DETECTION AND MODELING OF TRANSIENT AUDIO SIGNALS WITH PRIOR INFORMATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Harvey Thornburg
September 2005

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Julius O. Smith, III (Principal Adviser)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Robert M. Gray

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jonathan Berger

Approved for the University Committee on Graduate Studies.

Abstract

Many musical audio signals are well represented as a sum of sinusoids with slowly varying parameters. This representation has uses in audio coding, time and pitch scale modification, and automated music analysis, among other areas. Transients (events where the spectral content changes abruptly, or regions for which spectral content is best modeled as undergoing persistent change) pose particular challenges for these applications. We aim to detect abrupt-change transients, identify transient region boundaries, and develop new representations utilizing these detection capabilities to reduce perceived artifacts in time and pitch scale modifications. In particular, we introduce a hybrid sinusoidal/source-filter model which faithfully reproduces attack transient characteristics under time and pitch modifications.

The detection tasks prove difficult for sufficiently complex and heterogeneous musical signals. Fortunately, musical signals are highly structured – both at the signal level, in terms of the spectrotemporal structure of note events, and at higher levels, in terms of melody and rhythm. These structures generate context useful in predicting attributes such as pitch content, the presence and location of abrupt-change transients associated with musical onsets, and the boundaries of transient regions. To this end, a dynamic Bayesian framework is proposed for which contextual predictions may be integrated with signal information in order to make optimal decisions concerning these attributes. The result is a joint segmentation and melody retrieval for nominally monophonic signals. The system detects note event boundaries and pitches, also yielding a frame-level sub-segmentation of these events into transient/steady-state regions. The approach is successfully applied to notoriously difficult examples like bowed string recordings captured in highly reverberant environments.

The proposed transcription engine is driven by a probabilistic model of short-time Fourier transform peaks given pitch content hypotheses. The model proves robust to missing and spurious peaks as well as uncertainties about timbre and inharmonicity. The peaks' likelihood evaluation marginalizes over a number of observation-template linkages exponential in the number of observed peaks; to remedy this, a Markov-chain Monte Carlo (MCMC) traversal is developed which yields virtually identical results with greatly reduced computation.

Preface

This dissertation concerns the detection and modeling of transient phenomena in musical audio signals, and applications in audio segmentation, analysis-based sound transformation, and related areas. Since musical signals are often highly structured, at the signal level in terms of the spectrotemporal evolution of note events, and at higher levels, in terms of melody and rhythm, the primary focus is on how we can use this information to improve detection and modeling capabilities. This is not a mere academic exercise, since real-world musical recordings can be highly complex. One needs to make use of as many sources of information as possible.

The systematic integration of structural aspects with signal information is perhaps the key point of this dissertation. Everything else (while possibly interesting in its own right) plays a supporting role. Additional material may demonstrate applications (hence, situating the dissertation work in the greater context of past literature), or it may provide tools which are necessary to fully implement the proposed integration in the context of real-world signals.

I have organized this material in a linear fashion, which may not be the best choice for any particular reader. Nonetheless, it makes for the most concise presentation. Acknowledging this, I have also attempted to make each chapter self-contained, summarizing at the beginning of each the necessary information from previous chapters, although one must often take this information at face value.

Chapter 1 introduces the transient detection and modeling problems, surveys approaches from past literature, and (in light of this background) previews the contributions most specific to this dissertation. Chapter 2 details modeling applications and develops a set of detection requirements common to these applications. Chapter

3, perhaps the heart of the dissertation, develops a systematic approach for the use of signal-level and higher-level musical structures to improve the detection capabilities in light of the requirements discussed in Chapter 2. An application towards the joint segmentation and melody extraction for nominally monophonic recordings (which, however, may be corrupted by significant reverberation, note overlaps due to legato playing, and background instrumentation) is shown for a variety of piano and violin recordings. Chapter 4 discusses methods for robust pitch hypothesis evaluation which are vital towards implementing the methods covered in Chapter 3. Several appendices provide more details concerning the algorithms proposed in Chapter 3. These appendices can probably be skipped unless one is considering implementation issues.

Since the main focus is on the role of musical structure, I would encourage the beginning reader to skim Chapter 1 then read Chapter 3 as early as possible, taking the "transient detection requirements" stated at the beginning of that chapter at face value. Then if the reader desires further background on detection or modeling issues, a full development can be found in Chapter 2. If the reader is more interested in low-level implementation issues concerning the material in Chapter 3, Chapter 4 and the two appendices may immediately prove useful. However, the reader may be interested in robust pitch detection (and pitched/non-pitched classification) in more general scenarios, in which case Chapter 4 may be the best place to start. From that perspective, Chapter 3 serves as a way to adapt the pitch detection methods developed in Chapter 4 towards tracking pitch content over time, in a way that is robust to transients and nominally silent portions of the audio.

Acknowledgements

I would like to thank my principal advisor, Prof. Julius O. Smith III, for fostering the type of research environment which encourages one to take risks and rethink fundamental approaches, rather than pursue incremental improvements on existing ideas. He also provided tremendous help in the form of a continuous stream of signal processing insights delivered in his classes and during the DSP seminars. I am also indebted to my frequent collaborator Randal Leistikow who helped me tremendously with practical approaches and also in prompting me to clarify and refine my often "crazy" ideas in our many discussions. Next, I'd like to give special thanks to Prof. Jonathan Berger, who contributed much regarding music-theoretic ideas and perspectives from music cognition, and I especially appreciated his almost infinite patience as I attempted to learn the relevant material from music theory. Most importantly he brought to the table the mind of a composer, continually refreshing and illuminating the musical purpose behind many of these ideas. Next, Jonathan Abel provided a great sounding board in our many discussions and contributed much regarding general mathematical and estimation-theoretic insights. My educational experience as a whole was transformative; to this end I would especially like to thank again Julius O. Smith, also in particular Profs. Daphne Koller, Thomas Kailath and Thomas Cover, each through their coursework responsible for my completely changing the way I think about and approach problems. Lastly, I'd like to thank countless others both at and outside of CCRMA who helped and inspired me, especially Tareq Al-Naffouri, John Amuedo, Dave Berners, Fabien Gouyon, Arvindh Krishnaswamy, Yi-Wen Liu, Juan Pampin, Stefania Serafin, Tim Stilson, Steve Stoffels, and Caroline Traube.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The detection and modeling of transient phenomena in musical audio signals is a long-standing problem with applications in areas as diverse as analysis-based sound modification, lossy audio compression, and note segmentation for automated music analysis, transcription, and performance parameter extraction. We begin by defining "transient" in musical audio contexts and describing common transient phenomena which occur in these contexts. We review extensively the past literature on transient modeling, particularly in sound modification and compression applications which use sinusoidal models; additionally, we introduce a model for attack transients which hybridizes sinusoidal and source-filter modeling to facilitate novel, transient-specific processing methodologies.

Most of these modeling applications, we find, concern essentially two types of transient phenomena: *abrupt changes* in spectral information, usually associated with *musical onsets*, and *transient regions*, during which spectral information undergoes persistent, often rapid, change. To apply transient models, therefore, we must be able to detect abrupt changes and identify transient region boundaries. These detection tasks become quite challenging for real-world musical signals. For instance, consider the class of *nominally monophonic* recordings; here, each is considered to have been generated from a monophonic score. Nominally monophonic recordings often contain significant interference as well as effective polyphony due to reverberation, overlapping notes, and background instrumentation, all of which increase the possibility of

detection errors. On the other hand, musical signals are highly structured – both at the signal level, in terms of the spectrotemporal evolution of note events, and at higher levels, in terms of melody and rhythm. These structures generate context useful in predicting attributes such as pitch content, the presence and location of abrupt-change transients, and the boundaries of transient regions. Perhaps the key contribution of this dissertation is the *integration* of these contextual predictions with raw signal information in a Bayesian probabilistic framework, in order to minimize the expected costs associated with errors which arise in transient detection. We present not a single solution for one set of recording conditions, but an entire framework in which musical domain knowledge may be systematically encoded (via prior or transitional probability distributions) and adapted for a wide variety of applications and contexts.

1.1 Definition of "transient"

Both analysis-based sound modification and lossy audio compression make extensive use of *sinusoidal models*. Traditional approaches include the *phase vocoder* [41, 90], as well as methods based on short-time Fourier transform (STFT) analysis and peak-picking [81, 110, 106][1]. A primary reason for its widespread use is that the sinusoidal model offers an explicitly parametric representation of a sound's time-frequency evolution. The sinusoidal model for input $y_t, t \in 1\!:\!N$ is given as follows:

$$y_t \;=\; \sum_{k=1}^{p} A_k(t) \cos\left(\phi_k(t) + \sum_{s=0}^{t-1} \omega_k(s)\right) \tag{1.1}$$

Here $A_k(t)$ is the *amplitude* of the k^{th} sinusoid, $\omega_k(t)$ is the *frequency*, and $\phi_k(t)$ is the *phase offset*[2]. Since the time-frequency paradigm, at least to first approximation,

[1] The method proposed in [106] by Serra and Smith, called "spectral modeling synthesis" (SMS), is of particular interest because it represents also the part of the signal which is *not* well-modeled by sinusoids. This part, known as the *residual*, is obtained by subtracting the sinusoidal part from the original signal. For lossy compression purposes, unless absolute perceptual fidelity is necessary, this residual may be modeled via filtered white noise; see also [74, 75, 77] for related applications.

[2] Since frequency is the time difference of phase, it is redundant to represent both frequency and phase using time-varying functions. However, this redundancy becomes quite useful when we

reflects our "mental image" of sound [53, 44], sinusoidal models help us apply musical intuition towards designing interesting and meaningful sound modification schema. Furthermore, most regions in typical musical audio signals are considered *steady-state* with respect to the sinusoidal representation; in other words, these regions may be represented using either constant or slowly time-varying parameter trajectories. For compression applications, this facilitates significant reductions in bitrate with minimal perceptual distortion [81, 77, 91, 76].

Unfortunately, real-world musical signals contain many instances, called *transients*, which violate these steady-state conditions. Common instances include:

- *Abrupt changes* in amplitudes, phases, or frequencies: in recordings of acoustic material, these changes are often due to *energy inputs* on the part of the performer; hence, abrupt change transients often associate with *onsets* of note events or other phenomena that may be notated in the score
- *Rapid decays* in amplitudes, usually associated with *attack regions* following onsets of percussive sources
- *Fast transitions* in frequencies and amplitudes: musical examples include expressive pitch variations (portamento, vibrato, etc.) and timbral transitions (such as a rapid shift in the vocal formant structure)
- *Noise and chaotic regimes*, primarily responsible for *textural* effects: environmental sounds, such as rain or crackling fire, exhibit persistent textures which are important to preserve in resynthesis; textures can also arise from nonlinear feedback mechanisms in acoustic sources, e.g., bowed string and wind instruments [103, 99]; in most circumstances, the latter are likely to be found in short regions near onsets, as such regimes are often activated when the performer's energy input becomes large

What is considered "transient", however, depends on the details of the underlying sinusoidal model. More than one model may represent a particular signal. To cite

constrain the variation of either quantity. For instance, if frequency is modeled as piecewise-constant or piecewise-linear over short regions, the phase-offset trajectory may absorb the remainder of the local frequency variations which actually do occur.

an extreme case, the Fourier theorem guarantees that any signal of finite length, for instance a sinusoidal chirp sampled at 44100 Hz for which the pitch varies linearly from zero to 2000 Hz in 0.01 seconds, may be represented as a sum of sinusoids with *constant* amplitudes, frequencies, and phases (the chirp example requiring exactly 221 sinusoids). If one wants to warp a time-varying sinusoid's frequency trajectory, modifying the trajectories of each individual sinusoid in the "Fourier representation" will likely *not* have the desired effect. Figure 1.1 displays the results of such an experiment with the aforementioned chirp signal where the frequencies of all Fourier component sinusoids are doubled. Contrary to one's expectation, the result is no longer a single chirp, and will hence be heard as an artifact.

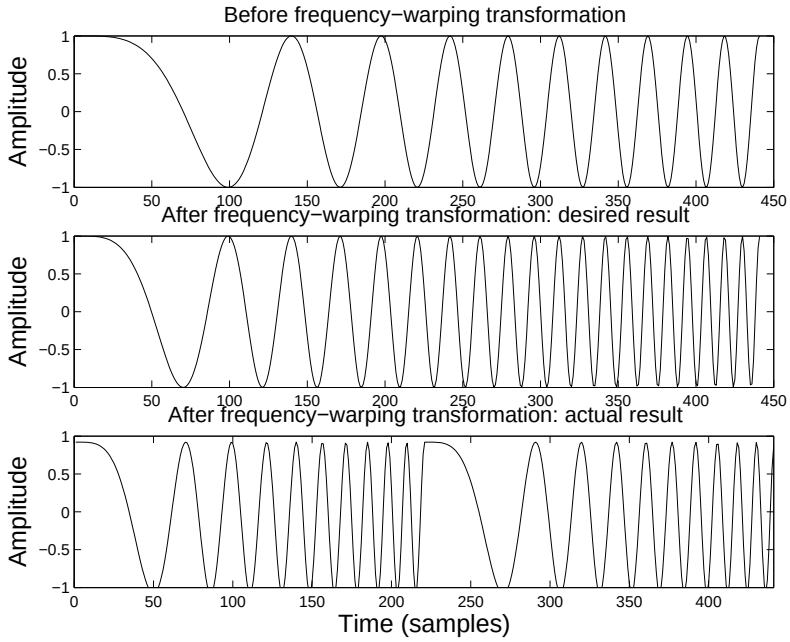


Figure 1.1: *Modification of sinusoidal chirp via stationary Fourier model*

The sinusoidal modeling ambiguity manifests in more common scenarios, such as amplitude and frequency modulation. For example, let y_t be a sinusoid with zero phase and constant frequency ω_1, and time-varying amplitude $A_t = 1 + \cos(\omega_1 t)$:

$$y_t = (1 + \cos \omega_1 t) \cos \omega_0 t \tag{1.2}$$

But y_t, as defined via (1.2), is equivalently the sum of three sinusoids with constant parameters:

$$y_t = \frac{1}{2}\cos(\omega_0 + \omega_1)t + \frac{1}{2}\cos(\omega_0 - \omega_1)t + \cos(\omega_0 t) \tag{1.3}$$

Which representation is heard depends on the relationships between ω_0, ω_1, and the integration time of the ear. Generally, if $|\omega_1 - \omega_0|$ is less than the critical bandwidth about ω_0, the result will be heard as time-varying, according to the representation (1.2).

1.2 Modeling and detection requirements

As the discussion throughout Chapter 2 attempts to motivate, the types of transient phenomena introduced in the previous section (abrupt changes, rapid decays, fast timbral transitions, and noise/chaotic regimes), may for the vast majority of modeling applications discussed in the literature, be combined into two types: *abrupt changes* and *transient regions* of nonzero width. The associated detection requirements become as follows.

- Detect the presence of all abrupt changes, and estimate their locations
- Detect the presence of all transient regions, and estimate their beginning and end points

Chapter 2 summarizes key applications of transient modeling in analysis-based sound modifications which use sinusoidal models (cf. [31, 81, 93, 67, 74, 75, 68, 39, 35], among others). In particular, *time and pitch scaling*[3] are addressed. Since pitch scaling is usually implemented by time scaling followed by sampling rate conversion [67], we focus on time scaling. Traditional time scaling methods assume a steady-state representation; as such, they focus on preserving the magnitudes and instantaneous

[3]Changing the playback speed of a recording modifies both duration and pitch; time and pitch scaling attempt to allow us independent control of each attribute. As such, time and pitch scaling are among the most well-known modification possibilities. Further definitions and relevant examples are given in Section 2.2.1.

frequencies of each sinusoidal component in the resynthesis. In the steady-state representation, the phase relationships become perceptually unimportant[4]. However, at abrupt-change transients, the situations become reversed: phase relationships instead play vital roles in the perception of these events whereas instantaneous frequency relationships become less important [93, 39]. Additionally, for high-fidelity applications, it becomes necessary to either preserve or guarantee appropriate scaling of instantaneous magnitude time differences [93]. Failure to preserve phase relationships (and to a lesser extent magnitude time differences) may generate audible artifacts in resynthesis. In Section 2.2.3, we illustrate the importance of phase relationships at the abrupt-change transient boundary using the simple example of a sub-audio impulse train. This impulse train is normally heard as a series of "ticks". Simply by modifying phase relationships, we can generate entirely different-sounding results ranging from sinusoidal chirps to noise textures (Figure 2.8), though the instantaneous frequency and magnitude content remains the same.

With transient regions, it becomes additionally necessary to maintain phase relationships throughout [39, 35]. By so doing, we preserve textures and other nonstationary phenomena which are otherwise difficult to model. A fundamental conflict exists between the maintenance of phase relationships throughout a contiguous region and the appropriate scaling of magnitude time differences at the beginning of that region, at least within the framework of existing methods; Section 2.2.4 discusses this conflict at length. It is usually resolved in favor of preserving phase relationships [74, 35], because perceptually, this is the more important goal [35]. However, significant portions of some signals (e.g., some percussion sources) consist entirely of transient regions. In this case, failure to appropriately modify the initial decay envelopes will cause the resynthesis to be perceived as "same instrument, different tempo" [35]. If one wishes to speed up a drum loop by a factor of, say, 25 percent, failure to shorten the decay envelopes by this amount may lead to an unnaturally "dense" resynthesis, leaving less room for other instruments in the mix.

[4]This fact has been well-known in even the earliest literature on modern psychoacoustics. The ear's insensitivity to absolute phase during steady-state portions was proposed by Ohm and given psychoacoustic verification by Helmholtz [98, 17].

On the other hand, these perceptual artifacts become less pronounced if transient regions are sufficiently short [74]. If the conflict between phase relationship preservation and magnitude time-difference scaling cannot be resolved within the framework of existing methods, one is hence motivated to seek an extended signal representation such that the transient regions (or, the signal information necessary to reconstruct these regions) become as short as possible. This leads down the path of source-filter modeling [116]. To this end, a hybrid sinusoidal/source-filter representation for attack transients is developed (Figure 1.2), as discussed in Section 2.3. The main idea

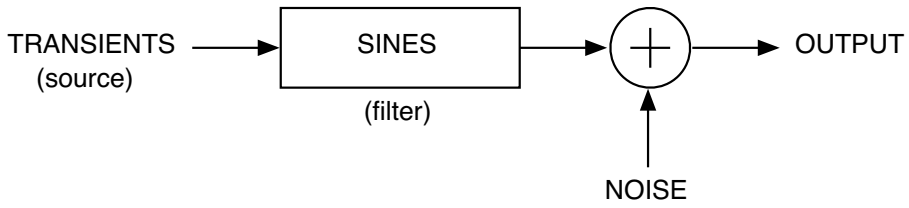


Figure 1.2: *Hybrid sinusoidal/source-filter representation for attack transients*

is that signals of effectively short duration called *input residuals* excite a bank of exponentially-decaying sinusoidal oscillators (Figure 2.11). Added to these oscillators is an *output residual* which for noise added during the recording process. Absent modification, the model is *perfect reconstruction*; i.e., the resynthesis is identical to the input.

A piano attack transient and the extracted input residuals associated with the first and 32^{nd} partials, respectively, are displayed in Figure 1.3. The effective temporal support of the input residuals appears substantially less than that of the input. Section 2.3.4 discusses the improved time and pitch scaling methods facilitated by this hybrid representation as well as some novel, "transient-specific" effect possibilities involving residual modifications.

In summary, the discussion in Chapter 2 establishes that a tremendous variety of transient modeling goals for analysis-based sound modification, especially those involving sinusoidal models, require the detection and location estimation of abrupt-change transients, and the identification of beginning and end points of transient regions. These detection capabilities find use as well in lossy audio compression. For

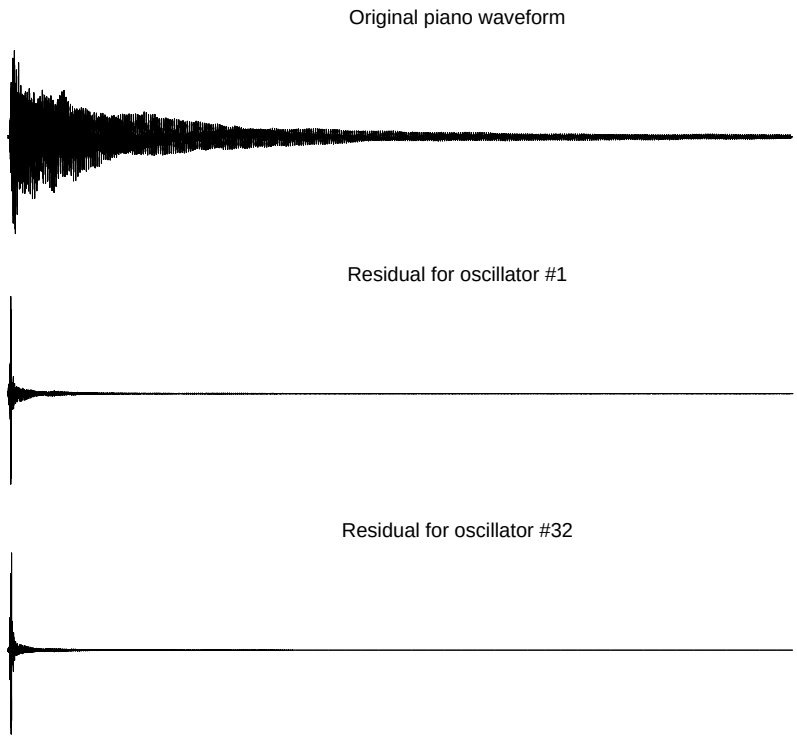


Figure 1.3: *Residuals vs. original attack transient for* ′D2′ *piano tone*

instance, *window switching* [36] has helped increase the efficiency and perceptual fidelity of transform audio codecs (e.g., MP3, AAC) in the reproduction of transient sound material [16, 15, 120, 14]. At least two reasons exist for the efficacy of window switching. First, the spectral content of transient regions is generally broadband and rapidly time-varying. Hence, it is appropriate to use shorter windows for these regions and longer windows for the steady-state regions, because shorter windows have less frequency resolution but more time resolution than longer windows. Second, the asymmetric nature of temporal masking about abrupt change transients [82] makes it necessary to limit the scope of *pre-echo* artifacts in reconstruction by applying shorter and possibly differently-shaped windows at these occurrences [14]. A further application concerns lossy compression schema which allow compressed-domain modifications [74, 75]. The spectrotemporal properties of transient regions as well as the need to preserve phase relationships throughout these regions after modification and resynthesis imply that different encodings *and* modification strategies must be used

for these regions [74].

Finally, the detection of abrupt-change transients and identification of transient regions both have direct applications in automated music analysis and performance parameter extraction[5]. The main reasons concern the spectrotemporal structures commonly associated with "note events". Most often in acoustic recordings, abrupt-change transients result from *energy inputs* or *decisions* on the part of the performer. Ideally, we would like to say that abrupt changes associate always with *musical onsets*, defined as the beginnings of note events, as this is often the case. Unfortunately, the level of detail provided by most traditional score-based representations may be too coarse to adequately represent all of the performer's energy inputs and decisions. For instance, consider a recording of an overblown flute. During a single *notated* event, multiple pitched regions may occur due to the different regimes of oscillation. Transient regions may exist between these pitched regions because of chaotic behaviors activated upon transitioning between oscillatory regimes [99]. Nevertheless, despite what may or may not be explicitly notated, the navigation between oscillatory regimes is under the performer's control, and may hence be characterized as a sequence of discrete decisions. Discovering these decision points provides valuable information for *performance parameter extraction*, which may be of use, for instance, in driving a physical model of the same instrument [52, 79, 29], or animating a virtual performer [104]. Since this low-level segmentation based on abrupt-change events and transient regions may err on the side of too much, rather than too little, detail for score extraction purposes, this information may be clustered in a subsequent pass. As Chapter 3 discusses, the transient detection problem may be considered *jointly* with note segmentation. Particularly in the violin examples analyzed in Section 3.9, ornamentations such as portamento and vibrato do not cause extraneous detail in the note segmentation.

[5]Perhaps the primary difference in detection requirements for automated music analysis and performance parameter extraction is that less temporal accuracy may be required for music analysis tasks when compared with applications in analysis-based sound modification and audio compression; see the beginning of Section 3.10 and also 3.10.2 for further details.

1.3 The role of musical structure in transient detection

With sufficiently complex musical signals, the transient detection tasks required for the modeling applications summarized in the previous section may be difficult to reliably perform. Even restricting to simpler cases such as *nominally monophonic* signals (which may be considered as *lead melodies*, arising from monophonic scores), we encounter difficulties such as noise, interference, and effective polyphony due to background instrumentation, overlapping notes, and reverberation. These difficulties may lead to false alarms or missed detections for both abrupt-change events and transient regions, as well as estimation errors in the locations of abrupt-change events and transient region boundaries.

On the other hand, musical signals are highly structured; both at the signal level, in terms of the spectrotemporal evolution of note events, and at higher levels, in terms of melody and rhythm. This structure manifests by constraining what is possible concerning attributes such as pitch content or the presence and location of abrupt-change events and transient region boundaries. These tendencies generate *contextual predictions* regarding these attributes; such predictions may be combined with raw signal information to improve detection and estimation capabilities in ways that are robust to uncertainties in this contextual knowledge and noise in the signal. For instance, Sections 3.3.2 and 3.3.3 demonstrate how the consistency of pitch information during steady-state regions of note events influences our ability to detect abrupt-change transients associated with note onsets. The beginning of Section 3.3 as well as Section 3.10.1 discusses the role of melodic expectations, while Section 3.10.2 addresses temporal expectations of note onsets due to the presence of rhythm.

Let us now demonstrate what is meant in a general sense by "the ability of contextual predictions to improve estimation capabilities" using the framework of a linear Gaussian model. This framework is useful because everything we wish to demonstrate follows in closed algebraic form. Suppose $y_{1:N}$ is an independent and identically distributed Gaussian sequence with unknown mean x and known variance σ_y^2, and consider the estimation of x. An estimate, $\hat{x}$, is derived as a function of $y_{1:N}$; we

want this estimate to be "best" in the sense that it minimizes the *expected squared error*, $E|\hat{x} - x|^2$.

A well-known lower bound on the expected squared error; i.e., the Cramer-Rao bound [26] applies in this case:

$$E|\hat{x} - x|^2 \geq \sigma_y^2/N \tag{1.4}$$

It is easily shown (in this example) that the Cramer-Rao bound is achieved by $\hat{x}_{MLE}$:

$$\begin{aligned} \hat{x}_{MLE} &= \underset{x}{\text{argmax}}\ p(y_{1:N}|x, \sigma_y^2) \\ &= \frac{1}{N}\sum_{t=1}^{N} y_t \end{aligned} \tag{1.5}$$

where $p(y_{1:N}|x, \sigma_y^2)$ is the conditional probability density function of the observations given x and σ_y^2.

