パワースペクトル上の無限オーダー畳みこみ型室内伝達モデルに基づく 実時間セミブラインド音声強調, 日本音響学会 春季研究報告会 研究報告, 2-1-3, 2013.
- 4) 前澤 陽, 奥乃 博: 楽曲パート混合オーディオ同士の楽譜なしアライメント手法, 情報処理学会 音楽情報科学研究会, 2013-MUS-100, 2013.
- 5) 前澤 陽: 無限オーダーの残響フィルタを持つ残響抑圧信号のベイズ推定手法, 日本音響学会 春季研究報告会 研究報告, 2-11-11, 2012.
- 6) 前澤 陽, 後藤 真孝, 尾形 哲也, 奥乃 博: 拍長の連続性を考慮した潜在的調波配分法に基づくスコアアライメント手法, 日本音響学会 春季研究報告会 研究報告, 3-1-16, 2011.

- 7) 前澤 陽, 後藤 真孝, 尾形 哲也, 奥乃 博: 潜在的調波配分法に基づく隠れセミマルコフモデルを用いたベイズ的スコアアライメント, 情報処理学会 第73回全国大会, 2-299-300, 5R-4, 2011.
- 8) 前澤 陽, 後藤 真孝, 奥乃 博: Query-by-Conducting: テンポ類似性に基づく同一楽曲における多様な解釈の検索インタフェース, 情報処理学会 音楽情報処理研究会, 2010-MUS-86, 2010.
- 9) 前澤 陽, 後藤 真孝, 高橋 徹, 駒谷 和範, 尾形 哲也, 奥乃 博: クラシック音楽理解力拡張インターフェイスを目指して: 解釈共通旋律と解釈相違旋律の推定, 情報処理学会 第72回全国大会, 3T-1, 2010.
- 10) 前澤 陽, 糸山 克寿, 高橋 徹, 尾形 哲也, 奥乃 博: 音響信号とコンテキスト制約を併用したバイオリンの演奏弦系列の推定, 日本音響学会 秋季研究報告会 研究報告, 2-5-15, 2009.
- 11) 前澤 陽, 糸山 克寿, 高橋 徹, 尾形 哲也, 奥乃 博: 音響信号と音楽的制約を統合したバイオリンの演奏弦系列の推定, 情報処理学会 音楽情報処理研究会, 2009-MUS-81, 2009.
- 12) Akira Maezawa, Katsutoshi Itoyama, Toru Takahashi, Tetsuya Ogata, Hiroshi G. Okuno: Probabilistic classification of monophonic musical instrument playing techniques. 情報処理学会 第71回全国大会, 4X-3, 2009.