If conditions are such that σ_y^2/N becomes unacceptably large, (1.4) indicates that nothing further can be done with the current set of observations, since no estimator exists with less mean square error. Nevertheless, many problems contain additional sources of information, which do not take the form of extra observations. Suppose a context is established, where we expect that x lies "close to" some value, say x_0. To be precise, suppose that x is Gaussian with mean x_0 and variance σ_x^2. Now construct the following estimator:

$$\begin{aligned} \hat{x}_{MAP} &= \underset{x}{\text{argmax}}\ p(x|y_{1:N}, \sigma_x^2, \sigma_y^2) \\ &= \frac{\sigma^{-2}x_0 + \sigma_y^{-2}\sum_{t=1}^{N} y_t}{\sigma_x^{-2} + N\sigma_y^{-2}} \end{aligned} \tag{1.6}$$

where $p(x|y_{1:N}, \sigma_x^2, \sigma_y^2)$ is the posterior density of x given the observations and variance parameters σ_x^2 and σ_y^2. Some algebra shows that the expected squared error, $E|\hat{x} - x|^2$,

is

$$\begin{aligned} E|\hat{x}_{MAP} - x|^2 &= (\sigma_x^{-2} + N\sigma_y^{-2})^{-1} \\ &< (N\sigma_y^{-2})^{-1} \\ &= \sigma_y^2/N \end{aligned} \tag{1.7}$$

The strict inequality in (1.7) holds provided that $\sigma_x^2 < \infty$. That is, we have constructed an estimator, *given an additional source of contextual knowledge as represented by a prior distribution on* x, with expected squared error less than that of the Cramer-Rao lower bound. Hence, this example demonstrates in concrete, quantitative terms, what is meant by prior contextual knowledge "extending our abilities" to estimate unknown attributes from data. Analogous properties for the signal-level structures encountered in musical audio signals (e.g., the consistency of pitch information during pitched portions of note events) are derived in Section 3.3.2.

Unfortunately, the vast majority of transient detection approaches in the music signal processing literature are fundamentally heuristic in nature. It is hence unclear how we can adapt them to exploit contextual knowledge from musical structure in ways which are robust to uncertainties in this knowledge. Most commonly, these methods threshold "novelty functions" [48] (usually filtered derivatives; cf. [9, 7]) based on signal characteristics such as amplitude [102], phase [10], combined phase and amplitude [33, 34], sinusoidal-model-residual level [74, 35], or automatically-weighted combinations of individual features [48], to detect abrupt-change transients. (This novelty-function approach may be adapted for the detection of transient regions; cf. [35].) While these heuristic methods may be easy to implement, they are often difficult to adapt to changing problem conditions (e.g., signal-to-noise ratio, the expected rates of change of the signal characteristic during nominally steady-state vs. transient regions, and so forth.) because they lack explicit models for *uncertainty* in these conditions. If a method fails under certain conditions, it is difficult to ascertain by what extent that method can be improved.

On the other hand, a variety of statistical methods have been applied to the problem of detecting abrupt changes in spectrotemporal structure. These methods

provide robustness to uncertainties; as well, they address portability and optimality concerns. Of note are the online (real-time) methods based on sequential hypothesis testing; e.g., the *divergence algorithm* [8], the *forward-backward method* [5], offline maximum-likelihood methods [111, 61], and integrated online-offline approaches [115]. Unfortunately, few applications of these techniques exist in musical audio; known exceptions being [56, 50, 115]. Perhaps the primary reason is that these methods fail to incorporate contextual predictions from musical structure, so that the limitations imposed by adverse problem conditions (i.e., poor signal-to-noise ratios, complex model structures, and limited amounts of data) may be overcome.

To this end, Chapter 3 proposes a Bayesian probabilistic framework for *joint* melody extraction and note segmentation of nominally monophonic signals for which steady-state regions have discernible pitch content[6]. This framework may be considered as a transcription system with additional features for transient detection. A block diagram is shown in Figure 3.13; objectives may be summarized:

- The recording is segmented into discrete note events, possibly punctuated by *null regions*. Null regions are gaps between note events containing only silence, recording noise, or spurious events such as the performer knocking the microphone, or clicks and pops from vinyl transfer. For each event, we identify its onset time, duration, and MIDI note value.

- Note events are further segmented into transient and steady-state regions, where applicable. Hence, we identify all abrupt-change transients which associate with musical onsets as well as all boundaries of transient regions. Transients resulting from spurious events are suppressed; this becomes a key robustness consideration when dealing with musical audio.

- The system makes efficient use of prior contextual knowledge from musical structure, both at the signal level and at the level of syntax (melody and rhythm).

[6]To conform to real-world cases involving instruments such as piano and marimba, inharmonicity and more generally, uncertainty in harmonic structure is tolerated; see Chapter 4 for further discussion of the evaluation of pitch hypotheses, in particular Section 4.3.3 which addresses the modeling of uncertainties in harmonic structure.

The system proposed in Chapter 3 operates on framewise short time Fourier transform (STFT) peak features. Use of STFT peak features substantially reduces computations when compared against sample-accurate methods, without sacrificing too much information relevant for note identification. Unfortunately, this limits the segmentation's temporal resolution to the frame rate[7]. A frame-accurate segmentation may suffice for automatic transcription, but finer resolutions may be required for sound transformation and compression applications. Nonetheless, a frame-accurate segmentation may facilitate subsequent sample-accurate processing. The frame-accurate method identifies local neighborhoods where abrupt-change events and transient region boundaries are likely to be found; moreover, it provides information regarding pitch content before and after the segment boundary locations. Section 3.10.4 discusses how the present methods may be extended to produce a sample-accurate segmentation.

Contextual knowledge from musical structure is incorporated at the signal level via consistency of pitch and amplitude information during steady-state (pitched) regions of note events. In conjunction we exploit prior knowledge that the signal arises from a monophonic score, according to a stochastic grammar governing the succession of transient, pitched, and null signal regions, null regions representing gaps between note events. Section 3.4 introduces the grammar while Section 3.6 provides its distributional specification. Since tempo, the amount of legato playing, and the relative presence of transient information in each note event (among other characteristics) vary from piece to piece, and this variation is otherwise difficult to model, we specify the grammar's transition distribution up to a number of *free parameters* which must be *estimated* from the observations. This estimation process, introduced in Section 3.7.2, is based on the expectation-maximization (EM) algorithm [28].

Additionally, the system enables higher-level, melodic structures to inform the segmentation, as introduced in Section 3.6.2. Here we represent *melodic expectations* (the predictive distributions for subsequent notes based on past information) using a first-order Markov note transition model. Unfortunately, the latter fails to capture

[7]For the examples shown in Section 3.9, the frame rate is 11.6 ms.

common melodic expectations which arise, e.g., in the context of Western tonal music. Forthcoming work by Leistikow [71], based on recent music cognition literature (cf. Narmour [85], Krumhansl [64], Schellenberg [101], and Larson and McAdams [69], among others) addresses the Markovian probabilistic modeling of melodic expectations. The resultant models may be integrated with the present signal-level framework. Section 3.10.1 summarizes these extensions.

To allow rhythmic structure to inform the segmentation, we may extend the stochastic grammar representing the succession of transient, pitched, and null regions; Section 3.10.2 discusses a proposed extension using *probabilistic phase-locking structures*. Previous approaches to modeling rhythmic onset patterns, from recent literature on tempo and beat induction from the audio signal (cf. [49, 51, 18, 65]) make suboptimal early decisions about onset locations as they use the detected onsets as "observations" for the higher-level tempo models. By contrast, the probabilistic phase-locking method introduced in Section 3.10.2 is *fully integrated* with signal-level observations, in the sense that onsets (and other transient boundaries) are identified *jointly* with tempo and beat information. Which is to say, not only do the detected onset (and region boundary) patterns inform the tempo and beat induction; a reverse path of influence is established between the tempo/beat layer and the onset detection via *temporal expectations*. Moreover, the use of probabilistic phase-locking structures in tempo and beat induction may find application in music cognition research, because each temporal expectation therein explicitly encodes the *anticipation* that a certain event is about to occur. One may investigate affective qualities: for instance, the buildup of tension from sustained anticipation.

Structurally, the proposed Bayesian framework for joint transient detection and region segmentation relates to recent work in automatic transcription; cf. Kashino *et al.* [60], Raphael [95, 96], Walmsley *et al.* [119], Godsill and Davy [46], Sheh and Ellis [107], Hainsworth [51], [20, 18], and Kashino and Godsill [59], among possibly others. Indeed, the use of Bayesian methods in automatic transcription is presently an emerging field. Regarding modeling aspects, perhaps the most similar work is that of Cemgil *et al.* [20, 18]. The authors therein propose a generative probabilistic

model for note identification in both monophonic and polyphonic cases[8]. Their model contains what can be interpreted as a simplified version of the stochastic grammar proposed in Section 3.6.2, in that a discrete (in this case binary) variable indicates if a note is sounding at a given time. However, [20] models the transient information in an *additive* sense, as filtered Gaussian noise superposed with the sinusoidal part, paralleling the "sines plus noise" approach of SMS [106]. This clearly fails to satisfy the detection requirements for the transient modeling applications in sound modification and lossy audio compression as previously discussed. These applications favor the explicit characterization of abrupt-change transients as well as the restriction of transient information to contiguous *regions* within each note event. By contrast, the stochastic grammar proposed in Section 3.6.2 yields not only a segmentation into individual note events, but also a sub-segmentation of each event into transient and steady-state regions.

A further innovation of the present method is the use of *cost functions* which adequately represent the *effects* of various types of transcription errors, rather than relying on byproducts of standard Bayesian filtering, smoothing, or Viterbi inference techniques. As an example, it is less problematic for the locations of note onsets to be shifted by small amounts than it is for notes to be missing or extra notes introduced. By using an appropriate cost function, the solution to the decision problem yields the transcription. Since one goal of Bayesian inference methods is to produce sufficient statistics for decision problems, this means that the inference results may be immediately converted into MIDI data without requiring complex heuristics in postprocessing. A straightforward conversion process is detailed in Section 3.8. Here two hidden variables associate with each STFT frame: M_t, which encodes the segmentation (i.e., an indication whether or not the current frame contains an onset, as

[8]In the framework proposed in Chapter 3, polyphonic extensions are not presently implemented. The primary reason is that the results would characterize all abrupt-change transients and transient regions for note events which overlap in time. To use these results in sound modification and compression appliations, the transient modeling would need to perform also the source separation and demixing of individual note events which is by no means an easy task. However, the polyphonic extensions are readily applicable in performance analysis and parameter extraction. The extensions are conceptually straightforward but may experience computational difficulties using the Bayesian inference methods discussed in Section 3.7. Section 3.10.3 provides a thorough discussion of these issues, suggesting approximate inference schema which may greatly reduce computational costs.

well as the type of region containing this frame), and S_t, which encodes hidden signal characteristics representing inherent spectral content (pitch, sinusoidal amplitude, and transient amplitude information). The result of standard Bayesian smoothing inference is the computation of the *smoothed posterior* $P(M_t, S_t|Y_{1:N})$ for all $t \in 1:N$, where Y_t is the vector of STFT peak observations (peak frequencies; amplitudes) associated with the t^{th} frame. From $P(M_t, S_t|Y_{1:N})$, $P(M_t|Y_{1:N})$ may be extracted by marginalizing out S_t. Now, via (3.38), the collection $\{P(M_t|Y_{1:N})\}_{t=1}^{N}$ is a sufficient statistic for the decision problem which minimizes the expected number of frames for which the detected M_t is in error. In practice, most segmentation errors arise from ambiguities concerning whether the onset boundary occurs in a given frame or the adjacent frame. Two errors are particularly common: first, the detected onset could occur in the wrong frame; second, onsets could be detected in both frames. Detecting an onset in the wrong frame results in a shift of the onset location by the frame resolution, which has only a slight effect, especially since onset times are quantized to this resolution. Detecting onsets in both frames, however, introduces an additional note event. This becomes disastrous for transcription-related purposes. Hence, minimizing the expected number of frames for which the detected M_t is in error is clearly *not* the proper cost objective for transcription.

Our solution is to preserve the integrity of the *entire* segmentation sequence $M_{1:N}$. That is, as described in Section 3.3.3, we estimate $\hat{M}_{1:N}$ to minimize the probability that *any* $\hat{M}_t$ is in error for the entire sequence $M_{1:N}$, which leads naturally to a Viterbi-type approach. Unfortunately, straightforward Viterbi inference chooses $\hat{M}_{1:N}$ and $\hat{S}_{1:N}$ *jointly* to minimize the corresponding error probability in $\{M_{1:N}, S_{1:N}\}$, This is clearly not the same thing as minimizing the probability that $M_{1:N}$ alone is in error because it avoids the implicit marginalization over $S_{1:N}$. Moreover, the estimated $\hat{S}_{1:N}$ should be *synchronous* with $\hat{M}_{1:N}$ in that $\hat{S}_{1:N}$ is chosen to satisfy some expected cost objective under which $Y_{1:N}$ and $M^*_{1:N}$ are both entered into evidence. Inference and estimation objectives which do satisfy these requirements are derived in Sections 3.3.3 and 3.5.2; Section 3.7.1 describes an approximate inference algorithm satisfying these requirements.

Lastly, the present method proves robust to interference from recording noise

and actual instances of polyphony resulting from background instrumentation, note overlaps from legato playing, and excessive reverberation. These results are demonstrated in Section 3.9. We find that this robustness is largely due to the integration of contextual predictions concerning the consistency of inherent pitch and amplitude characteristics during pitched regions of note events with STFT peak observations. For instance, suppose that a frame belonging to a pitched region of a note event is occluded by interference. The method in this case automatically relies on the surrounding frames within this region to estimate the instantaneous pitch and amplitude characteristics for this frame, as demonstrated in Section 3.3.3. However, this robustness is also partially due to the way pitch and amplitude information is extracted from STFT peak observations, via the distributional model $P(Y_t|S_t)$. The quality and robustness of this evaluation may be assessed by embedding it in a single-frame maximum-likelihood pitch estimator, as the latter does not use information from surrounding frames.

Chapter 4 introduces a model for evaluating $P(Y_t|S_t)$ based on a harmonic template, demonstrating its use in robust maximum-likelihood pitch detection under moderately adverse interference conditions. The harmonic template idea is introduced in Section 4.2.2 and may be summarized as follows. Consider a pitch hypothesis[9] $\{f_0, A_0\}$ generated from one of the possibilities for S_t: here f_0 represents the pitch value and A_0 the corresponding (pitched) reference amplitude. The probabilistic model of Chapter 4 generates a joint distribution over all frequency and amplitude peak values *potentially observed* in the STFT. This model accounts for additive Gaussian noise in the time domain plus uncertainties in harmonic structure resulting from inharmonicity and other timbral variations. It may be considered an extension of the template model in Goldstein's probabilistic pitch detector [47], although Goldstein's approach ignores amplitude information.

Unfortunately, thanks to interference, we do not know which template peaks correspond to peaks *actually observed* in the STFT. Without this linkage we cannot evaluate $P(Y_t|f_0, A_0)$ via the template distributions described above. Our solution

[9]The necessary extension to *non-pitch* hypotheses, represented by the reference amplitude A_0^Q, is discussed in Section 4.1.

is to marginalize over the unknown linkage possibility with respect to a prior (see Section 4.3.2) favoring the survival of template peaks with a low harmonic index. The exact marginalization, however, proves computationally intractable because the number of linkage possibilities grows exponentially with N_p where N_p is the minimum of the number of template peaks and the number of observed STFT peaks. Nevertheless, we recognize that in practice, virtually all but a few possibilities contribute negligibly to the likelihood evaluation (see Section 4.4 for examples and further discussion). This motivates a fast Markov-chain Monte Carlo (MCMC) approximate evaluation, developed in Section 4.5, which obtains virtually identical results for a noisy (single-frame) piano example when compared against the exact evaluation, at a small fraction of the computational cost. In either case, MCMC evaluation vs. exact evaluation, maximum-likelihood pitch estimation yields acceptable results under these conditions (as shown in Sections 4.4 and 4.5). On the other hand, the MCMC evaluation may still be too slow for some applications. Alternatively, we derive a less exact, but (in most circumstances) faster, determinstic approximation, as discussed in Section 4.6. The computational cost of the deterministic approximation is quadratic in N_p, as opposed to the exponential cost of the exact method. This deterministic approximation is used to evaluate $P(Y_t|S_t)$ for the joint melody extraction and transient detection results shown in Section 3.9.

1.4 Conclusion

In conclusion, the main contribution of this dissertation appears to be the introduction of prior information from musical structure towards the transient detection problems outlined above, which arise repeatedly in both established and newly introduced transient modeling contexts. Structural information is introduced both at the signal level, in terms of the "standard note evolution" grammar, and at the level of syntax, in terms of melodic structure. As the results of Section 3.9 demonstrate, the resultant system for melody tracking, note onset identification and note sub-segmentation (revealing both transient and steady-state regions within a particular note event) for nominally monophonic musical audio appears robust to real-world interference phenomena and

actual instances of polyphony; e.g., reverberation, overlapping notes, and background instrumentation. Moreover, because the relevant structural information is explicitly represented using conditional probability distributions, it becomes straightforward to adapt this system across varying musical contexts. Secondary contributions include the robust evaluation of pitch hypotheses using a highly reduced feature set, that of STFT peak data. This evaluation becomes useful in scenarios (e.g., maximum likelihood pitch estimation) where prior structural information may not be readily available, and it is easily extended to the polyphonic case as described in [72]. Extensions and further applications are discussed in Sections 3.10.1 (incorporation of more sophisticated models of musical expectation), 3.10.2 (incorporation of temporal expectations from rhythm via probabilistic phase-locking networks), 3.10.3 (extension to the polyphonic case), and 3.10.4 (extension to sample-accurate segmentation and applications in interactive audio editing).

Chapter 2

Modeling and detection requirements

2.1 Introduction

Sinusoidal modeling is readily applicable to the analysis, transformation and resynthesis of recorded sound. The main reason is that the sinusoidal model offers an explicitly parametric representation of a sound's time-frequency evolution. Since the time-frequency paradigm, at least to first approximation, reflects our mental image of sound, one may readily apply musical intuition towards specific strategies for sound transformation.

When the realities of the signal model work contrary to musical intuition, the result after transformation is not as expected. Here we say that *artifacts* occur. A typical sinusoidal model is usually given as follows:

$$y_t = \sum_{k=1}^{p} A_k(t) \cos\left(\phi_k(t) + \sum_{s=0}^{t-1} \omega_k(s)\right) \tag{2.1}$$

where $A_k(t)$ is the *amplitude* of the k^{th} sinusoid, $\phi_k(t)$ is the *phase*, and $\omega_k(t)$ is the *frequency* at time t, where $t \in 1\!:\!N$.

Figure 2.1 depicts the usual "analysis-synthesis" framework for transforming sounds

via the model (2.1). In the figure, the labeling of canonical blocks *analysis*, *transfor-*

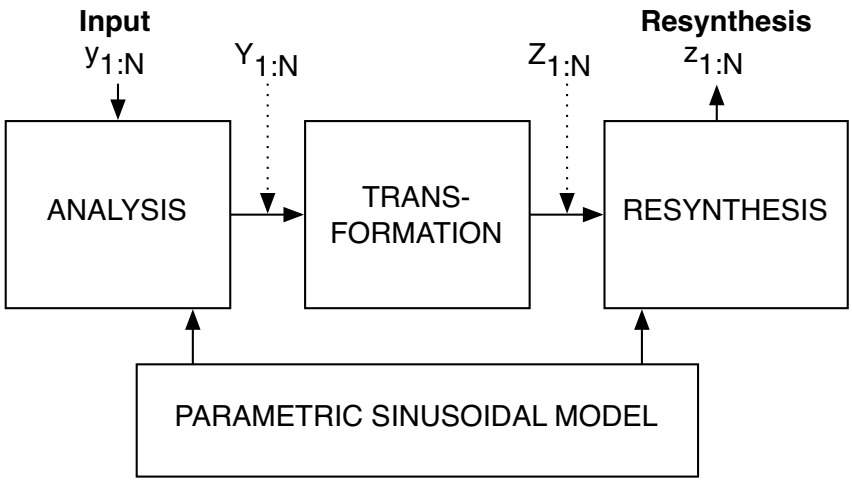


Figure 2.1: *Analysis, transformation, and resynthesis*

mation, and *resynthesis*, is inspired by Serra [105]; also Pampin [86]. Analysis means the estimation of the amplitude, phase, and frequency trajectories from the input $y_{1:N}$; in the figure, we denote these trajectories collectively as $Y_{1:N}$. Transformation modifies these trajectories, producing $Z_{1:N}$. The output, $z_{1:N}$, is then resynthesized from $Z_{1:N}$, again using (2.1). We also refer to $z_{1:N}$ as the *resynthesis*.

The canonical assumption regarding the model (2.1) is that it is *steady-state*, meaning that the amplitude, phase, and frequency trajectories do not vary rapidly with time. In this way, a short time Fourier transform may be used as a front end for the analysis, as originally proposed by Gabor [44] and adapted for digital implementation by Portnoff [90][1].

However, musical signals contain many instances or time intervals, called *transients*, which violate the steady-state assumption. Transients are hence a common source of resynthesis artifacts. We recall the types of transients defined in Section 1.1:

- *Abrupt changes* in amplitudes, phases, or frequencies: in recordings of acoustic material, these changes are often due to the energy input on the part of the

[1]Among others, see also [31]. For a thorough overview of contemporary applications of the short time Fourier transform in sinusoidal modeling and music signal processing, see [108].

performer; hence, abrupt change transients often associate with *onsets* of note events or other phenomena that may be notated in the score

- *Rapid decays* in amplitudes, usually associated with *attack regions* following onsets of percussive sources
- *Fast transitions* in frequencies and amplitudes: musical examples include expressive pitch variations (portamento, vibrato, etc.) and timbral transitions (such as a rapid shift in the vocal formant structure)
- *Noise and chaotic regimes*, primarily responsible for *textural* effects: environmental sounds, such as rain or crackling fire, exhibit persistent textures which are important to preserve in resynthesis; textures can also arise from nonlinear feedback mechanisms in acoustic sources, e.g., bowed string and wind instruments [103, 99]; in most circumstances, the latter are likely to be found in short regions near onsets, as such regimes are often activated when the performer's energy input becomes large

What is considered "transient" depends greatly on the underlying signal model: numerous examples are presented in Section 1.1.

2.2 Transient processing in the phase vocoder

2.2.1 Time and pitch scaling

Some of the most widespread applications of sinusoidal modeling (in the sense of analysis-synthesis transformations) consist of *time and pitch scaling* and variants. It is well known that changing the playback speed of a sound may be accomplished in digital systems by a sampling-rate alteration; unfortunately, this operation modifies both pitch and duration. Often we desire *independent* control, over these attributes. In *time scaling*, the goal is to modify the sound's duration while preserving its pitch. This means that the amplitude and frequency trajectories for each sinusoidal component (the parameters $A_k(t)$ and $\omega_k(t)$ in (2.1)) are interpolated over the resynthesis

time base, and $\phi_k(t)$ is adjusted to preserve instantaneous frequency relationships between analysis and resynthesis. In *pitch scaling*, the goal is to modify the frequencies of each sinusoidal component; specifically, in *transposition*, each frequency is multiplied by a fixed amount. The ideal effect of each operation (playback speed alteration, time scaling, and transposition pitch scaling) is displayed in Figure 2.2. Since transposition is usually implemented by time scaling followed by playback speed alteration [67], we consider only time scaling.

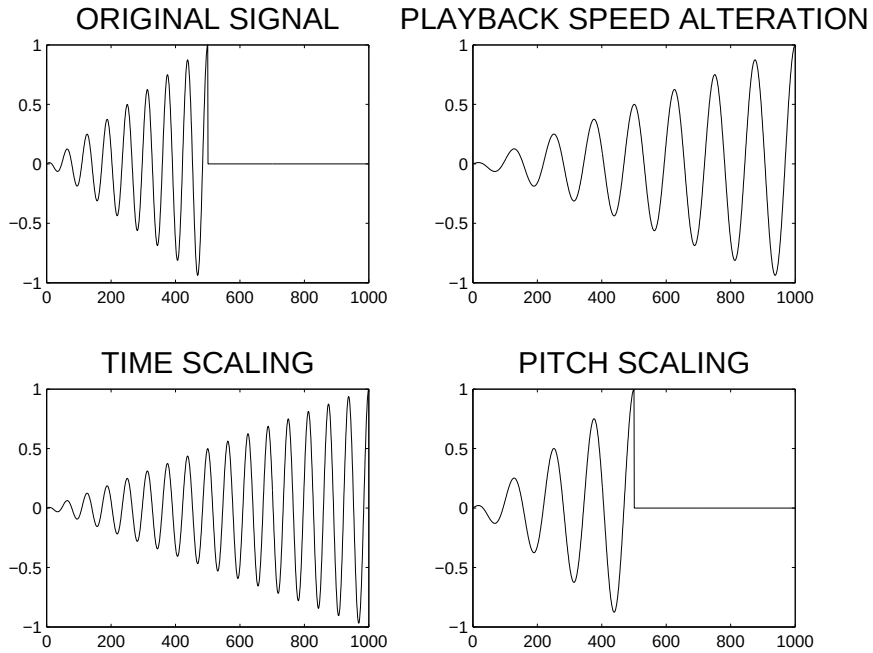


Figure 2.2: *Ideal resyntheses for playback speed alteration, time scaling, and pitch scaling operations*

2.2.2 Phase vocoder time scaling

A common method for high quality time scaling makes use of a heterodyned filterbank called the *phase vocoder*, originally developed for speech coding by Flanagan and Golden [41], and adapted for digital implementation by Portnoff [90]. A schematic is displayed in Figure 2.3. In the figure, $j \triangleq \sqrt{-1}$.

Ideally, each component sinusoid of y_t is isolated in exactly one analysis channel.

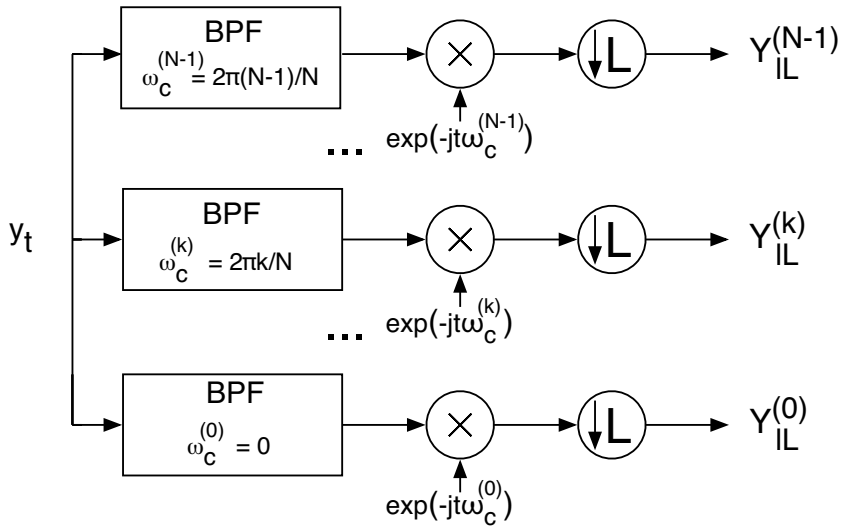


Figure 2.3: *Phase vocoder analysis section*

This enables the time scaling process to proceed on a sinusoid-by-sinusoid basis. Now, suppose the bandpass filters are ideal. This means, letting $H^{(k)}(\omega)$ denote the response of the bandpass filter for the k^{th} channel:

$$H^{(k)}(\omega) \;=\; \begin{cases} 1, & |\omega - \omega_c^{(k)}| < \frac{\pi k}{N} \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

where $\omega_c^{(k)}$, the channel center frequency, equals $2\pi k/N$. Then, each channel's output may be reconstructed after heterodyning by $e^{-jt\omega_c^{(k)}}$ and downsampling by N, by means of ideal sinc interpolation and subsequent modulation by $e^{jt\omega_c^{(k)}}$. Since the bandpass filters are generally non-ideal, their bandwidth will exceed $2\pi k/N$ and hence a more conservative downsampling by factor $L < N$ is advised.

To achieve time expansion by factor α, we reconstruct each $Y_{lL}^{(k)}$ at instants $t = lL'$, where $L' = \alpha L$, to produce the modified channel output $Z_{lL'}^{(k)}$. If the component is perfectly isolated by $H^{(k)}(\omega)$ and the latter produces no phase distortion, this component may be recovered at the frame boundaries $t = lL$, as $\tilde{Y}_{lL}^{(k)}$:

$$\tilde{Y}_{lL}^{(k)} \;\triangleq\; e^{jlL\omega_c^{(k)}} Y_{lL}^{(k)} \tag{2.3}$$

according to the preceding discussion. Hence, if we define:

$$\tilde{Z}_{lL'}^{(k)} \triangleq e^{jlL'\omega_c^{(k)}} Z_{lL}^{(k)} \tag{2.4}$$

then, absent modification, the resynthesis may be taken at $t = lL'$ to be $\tilde{Z}_{lL'}^{(k)}$. Between these times, both the amplitude and phase of $\tilde{Z}_{lL'}^{(k)}$ may be interpolated to obtain $\tilde{Z}_t^{(k)}$. This is of course assuming the phase of $\tilde{Z}_{lL'}^{(k)}$ is appropriately unwrapped, which, as we will see, is facilitated by the heterodyning process.

The resynthesis procedure is diagrammed in Figure 2.4, where the magnitude/phase interpolation, detailed in Figure 2.5, proceeds according to the approach of McAulay and Quatieri [81], which uses linear interpolation for the log amplitude and cubic interpolation for the unwrapped phase[2].

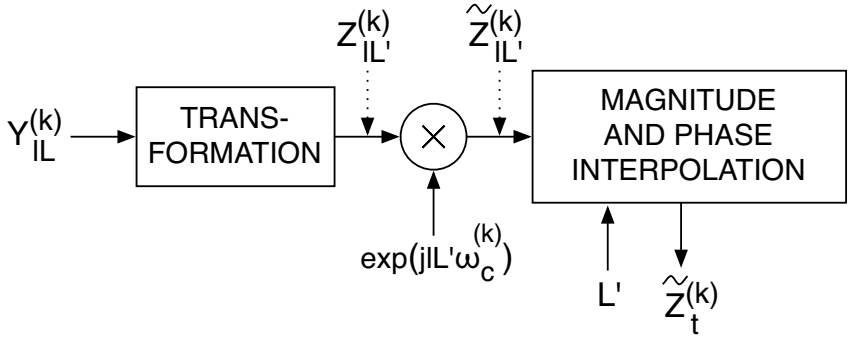


Figure 2.4: *Resynthesis from single channel of phase vocoder analysis*

It remains to determine the mapping $Y_{lL}^{(k)} \rightarrow Z_{lL'}^{(k)}$, such that the resyntheses, $\tilde{Z}_t^{(k)}$ and $\tilde{Y}_t^{(k)}$, maintain desired relationships at frame boundaries. These relationships are as follows [31, 68]:

- *Preservation of magnitudes*:

$$|\tilde{Z}_{lL'}^{(k)}| = |\tilde{Y}_{lL}^{(k)}| \ \ \forall k \in 0\!:\!M-1,\ l \in 1\!:\!N_l \tag{2.5}$$

[2]Fitz *et al.* summarize the benefits of cubic phase interpolation for coding purposes (unmodified reconstruction) as follows: "In unmodified reconstruction, cubic interpolation prevents the propagation of phase errors introduced by unreliable parameter estimates, maintaning phase accuracy in transients, where the temporal envelope is important" [39].

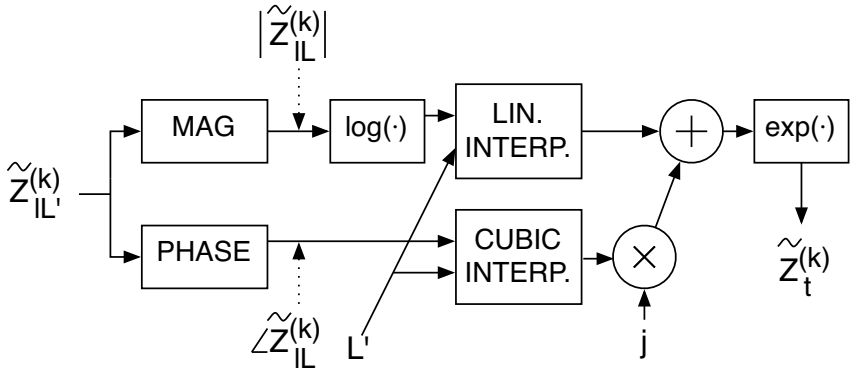


Figure 2.5: *Magnitude and phase interpolation for phase vocoder resynthesis*

where N_l is the number of frames.

- *Preservation of frequencies*:

$$\tilde{\omega}_{lL'}^{(k,Z)} \;=\; \tilde{\omega}_{lL'}^{(k,Y)} \quad \forall k \in 0\!:\!M\!-\!1,\; l \in 1\!:\!N_l \tag{2.6}$$

where each instantaneous frequency is defined as the average per-sample change in the unwrapped phase:

$$\begin{aligned} \tilde{\omega}_{lL}^{(k,Y)} &\triangleq \frac{1}{L}\left(\angle\tilde{Y}_{(l+1)L}^{(k)} - \angle\tilde{Y}_{lL}^{(k)}\right) \\ \tilde{\omega}_{lL'}^{(k,Z)} &\triangleq \frac{1}{L'}\left(\angle\tilde{Z}_{(l+1)L'}^{(k)} - \angle\tilde{Z}_{lL'}^{(k)}\right) \end{aligned} \tag{2.7}$$

- *Maintenance of phase continuity at frame boundaries*

Figure 2.6 displays the time scaling of a sinusoid with linearly increasing frequency and exponentially increasing amplitude. In the figure we observe the matching of sinusoidal magnitudes and instantaneous frequencies across frame boundaries, as well as the continuity of the phase in both analysis and resynthesis.

The standard *phase propagation* approach [83, 89, 31] maps $Y_{lL}^{(k)} \rightarrow Z_{lL}^{(k)}$ (see Figure 2.4) in order to preserve the desired relations between $\tilde{Y}_{lL}^{(k)}$ and $\tilde{Z}_{lL}^{(k)}$. Magnitudes and phases are treated separately. By the definitions (2.3 - 2.4) and the magnitude

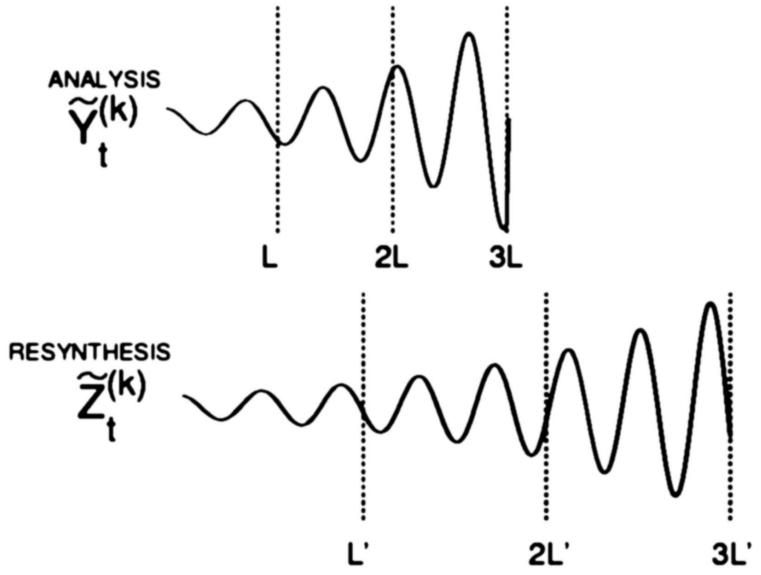


Figure 2.6: *Time scaling of single sinusoid with increasing frequency and amplitude*

preservation criterion (2.5), it becomes equivalent to specify:

$$|Z_{lL'}^{(k)}| \;=\; |Y_{lL}^{(k)}| \;\; \forall k \in 0\!:\!N\!-\!1,\; l \in 1\!:\!N_l \tag{2.8}$$

From (2.6), we see that instantaneous frequency preservation and phase continuity are satisfied if we maintain:

$$\angle \tilde{Z}_{(l+1)L'}^{(k)} \;=\; \angle \tilde{Z}_{lL'}^{(k)} + L'\hat{\omega}_{lL} \;\; \forall k \in 0\!:\!M\!-\!1,\; l \in 1\!:\!N_l \tag{2.9}$$

where $\hat{\omega}_{lL}$, the common instantaneous frequency, is derived:

$$\hat{\omega}_{lL} \;\triangleq\; \frac{\angle \tilde{Y}_{(l+1)L}^{(k)} - \angle \tilde{Y}_{lL}^{(k)}}{L} \tag{2.10}$$

Now, from (2.3 - 2.4):

$$\begin{aligned} \angle\tilde{Y}_{k,lL} &= \angle Y_{k,lL} + \frac{2\pi klL}{M} \\ \angle\tilde{Z}_{k,lL'} &= \angle Z_{k,lL'} + \frac{2\pi klL'}{M} \end{aligned} \tag{2.11}$$

Substituting (2.10) and $\alpha = L'/L$ into (2.9), then applying (2.11) obtains:

$$\angle Z^{(k)}_{(l+1)L'} = \angle Z^{(k)}_{lL'} + \alpha\left(\angle Y^{(k)}_{(l+1)L} - \angle Y^{(k)}_{lL}\right) \tag{2.12}$$

Since analysis phases are sampled only at the frame boundaries, the role of heterodyning in the phase vocoder analysis becomes clear: the heterodyned phase difference $\angle Y^{(k)}_{(l+1)L} - \angle Y^{(k)}_{lL}$ used in the transformation (2.12) is likely to be small compared with the actual phase difference $\angle\tilde{Y}^{(k)}_{(l+1)L} - \angle\tilde{Y}^{(k)}_{lL}$; the actual difference is exactly $2\pi klL/M$ greater than the heterodyned difference. As such, heterodyning facilitiates the requisite phase unwrapping task implicit in the instantaneous frequency determination (2.10).

2.2.3 Phase locking at the transient boundary

Unfortunately, the requirements for sound reproduction at the transient boundary [93, 39] differ somewhat with respect to the generic requirements proposed in the previous section; i.e., instantaneous frequency/magnitude preservation and maintenance of phase continuity at frame boundaries. For instance, suppose that frame l^* contains an abrupt-change transient, such as the onset of a new note event. Quatieri *et al.* suggest that the following qualities of the transient's *instantaneous temporal envelope* be maintained in resynthesis:

- *Preservation of magnitudes*

$$|\tilde{Z}^{(k)}_{lL'}| = |\tilde{Y}^{(k)}_{lL}| \;\; \forall k \in 0\!:\!M-1,\; l \in 1\!:\!N_l \tag{2.13}$$

- *Preservation of phase relationships* For all $j, k \in 0 : M-1$, wrapped phase

differences must be identical:

$$\mathrm{mod}\left(\angle\tilde{Z}^{(k)}_{l^*L'} - \angle\tilde{Z}^{(j)}_{l^*L'}, [\pi,\pi)\right) = \mathrm{mod}\left(\angle\tilde{Y}^{(k)}_{l^*L} - \angle\tilde{Y}^{(j)}_{l^*L}, [\pi,\pi)\right) \quad (2.14)$$

- *Appropriate scaling of magnitude time differences* If one time-scales a percussive event by a factor of two, we expect that the event will decay twice as slowly, even initially. Hence, under scaling factor α, we desire that the per-sample time difference of the resynthesis amplitude envelope be scaled by $1/\alpha$, immediately after the transient boundary. In other words, we desire:

$$\frac{1}{L'}\left(|\tilde{Z}^{(k)}_{(l^*+1)L'}| - |\tilde{Z}^{(k)}_{l^*L'}|\right) = \frac{1}{\alpha L}\left(|\tilde{Y}^{(k)}_{(l^*+1)L}| - |\tilde{Y}^{(k)}_{l^*L}|\right) \ \forall k \in 0\!:\!M-1 \quad (2.15)$$

The importance of preserving phase relationships as opposed to instantaneous frequencies is demonstrated by the following example. Consider a bandlimited impulse train at some sub-audio fundamental frequency, say 4 Hz. As this fundamental is sufficiently low, the result is heard as a periodic repetition of individual "ticks", each comprising a distinct transient event. The impulse train may be synthesized using a bank of sinusoidal oscillators for which each frequency is an integer multiple of the fundamental, and all amplitudes and phases are the same, i.e.,

$$y_t = A_0 \sum_{k=1}^{p(\omega)} \cos(k\omega t + \phi_0) \quad (2.16)$$

The number of sinusoids, $p(\omega)$, is chosen such that the frequency, $k\omega$, is always less than the Nyquist limit π rad/sample, i.e.,

$$p(\omega) = \lceil \pi/\omega \rceil - 1 \quad (2.17)$$

With $\phi_0 = 0$, $\omega = 5.699 \cdot 10^{-4}$ rad/sample establishing a 4.0 Hz fundamental at a sampling rate of 44.1 kHz , and A_0 establishing a peak amplitude of 1.0, the first 441 samples of the bandlimited impulse train are plotted in the top section of Figure 2.7. The resyntheses displayed in the bottom sections of the figure have

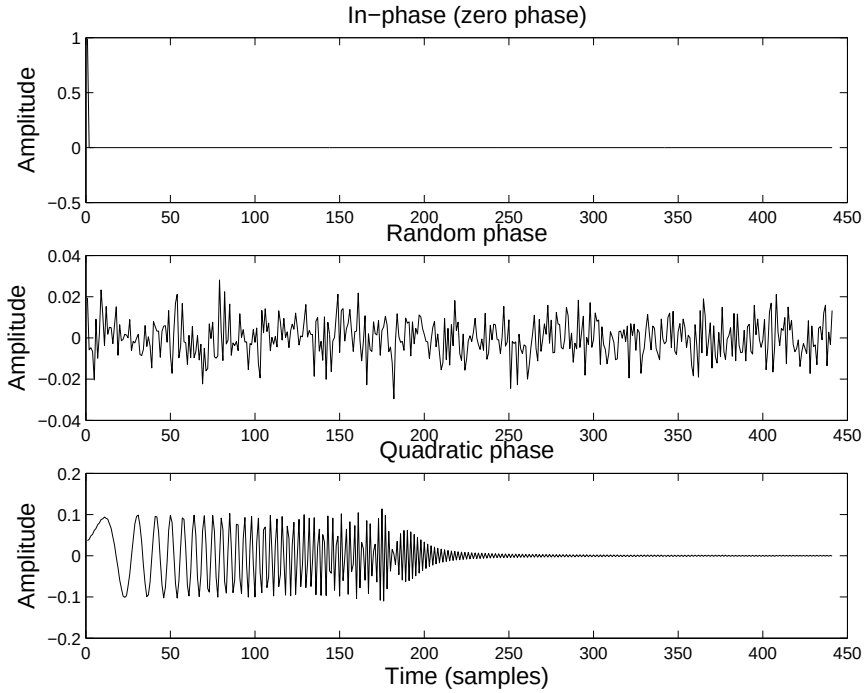


Figure 2.7: *Effect of phase relationships on transient reproduction*

identical amplitudes and frequencies for all sinusoidal components, but different phase relationships:

$$y_t = A_0 \sum_{k=1}^{p(\omega)} \cos(k\omega t + \phi_k) \tag{2.18}$$

In the middle section of the figure, ϕ_k is random, following a uniform distribution over $[-\pi, \pi)$. In the bottom section, $\phi_k = -1.0 \cdot 10^{-5} k^2$, producing a chirp with rapidly increasing frequency. This example demonstrates the role of phase relationships towards the perceived character of the transient reproduction.

To analyze the phase propagation algorithm with respect to the instantaneous temporal envelope criteria outlined above, we recall that the magnitude preservation is immediate from (2.8). As for the scaling of magnitude time differences, if we

multiply both sides of (2.15) by L' and substitute the definition $\alpha = L'/L$, we obtain:

$$|\tilde{Z}^{(k)}_{(l^*+1)L'}| - |\tilde{Z}^{(k)}_{l^*L'}| \;=\; |\tilde{Y}^{(k)}_{(l^*+1)L}| - |\tilde{Y}^{(k)}_{l^*L}| \;\; \forall k \in 0\!:\!M-1 \tag{2.19}$$

But (2.19) is immediate from the magnitude preservation criterion (2.13).

Unfortunately, the phase propagation generally fails to preserve phase relationships in the sense of (2.14). Even if (2.14) were true for a specific l^*, there is no guarantee, unless $\alpha = 1$, that this criterion will hold for subsequent frames. For instance, suppose the first transient boundary occurs when $t = 0$ (frame $l = 0$), and analysis phases are identically zero at this point. For this frame we may choose the resynthesis phases to match the analysis phases, hence preserving phase relationships. Now, suppose that the k^{th} sinusoid has constant frequency $\omega^{(k)}$. Suppose then at $t = l^*L$, a second transient occurs, for which amplitudes and frequencies experience a sudden discontinuity but the phases remain continuous. In this example, the analysis phases are as follows:

$$\angle\tilde{Y}^{(k)}_{l^*L} \;=\; \omega^{(k)} l^* L \;\; \forall k = 0\!:\!M-1 \tag{2.20}$$

Due to the phase propagation (2.9), the resynthesis phases obtain:

$$\angle\tilde{Z}^{(k)}_{l^*L} \;=\; \omega^{(k)} l^* L' \;\; \forall k = 0\!:\!M-1 \tag{2.21}$$

From (2.20) and (2.21), it follows that analysis phase relationships are not preserved in resynthesis. For $j \neq k$, the difference between analysis phases is $(\omega^{(k)} - \omega^{(j)})l^*L$; the corresponding difference between resynthesis phases is $(\omega^{(k)} - \omega^{(j)})l^*L'$. Unless $L' = L$, meaning that there is no modification, the phase differences will fail to match for arbitrary $\omega^{(j)}$, $\omega^{(k)}$, and l^*.

To remedy this, Quatieri *et al.* [93] propose *locking* resynthesis to analysis phases at the transient boundary[3]; i.e.,

$$\angle Z^{(k)}_{l^*L'} \quad = \quad \angle Y^{(k)}_{l^*L'} \tag{2.22}$$

While resetting the resynthesis phases modifies instantaneous frequencies for $t \in (l^*-1)L' : l^*L'$, the latter becomes less problematic than modifying phase relationships in the immediate vicinity of the transient boundary. For instance, consider the impulse signal plotted in the top section of Figure 2.7. This signal is synthesized via (2.16) using a fundamental frequency of 4 Hz. If instead the fundamental is 6 Hz and all other parameters are unchanged[4], the transient characteristics remain qualitatively similar despite the 50% increase in all component frequencies. A comparison is shown in Figure 2.8. Finally, it is important to emphasize that the phase locking at the transient boundary, while an effective solution for reducing artifacts due to abrupt-change transients, requires the *detection* of the frame l^* in which the transient occurs.

2.2.4 Phase locking throughout transient regions

A problem with phase locking only at transient boundaries is that the lock is not maintained during transient regions of nonzero duration unless $\alpha = 1$. This is clear from the discussion in the previous section surrounding (2.20 - 2.21). Maintaining phase relationships throughout transient regions becomes especially important in the resynthesis of textural sounds. Particularly problematic are textures composed of a large collection of superposed, randomly spaced impulsive events, such as rain, crackling fire, and so forth. Figure 2.7 displays the effects of various phase distortions on a single impulsive event.

To this end, a number of authors, for instance Levine [74, 75], and later Duxbury

[3]The actual scheme is more general: it involves detecting specific groups of sinusoids which undergo abrupt changes in amplitude, phase, or frequency characteristics. Phase locking is then applied individually to each group. In this way, the method can deal with more complex sounds where transient phenomena may overlap significantly in time, but become more sparse throughout time when restricted to particular subbands.

[4]The number of sinusoids, $p(\omega_0)$, is also adjusted via (2.17) to avoid aliasing.

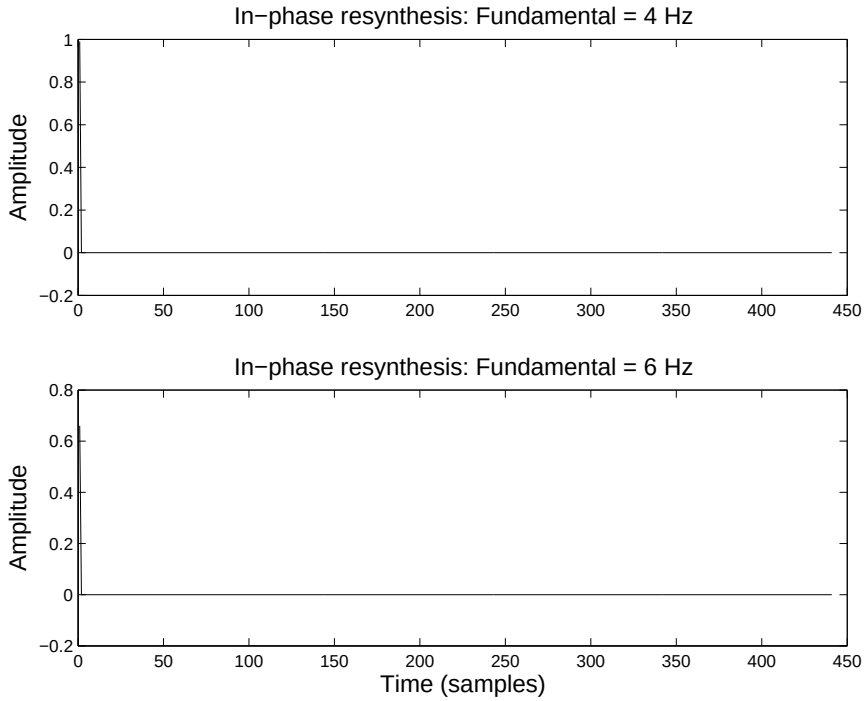


Figure 2.8: *Effect of frequency relationships on transient reproduction. The top figure uses a fundamental frequency of* 4 *Hz, the bottom uses* 6 *Hz. Despite the 50 % increase in all oscillator frequencies, little qualitative difference can be seen or heard*

et al. [35] propose the locking of resynthesis to analysis phases at the beginning of the transient region, as well as setting $\alpha = 1$ to maintain phase locking throughout the entire transient region. The scaling factor may be adjusted during steady-state regions to achieve the desired resynthesis tempo which equals α times the analysis tempo. For instance, if the input signal's duration is 5000 samples and the desired stretch factor equals 2.0, and the initial 1000 samples are designated as a transient region, one specifies $\alpha = 1$ for the first 1000 samples and $\alpha = 2.25$ for the remainder.

One problem with this method of locking resynthesis phases to analysis phases during transient regions is that the magnitude time differences are no longer scaled by the inverse of the scaling factor throughout these regions. Instead, the resynthesis's initial decay envelope becomes identical to that for the analysis. If transient regions are sufficiently long, the result will begin to sound like the same instrument, but

played at a different tempo. Duxbury *et al.* claim this as desirable: "...despite being an ill-posed problem, it is generally agreed that when time scaling audio, the aim is for the resulting signal to sound as if the piece is being played at a different tempo" [35]. However, this approach severely restricts the user's ability to effect timbral changes. Furthermore, it may generate artifacts in pitch scaling if the latter is implemented by time scaling followed by sampling rate conversion. In pitch scaling, we expect the initial decay rates of the resynthesis to match those of the original signal. If, instead, these rates match after time scaling, they will no longer match after the sampling rate conversion.

To this end, we seek a more flexible representation of transient regions within the context of sinusoidal modeling in which the temporal support of the raw information necessary to reconstruct these regions is as short as possible. One such representation, introduced by the author and Leistikow [116], effectively hybridizes source-filter and sinusoidal modeling to achieve this task. This approach relates to aspects of the nonlinear parameter estimation by Wold [122], the Prony modeling by Laroche [66], earlier transient modeling work by the author and Gouyon [115], spectral estimation work by Qi *et al.* [92], as well as the signal-level models used in the transcription methods of Cemgil *et al.* appearing around the same time [20, 19]. Section 2.3 presents a brief overview of this hybrid sinusoidal/source-filter approach to time scaling, as well as detailing new kinds of delay-based effects based on splitting the transient information among different sinusoidal components.

In conclusion, essentially two types of detection are required to reduce time/pitch scaling artifacts for sounds with significant transient content: first, the detection of abrupt-change phenomena; second, the identification of transient *regions* of nonzero width (meaning the determination of beginning and end points for these regions). Furthermore, as the following section demonstrates, applications are by no means limited to time and pitch scaling.

2.3 Improved transient region modeling via hybrid sinusoidal/source-filter model

One may recall the sinusoidal modeling approaches of Levine and Smith [75], commonly called "transients + sines + noise", for which the signal is segregated in time into regions containing *either* transient information *or* "sines plus noise". Figure 2.9 displays a schematic for this representation.

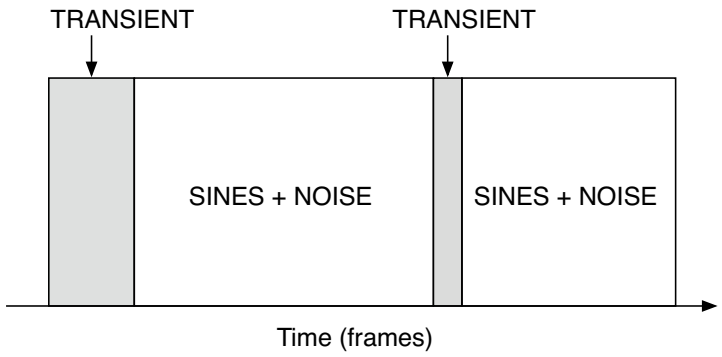


Figure 2.9: *"Transients + sines + noise" representation, after [75]*

By contrast, [116] proposes a *convolutive* representation, which may be summarized as "transients $\star$ sines + noise". Here each sinusoid consists of an exponentially damped, quadrature oscillator which is *driven* by the information necessary to reconstruct the transient region. A block diagram of this approach is displayed in Figure 2.10.

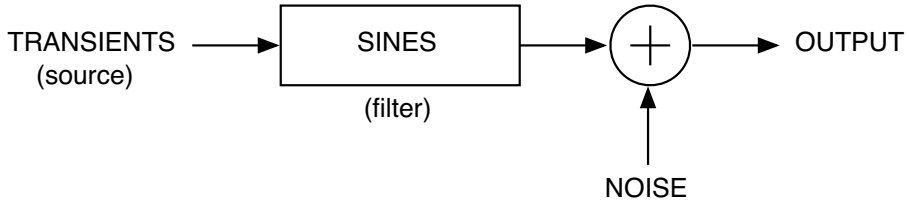


Figure 2.10: *"Transients $\star$ sines + noise", or convolutive representation*

The "transients $\star$ sines + noise" representation facilitates the modeling of *attack* transients, which consist of an abrupt-change event signifying the onset of a new note,

followed by a transient region where the sinusoidal amplitudes undergo a rapid, quasi-exponential decay. Attack transients may also exhibit textural characteristics which are difficult to represent by a direct sum of exponentially damped sinusoids. As later demonstrated, the source-filter representation facilitates time-scaling modifications in such a way that preserves textural characteristics as well as guarantees appropriate scaling of the decay rate by the inverse of the time expansion factor, following (2.15), because the effective temporal support of the "source" is greatly reduced with respect to that of the original signal.

2.3.1 The driven oscillator bank

The filter (sines) component in Figure 2.10 consists of a *driven oscillator bank*, displayed in Figure 2.11. In the figure, $s_t^{(I)}(k)$ denotes the in-phase component and

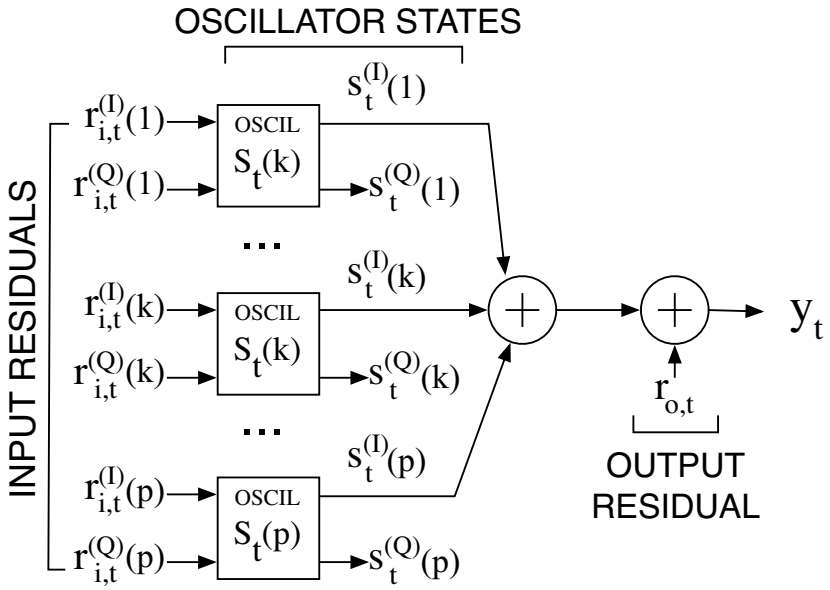


Figure 2.11: *Driven oscillator bank*

$s_t^{(Q)}(k)$ the quadrature component of the k^{th} oscillator at time t. The amplitude and

phase of this oscillator may be retrieved:

$$\begin{aligned} A_t(k) &= \sqrt{\left[s_t^{(I)}(k)\right]^2 + \left[s_t^{(Q)}(k)\right]^2} \\ \phi_t(k) &= \tan^{-1}\left[\frac{s_t^{(Q)}(k)}{s_t^{(I)}(k)}\right] \end{aligned} \tag{2.23}$$

The in-phase and quadrature *input residuals* associated with the k^{th} oscillator are $r_{i,t}^{(I)}(k)$ and $r_{i,t}^{(Q)}(k)$, which drive the respective oscillator states, $s_t^{(I)}(k)$ and $s_t^{(Q)}(k)$. Starting from zero initial state for $t \leq 0$, the residuals must supply the excitation for subsequent oscillation. However, suppose $r_t^{(I)}(k)$ and $r_t^{(Q)}(k)$ are identically zero for all $t \geq T$, where T is subsequent to the excitation, then the k^{th} oscillator's contribution becomes for $t \geq T$, a pure, exponentially decaying sinusoid. Residual contributions which persist after the onset time contribute to non-sinusoidal qualities, such as the perceived "texture" of the attack.

For the k^{th} oscillator, the relation between the current oscillator state, the previous oscillator state, and the residual at time t may be represented by the following (linear) recursion:

$$\begin{bmatrix} s_t^{(I)}(k) \\ s_t^{(Q)}(k) \end{bmatrix} = e^{\gamma_t(k)} \begin{bmatrix} \cos\omega_t(k) & -\sin\omega_t(k) \\ \sin\omega_t(k) & \cos\omega_t(k) \end{bmatrix} \begin{bmatrix} s_{t-1}^{(I)}(k) \\ s_{t-1}^{(Q)}(k) \end{bmatrix} + \begin{bmatrix} r_t^{(I)}(k) \\ r_t^{(Q)}(k) \end{bmatrix} \tag{2.24}$$

The output, y_t, sums over the in-phase oscillator states, adding a scalar *output residual*, $r_{o,t}$:

$$y_t = \sum_{k=1}^{p} s_t^{(I)}(k) + r_{o,t} \tag{2.25}$$

This output residual accounts for additive noise due to the recording process. It becomes important to distinguish additive noise from the possibly noise-like transient information responsible for non-sinusoidal qualities of the attack, the latter encoded by input residuals. In this way, the driven oscillator bank effectively generalizes the

canonical "sines + noise" model introduced by Serra and Smith, also known as "spectral modeling synthesis" (SMS) [106], although it specializes this approach as well, not allowing for arbitrary envelope shapes. In SMS, a single residual is obtained by subtracting the sinusoidal resynthesis (absent modification) from the original signal. If all input residuals are identically zero except for the initial excitation, the SMS residual equals $r_{o,t}$; the present method augments this by separating residual information inherent to the acoustic source ($r_{i,t}$) from information inherent to the recording process ($r_{o,t}$). Furthermore, the association of input residuals with individual oscillators generates novel resynthesis possibilities which go beyond the canonical time/pitch scaling paradigm; e.g., oscillator-variable delay effects. Further details concerning these effects are discussed in Section 2.3.4.

The oscillator bank also may be viewed as a collection of second-order resonant filters of bandpass/formant type, excited by input residuals: hence the "source-filter" interpretation of Figure 2.10. This interpretation results from analyzing transfer relations between $r_{i,t}^{(I)}(k)$ and $s_t^{(I)}(k)$, and between $r_{i,t}^{(Q)}(k)$ and $s_t^{(I)}(k)$, since only $s_t^{(I)}(k)$ is observed in the output. Assuming $\omega_t(k)$ and $\gamma_t(k)$, are constant with respect to t, taking z-transforms of both sides of (2.24) obtains as follows.

$$S^{(I)}(z;k) \;=\; H^{(I\rightarrow I)}(z;k)R^{(I)}(z;k) + H^{(Q\rightarrow I)}(z;k)R^{(Q)}(z;k) \tag{2.26}$$

where $R^{(I)}(z;k)$, $R^{(Q)}(z;k)$, and $S^{(I)}(z;k)$, assuming appropriate convergence of the ensuing summations[5], are defined as follows.

$$\begin{aligned} S^{(I)}(z;k) &\triangleq \sum_{t=-\infty}^{\infty} s_t^{(I)}(k)z^{-t} \\ R^{(I)}(z;k) &\triangleq \sum_{t=-\infty}^{\infty} r_t^{(I)}(k)z^{-t} \\ R^{(Q)}(z;k) &\triangleq \sum_{t=-\infty}^{\infty} r_t^{(Q)}(k)z^{-t} \end{aligned} \tag{2.27}$$

[5]In other words, we consider $r_t^{(I)}(k)$, $r_t^{(Q)}(k)$, and $s_t^{(I)}(k)$ bounded and causal and $z \in \mathcal{C}$; $|z| \leq 1$. The boundedness of $s_t^{(I)}(k)$ is guaranteed when $\gamma(k)$, the assumed constant value of $\gamma_t(k)$, is less than 0.

and

$$\begin{aligned} H^{(I\to I)}(z;k) &= \frac{1 - e^{\gamma(k)}\cos\omega(k)z^{-1}}{1 - 2e^{\gamma(k)}\cos\omega(k)z^{-1}e^{2\gamma(k)}z^{-2}} \\ H^{(Q\to I)}(z;k) &= \frac{e^{\gamma(k)}\sin\omega(k)z^{-1}}{1 - 2e^{\gamma(k)}\cos\omega(k)z^{-1} + e^{2\gamma(k)}z^{-2}} \end{aligned} \tag{2.28}$$

In (2.28), $\omega(k)$ is the (assumed) constant value of $\omega_t(k)$ and $\gamma(k) < 0$ is the constant value of $\gamma_t(k)$. Both transfer functions share the same denominator; common poles are $z = e^{\gamma(k)\pm j\omega(k)}$.

Figure 2.12 plots magnitude responses $|H^{(I\to I)}(\omega;k)|$ and $|H^{(Q\to I)}(\omega;k)|$ as a function of radian frequency ω, for $\gamma(k) = -0.5$ and $\omega(k) \in \{\pi/10, \pi/5, \pi/2\}$. The

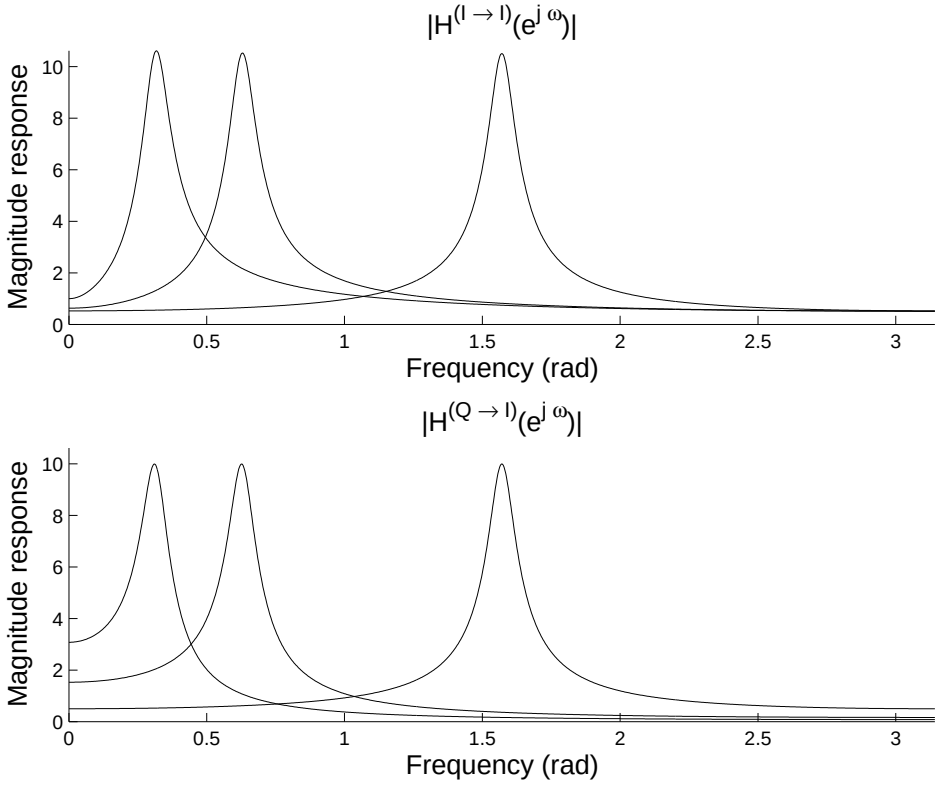


Figure 2.12: *Magnitude responses of oscillator components viewed as filters*

in-phase response, $|H^{(I\to I)}(\omega;k)|$, obtains a pure bandpass characteristic while the

quadrature response, $|H^{(Q\rightarrow I)}(\omega;k)|$, obtains more of a formant/lowpass characteristic at low values of $\omega(k)$, changing to a bandpass characteristic at high values.

2.3.2 State space representation, Kalman filtering and residual extraction

If we concatenate oscillator and residual states into the following vectors

$$\begin{aligned} s_t &\triangleq \left[s_t^{(I)}(1)\ s_t^{(Q)}(1)\ s_t^{(I)}(2)\ s_t^{(Q)}(2)\ \ \ldots\ s_t^{(I)}(p)\ s_t^{(Q)}(p) \right]^T \\ r_{i,t} &\triangleq \left[r_{i,t}^{(I)}(1)\ r_{i,t}^{(Q)}(1)\ r_{i,t}^{(I)}(2)\ r_{i,t}^{(Q)}(2)\ \ \ldots\ r_{i,t}^{(I)}(p)\ r_{i,t}^{(Q)}(p) \right]^T \end{aligned} \tag{2.29}$$

the model (2.24 - 2.25) may be expressed in state-space form:

$$\begin{aligned} s_t &= F_t s_{t-1} + r_{i,t} \\ y_t &= H s_t + r_{o,t} \end{aligned} \tag{2.30}$$

where $F_t \in \mathbb{R}^{2p\times 2p}$ is block diagonal with 2×2 blocks $F_t(k)$:

$$F_t(k) \triangleq e^{\gamma_t(k)} \begin{bmatrix} \cos\omega_t(k) & -\sin\omega_t(k) \\ \sin\omega_t(k) & \cos\omega_t(k) \end{bmatrix} \tag{2.31}$$

and $H \in \mathbb{R}^{1\times 2p}$ sums over the in-phase components of s_t:

$$H \triangleq \begin{bmatrix} 1 & 0 & 1 & 0 & \ldots & 1 & 0 \end{bmatrix} \tag{2.32}$$

If we model input and output residuals as independent Gaussian vectors, i.e.,

$$\begin{aligned} r_{i,t} &\sim \mathcal{N}\left(\mathbf{0}_{2p\times 1}, q\mathbf{I}_{2p}\right) \\ r_{o,t} &\sim \mathcal{N}(0,r) \end{aligned} \tag{2.33}$$

where, additionally, $r_{o,t}$ and $r_{i,t}$ are independent, one may estimate the oscillator state

recursively for all $t \in 1{:}N$ using a Kalman filter [57]. This estimate is

$$\hat{s}_t \triangleq E(s_t|y_{1:t}) \tag{2.34}$$

From [57], the Kalman filtering recursions are:

- **Time Update**

$$\begin{aligned} \hat{s}_{t+1|1:t} &= F_{t+1}\hat{s}_{t|1:t} \\ P_{t+1|1:t} &= F_{t+1}P_{t|1:t}F_{t+1}^T + qI \end{aligned} \tag{2.35}$$

- **Measurement Update**

$$\begin{aligned} K_{f,t+1} &= P_{t+1|1:t}H^T(HP_{t+1|1:t}H^T + r)^{-1} \\ \hat{s}_{t+1|1:t+1} &= \hat{s}_{t+1|1:t} + K_{f,t+1}\left(y_{t+1} - H\hat{s}_{t+1|1:t}\right) \\ P_{t+1|1:t+1} &= \left(I - K_{f,t+1}H\right)P_{t+1|1:t} \end{aligned} \tag{2.36}$$

where, for $1 < r, t < N$,

$$\begin{aligned} \hat{s}_{t|1:r} &\triangleq E(s_t|y_{1:r}) \\ P_{t|1:r} &\triangleq Cov(s_t|y_{1:r}) \end{aligned} \tag{2.37}$$

These recursions, which run for $t \in 1:N$, are initialized:

$$\begin{aligned} \hat{s}_0 &= \mathbf{0}_{2p\times 1} \\ P_0 &= \epsilon^{-1}\mathbf{I}_{2p} \end{aligned} \tag{2.38}$$

and the limit is taken as $\epsilon \to 0$. The estimate $\hat{s}_t$, as defined in (2.34), is taken to be $\hat{s}_{t|1:t}$; the latter is commonly referred to as the *filtered estimate*. Combined with the original state-space model (2.30), the filtered estimates are used to extract residual

quantities, as follows.

$$\begin{aligned} \hat{r}_{i,t} &= \hat{s}_t - F_t \hat{s}_{t-1} \\ \hat{r}_{o,t} &= y_t - H \hat{r}_{i,t} \end{aligned} \tag{2.39}$$

Resynthesis proceeds by substituting $\hat{r}_{i,t}$ and $\hat{r}_{o,t}$ in place of $r_{i,t}$ and $r_{o,t}$ in the original state-space model. Absent modification to the *state dynamics* parameters $\gamma(k)_t, \omega(k)_t$, or the Kalman parameters q, r, the resynthesis will be *perfect reconstruction*, producing exactly y_t. In general one may modify the residuals $\hat{r}_{i,t}$ and $\hat{r}_{o,t}$, and the state dynamics parameters to produce a variety of transformations; e.g., time and pitch scaling, and other novel effects. Section 2.3.4 explores these possibilities in greater detail.

Since the bulk of the residual energy accounts for the excitation, the input residuals' effective temporal support becomes significantly reduced with respect to the original attack transient. Figure 2.13 compares the temporal supports of a $'\mathrm{D2}'$ piano tone with those of the in-phase residuals corresponding to the first and 32^{nd} partials, using a 32-oscillator harmonic-comb model fit via the methods of [117]. Indeed, each residual appears quite similar despite the difference in oscillator frequencies. The excitation part where the energy of each residual is most significant occupies at most a few hundred samples. As indicated via informal listening tests, one may truncate the remainder without affecting the recognizable character of the original piano tone.

2.3.3 Tuning of the residual covariance parameters

If frequency and decay trajectories are modified, the residual covariance parameters q and r, defined via (2.33), exert a considerable influence on the resynthesis. However, we show that the parameterization $\{q, r\}$ is redundant; assuming $0 < q, r < \infty$, only the *ratio* $\rho \triangleq r/q$ affects the resynthesis. Furthermore, the input residual becomes identically zero when $\rho \to \infty$, and the output residual vanishes when $\rho \to 0$. Implications for intermediate values of ρ are as follows.

- A large ρ favors a small input residual and a large output residual. Here the

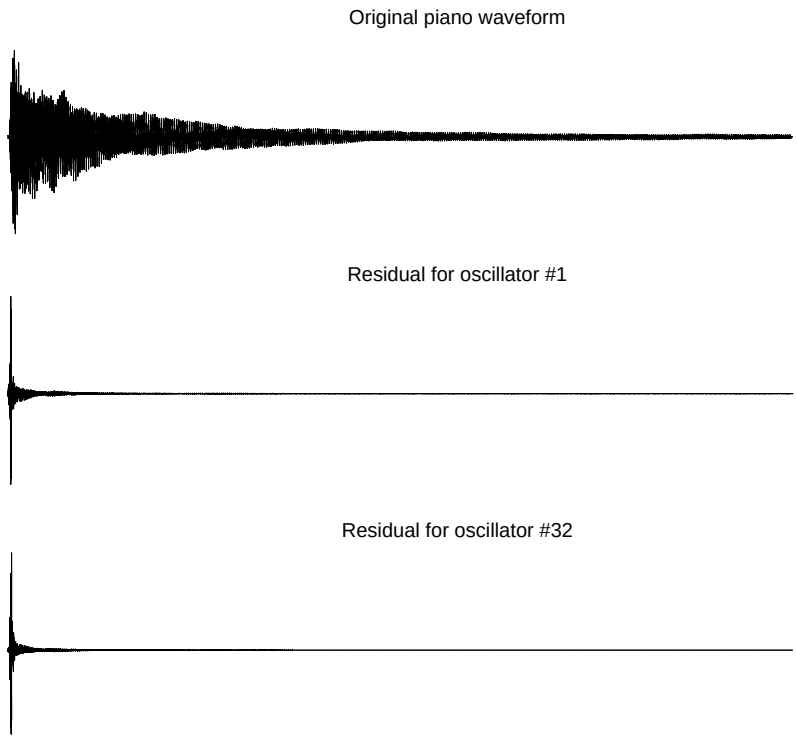


Figure 2.13: *Residuals vs. original attack transient for* $'\mathrm{D2}'$ *piano tone*

state estimation becomes more robust to additive noise, but it loses the ability to track actual variations in state dynamics if $\gamma_t(k)$ and $\omega_t(k)$ are assumed constant for any length of time.

- A small ρ favors a large input residual and a small output residual, yielding increased ability to track variations in state dynamics at the expense of greater sensitivity to additive noise. Furthermore, an excessively small ρ may lead to envelope distortion artifacts in resynthesis. If input residuals are large, the individual state resyntheses (each $2k-1$ component of $\hat{s}_t$ for $k \in 1:p$) may also be large, even with respect to y_t. Since output residuals are small, however, the sum of these resyntheses (before the output residual is added) must be close to y_t. This suggests that the individual state resyntheses must undergo phase cancellation in the summation producing $y_t - r_{o,t}$. With modifications (e.g., pitch scaling), for sufficiently large t the individual state resyntheses will

begin to deviate from the specific phase relationships responsible for the cancellation. Even though the individual state resyntheses may decay over time, their summation may grow substantially over time, leading to the perception of an unnaturally soft attack. Such envelope distortion artifacts, if they exist only at moderate levels, may be corrected in postprocessing by applying envelope corrections. Nonetheless, one should avoid specifying ρ too small.

To analyze the effect of ρ on the residual extraction, we first establish that the filtered estimates depend on q and r only through ρ. In other words, if for any $c > 0$, we replace $q \to cq$ and $r \to cr$ in the Kalman recursions (2.35, 2.36), and replace P_0 in the initialization (2.38) by cP_0, an identical expression for $\hat{s}_{t|1:t}$ should result.

To begin, (2.35) and (2.36) obtain the following identities:

$$P_{t+1|1:t+1} = \left[\left(F_{t+1}P_{t|1:t}F_{t+1}^T + qI\right)^{-1} + r^{-1}H^TH\right]^{-1} \tag{2.40}$$

$$K_{f,t+1} = \left(F_{t+1}P_{t|1:t}F_{t+1}^T + qI\right)H^T\left[H\left(F_{t+1}P_{t|1:t}F_{t+1}^T + qI\right)H^T + r\right]^{-1} \tag{2.41}$$

$$\hat{s}_{t+1|1:t+1} = \left(I - K_{f,t+1}H\right)F\hat{s}_{t|1:t} + K_{f,t+1}y_{t+1} \tag{2.42}$$

Now, define:

$$\begin{aligned} P_{t+1|1:t+1}^{(c)} &\triangleq \left[\left(F_{t+1}P_{t|1:t}F_{t+1}^T + cqI\right)^{-1} + (cr)^{-1}H^TH\right]^{-1} \\ &= c\left[\left(F_{t+1}(c^{-1}P_{t|1:t})F_{t+1}^T + qI\right)^{-1} + r^{-1}H^TH\right]^{-1} \end{aligned} \tag{2.43}$$

Hence, if q is replaced by cq and r by cr in (2.40),

$$P_{t+1|1:t+1}^{(c)} = cP_{t+1|1:t+1} \quad \forall t \in 1:N \tag{2.44}$$

To achieve (2.44) for all $t \in 1:N$, it suffices to set $P_0^{(c)} = cP_0$ in the initialization (2.38). In the limit as $\epsilon \to 0$, however, these initializations each tend to the same result.

Similarly, define $K^{(c)}_{f,t+1}$ by replacing q by cq and $P_{t|1:t}$ by $P^{(c)}_{t|1:t}$ on the r.h.s. of (2.42). With some algebra, it is easily shown:

$$K^{(c)}_{f,t+1} \quad = \quad K_{f,t+1} \tag{2.45}$$

Since no other term in (2.42) besides $K^{(c)}_{f,t+1}$ depends on c, it follows that $\hat{s}_{t|1:t}$ remains unchanged, as was to be shown.

To justify the assertions made at the beginning of this section concerning the effect of ρ on the state estimates, we consider the limiting cases, $\rho \to \infty$ and $\rho \to 0$. As established previously, no loss of generality results by fixing $q = 1$ and $r = \rho$. If $r \to \infty$, the term $\left[H\left(F_{t+1}P_{t|1:t}F^T_{t+1} + qI\right)H^T + r\right]^{-1}$ vanishes; by (2.41), all elements of $K_{f,t+1}$ converge to 0. By (2.42):

$$\hat{s}_{t+1|1:t+1} \quad \to \quad F\hat{s}_{t|1:t} \tag{2.46}$$

According to (2.39), (2.46) implies that $r_{i,t} \to \mathbf{0}_{2p\times 1}$, as was to be shown.

On the other hand, multiplying both sides of (2.42) on the left by H obtains:

$$H\hat{s}_{t+1|1:t+1} \quad = \quad (H - HK_{f,t+1}H^T)F_{t+1}\hat{s}_{t|1:t} + HK_{f,t+1}y_{t+1} \tag{2.47}$$

As $\rho \to 0$, it follows from (2.41), that $HK_{f,t+1} \to 1$. Substituting this limit into (2.47) obtains $H\hat{s}_{t|1:t+1} \to y_{t+1}$. As a result, (2.39), implies that $r_{o,t} \to 0$, as was to be shown.

2.3.4 Analysis, transformation and resynthesis

The general analysis-transformation-resynthesis process is summarized by Figure 2.14.

- **Analysis**: The input signal $y_{1:N}$ is analyzed to extract frequency and decay trajectories $\omega_{1:N}(k)$ and $\gamma_{1:N}(k)$ for $k \in 1:p$. These trajectories are converted into the state transition matrix sequence $F_{1:N}$ by repeated application of (2.31). Then $y_{1:N}$ and $F_{1:N}$ are passed to the Kalman filter consisting of the recursions

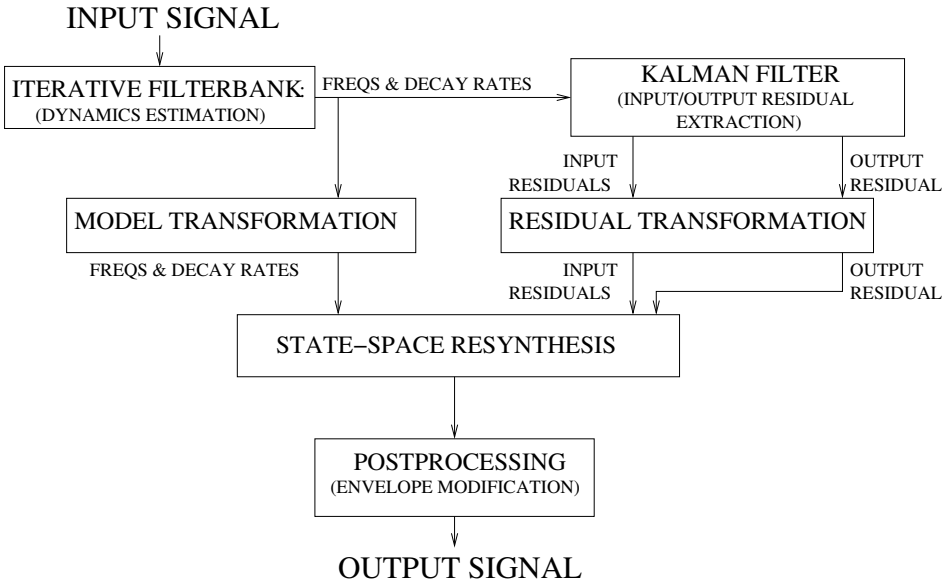


Figure 2.14: *Block diagram for analysis-transformation-resynthesis using the hybrid sinusoidal/source-filter model*

(2.35, 2.36), initialized by (2.38). The Kalman filter produces the sequence of filtered state estimates $\hat{s}_{1:N}$ (defined via (2.34)), from which, given $y_{1:N}$, the residual sequences $r_{i,1:N}$ and $r_{o,1:N}$ are extracted via (2.39).

- **Transformation** The frequency and decay trajectories may be modified, along with the residual sequences, to produce new versions of $F_{1:N}$, $r_{i,1:N}$, and $r_{o,1:N}$. If storage is at a premium, all but the initial *excitation* part of these residuals may be discarded without too much effect on the quality of the resynthesis.

- **Resynthesis** The modified sequences: $F_{1:N}$, $r_{i,1:N}$, and $r_{o,1:N}$, are presented to the state-space model (2.30) which synthesizes a preliminary output signal. If needed, envelope distortion artifacts caused by underspecification of the ratio $\rho \stackrel{\Delta}{=} r/q$ (see Section 2.3.3) may be addressed in postprocessing which yields the final output signal.

Extraction of the frequency and decay trajectories, $\gamma_t(k)$ and $\omega_t(k)$, is in general a quite difficult problem for which the literature remains incomplete. Nevertheless, there exist many special cases concerning acoustic sources for which feasible extraction

methods have been developed. For attack transients originating from quasi-harmonic sources, for instance, the iterative filterbank method of [117] may be used. A *quasi-harmonic* source obeys the following criteria [117]:

1. Frequency and decay trajectories are modeled as constant over frames. However, variations in amplitude and phase characteristics, as encoded by the oscillator state, may proxy for small, local variations in frequencies and decays.

2. The frequency distribution of spectral components admits a hierarchy in which components cluster about *principal harmonics*. Figure 2.15 displays an example frequency distribution. The frequency associated with the principal harmonic

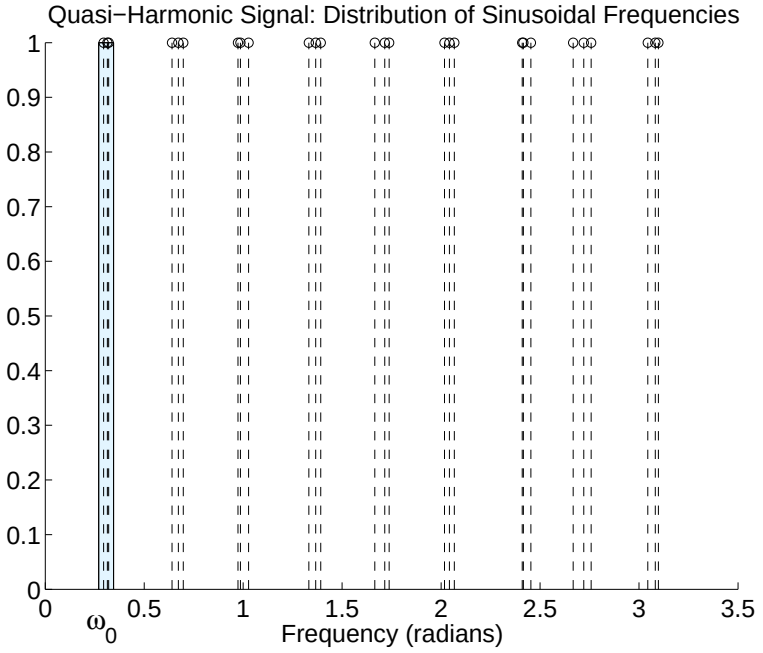


Figure 2.15: *Sample frequency distribution for quasi-harmonic source*

is defined as the amplitude-weighted average of all frequencies within its associated cluster. Component frequencies assigned to a particular cluster may be arbitrarily close.

3. Principal harmonic frequencies exist roughly about integer multiples of some

fundamental, hence the term *quasi-harmonic*. Even moderate amounts of inharmonicity are allowed, as long as the spacing between principal harmonics varies smoothly over the entire frequency range.

Many attack transients from acoustic sources (piano, marimba, bowed string, some bell/chime tones etc.) may be modeled as quasi-harmonic; some cannot, such as cymbals and gongs. For more general sources the literature is by no means complete, however Prony-based methods such as [66, 115] may be useful as long as the number of sinusoidal components does not become too large. The advantage of methods such as [117] as well as related "frequency zooming" work of Karjalainen *et al.* [58] and Esquef *et al.* [37] is the use of *spectral hierarchies* to decompose the frequency/decay estimation problem into a collection of smaller problems, each involving only a few sinusoidal components.

Since input and output residuals as well as frequencies and decay trajectories may be transformed, an almost limitless variety of resyntheses are possible using the framework of Figure 2.14. We briefly discuss a few options relating to canonical analysis-synthesis tasks (e.g., time and pitch scaling, cross-synthesis), as well as introduce several types of effects specific to this framework, which process each residual by different means.

- **Time scaling** The identified frequency and decay trajectories are resampled via bandlimited sinc interpolation[6] [109]. Each decay trajectory is then multiplied by the inverse of the stretch factor to achieve the magnitude time difference scaling indicated by (2.15). Input and output residuals are time scaled according to the method discussed in Section 2.2.4, keeping in mind that the transient regions for each residual are considerably shorter than the transient region for the original signal. Alternatively, an *ad hoc* residual processing method following [116] may be used; this method works especially well for quasi-harmonic sources. This method defines the excitation region as the first M samples after the onset, where M is chosen so that, averaging across all residuals, a certain fraction of the overall residual energy is captured within the excitation region.

[6] If these trajectories are constant, they are extended to cover the new signal duration.

The excitation region remains unmodified, paralleling the "region locking" approach of Section 2.2.4, while the remainder of the residual is processed by some adaptive pitch-synchronous overlap-add technique. In [116], the authors find that the WSOLA (wave-synchronous overlap-add) protocol of Verhelst and Roelands [118] achieves excellent results.

- **Pitch scaling** Transposition pitch scaling may be implemented by time scaling followed by sampling rate conversion. However, the present method allows direct modification of the frequency trajectories. For transposition, each trajectory is multiplied by the transposition factor while residuals and decay trajectories are preserved. This leads to more general types of pitch scaling effects; e.g., inharmonic scaling, timbre superposition (reassigning frequency components to those obtained from a different source), and time-varying scaling, even at audio rates.

- **Cross synthesis** If the analysis is performed on several sounds, residuals and models (meaning the frequency and decay trajectories) may be interchanged. Hybrid cross-syntheses become possible where some residuals come from one source, and the rest from another. Furthermore, residuals extracted using different source-filter analyses; e.g., linear predictive coding (LPC; see [6]) may replace the input residuals.

- **Residual modifications** Each input residual or groups of such may be processed by independent means. A simple approach is to feed each pair of in-phase and quadrature residuals corresponding to a single oscillator through an independent delay line. If different delay times are set, this results in a splitting of the excitation among the various harmonics, as if each harmonic were "plucked" by a different excitation. If the delay time varies directly or inversely with frequency, a "strumming" sound may be achieved. Moreover, if the independent delay lines become regenerative, polyrhythmic textures may be superimposed upon or seem to emerge from the original sound, creating quite striking effects.

Chapter 3

The role of musical structure

3.1 Introduction

In Chapter 1, two primary objectives are introduced for the transient detection as applied to musical audio signals:

- First, the identification of *abrupt changes* in spectral content. These often arise from a performer's action, associating with *musical onsets*.
- Second, the identification of *transient regions* of nonzero width. Throughout these regions, the signal fails to be "well modeled" by representations indicating constant or slowly changing spectral content[1]. Transient regions are often found near onset boundaries, for instance, during attack portions of note events.

Satisfying these objectives goes a long way towards obtaining high-quality, low-storage analysis-based sound transformations for recordings with significant transient

[1]Here as in the abrupt-change designation, we are assuming some parsimonious, meaningful signal representation, such as the superposition of a limited number of sinusoids with time-varying parameters. It is these parameters, (i.e., amplitudes, frequencies, and phases), which we expect to undergo abrupt change; during transient regions we say only that no such parsimonious representation may be found for which the parameters are constant or slowly varying. Of course, via the discrete Fourier transform (DFT), one may represent *any* real-valued signal of finite length N with $\lceil N+1 \rceil/2$ complex sinusoidal components of constant frequencies and amplitudes. However, this representation lacks parsimony; it overfits noise elements; it is psychoacoustically and cognitively irrelevant and thus perceived *artifacts* result from most analysis-based sound transformations. See Chapter 2 for a more involved discussion of these artifacts.

content. Of particular interest are time and pitch scaling transformations. In Chapter 2, we discuss phase-vocoder and sinusoidal-modeling approaches as proposed by Quatieri et al. [93], Levine and Smith [75], Duxbury et al. [35]; and the hybrid sinusoidal/source-filter representation of the author and Leistikow [117, 116]. The latter is discussed at length in Section 2.3. Common to all these methods is the use of abrupt-change detection or region identification to reduce perceived artifacts under transformation.

3.2 The role of musical structure

Unfortunately, it becomes difficult to identify transient regions and points of abrupt change for complex, heterogeneous musical sounds. Even near-optimal statistical methods monitoring spectral change via piecewise constant autoregressive models, for instance the online approaches of Basseville and Benveniste's divergence algorithm [8] and Andre-Obrecht's forward-backward method [5], the offline approach of Svendsen and Soong [111], and the integrated online-offline approach of the author and Gouyon [115], may experience difficulties when presented with sufficiently complex signals. Irrespective of algorithm quality, theoretical limits (e.g., Cramer-Rao bound [26, 88, 78]) exist as to how well we can estimate signal characteristics given certain noise levels, model complexities, and amounts of data.

To surpass these limits, we restrict the application to musical signals. Fortunately, musical signals are highly structured, both at the signal level, in terms of the expected timbral evolution of *note events*, and at higher levels, in terms of melodic and rhythmic tendencies. These structures constrain relationships among signal parameters and restrict their variation over time. For instance, musical signals contain many regions exhibiting significant pitch content. Throughout these regions sinusoidal component frequencies are close to integer multiples of some fundamental. Neglecting variations due to inharmonicity, the frequency parameterization reduces from one parameter per sinusoid (say, on the order of 20–60 parameters) to a single parameter encoding fundamental frequency[2]. Now let us consider the general problem of estimating a

[2]Advantages in terms of the Cramer-Rao bound are discussed in [112].

signal corrupted by noise. If we have a variety of nested model structures, all of which are able to fit the signal in the absence of noise, it is well known [2] that the model with fewest parameters exhibits the least variance in its estimation of the signal in the presence of noise[3].

As an example, we consider the estimation of a linear trend with a succession of nested polynomial models. Let observations $\{Y_t\}, t \in 1:N$ be generated as follows:

$$\begin{aligned} X_t &= \begin{bmatrix} 1 & t \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \\ Y_t &\sim \mathcal{N}(X_t, 1) \end{aligned} \tag{3.1}$$

The objective is to estimate the "signal" X_t. The true model structure (3.1) is unknown; hence, we postulate a variety of polynomial models for X_t:

$$\begin{aligned} X_{p,t} &= H_{p,t}\theta_p \\ Y_t &\sim \mathcal{N}\left(X_{p,t}, \sigma^2\right) \end{aligned} \tag{3.2}$$

where σ^2 is *known*, and

$$H_{p,t} = \begin{bmatrix} 1 & t & t^2 & \dots & t^{p-1} \end{bmatrix} \tag{3.3}$$

and $\theta \in \mathcal{R}^p$. Here p represents the number of free parameters, the degree of the polynomial fit being $p-1$.

We estimate X_t by substituting the maximum-likelihood estimate of θ_p into (3.2). It becomes convenient to define the vector quantities $X \in \mathbb{R}^N \triangleq X_{1:N}$, $Y \in \mathbb{R}^N \triangleq Y_{1:N}$, and $X_p \in \mathbb{R}^N \triangleq X_{p,1:N}$, as well as the matrix $H_p \in \mathcal{R}^{N \times p}$:

$$H_p \triangleq \begin{bmatrix} H_{p,1}^T & H_{p,2}^T & \dots & H_{p,N}^T \end{bmatrix}^T \tag{3.4}$$

[3]This is essentially a restatement of Ockham's razor: *Pluralitas non est ponenda sine necessitate*; "Plurality should not be posited without necessity."[121]

Under the conditionally Gaussian model with known variance (3.2), the maximum-likelihood estimate is nothing but the least squares estimate $\hat{\theta}_p$:

$$\hat{\theta}_p = (H_p^T H_p)^{-1} H_p^T Y \tag{3.5}$$

Hence the estimate of X, denoted as $\hat{X}_p$, is $P_p Y$, where the *projection matrix* P_p is defined:

$$P_p \triangleq H_p^T (H_p^T H_p)^{-1} H_p^T \tag{3.6}$$

Similarly, we define $P_p^\perp = \mathbf{I}_N - P_p$ and note that P_p is the projection matrix onto the column space of H_p; $P_p^\perp$ is the projection matrix onto the orthogonal complement of this space. The following, easily verified, properties become useful: $P_p P_p = P_p$; $\mathrm{Tr} P_p = p$; $P_p^\perp P_p^\perp = P_p^\perp$; $\mathrm{Tr} P_p^\perp = N - p$; for $p \geq 2$; $P_p X = X \rightarrow P_p^\perp X = \mathbf{0}_{N\times 1}$.

Consider now the expected fits: both to the observed data Y_t, and to the signal X_t, as a function of $p \in 2:N$. The simplest model able to capture the linear trend has $p = 2$; the most complex model has $p = N$. Let the quality of fit to tbe observed data be measured by the squared error with regards to Y_t, summed over all samples. Taking expectation with respect to the "true model" (3.2) yields

$$\begin{aligned} E \parallel Y - \hat{X}_p \parallel^2 &= \mathrm{Tr} E \left(Y - \hat{X}_p\right) \left(Y - \hat{X}_p\right)^T \\ &= \mathrm{Tr} E \left(P_p^\perp Y Y^T P_p^\perp\right) \\ &= \mathrm{Tr} P_p^\perp \left(X X^T + \sigma^2 \mathbf{I}_N\right) P_p^\perp \\ &= \mathrm{Tr} P_p^\perp X X^T P_p^\perp + \sigma^2 \mathrm{Tr} P_p^\perp \\ &= \sigma^2 (N - p) \end{aligned} \tag{3.7}$$

The final step follows from $P_p^\perp X = \mathbf{0}$ and $\mathrm{Tr} P_p^\perp = N - p$.

Figure 3.1 on the left side illustrates fits of the observed data for $p = 2$ and for $p = N$, where $N = 7$. With $p = N$, the fit is error free as predicted by (3.7). However, the lack of error is clearly due to *overfitting* noise elements in the data. The fit of the underlying trend X_t, as shown on the right side of Figure 3.1, seems better for the

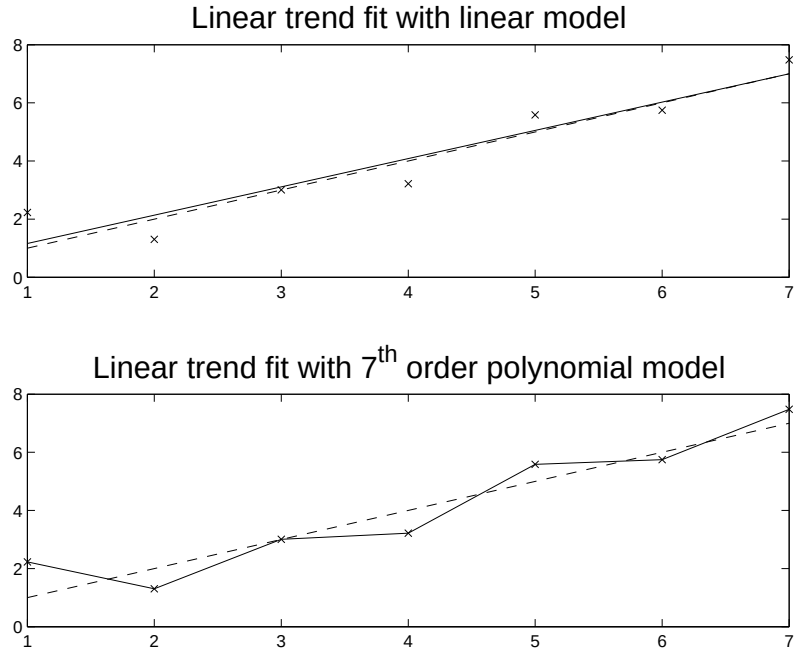


Figure 3.1: *Linear vs. maximal degree polynomial fits for linear trend*

$p = 2$ case. We can verify this analytically: the expected squared error for the fit of X_t under $p \geq 2$ is as follows:

$$\begin{aligned} E \parallel X - \hat{X}_p \parallel^2 &= \mathrm{Tr} E \left(XX^T - XX_p^T - X_p X^T + X_p X_p^T \right) \\ &= \mathrm{Tr} \left(XX^T - XX^T P_p + P_p XX^T + P_p XX^T P_p + \sigma^2 P_p \right) \\ &= \sigma^2 \mathrm{Tr} P_p \\ &= \sigma^2 p \end{aligned} \tag{3.8}$$

From (3.8), we see the expected error in fitting the underlying trend actually *increases* with p. The best results are achieved when $p = 2$; this model has the fewest parameters and is hence most constrained among all models able to fit the linear trend. In other words, assuming nothing about the model structure when correct assumptions can be made yields an overly complex model. The latter detracts from our ability to extract meaningful information from noisy data.

We can see the analogy to the multiple sinusoids case: for perfectly harmonic

signals, the "best" model recognizes that all component frequencies are integer multiples of some fundamental, using one parameter to encode all frequency values. Unfortunately, parameter constraints implied by real-world musical signals are seldom representable by "hard" restrictions on the parameter space (i.e., stipulations that parameters belong to certain subsets of the nominal space). For example, many pitched acoustic signals, such as piano and marimba, contain significant amounts of inharmonicity. The degrees and qualities of inharmonicity vary from instrument to instrument. Consequently, we desire a model which encodes the general trend (that the frequencies of sinusoidal components lie close to integer multiples of some fundamental), while maintaining robustness to uncertain, incompletely specified *deviations* from this trend. In general, when proposing model structures for real-world signals, we must plan for deviations and uncertainties in these structures, while controlling the range of allowed variation. We later demonstrate that a Bayesian decision theoretic framework proves most amenable to these considerations, especially in its capacity to represent uncertain prior or structural knowledge.

3.3 Integrating context with signal information

In a broader sense, we recognize that musical structure creates *context* which is useful in *predicting* attributes of interest; i.e., pitch content, the presence and location of musical onsets, and the locations of transient regions. The main goal becomes to integrate these contextual predictions with information from the signal to make optimal decisions concerning these attributes. A schematic is shown in Figure 3.2.

What means an "optimal decision" takes on a necessarily probabilistic formulation: in the simplest case, we aim to minimize the probability of decision error. More generally, we aim to minimize expected costs (Bayes risks) arising from the various types of hypotheses that are confused for one another.

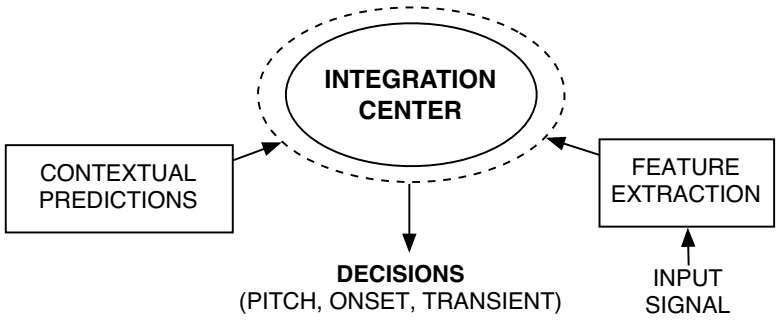


Figure 3.2: *Integration of contextual predictions with signal information*

3.3.1 Integrating a single predictive context

To illustrate the integration of a single predictive context with signal information in decision frameworks, consider a hypothetical example of piano recording containing one note. Suppose the recording is extremely noisy and the piano content barely audible. The task is to decide whether the current note, represented by N_k, equals $'\mathrm{C4}'$ or $'\mathrm{C5}'$. No other possibilities are considered.

Context arises from the fact the previous note N_{k-1}, equals $'\mathrm{B4}'$; considering the composer, it is nine times as likely the $'\mathrm{B4}'$ will step into $'\mathrm{C5}'$ than leap down to $'\mathrm{C4}'$[4]:

$$\begin{aligned} P(N_k = {'\mathrm{C5}'} | N_{k-1} = {'\mathrm{B4}'}) &= 0.9 \\ P(N_k = {'\mathrm{C4}'} | N_{k-1} = {'\mathrm{B4}'}) &= 0.1 \end{aligned} \tag{3.9}$$

Signal information is summarized in the *feature* Y_k, an estimate of the fundamental pitch. $Y_k = 350$ Hz is observed. From offline experiments, it is determined:

$$P(Y_k | N_k) = \mathcal{N}\left(\mathrm{toHz}(N_k), 8000\ \mathrm{Hz}^2\right) \tag{3.10}$$

Here $\mathrm{toHz}(\cdot)$ converts the symbolic note value into a corresponding Hz pitch value.

[4]Such a statement could be verified by taking all scores produced by this composer, counting the number of $'\mathrm{B4}' \rightarrow '\mathrm{C5}'$ transitions, and dividing by the number of $'\mathrm{B4}' \rightarrow '\mathrm{C4}'$ plus $'\mathrm{B4}' \rightarrow '\mathrm{C5}'$ transitions.

In particular, toHz($'$C4$'$) = 262 Hz, and toHz($'$C5$'$) = 523 Hz.

The interaction of contextual and signal information is shown via the directed acyclic graph of Figure 3.3, representing the factorization of the joint distribution $P(N_{k-1}, N_k, Y_k)$.

$$P(N_{k-1}, N_k, Y_k) \quad = \quad P(N_{k-1})P(N_k|N_{k-1})P(Y_k|N_k) \tag{3.11}$$

While Y_k alone seems to indicate that $N_k = {}'$C4$'$, melodic context predicts $N_k = {}'$C5$'$.

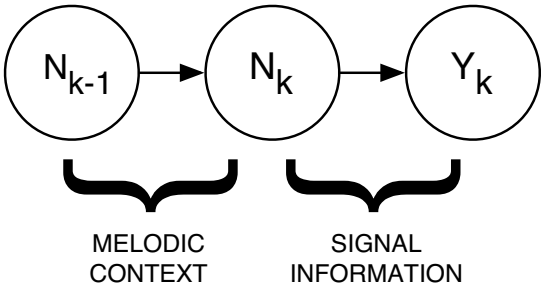


Figure 3.3: *Integration of melodic context with signal information*

Which tendency wins out? The objective becomes to minimize the *probability of error*, $P(\hat{N}_k \neq N_k)$, where $\hat{N}_k$ is the estimate of the current note based on N_{k-1} and Y_t. It is easily shown [13] that the error probability minimization is equivalent to the maximization

$$\hat{N}_k \quad = \quad \underset{N_k}{\text{argmax}} \; P(N_k|N_{k-1}, Y_k) \tag{3.12}$$

The *posterior probability* $P(N_k|N_{k-1}, Y_k)$, is computed via *Bayes' rule*:

$$P(N_k|N_{k-1}, Y_k) \quad = \quad \frac{P(N_k|N_{k-1})P(Y_k|N_k)}{\sum_{N_k} P(N_k|N_{k-1})P(Y_k|N_k)} \tag{3.13}$$

Essentially, (3.13) states that the posterior is proportional to the product of the

contextual and signal dependences[5]. As a result:

$$\begin{aligned} P(N_t = \text{'C5'}|N_{t-1}, Y_t) &= 0.692 \\ P(N_t = \text{'C4'}|N_{t-1}, Y_t) &= 0.308 \end{aligned} \tag{3.14}$$

Hence 'C5' is the correct decision. Although the signal provides evidence to the contrary, the uncertainty in this evidence as represented by $P(Y_k|N_k)$ is so great that it is overridden by the contextual dependence. The signal information does shift the balance somewhat: while 'C5' is *a priori* nine times as likely than 'C4', after observing the signal, 'C5' is only $0.692/0.308 \approx 2.25$ times as likely.

3.3.2 Integrating information across time

Often, the context inherent in musical signals manifests not as a single source of prior information, but in the *consistency* of a given attribute (e.g., pitch) over time. While we may lack prior information concerning the attribute at a specific point in time, the fact that attributes evolve consistently suggests that we may combine features observed at different points in time, to improve the estimation of a given attribute at any point in time.

Consider, for instance, a recording of one note of a vocal passage, where the vocalist exhibits slight, uncertain fluctuations in pitch. The inherent pitch of the vocalist is consistent over time, while not being exactly the same. Let the recording be dissected into N frames of equal length; for each frame, a pitch estimate, Y_t, is extracted. The inherent pitch of the vocalist is represented by the trajectory $S_{1:N}$. We model Y_t as a noisy version of S_t, via

$$P(Y_t|S_t) \sim \mathcal{N}(S_t, \lambda_y) \tag{3.15}$$

The noise variance, λ_y, may be determined via offline experiments.

For any given frame, nothing is known *a priori* about its pitch. This lack of

[5]The purpose of the denominator is to renormalize the product such that it sums to unity and is hence a valid probability distribution.

knowledge in the framework of Gaussian dependences, may be represented [57]:

$$P(S_t) \sim \mathcal{N}\left(0, \epsilon^{-1}\right) \tag{3.16}$$

where $\epsilon > 0$ is arbitrarily small.

The consistency of inherent pitch over time is represented by $S_{t+1} \approx S_t$. As a Gaussian dependence, this is modeled:

$$P(S_{t+1}|S_t) \sim \mathcal{N}\left(S_t, \lambda_s\right) \tag{3.17}$$

It can be shown that for any $\lambda_s \geq 0$, if $P(S_1)$ is specified via (3.16), and $\epsilon \downarrow 0$, there exists an equivalent representation for $P(S_t)$:

$$P(S_t) \sim \mathcal{N}\left(0, \epsilon_t^{-1}(\lambda_s, \epsilon)\right) \tag{3.18}$$

where $\epsilon_t^{-1}(\lambda_s, \epsilon) \downarrow 0$.

A complete specification of the joint distribution satisfying (3.15 - 3.17) follows:

$$P(S_{1:N}, Y_{1:N}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{N} P(S_t|S_{t-1}) \tag{3.19}$$

The factorization (3.19) is represented by the directed acyclic graph of Figure 3.4.

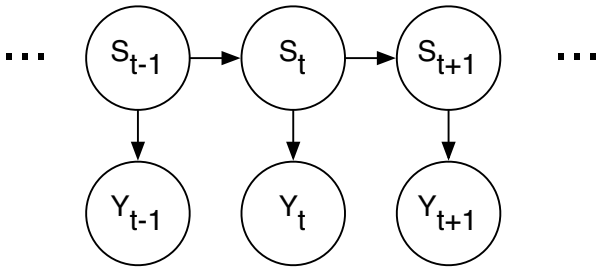


Figure 3.4: *Directed acyclic graph for pitch consistency model across time*

Let S_t^*, $t \in 1:N$ be the maximized posterior for each frame, which serves as an

estimated pitch trajectory:

$$S_t^* \;=\; \underset{S_t}{\operatorname{argmax}}\; P(S_t|Y_{1:N}) \;\; \forall t \in 1:N \tag{3.20}$$

If the set of possibilities for S_t were countable, this choice of S_t^* would minimize the expected number of frame errors. Since $S_{1:N}$ and $Y_{1:N}$ are jointly Gaussian, $P(S_t|Y_{1:N})$ is Gaussian and thus peaks at its mean; hence, $S_t^* = E(S_t|Y_{1:N})$, also minimizes the mean squared error, $E|S_t^* - S_t|^2$. It is well known [57] that for linear Gaussian models, the minimum mean squared error estimator becomes a weighted linear combination of the observations; i.e.:

$$S_t^* \;=\; \sum_{\tau=1}^{N} w_{t,\tau} Y_\tau \;\; \forall t, \tau \in 1\!:\!N \tag{3.21}$$

Moreover, S_t^* via (3.20) depends only on the ratio $\rho \stackrel{\Delta}{=} \lambda_s/\lambda_y$, irrespective of the actual values of λ_s and λ_y. If $\rho = 0$, each S_t becomes an identical copy of the same underlying parameter: in this case we expect S_t^* to be the *unweighted* average of Y_t; i.e.,

$$w_{t,\tau} \;=\; 1/N \;\; \forall t, \tau \in 1\!:\!N. \tag{3.22}$$

As $\rho \to \infty$, the dependence among $\{S_t, Y_t\}$ across different values of t vanishes; we expect that S_t^* depends only on the current observation Y_t. For intermediate values of ρ, we expect $w_{t,\tau}$ to peak about $\tau = t$, emphasizing observations in the immediate neighborhood of t while discounting observations that are further away. As ρ becomes small, we expect $w_{t,\tau}$ to decay more gradually on both sides of t.

These assertions may be verified by obtaining a closed form expression for the posterior $P(S_t|Y_{1:N})$ in terms of $Y_{1:N}$, λ_s, and λ_y. This posterior is computed recursively over time in two passes: the *filtering* pass updates $P(S_{t+1}|Y_{1:t+1})$ given $P(S_t|Y_{1:t})$, and the *smoothing* pass updates $P(S_t|Y_{1:N})$ given $P(S_{t+1}|Y_{1:N})$.

Using conditional independence relations implicit in the factorization (3.19), filtering and smoothing recursions can be derived. The filtering recursion begins with

the *time update* step, updating $P(S_t|Y_{1:t}) \rightarrow P(S_{t+1}|Y_{1:t})$:

$$\begin{aligned} P(S_{t+1}|Y_{1:t}) &= \int P(S_t, S_{t+1}|Y_{1:t})\ dS_t \\ &= \int P(S_t|Y_{1:t})P(S_{t+1}|S_t, Y_{1:t})\ dS_t \\ &= \int P(S_t|Y_{1:t})P(S_{t+1}|S_t)\ dS_t \end{aligned} \tag{3.23}$$

The *measurement update* follows, updating $P(S_{t+1}|Y_{1:t}) \rightarrow P(S_{t+1}|Y_{1:t+1})$.

$$\begin{aligned} P(S_{t+1}|Y_{1:t+1}) &= \frac{P(S_{t+1}, Y_{t+1}|Y_{1:t})}{\int P(S_{t+1}, Y_{t+1}|Y_{1:t})\ dS_{t+1}} \\ P(S_{t+1}, Y_{t+1}|Y_{1:t}) &= P(S_{t+1}, Y_{1:t})P(Y_{t+1}|S_{t+1}, Y_{1:t}) \\ &= P(S_{t+1}|Y_{1:t})P(Y_{t+1}|S_{t+1}) \end{aligned} \tag{3.24}$$

The filtering recursion is initialized after the first time update step with $P(S_1)$, via (3.18). The final stage of the filtering recursion obtains $P(S_N|Y_{1:N})$, which is used to initialize the smoothing recursion. The smoothing recursion updates $P(S_{t+1}|Y_{1:N}) \rightarrow P(S_t|Y_{1:N})$:

$$\begin{aligned} P(S_{t+1}|Y_{1:t}) &= \int P(S_t, S_{t+1}|Y_{1:N})\ dS_{t+1} \\ &= \int P(S_{t+1}|Y_{1:N})P(S_t|S_{t+1}, Y_{1:N})\ dS_{t+1} \\ &= \int P(S_{t+1}|Y_{1:N})P(S_t|S_{t+1}, Y_{1:t})\ dS_{t+1} \\ &= P(S_t|Y_{1:t}) \int \frac{P(S_{t+1}|Y_{1:N})}{P(S_{t+1}|Y_{1:t})} P(S_{t+1}|S_t)\ dS_{t+1} \end{aligned} \tag{3.25}$$

where $P(S_t|Y_{1:t})$ and $P(S_{t+1}|Y_{1:t})$ are precomputed and stored in the filtering pass.

Since all of the intermediate computations in (3.15-3.17) involve multiplication, conditioning, and marginalization operations on Gaussian distributions, all intermediate quantities encountered in the filtering and smoothing recursions remain Gaussian. As such, these quantities are completely specified by mean and variance parameters and we may write filtering and smoothing recursions in terms of these parameters. To

this end, we use the "Gaussian potential" algebra introduced by Lauritzen [70] and generalized to the multivariate case by Murphy [84]. Letting $\epsilon \to 0$ in (3.16) obtains the standard *Kalman filter* and *Rauch-Tung-Striebel smoother* discussed in [84, 57]:

$$\begin{aligned} P_1 &= 1/\lambda_y \\ S_1^{*,(f)} &= y_1 \\ P_{t+1} &= \frac{\lambda_y (\lambda_s + P_t)}{\lambda_y + \lambda_s + P_t} \\ S_{t+1}^{*,(f)} &= P_{t+1} \left(\frac{S_t^{*,(f)}}{\lambda_s + P_t} + \frac{y_{t+1}}{\lambda_y} \right) \\ S_N^* &= S_N^{*,(f)} \\ S_t^* &= \frac{\lambda_s}{\lambda_s + P_t} S_t^{*,(f)} + \frac{P_t}{\lambda_s + P_t} S_{t+1}^* \end{aligned} \tag{3.26}$$

where

$$\begin{aligned} S_t^{*,(f)} &\triangleq E(S_t | Y_{1:t}) \\ P_t &\triangleq Var(S_t | Y_{1:t}) \\ S_t^* &\triangleq E(S_t | Y_{1:N}) \end{aligned} \tag{3.27}$$

Defining $P_t' = \lambda_y^{-1} P_t$, (3.26) may be rewritten in terms of $\rho \triangleq \lambda_s / \lambda_y$:

$$\begin{aligned} P_1^{'} &= 1 \\ S_1^{*,(f)} &= y_1 \\ P_{t+1}^{'} &= \frac{\rho + P_t^{'}}{1 + \rho + P_t^{'}} \\ S_{t+1}^{*,(f)} &= P_{t+1}^{'} \left(\frac{S_t^{*,(f)}}{\rho + P_t^{'}} + y_{t+1} \right) \\ S_N^* &= S_N^{*,(f)} \\ S_t^* &= \frac{\rho}{\rho + P_t^{'}} S_t^{*,(f)} + \frac{P_t^{'}}{\rho + P_t^{'}} S_{t+1}^* \end{aligned} \tag{3.28}$$

By induction, it is easily shown that there exist weights $w_{t,\tau}^{(f)}$ and $w_{t,\tau}$ satisfying

$S_t^{*,(f)} = \sum_{t=1}^{N} w_{t,\tau}^{(f)} Y_\tau$ and $S_t^{*,(f)} = \sum_{t=1}^{N} w_{t,\tau} Y_\tau$. Substituting these expressions into (3.28) and equalizing coefficients obtains:

$$\begin{aligned}
P_1^{'} &= 1 \\
w_{1,\tau}^{(f)} &= \delta_{1,\tau},\ \tau \in 1:N \\
P_{t+1}^{'} &= \frac{\rho + P_t^{'}}{1 + \rho + P_t^{'}} \\
w_{t+1,\tau}^{(f)} &= P_{t+1}^{'} \left(\frac{w_{t,\tau}^{(f)}}{\rho + P_t^{'}} + \delta_{t+1,\tau} \right), \quad \forall \tau \in 1:N \\
w_{N,\tau} &= w_{N,\tau}^{(f)} \\
S_t^{*} &= \frac{\rho}{\rho + P_t^{'}} S_t^{*,(f)} + \frac{P_t^{'}}{\rho + P_t^{'}} S_{t+1}^{*}
\end{aligned} \tag{3.29}$$

where $\delta_{t,\tau}$ is the Kronecker delta function; i.e.,

$$\delta_{t,\tau} = \begin{cases} 1, & t = \tau \\ 0, & \text{otherwise} \end{cases} \tag{3.30}$$

Figure 3.5 displays the behavior of $w_{t,\tau}$, as ρ ranges from 0 to ∞. If $\rho = 0$, the weights appear uniform; this may be verified by simple substitution of $\rho = 0$ into (3.29):

$$\begin{aligned}
P_t^{'} &= 1/t \\
w_{t,\tau}^{(f)} &= \begin{cases} 1/t, & \tau \leq t \\ 0, & \tau > t \end{cases} \\
w_{t,\tau} &= 1/t, \quad \forall t, \tau \in 1:N
\end{aligned} \tag{3.31}$$

Assuming $\lambda_y > 0$, $\rho = 0$ implies $\lambda_s = 0$; in other words, the trajectory $S_{1:N}$ is constant. Here it makes sense to weight all observations equally since each Y_t is a conditionally independent noisy observation of the same underlying parameter.

Likewise, as $\rho \to \infty$, it is readily shown that $w_{t,\tau}^{(f)}$ converges for each $t, \tau \in 1:N$

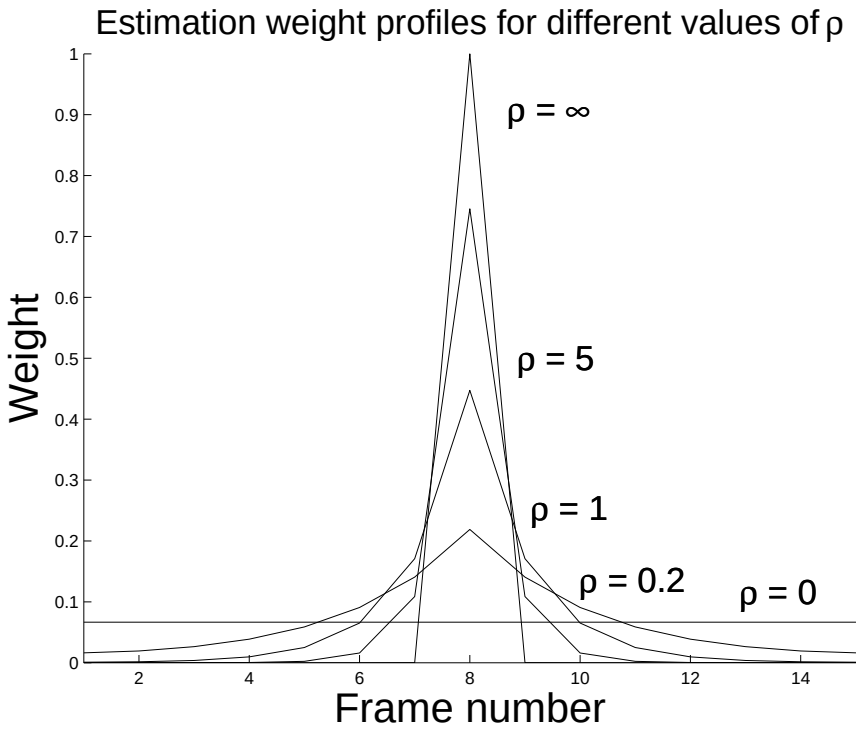


Figure 3.5: *Estimation weight profiles for different values of* ρ

to $\delta_{t,\tau}$, and that $w_{t,\tau}$ converges to $w_{t,\tau}^{(f)}$. As a result:

$$w_{t,\tau} \overset{\rho\to\infty}{\to} \delta_{t,\tau}, \quad \forall t, \tau \in 1:N \tag{3.32}$$

If $\lambda_y > 0$ is finite, $\rho \to \infty$ implies $\lambda_s \to \infty$, meaning that the S_t become mutually independent. The best estimate of S_t given *all* observations depends only on the observation at time t.

Real-world scenarios require robustness to small drifts in pitch, while maintaining consistency of pitch over time. Consequently, a good weight function emphasizes observations in the neighborhood of t and discounts observations that are further away. Figure 3.5 shows that this type of weight function is guaranteed via the probabilistic model (3.15 - 3.17; 3.19).

3.3.3 Temporal integration and abrupt change detection

Now we consider an example where pitch is consistent, but punctuated by points of abrupt change: the "legato" model shown in Figure 3.6. The change points are

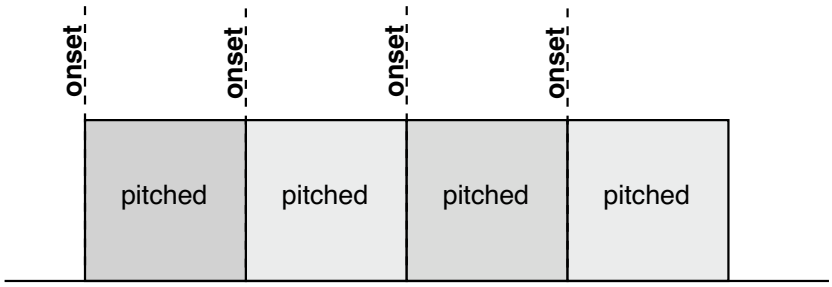


Figure 3.6: *"Legato" model for pitch consistency with points of abrupt change*

interpreted as *note onsets*; they occur at unknown times. Each pair of onsets bounds a *note event*. Throughout note events, pitch information is salient and consistent; this corresponds to legato playing. We note that the legato model may be considered as a reduction of the *nominally monophonic* model introduced in Section 3.4. Adding transient regions within note events and "null" regions, which represent gaps between note events, extends the legato model to the nominally monophonic situation.

Given the segmentation (onset times), a heuristic may be devised in which all observations within a note event are used to estimate the pitch trajectory for any time along that event. The observations' weighting would follow the profiles shown in Figure 3.5; however, zero weight would be given to observations from other note events. Unfortunately, this segmentation is unknown. If instead we had a good estimate of the inherent pitch trajectory, we could monitor jumps in this trajectory to determine onset times. However, at any *fixed* point in time, nothing is known *a priori* about the inherent pitch. As a result, we have a classic "chicken/egg" dilemma (Figure 3.7) characteristic of such segmentation problems [50].

To resolve this situation, we encode the unknown segmentation in a hidden binary *mode variable* $M_t \in \{'\mathrm{O}', '\mathrm{C}'\}$ which indicates whether or not frame t contains an onset. If $M_t = '\mathrm{O}'$ (onset), an onset occurs in frame t, meaning that frames $t-1$ and t belong to different note events. In this case we do not expect S_t and S_{t-1} to

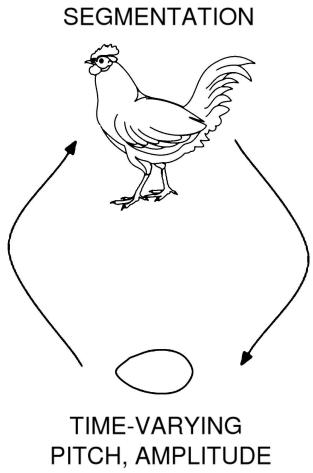


Figure 3.7: *Canonical chicken-egg situation for segmentation applications*

be related. Otherwise, $M_t = \text{'C'}$ (continuation), indicating that frames $t-1$ and t belong to the same note event. In this case $S_t \approx S_{t-1}$. The directed acyclic graph of Figure 3.8 displays the complete factorization of the joint $P(M_{1:N}, S_{1:N}, Y_{1:N})$:

$$\begin{aligned} P\left(M_{1:N}, S_{1:N}, Y_{1:N}\right) &= P(M_1)P(S_1|M_1)P(Y_1|S_1) \\ &\times \prod_{t=2}^{N} P(M_t|M_{t-1})P(S_t|S_{t-1}, M_t)P(Y_t|S_t) \end{aligned} \tag{3.33}$$

Here $P(M_{t+1}|M_t)$ follows the stochastic grammar and state transition diagram displayed in Figure 3.9. Here the main restriction is that onsets cannot occur in adjacent frames. The $\text{'C'} \rightarrow \text{'O'}$ transition probability p, $0 < p \ll 1$ models the

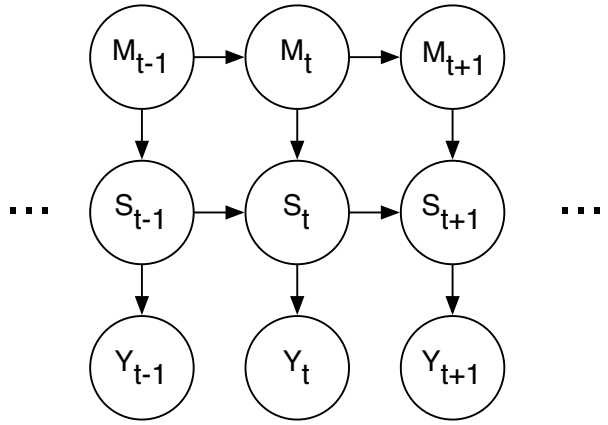


Figure 3.8: *Factorization of joint distribution for legato model*

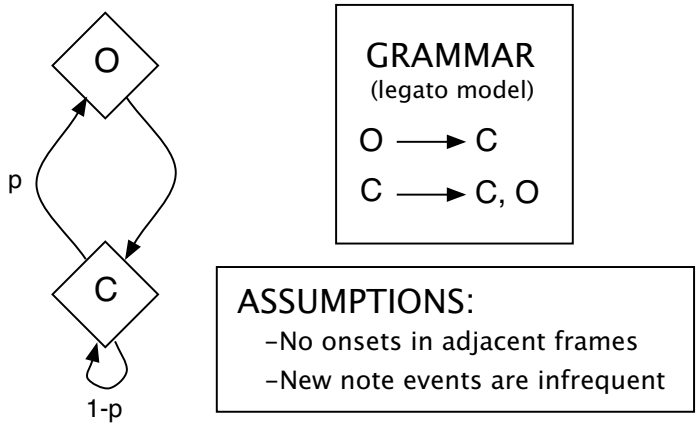


Figure 3.9: *Stochastic grammar for mode variables, legato model*

expectation that onsets are sparse[6]. $P(M_{t+1}|M_t)$ obtains in closed form:

$$P(M_{t+1}|M_t) \quad = \quad \begin{cases} 0, & M_t = \text{'O'}, M_{t+1} = \text{'O'} \\ 1, & M_t = \text{'O'}, M_{t+1} = \text{'C'} \\ p, & M_t = \text{'C'}, M_{t+1} = \text{'O'} \\ 1-p, & M_t = \text{'C'}, M_{t+1} = \text{'C'} \end{cases} \tag{3.34}$$

Consistency of pitch is governed by $P(S_{t+1}|S_t, M_{t+1})$. If $M_{t+1} = \text{'C'}$, S_{t+1} depends

[6]The actual value of p controls the mean note event length in the legato case. The latter is approximately (slightly greater than) $1/p$ frames.

on S_t in the manner of (3.17), encoding the consistency hypothesis $S_{t+1} \approx S_t$; otherwise, if $M_{t+1} = \text{'O'}$, S_{t+1} is statistically independent of S_t. Furthermore, the fact that $Var(S_{t+1}|M_{t+1} = \text{'C'}) = \epsilon^{-1}$ indicates that there lacks additional information concerning S_{t+1}, which parallels the situation in (3.18). Hence, $P(S_{t+1}|S_t, M_{t+1})$ is modeled as follows:

$$\begin{aligned} P\left(S_{t+1}|S_t, M_{t+1} = \text{'C'}\right) &\sim \mathcal{N}\left(S_t, \lambda_s\right) \\ P\left(S_{t+1}|S_t, M_{t+1} = \text{'O'}\right) &\sim \mathcal{N}\left(0, \epsilon^{-1}\right) \end{aligned} \tag{3.35}$$

The prior dependences are specified as noninformative: $P(M_1)$ uniform; $P(S_1|M_1)$ via (3.16). Finally, the observation dependence, $P(Y_t|S_t)$, is specified via (3.15).

The joint segmentation and pitch tracking amounts to estimating all values of the hidden variables $M_{1:N}$ and $S_{1:N}$, given observations $Y_{1:N}$. We aim to preserve the integrity of the entire mode sequence, $M_{1:N}$, by minimizing the probability that *any* such error in this sequence may occur:

$$M^*_{1:N} = \underset{\hat{M}_{1:N}}{\text{argmin}} \; P(\hat{M}_{1:N} \neq M_{1:N}) \tag{3.36}$$

Once again, following (3.12), the optimal sequence satisfying (3.36) maximizes the joint posterior:

$$M^*_{1:N} = \underset{M_{1:N}}{\text{argmax}} \; P(M_{1:N}|Y_{1:N}) \tag{3.37}$$

We note that the objective (3.36) comprises an "all or nothing" approach, penalizing estimates which differ from the true sequence $M_{1:N}$ in only one frame the same as estimates which have nothing to do with $M_{1:N}$. Furthermore, for sufficiently long sequences and under sufficiently noisy conditions, $P(M^*_{1:N} \neq M_{1:N})$ may approach unity. As such, it may be preferable to minimize the expected number of frame errors, following the criterion used for $S_{1:N}$ (3.20) in the previous section. However, a common source of error is when the true onset sample location lies infinitesimally

close to a frame boundary, though just preceding it; here, considerable ambiguity exists concerning whether to assign the onset to the frame in which it actually occured, or to the subsequent frame, for which the change in the spectral content regarding the previous frame is most salient. If $\hat{M}_{1:N}$ declares onsets in both frames, this incurs at most one frame error; if a single onset is declared in the wrong frame, two frame errors result. But detecting onsets in adjacent frames is disastrous for the segmentation objective, especially if the results are to be used in transcription, because this detection introduces an extra note event. By contrast, shifting the onset location by one frame is far less objectionable. A more striking consideration is that the joint posterior $P\left(M_{1:N}|Y_{1:N}\right)$ vanishes over sequences containing onsets in adjacent frames because $P\left(M_{t+1}|M_t\right)$ assigns these instances zero probability (3.34). Any particular decision $M^*_{1:N}$ exhibiting adjacent onsets will be *invalid* with respect to the generative model, thus the global maximum *a posteriori* criterion, (3.36) guarantees the validity of all detected mode sequences. As such, (3.37) is validated as the proper segmentation criterion.

Regarding the estimated pitch trajectory, $S^*_{1:N}$, the objective remains to minimize the expected number of frame errors, following (3.20). However, we require additionally that $S^*_{1:N}$ synchronize with $M^*_{1:N}$. We cannot tolerate, for instance, a sudden jump in S^*_{t+1} from S^*_t if frames t and $t+1$ belong to the same note event. The solution is to choose $S^*_{1:N}$ minimizing expected frame error rate *given* $M^*_{1:N}$:

$$S^*_t \;=\; \underset{S_t}{\operatorname{argmax}}\; P\left(S_t|Y_{1:N}, M^*_{1:N}\right) \tag{3.38}$$

where $M^*_{1:N}$ follows via (3.36).

3.4 Nominally monophonic signals and segmentation objectives

Unfortunately, real-world musical audio signals contain regions for which steady-state pitch content is not salient. Such regions include *transients*, which associate with note events, and *null regions*, or gaps, which separate note events. Ideally null

regions would contain nothing but silence. More realistically, these regions consist of recording noise, background instrumentation, reverberation tails from past events, and so forth.

To this end, we extend the "legato model" introduced in Section 3.3.3 to incorporate transient and null regions. The result we denote as the *nominally monophonic* model, represented by the cyclic succession (transient → pitched → null) depicted in Figure 3.10. Here each *note event* comprises a transient region followed by a pitched,

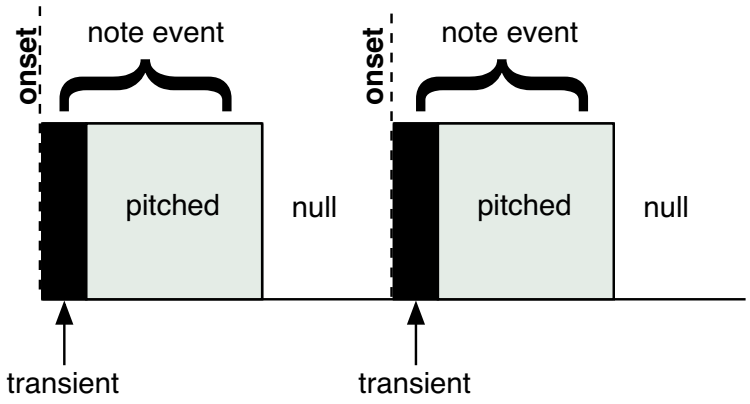


Figure 3.10: *Region characterization for nominally monophonic signals*

steady-state region. Null regions separate note events. The lengths of transient and null regions may be zero, to encode a certain flexibility towards cases where these regions may be absent: if both transient and null lengths are zero, the legato model results. A new note event is instantiated upon *crossing of the transient boundary*: either null → transient, pitched → transient, or null → pitched transitions trigger onsets.

We note that the nominally monophonic model fails to allow multiple pitched and transient regions within a single note event. This becomes problematic, for instance, in overblown flute recordings, where multiple pitched regions may occur due to the different oscillatory regimes, and where multiple transient regions may occur due to chaotic behavior [99]. However, the lack of an explicit model for multiple transient/pitched regions does not cause problems in practice. First, multiple-region instances rarely occur when dealing with single acoustic sources. Second, when these

instances do occur, the resultant segmentation using the nominally monophonic model retains all information concerning the locations of abrupt-change events and transient region boundaries, the only difference being that this model declares a new note event upon the crossing of *each* transient boundary within the actual event. This information may be aggregated in postprocessing, to form the actual note events (see Figure 3.11):

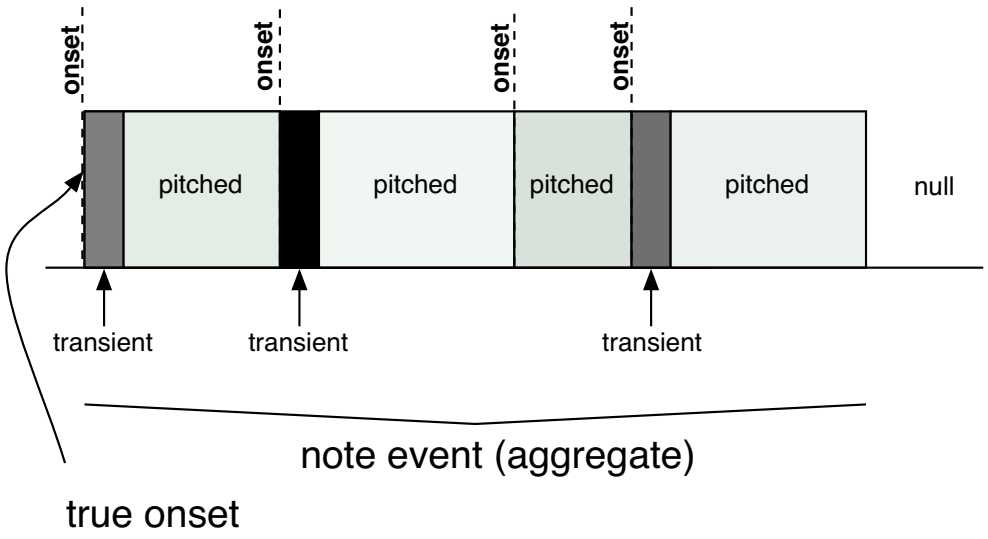


Figure 3.11: *Aggregation of note events*

While the nominally monophonic model represents signals which may arise from a monophonic "score", the intent is to be robust to various types of polyphony; e.g., note overlaps due to legato playing, reverberation, or background instrumentation. These instances occur in virtually all real-world recordings regardless of whether or not the performance originates from a monophonic score. Such robustness considerations, given limited computational resources, suggest the use of framewise short-time Fourier transform (STFT) peaks as a feature set. This stands in opposition to time domain approaches which do not attempt feature reduction, such as the methods summarized in Chapter 1 [20, 51, 18], among possibly others. The latter methods deliver a global, sample-accurate segmentation, but at considerable computational expense. Even in

some offline applications such as interactive audio editing (Section 3.10.4), a sample-accurate response may be desired.

To summarize, we obtain a robust segmentation and melody transcription for nominally monophonic signals, defined in the sense of the transient → pitched → null cyclic succession (Figure 3.10). The segmentation operates on a quite limited yet psychoacoustically relevant feature set (framewise sequences of STFT peaks). The result amounts to a *transcription*, but is more general: not only do we detect all note onsets, values, and durations, we provide a sub-segmentation of each event indicating the locations of transient regions as well as steady-state regions containing pitch content. While this method fails to provide a sample accurate segmentation, it facilitates the latter in postprocessing, by isolating region boundaries to frame neighborhoods and identifying possible pitch content before and after the true boundary sample location. (Given the fact that a single abrupt-change event occurs, as well as signal models before and after change, a simple offline likelihood maximization may be used to estimate the changepoint location; see [61], chapter 11 for further details.)

3.5 Probabilistic model

3.5.1 Variable definitions

To encode the cyclic succession (Figure 3.10) as well as label onsets at the frame level, we introduce the *mode* variable M_t for frames $t \in 1{:}N$. M_t takes on any of the following values:

- $'\mathrm{OT}'$ – the beginning frame of a transient region, of which there can be at most one per note event.
- $'\mathrm{OP}'$ – the beginning frame of a note event in which the first frame already contains salient pitch content, of which there can be at most one per note event.
- $'\mathrm{CT}'$ – the continuation of a transient region in the event the region occupies more than one frame; must follow a previous $'\mathrm{CT}'$ or $'\mathrm{OT}'$.

- $'\mathrm{CP}'$ – the continuation of a pitched region; must follow an $'\mathrm{OP}'$ or $'\mathrm{CP}'$.
- $'\mathrm{N}'$ – a null frame which occurs anytime after the last frame of a note event. A null frame is followed by another null frame, or an onset ($'\mathrm{OT}'$ or $'\mathrm{OP}'$).

Table 3.1 defines special groupings of modes with common characteristics. For

Symbol	Definition	Description
$\mathcal{P}$	$\{'\mathrm{OP}', '\mathrm{CP}'\}$	Pitched modes
$\mathcal{Q}$	$\{'\mathrm{OT}', '\mathrm{CT}', '\mathrm{N}'\}$	Non-pitched modes
$\mathcal{T}$	$\{'\mathrm{OT}', '\mathrm{CT}'\}$	Transient modes
$\mathcal{O}$	$\{'\mathrm{OT}', '\mathrm{OP}'\}$	Onset modes
$\mathcal{C}$	$\{'\mathrm{CT}', '\mathrm{CP}'\}$	Continuation modes

Table 3.1: *Definitions of mode groupings*

instance, we represent an onset by $M_t \in \mathcal{O}$, regardless of whether this onset occurs in a transient or pitched frame. Additionally, it becomes convenient to define $\mathcal{M}$ as the set of all modes:

$$\begin{aligned} \mathcal{M} &\triangleq \mathcal{P} \cup \mathcal{Q} \\ &= \mathcal{O} \cup \mathcal{C} \cup \{'\mathrm{N}'\} \end{aligned} \tag{3.39}$$

We introduce the *state* variable, S_t, to represent inherent signal characteristics known to be consistent during steady-state portions of note events (and changing abruptly across event boundaries) as these characteristics are primarily informative for the segmentation. In the legato example of Section 3.3.3, S_t represents the inherent pitch of the t^{th} frame. In real-world acoustic signals, amplitude, as well as pitch, is expected to be consistent in pitched regions. Depending on the instrument, the amplitude may exhibit a downward bias over time (e.g., percussive instruments such as piano and marimba), or it may exhibit equal tendencies to become softer or louder at any point in time (wind, brass, bowed string). Hence, for the nominally monophonic model, we encode both pitch and amplitude characteristics in S_t.

The pitch encoding is split among two quantities: an integer note value, plus a fractional tuning offset. There are two advantages: first, it becomes convenient for

the transcription task to marginalize out all but the note value, as the latter is usually what is notated in the score. Handling note values separately also facilitates the incorporation of melodic expectations, as Section 3.10.1 details. Moreover, characterizing the tuning offset as the main source of pitch drift allows the exploitation of structural tendencies which are difficult to model otherwise. Several factors contribute to pitch drift: first, global factors, such as the mistuning of an instrument or the playback of the recording at a different speed. These global factors are likely to exhibit a high degree of consistency throughout the entire recording, and they are largely independent of the audio source and hence the region boundaries as indicated by M_t. Second, local drifts may occur which are inherent to the acoustic source, hence responding to note event boundaries. Examples include the pitch bend in the attack portions of tuned percussion sources (e.g., timpani), or expressive variations such as portamento and vibrato, found in vocals, bowed strings, and other acoustic sources.

Individual state components of note value, tuning, and amplitude are encoded as follows:

- $N_t \in \mathcal{S}_N = \{N_{min}, N_{min}+1, \ldots, N_{max}\}$, where each element of $\mathcal{S}_N$ is an integer representing the MIDI note value (e.g., the note C4 corresponds to $N_t = 60$).
- $T_t \in \mathcal{S}_T$, where $\mathcal{S}_T$ is a uniformly spaced set of tuning values in $[-0.5, 0.5)$, with the minimum value equal to -0.5.
- $A_t \in \mathcal{S}_A$, where $\mathcal{S}_A$ is an exponentially spaced set of reference amplitude values active when $M_t \in \mathcal{P}$.
- $A_t^Q \in \mathcal{S}_{A^Q}$, where $\mathcal{S}_{A^Q}$ is an exponentially spaced set of reference amplitudes active when $M_t \in \mathcal{Q}$.

S_t denotes the collection of valid possibilities for these components:

$$S_t \in \mathcal{S} \;\; = \;\; \mathcal{S}_N \otimes \mathcal{S}_T \otimes (\mathcal{S}_A \cup \mathcal{S}_{A^Q}) \,. \tag{3.40}$$

which is to say, *either* $S_t = \{N_t, T_t, A_t\}$ if $M_t \in \mathcal{P}$ *or* $S_t = \{N_t, T_t, A_t^Q\}$, if $M_t \in \mathcal{Q}$.

The STFT peak observations are represented by Y_t, which consists of parallel lists of peak frequencies and amplitudes. We represent $Y_t = \{F, A\}$, where:

$$\begin{aligned} F &\triangleq \{F(1), F(2), \ldots, F(N_o)\} \\ A &\triangleq \{A(1), A(2), \ldots, A(N_o)\} \end{aligned} \tag{3.41}$$

where $F(k)$ denotes the frequency of the k^{th} lowest-frequency STFT peak, $A(k)$ the corresponding amplitude, and N_o the number of observed peaks. Peaks are chosen from overlapping, Hamming windowed, zeropadded frames following the quadratic interpolation methods described in [110, 1]; see Section 4.2.1 for further details.

The joint distribution over all variables of interest, $P(M_{1:N}, S_{1:N}, Y_{1:N})$, factors over the directed acyclic graph of Figure 3.12; i.e.:

$$\begin{aligned} P(M_{1:N}, S_{1:N}, Y_{1:N}) &= P(M_1)P(S_1|M_1)P(Y_1|S_1) \\ &\times \prod_{t=2}^{N} P(M_t|M_{t-1})P(S_t|S_{t-1}, M_{t-1}, M_t)P(Y_t|S_t) \end{aligned} \tag{3.42}$$

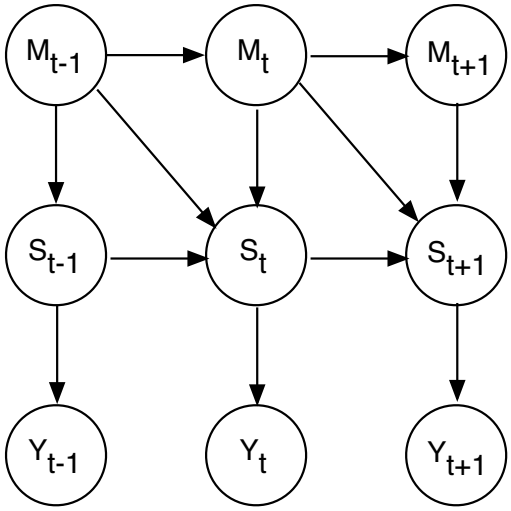


Figure 3.12: *Directed acyclic graph for nominally monophonic signal model*

The essential difference between the factorization (3.42) and that of the legato model (3.33) is the additional dependence on M_{t-1} in $P(S_t|S_{t-1}, M_{t-1}, M_t)$. The necessity of this dependence is illustrated by the following example: if $M_t = \text{'CP'}$, either $M_{t-1} \in \mathcal{T}$ or $M_{t-1} \in \mathcal{P}$. In this example $t-1$ and t belong to the same pitched region of a note event; hence, we expect consistency of the corresponding pitch values: $S_t \approx S_{t-1}$. If $M_{t-1} \in \mathcal{T} \subset \mathcal{Q}$, frame $t-1$ belongs to a transient region, while frame t belongs to a pitched region; in this case we do not expect the consistency $S_t \approx S_{t-1}$ to hold. We note that only the onset frame is explicitly encoded by M_t; in general, one must examine the pair $\{M_{t-1}, M_t\}$ to determine whether frame t crosses a particular region boundary.

3.5.2 Inference and estimation goals

Again, recalling the legato model (Section 3.3.3), we pursue identical objectives in terms of the segmentation. The optimal mode sequence is nothing but the global maximum *a posteriori* trajectory, obtained following (3.36):

$$M_{1:N}^* = \underset{M_{1:N}}{\text{argmax}} \; P\left(M_{1:N}|Y_{1:N}\right) \tag{3.43}$$

We recall that $M_{1:N}^*$ chosen via (3.43) preserves the integrity of the entire mode sequence, because it minimizes the probability that $M_{1:N}^*$ differs *anywhere* from the true $M_{1:N}$, regardless of how many frames make up the difference.

Individual state components, i.e., $N_{1:N}$, $T_{1:N}$, and ($A_{1:N}$ or $A_{1:N}^Q$) are chosen to minimize the expected number of frame errors given $M_{1:N}^*$. That is, if Z_t represents a particular state component for the t^{th} frame, we choose:

$$Z_t^* = \underset{Z_t}{\text{argmax}} \; P(Z_t|Y_{1:N}, M_{1:N}^*) \tag{3.44}$$

The primary inference and estimation goals consist of the computation of $M_{1:N}^*$ and $Z_{1:N}^*$. Secondary goals include estimating free parameters in the distributional specification, in particular those concerning the transition dependence $P(S_{t+1}, M_{t+1}|S_t, M_t)$, and postprocessing for purposes of transcription, meaning the transcription of desired

note event and transient region boundaries from $M^*_{1:N}$ and the assignment of pitch information to each note via $N^*_{1:N}$. Figure 3.13 summarizes the overall transcription process.

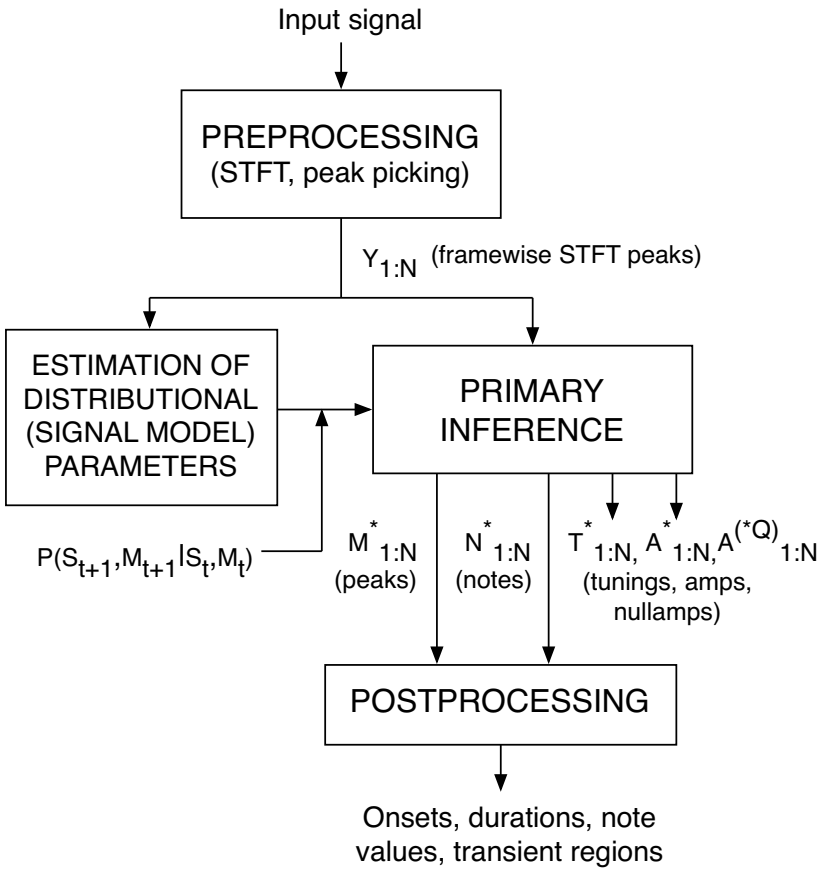


Figure 3.13: *Block diagram of overall transcription process*

The distributional specification is discussed below, in Section 3.6. Primary inference and parameter estimation methodologies are described in Section 3.8. However, discussion of preprocessing stages (STFT; peak picking) is deferred to Section 4.2.1 because certain details regarding these stages relate to the modeling of peak frequency and amplitude distributions, which is the central theme of that chapter.

3.6 Distributional specifications

For the model in Figure 3.12, we must specify the *priors* $P(M_1)$, and $P(S_1|M_1)$, the *transition dependence* across frames: $P(S_{t+1}, M_{t+1}|S_t, M_t)$, and the *observation likelihood* $P(Y_t|S_t)$.

3.6.1 Prior

The role of the prior is to encode information about the first frame of the recording. If it is known, for instance, that the recording begins with a note onset, we concentrate $P(M_1)$ on $\mathcal{O}$, the set of onset possibilities. In the most general case, however, such knowledge is absent. Hence, we specify $P(M_1)$ as uniform and $P(S_1|M_1)$ as factoring independently among the components of S_0:

$$\begin{aligned} P(S_0|M_0 \in \mathcal{P}) &= P(T_0)\,P(N_0)\,P(A_0) \\ P(S_0|M_0 \in \mathcal{Q}) &= P(T_0)\,P(N_0)\,P(A_0^{\mathcal{Q}}) \end{aligned} \tag{3.45}$$

where $P(T_0)$, $P(N_0)$, $P(A_0)$, and $P(A_0^{\mathcal{Q}})$ are uniform.

3.6.2 Transition dependence

The transition dependence factors accordingly:

$$P(S_{t+1}, M_{t+1}|S_t, M_t) = P(M_{t+1}|M_t)\,P(S_{t+1}|M_t, M_{t+1}, S_t) \tag{3.46}$$

It remains to specify mode and state dependences; respectively: $P(M_{t+1}|M_t)$, $P(S_{t+1}|S_t, M_t, M_{t+1})$.

The mode transition dependence, $P(M_{t+1}|M_t)$, is based around the following *standard note evolution* grammar encoding the cyclic succession depicted in Figure 3.10.

$$\begin{aligned} \text{'OT'} &\rightarrow \text{'CT'}, \text{'CP'} \\ \text{'OP'} &\rightarrow \text{'CP'}, \text{'N'} \\ \text{'CT'} &\rightarrow \text{'CT'}, \text{'CP'} \\ \text{'CP'} &\rightarrow \text{'CP'}, \text{'N'}, \text{'OT'}, \text{'OP'} \\ \text{'N'} &\rightarrow \text{'OT'}, \text{'OP'} \end{aligned} \tag{3.47}$$

The rationale behind this grammar is as follows. A primary governing princple is that onsets, as they indicate the beginnings of note events, may not occur in adjacent frames. In other words, an onset mode must be followed immediately by a continuation or null mode:

$$P\left(M_{t+1} \in \mathcal{C} \cup \mathcal{N} | M_t \in \mathcal{O}\right) = 1 \tag{3.48}$$

The latter ensures a well defined segmentation, especially when attack transients occupy more than one frame. Additionally, each note event must have at least one frame containing pitch content. The transition behavior adheres otherwise to the cyclic succession (transient $\rightarrow$ pitched $\rightarrow$ null). where region lengths are modeled as continuous valued random variables. Transient and null region lengths can be zero whereas the pitched region length must be at least one frame. Since more than one region may exist within a given frame, by convention we assign the mode M_t to be the label of the *rightmost* region contained within that frame: see Figure 3.14.

The Markov transition diagram, displayed in Figure 3.15, encodes the standard note evolution grammar, with additional tolerances for spurious incidents (e.g., an attack transient followed immediately by silence). The latter may arise, for instance, from the sudden "splicing" of recordings in the middle of note events. Solid lines in Figure 3.15 represent transitions due to the standard note evolution grammar while dotted lines represent transitions arising from spurious behaviors. The latter transitions are assigned a small, fixed probability (on the order of $1 \cdot 10^{-3}$).

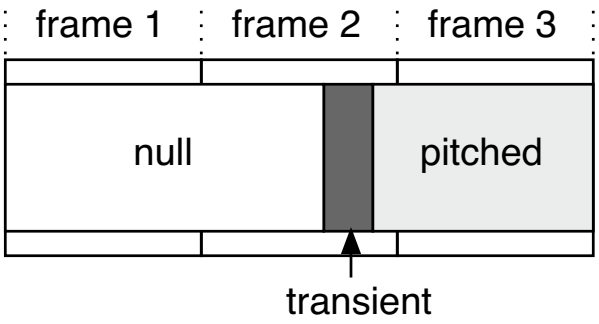


Figure 3.14: *Schema for labeling frames according to the rightmost region assignment. In this example, frame 2 is labeled* $'\mathrm{OP}'$ *even though the majority of this frame is occupied by a null region, and this frame also contains a transient region*

The free parameters of $P\left(M_{t+1}|M_t\right)$ are the transition probabilities for the standard note evolution grammar. Define:

$$p_{k|j} \triangleq P\left(M_{t+1}=k|M_t=j\right) \ \forall j,k \in \mathcal{M} \tag{3.49}$$

Then, let for each $j \in \mathcal{M}$, $\mathcal{S}_j \subset \mathcal{M}$ denote the set of possibilities for k for which $p_{k|j}$ represents a transition probability in the standard note evolution grammar. Define the vector θ_M as the collection of free parameters of $P\left(M_{t+1}|M_t\right)$:

$$\theta_M \triangleq \mathrm{Vec}\left(\bigcup_{j\in M}\bigcup_{k\in\mathcal{S}_j}\left\{p_{k|j}\right\}\right) \tag{3.50}$$

We estimate θ_M via the expectation-maximization (EM) algorithm [28]. Ideally, θ_M is chosen to maximize the likelihood $P(Y_{1:N}|\theta_M)$; however, the latter is generally intractable due to the marginalization over the hidden trajectories $M_{1:N}$, $S_{1:N}$, of which the number of possibilities grows exponentially with N. As such, the EM algorithm encompasses iterations $\theta_M^{(i)} \rightarrow \theta_M^{(i+1)}$ for which $\left\{\theta_M^{(i)}, i \geq 0\right\}$ converges to a local maximum of this likelihood; algorithm details are supplied in Section 3.7.2, with additional derivations in Appendix B. A favorable initialization ensures rapid convergence to the global likelihood maximum; we denote the latter as θ_M^*.

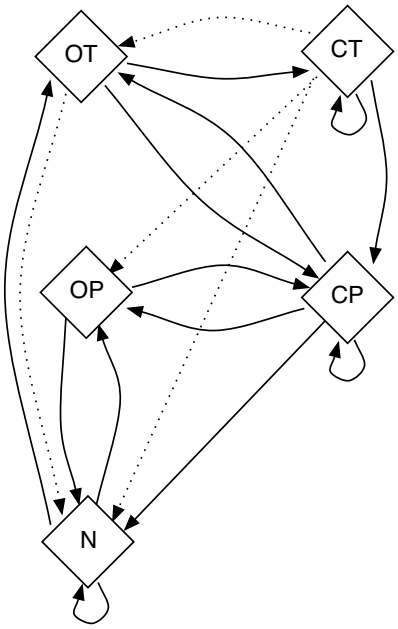


Figure 3.15: *Markov transition diagram for* $P(M_{t+1}|M_t)$.

To initialize the EM algorithm, we introduce a heterogeneous Poisson process model representing the cyclic succession of a transient region of expected length N_T, followed by a pitched region of expected length N_P, followed by a null region of expected length N_N. Individual lengths are modeled as independent, exponentially distributed random variables. Given the "rightmost region" assignment (Figure 3.14), one may determine any transition probability in the initial value $\theta_M^{(0)}$, as represented by the nonzero elements in Table 3.2. Here, each term $p_{j,k}^{(k)}$ denotes the probability that the beginning of the next frame lies in a region of type y of the m^{th} subsequent cycle given that the beginning of the current frame lies in a region of type j, where $j, k \in \{T, P, N\}$, and where T corresponds to a transient, P corresponds to a pitched, and N corresponds to a null region. For example, if the current frame corresponds to a pitched region, the probability that no transition occurs in the next frame is $p_{P,P}^{(0)}$. The probability that the boundary of the next frame lies within the pitched region of the *subsequent* note is $p_{P,P}^{(1)}$. Lastly, the probability of spurious transition, denoted as p_s, is set to some small, nonzero value; for instance, $p_s = 0.001$ for the results of

Section 3.9

	$M_{t+1}='\mathrm{OT}'$	$M_{t+1}='\mathrm{OP}'$	$M_{t+1}='\mathrm{CT}'$
$M_t='\mathrm{OT}'$	0	0	$(1-p_s)p^{(0)}_{T,T}$
$M_t='\mathrm{OP}'$	0	0	0
$M_t='\mathrm{CT}'$	p_s	p_s	$(1-3p_s)p^{(0)}_{T,T}$
$M_t='\mathrm{CP}'$	$p^{(0)}_{P,T}$	$p^{(1)}_{P,P}$	0
$M_t='\mathrm{N}'$	$p^{(0)}_{N,T}$	$p^{(1)}_{N,P}$	0
	$M_{t+1}='\mathrm{CP}'$	$M_{t+1}='\mathrm{N}'$	
$M_t='\mathrm{OT}'$	$(1-p_s)(1-p^{(1)}_{T,T})$	p_s	
$M_t='\mathrm{OP}'$	$p^{(0)}_{P,P}$	$1-p^{(0)}_{P,P}$	
$M_t='\mathrm{CT}'$	$(1-3p_s)p^{(1)}_{T,T}$	p_s	
$M_t='\mathrm{CP}'$	$p^{(0)}_{P,P}$	$p^{(0)}_{P,N}$	
$M_t='\mathrm{N}'$	0	$p^{(0)}_{N,N}$	

Table 3.2: *Generative Poisson model for the initialization of* θ_M .

The state transition dependence, $P\left(S_{t+1}|M_t, M_{t+1}, S_t\right)$, governs the expected consistency between S_t and S_{t+1} as a function of M_t and M_{t+1}. For instance, recalling the legato model of Section 3.3.3, we expect pitch content (as represented by N_t and T_t) to be highly consistent when frames t and $t+1$ belong to pitched regions within the same note event ($M_t, M_{t+1} \in \mathcal{P}$). In general, $P\left(S_{t+1}|M_t, M_{t+1}, S_t\right)$ depends on M_t at least through $M_t \in \mathcal{P}$ or $M_t \in \mathcal{Q}$, as the relation between two temporally adjacent pitched states is fundamentally different than the relation between a pitched state following a non-pitched state[7]. No further dependence on M_t is assumed.

For fixed M_t, the variation of $P(S_{t+1}|M_t, M_{t+1}, S_t)$ with respect to M_{t+1} yields the primary consideration for the detection of note region boundaries. Given $M_t \in \mathcal{P}$, $M_{t+1} = '\mathrm{CP}'$ indicates that frames t and $t+1$ belong to the same note event; hence N_{t+1}, T_{t+1}, and A_{t+1} are expected to be close to N_t, T_t, and A_t, respectively. On the other hand, $M_{t+1} = '\mathrm{OP}'$ signifies that frame $t+1$ corresponds to the onset of a new note event. Here, A_{t+1} is independent of A_t, and N_{t+1} depends only on N_t through the probabilistic relation between the values of adjacent notes.

[7]This fact alone accounts for the diagonal links in Figure 3.12 which do not appear in the corresponding graph (Figure 3.6) for the legato model, because in the latter every state is pitched.

For fixed $M_t, M_{t+1} \in \mathcal{P}$, the transition behavior factors independently over the components of S_t:

$$\begin{aligned} P\left(S_{t+1}|S_t, M_{t+1} \in \mathcal{P}, M_t \in \mathcal{P}\right) &= P\left(T_{t+1}|T_t, M_{t+1}, M_t \in \mathcal{P}\right) \\ &\times P\left(N_{t+1}|N_t, M_{t+1} \in \mathcal{P}, M_t \in \mathcal{P}\right) P\left(A_{t+1}|A_t, M_{t+1} \in \mathcal{P}, M_t \in \mathcal{P}\right) \end{aligned} \tag{3.51}$$

Similar expressions result for $M_t \in \mathcal{Q}$ and $M_{t+1} \in \mathcal{Q}$, except in these cases A_{t+1} is replaced by $A^{\mathcal{Q}}_{t+1}$, and A_t is replaced by $A^{\mathcal{Q}}_t$. We observe that the factorization (3.51) assumes no *a priori* interdependence between state components when M_t and M_{t+1} are in evidence. In practice, such interdependence exists: for instance, $\{T_t = 0.4\bar{9}, N_t = 60\}$ and $\{T_t = 0.5, N_t = 61\}$ refer to the same pitch hypothesis. The latter ambiguity occurs upon portamento/legato transitions between notes as evidenced, for instance, in the violin recording analyzed in Section 3.9.1. Despite these difficulties, the system correctly detects the onset of the second note in the transition, and identifies the portamento pitch contour.

We discuss now the individual distributions on the r.h.s. of (3.51), considering note, tuning, and amplitude, in that order. To begin, if $M_{t+1} = {}'\mathrm{CP}'$, $M_t \in \mathcal{P}$ with probability one; hence, frames t and $t+1$ belong to the same note event, and $N_{t+1} \approx N_t$. In these cases, we choose the conditional distribution of N_{t+1} given N_t to concentrate about N_t. To express this concentration, we define the double-sided exponential distribution:

$$\mathrm{E2}\left(X_1, \alpha_+, \alpha_-|X_0\right) = \begin{cases} c, & X_1 = X_0 \\ c\alpha_+^{K(X_1)-K(X_0)}, & X_1 > X_0 \\ c\alpha_-^{K(X_0)-K(X_1)}, & X_0 > X_1 \end{cases} \tag{3.52}$$

where c is chosen such that the distribution sums to unity, and $K(X) = k$ means that X is the k^{th} smallest element in the finite set of values for X. For N_{t+1} given N_t, the dependence is symmetric:

$$P\left(N_{t+1}|N_t, M_{t+1} = {}'\mathrm{CP}', M_t \in \mathcal{P}\right) \sim \mathrm{E2}\left(N_{t+1}|N_t, \alpha_N, \alpha_N\right) \tag{3.53}$$

Ideally $\alpha_N = 0$, but we must allow some small deviation for robustness to the case where the tuning offset approaches ± 0.5, as here some ambiguity may result as to the note value. Now, if $M_{t+1} \in \mathcal{Q}$, no information about the note is reflected in the observations. Here we adopt the convention that N_{t+1} refers to the value of the most recent note event; upon transition to a new event, we will have memorized the value of the previous event and can thus apply knowledge from melodic expectations insofar as note-to-note dependences are concerned[8]. Hence

$$P\left(N_{t+1}|N_t, M_{t+1} \in \mathcal{Q}, M_t \in \mathcal{M}\right) \;\sim\; \mathrm{E2}\left(N_{t+1}|N_t, \alpha_{N-}, \alpha_{N+}\right) \tag{3.54}$$

where $\alpha_{N+} = \alpha_{N-} \stackrel{\Delta}{=} \alpha_N$. Finally, we let $P_{note\_trans}(N_1|N_0)$ be the dependence where N_0 and N_1 are the values of adjacent note events; if such information is absent, the dependence is uniform over N_1, independent of N_0. If $M_{t+1} = \text{'OP'}$, the frame $t+1$ is the first frame where the value of the new note event can be observed. Since N_t memorizes the value of the previous note, the conditional distribution of N_{t+1} must follow $P_{note\_trans}(N_1|N_0)$:

$$P\left(N_{t+1}|N_t, M_{t+1} = \text{'OP'}, M_t \in \mathcal{M}\right) \;\sim\; P_{note\_trans}\left(N_1|N_0\right) \tag{3.55}$$

The remaining cases involve certain (M_t, M_{t+1}) combinations which occur with zero probability due to the mode transition dependence (Table 3.2). These distributions do not affect the inference outcome, so we specify them to minimize computational effort (see Table 3.3).

The conditional distribution of the tuning reflects the assumption that tuning is

[8]Of course, the implied first-order Markov characterization of the prior melodic structure becomes severely inadequate when considering the structural forms implied by most musics, specifically Western tonal music. In the latter case, more sophisticated Markov models may be developed, such as the model of Leistikow [71], which effectively augments N_t with past notes and intervals, as well as higher-level contextual attributes such as key, harmony, meter, harmonic rhythm, and beat position. Leistikow's model translates melodic expectations derived from the music cognition research of Narmour [85], Krumhansl [64], Schellenberg [101], Lerdahl [73] Larson and McAdams [69], and others into variable first-order Markov dependences of the form $P_{note\_trans}(N_1|N_0)$, where N_0, N_1 are the augmented states corresponding to successive note events; see Section 3.10.1 for further details. Whereas the simple note-to-note model is not very useful in practice, the key innovation in the memorization of N_t in non-pitched states is to demonstrate how contextual predictions on the level of syntax (i.e., melodic expectations) may inform detection capabilities at the signal level.

expected to be constant, or vary only slightly throughout the recording, independently of onsets, offsets and note events. Of course, this is not entirely true, as some instruments exhibit a dynamic pitch envelope. Hence

$$P\left(T_{t+1}|T_t, M_{t+1} \in \mathcal{M}, M_t \in \mathcal{M}\right) \;\sim\; \mathrm{E2}\left(T_{t+1}|T_t, \alpha_{T+}, \alpha_{T-}\right) \tag{3.56}$$

where $\alpha_{T+} = \alpha_{T-} \stackrel{\Delta}{=} \alpha_T$ indicates symmetry of the expected tuning variation. If it is known that the pitch will decay, such as in tuned percussion sources, we may adjust $\alpha_{T+} < \alpha_{T-}$ to reflect this possibility.

Finally, we consider the conditional distribution of both pitched and null amplitudes. The case $(M_{t+1} = {}'\mathrm{CP}', M_t \in \mathcal{P})$ implies that A_t and A_{t+1} belong to the same note event, A_{t+1} concentrating about A_t as follows:

$$P\left(A_{t+1}|A_t, M_{t+1} = {}'\mathrm{CP}', M_t \in \mathcal{P}\right) \;\sim\; \mathrm{E2}\left(A_{t+1}|A_t, \alpha_{A+}, \alpha_{A-}\right) \tag{3.57}$$

where $\alpha_{A+} \leq \alpha_{A-}$. Setting $\alpha_{A+} < \alpha_{A-}$ indicates a decaying amplitude evolution throughout the note duration, best adapted to percussive tones like piano and marimba. On the other hand, setting $\alpha_{A+} = \alpha_{A-}$ may be more appropriate for violin, voice, and other sustained tones, or instruments with lengthy attack regions. In all other cases, A_{t+1} is independent of A_t (or $A_t^{\mathcal{Q}}$). Where $M_{t+1} = {}'\mathrm{OP}'$, A_{t+1} corresponds to the pitch amplitude of the onset of a note event. In these cases, A_{t+1} resamples from a distribution favoring higher amplitudes:

$$\begin{aligned} P(A_{t+1}|A_t, M_{t+1} = {}'\mathrm{OP}', M_t \in \mathcal{P}) &\sim \mathrm{E1}(A_{t+1}, \beta_{A,{}'\mathrm{OP}'}) \\ P(A_{t+1}|A_t^{\mathcal{Q}}, M_{t+1} = {}'\mathrm{OP}', M_t \in \mathcal{Q}) &\sim \mathrm{E1}(A_{t+1}, \beta_{A,{}'\mathrm{OP}'}) \end{aligned} \tag{3.58}$$

where, using the notation of (3.52),

$$\mathrm{E1}(X, \beta) \;=\; c^{\beta K(X)} \tag{3.59}$$

The constant c is chosen such that for fixed β, $\mathrm{E1}(X, \beta)$ sums to unity over values of X. Setting $\beta_{A,OP} > 1$ means that the pitched onset amplitude concentrates

about higher amplitudes. Where $M_{t+1} \in {}'\mathrm{OT}'$ or $M_{t+1} \in {}'\mathrm{CT}'$ (i.e., $M_{t+1} \in \mathcal{T}$), the distribution is similar, but it concerns A^Q_{t+1} instead of A_{t+1}:

$$\begin{aligned} P(A^Q_{t+1}|A_t, M_{t+1} \in \mathcal{T}, M_t \in \mathcal{P}) &\sim \mathrm{E1}(A^Q_{t+1}, \beta_{A,\mathcal{T}}) \\ P(A^Q_{t+1}|A^Q_t, M_{t+1} \in \mathcal{T}, M_t \in \mathcal{Q}) &\sim \mathrm{E1}(A^Q_{t+1}, \beta_{A,\mathcal{T}}) \end{aligned} \tag{3.60}$$

where $\beta_{A,\mathcal{T}} > 1$. Where $M_{t+1} = {}'\mathrm{N}'$, the distribution of A^Q_{t+1} follows either line of (3.60), depending on $M_t \in \mathcal{P}$ or $\mathcal{Q}$, but with $\beta_{A,'N'} < 1$ in place of $\beta_{A,\mathcal{T}}$, since the null mode favors lower amplitudes. Table 3.3 summarizes the aforementioned state transition behavior, filling in "don't-care" possibilities.

M_{t+1}	$M_t \in \mathcal{Q}$
$'\mathrm{OT}'$	$P(N_{t+1}\|N_t) \sim \mathrm{E2}(N_{t+1}\|N_t, \alpha_N, \alpha_N)$ $P(T_{t+1}\|T_t) \sim \mathrm{E2}(T_{t+1}\|T_t, \alpha_T, \alpha_T)$ $P(A^Q_{t+1}\|A^Q_t) \sim \mathrm{E1}(A^Q_{t+1}, \beta_{A,\mathcal{T}})$
$'\mathrm{OP}'$	$P(N_{t+1}\|N_t) \sim P_{note\_trans}(N_{t+1}\|N_t)$ $P(T_{t+1}\|T_t) \sim \mathrm{E2}(T_{t+1}\|T_t, \alpha_T, \alpha_T)$ $P(A_{t+1}\|A^Q_t) \sim \mathrm{E1}(A_{t+1}, \beta_{A,'\mathrm{OP}'})$
$'\mathrm{CT}'$	$P(N_{t+1}\|N_t) \sim \mathrm{E2}(N_{t+1}\|N_t, \alpha_N, \alpha_N)$ $P(T_{t+1}\|T_t) \sim \mathrm{E2}(T_{t+1}\|T_t, \alpha_T, \alpha_T)$ $P(A^Q_{t+1}\|A^Q_t) \sim \mathrm{E1}(A^Q_{t+1}, \beta_{A,\mathcal{T}})$
$'\mathrm{CP}'$	$P(N_{t+1}\|N_t) \sim P_{note\_trans}(N_{t+1}\|N_t)$ $P(T_{t+1}\|T_t) \sim \mathrm{E2}(T_{t+1}\|T_t, \alpha_T, \alpha_T)$ $P(A_{t+1}\|A^Q_t) \sim \mathrm{E1}(A_{t+1}, \beta_{A,'\mathrm{OP}'})$
$'\mathrm{N}'$	$P(N_{t+1}\|N_t) \sim \mathrm{E2}(N_{t+1}\|N_t, \alpha_N, \alpha_N)$ $P(T_{t+1}\|T_t) \sim \mathrm{E2}(T_{t+1}\|T_t, \alpha_T, \alpha_T)$ $P(A^Q_{t+1}\|A^Q_t) \sim \mathrm{E1}(A^Q_{t+1}, \beta_{A,'\mathrm{N}'})$
M_{t+1}	$M_t \in \mathcal{P}$
$'\mathrm{OT}'$	$P(N_{t+1}\|N_t) \sim \mathrm{E2}(N_{t+1}\|N_t, \alpha_N, \alpha_N)$ $P(T_{t+1}\|T_t) \sim \mathrm{E2}(T_{t+1}\|T_t, \alpha_T, \alpha_T)$ $P(A^Q_{t+1}\|A_t) \sim \mathrm{E1}(A^Q_{t+1}, \beta_{A,\mathcal{T}})$
$'\mathrm{OP}'$	$P(N_{t+1}\|N_t) \sim P_{note\_trans}(N_{t+1}\|N_t)$ $P(T_{t+1}\|T_t) \sim \mathrm{E2}(T_{t+1}\|T_t, \alpha_T, \alpha_T)$ $P(A_{t+1}\|A_t) \sim \mathrm{E1}(A_{t+1}, \beta_{A,'\mathrm{OP}'})$
$'\mathrm{CT}'$	$P(N_{t+1}\|N_t) \sim \mathrm{E2}(N_{t+1}\|N_t, \alpha_N, \alpha_N)$ $P(T_{t+1}\|T_t) \sim \mathrm{E2}(T_{t+1}\|T_t, \alpha_T, \alpha_T)$ $P(A^Q_{t+1}\|A_t) \sim \mathrm{E1}(A^Q_{t+1}, \beta_{A,\mathcal{T}})$
$'\mathrm{CP}'$	$P(N_{t+1}\|N_t) \sim \mathrm{E2}(N_{t+1}\|N_t, \alpha_N, \alpha_N)$ $P(T_{t+1}\|T_t) \sim \mathrm{E2}(T_{t+1}\|T_t, \alpha_T, \alpha_T)$ $P(A_{t+1}\|A_t) \sim \mathrm{E2}(A_{t+1}\|A_t, \alpha_{A+}, \alpha_{A-})$
$'\mathrm{N}'$	$P(N_{t+1}\|N_t) \sim \mathrm{E2}(N_{t+1}\|N_t, \alpha_N, \alpha_N)$ $P(T_{t+1}\|T_t) \sim \mathrm{E2}(T_{t+1}\|T_t, \alpha_T, \alpha_T)$ $P(A^Q_{t+1}\|A_t) \sim \mathrm{E1}(A^Q_{t+1}, \beta_{A,'\mathrm{N}'})$

Table 3.3: *State transition table for component distributions of* $P(S_{t+1}|S_t, M_{t+1}, M_t)$

3.6.3 Frame likelihood

We wish to evaluate the likelihoods for frames with pitch content: $P(Y_t|N_t, T_t, A_t, M_t \in \mathcal{P})$, and for frames without: $P(Y_t|A_t^Q, M_t \in \mathcal{Q})$. For frames with pitch content, $P(Y_t|N_t, T_t, A_t, M_t \in \mathcal{P})$ is computed by a modification of the method developed in Chapter 4, to which we henceforth refer as the *canonical evaluation*. The latter evaluates $P(Y_t|f_{0,t}, A_{0,t})$ where $f_{0,t}$ is the radian fundamental frequency and $A_{0,t}$ the reference amplitude for the t^{th} frame, and Y_t consists of the joint collection of peak frequencies and amplitudes; i.e., $Y_t = \{F, A\}$, where

$$\begin{aligned} F &\triangleq \{F(1), F(2), \ldots, F(N_o)\} \\ A &\triangleq \{A(1), A(2), \ldots, A(N_o)\} \end{aligned} \tag{3.61}$$

Here $F(k)$ denotes the frequency of the k^{th} lowest-frequency STFT peak and $A(k)$ the corresponding amplitude. We henceforth denote the canonical evaluation as $P_{can}(Y_t|f_{0,t}, A_{0,t})$ to distinguish it from the subsequent modification.

The canonical evaluation proves robust to real-world phenomena such as inharmonicity, undetected peaks, and spurious peaks due to noise and other interference phenomena, as indicated by the results shown in Section 4.4; see Furthermore, straightforward extensions exist for the polyphonic case; as implemented in the Bayesian chord recognizer of Leistikow et al. [72]. However, care must be taken in the association of the hypotheses $(f_{0,t}, A_{0,t})$ with those of the state (N_t, T_t and A_t). While $f_{0,t}$ is uniquely determined by N_t and T_t, the relation between the reference amplitude, $A_{0,t}$, and A_t becomes more involved. In the canonical evaluation, the reference amplitude is estimated as the maximum amplitude over all peaks in the frame, denoted as $A_{max,t}$. The latter yields favorable psychoacoustic properties in the context of many real-world signals which are assumed to be monophonic, but are actually polyphonic. For instance, consider a recording of the introductory motive of Bach's Invention 2 in C minor (BWV 773) by Glenn Gould. Here the pianist hums two octaves below the piano melody. The humming can barely be heard in most frames; nevertheless, the likelihood evaluation sometimes favors the voice's fundamental rather than that of the piano, especially when these fundamentals are in

an exact harmonic relationship. While this result may be technically correct in the absence of explicit timbral models, it fails to represent what is heard as salient. Now, one may argue that the perceived salience of the piano melody arises from the consistency of pitch and amplitude information across long segments of frames, as the voice tends to fade in and out over these regions. We find, nevertheless, from informal listening tests, that the perceived salience of the piano tone persists even in the absence of contextual cues; for instance, when a single frame is extracted and repeated for any given duration. A plausible explanation is that in the absence of other contextual cues, we focus on the loudest of multiple pitch components, hence the choice $A_{0,t} = A_{max,t}$.

Unfortunately, use of $P_{can}(Y_t|f_{0,t}, A_{0,t})$ with $A_{0,t} = A_{max,t}$ ignores the state variable A_t, thus preventing the conditional distribution of A_t from being influenced by the signal, except indirectly via M_t. This in turn diminishes the capacity of jumps in the signal's amplitude envelope to inform the segmentation, which can be a critical issue when detecting onsets of repeated notes. Our solution is to take $A_{0,t} = A_t$ while introducing $A_{max,t}$ as an independent noisy observation[9] of A_t, as shown in Figure 3.16. By so doing, we blend the strategy which derives $A_{0,t}$ from the state ($A_{0,t} = A_t$) with the strategy incorporating psychoacoustic salience ($A_{0,t} = A_{max,t}$). The conditional distribution for the observation layer becomes

$$P\left(Y_t, A_{max,t}|N_t, T_t, A_t\right) = P\left(Y_t|N_t, T_t, A_t\right) P\left(A_{max,t}|A_t\right) \tag{3.62}$$

Here, $P(Y_t|N_t, T_t, A_t)$ is modeled by the canonical evaluation with $A_{0,t} = A_t$ and $f_{0,t}$ the fundamental frequency corresponding to the pitch hypothesis indicated by N_t and T_t, and $A_{max,t}$ is modeled as A_t plus Gaussian noise:

$$\begin{aligned} P\left(Y_t|N_t, T_t, A_t\right) &= P_{can}\left(Y_t|f_0(N_t, T_t), A_{0,t} = A_t\right) \\ P\left(A_{max,t}|A_t\right) &= \mathcal{N}\left(A_t, \sigma_A^2\right) \end{aligned} \tag{3.63}$$

[9]It may seem counterintuitive to model $A_{max,t}$ and Y_t as conditionally independent of A_t since, *unconditionally* speaking, $A_{max,t}$ is a *deterministic* function of Y_t. However, we wish not to introduce bias by assuming *specific* dependences between the noise on $A_{max,t}$ and the amplitude/frequency noises on other peaks of Y_t.

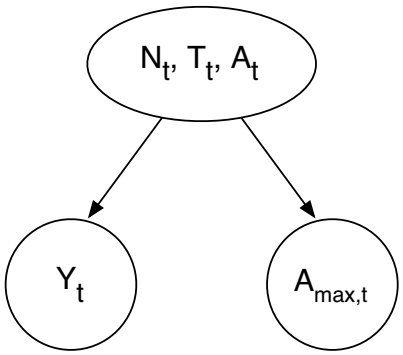


Figure 3.16: *Observation layer dependence with* $A_{max,t}$

We may interpolate between the rigid cases ($A_{0,t} = A_{max,t}$ vs. $A_{0,t} = A_t$) by varying σ_A^2 between 0 and ∞. Assuming $A_t \in \mathbb{R}^+$, as $\sigma_A^2 \to 0$, the pitch inference $P(N_t, T_t|Y_t, A_{max,t})$, becomes identical to the inference $P'(N_t, T_t|Y_t)$ where $P'(Y_t|N_t, T_t, A_t)$ equals the canonical evaluation, $P^*(Y_t|f_0(N_t, T_t), A_{0,t} = A_t)$, with $A_{0,t} = A_{max,t}$. On the other hand, as $\sigma_A^2 \to \infty$, the pitch inference $P(N_t, T_t|Y_t, A_{max,t})$ converges to $P(N_t, T_t|Y_t)$, which is the canonical evaluation using $A_{0,t} = A_t$, $A_{max,t}$ being ignored.

To show, we first consider $\sigma_A^2 \to \infty$; here the dependence on A_t vanishes: $P(A_{max,t}|A_t) \to P(A_{max,t})$. As a result, $A_{max,t}$ and the collection $\{Y_t, N_t, T_t, A_t\}$ become mutually independent. Then $P(N_t, T_t|Y_t, A_{max,t}) \to P(N_t, T_t|Y_t)$, as was to be shown.

Next we consider $\sigma_A^2 \to 0$; to begin, we note that in this case, $P(A_{max,t}|A_t)$ becomes impulsively concentrated about A_t; i.e.:

$$P(A_{max,t}|A_t) \sim \delta(A_{max,t}, A_t) \tag{3.64}$$

It suffices to show that, given (3.64), $P(N_t, T_t|Y_t, A_{max,t})$ becomes identical to the inference $P'(N_t, T_t|Y_t)$ where

$$P'(Y_t|N_t, T_t, A_t) = P_{can}(Y_t|f_0(N_t, T_t), A_{0,t} = A_{max,t}) \tag{3.65}$$

Expanding $P(N_t, T_t|Y_t, A_{max,t})$ according to Bayes' rule yields the following:

$$P(N_t, T_t|Y_t, A_{max,t}) = \frac{\sum_{\nu,\tau} \int_{A_t} \pi(A_t, \nu, \tau, Y_t, A_{max,t})\, dA_t}{\int_{A_t} \pi(A_t, N_t, T_t, Y_t, A_{max,t})\, dA_t} \;\forall\; N_t, T_t \quad (3.66)$$

where

$$\pi(A_t, N_t, T_t, Y_t, A_{max,t}) \triangleq P(A_t, N_t, T_t, Y_t, A_{max,t}) \quad (3.67)$$

and

$$\begin{aligned} P(A_t, N_t, T_t, Y_t, A_{max,t}) &= P(A_t)\, P(N_t, T_t|A_t) \\ &\times P_{can}(Y_t|f_0(N_t, T_t), A_{0,t} = A_t)\, \delta(A_{max,t}, A_t) \end{aligned} \quad (3.68)$$

Substituting (3.68) into (3.66) results in integral expressions with impulsive terms. These expressions, and hence (3.66), simplify to

$$\begin{aligned} \int_{A_t} P(A_t, N_t, T_t, Y_t, A_{max,t})\, dA_t &= P(N_t, T_t|A_t = A_{max,t}) \\ &\times P_{can}(Y_t|f_0(N_t, T_t), A_{0,t} = A_{max,t}) \end{aligned} \quad (3.69)$$

Now, since A_t and $\{N_t, T_t\}$ are *a priori* independent, (3.69) simplifies further:

$$\begin{aligned} \int_{A_t} P(A_t, N_t, T_t, Y_t, A_{max,t})\, dA_t &= P(N_t, T_t) \\ &\times P_{can}(Y_t|f_0(N_t, T_t), A_{0,t} = A_{max,t}) \end{aligned} \quad (3.70)$$

It follows that the substitution of (3.70) into (3.66) obtains the same relation as the expansion of (3.65) via Bayes' rule, in parallel fashion to (3.66). Hence

$$P(N_t, T_t|Y_t, A_{max,t}) = P'(N_t, T_t|Y_t, A_{max,t}) \quad (3.71)$$

as was to be shown.

In the preceding development, the space of A_t was assumed to be $\mathbb{R}^+$ which is an uncountably infinite space. In actuality the domain of A_t is limited and the space

discretized to a finite set of possibilities. Nevertheless, provided the domain's extent is sufficiently large, and σ_A^2 considerably exceeds the square of the largest spacing between A_t values, the results realized "in practice" become virtually identical to the analyzed situation where $A_t \in \mathbb{R}^+$.

As a final note, some frames may lack pitch content altogether; these correspond to purely transient effects (e.g., percussion), background noise, or silence. In these cases $M_t \in \mathcal{Q}$. Since we still wish to model a general amplitude characteristic associated with these frames, in order to distinguish transients from silence, for instance, we model the frame via $P_{can}(Y_t|f_0(N_t, T_t), A_t^{\mathcal{Q}})$ under the restriction that all peaks are spurious.

3.7 Inference methodology

3.7.1 Primary inference

The primary inference goals for the joint onset detection, transient region identification, and melody transcription, as discussed in Section 3.5.2, are the determination of the maximum *a posteriori* mode sequence $M_{1:N}^*$; i.e.,

$$M_{1:N}^* = \underset{M_{1:N}}{\operatorname{argmax}} \; P\left(M_{1:N}|Y_{1:N}\right) \tag{3.72}$$

and the computation of the smoothed state posterior given $M_{1:N}^*$; denoted as $\sigma_{1:N}^*$; i.e.,

$$\sigma^*(S_t) = P(S_t|M_{1:N}^*, Y_{1:N}), \forall t \in 1:N \tag{3.73}$$

If, for any $t \in 1:N-1$, Y_{t+1} is conditionally independent of $Y_{1:t}$ and $M_{1:t}$ given M_{t+1}, the Viterbi algorithm [94] may be used to identify $\hat{M}_{1:N}^*$. Unfortunately, the implicit marginalization of $S_{1:N}$ in $P\left(M_{1:N}|Y_{1:N}\right)$ precludes this possibility. As the complexity of the Viterbi approach is linear in the number of frames and quadratic

in the number of modes, we seek an approximate Viterbi inference with similar computational cost. To this end, we approximate:

$$P(Y_{t+1}|M_{1:t+1}, Y_t) \approx P(Y_{t+1}|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t}) \tag{3.74}$$

where

$$M^*_{1:t-1}(M_t) \approx \operatorname*{argmax}_{M_{1:t-1}} P(M_{1:t-1}|M_t, Y_{1:t}) \tag{3.75}$$

We refer to $M^*_{1:t-1}(M_t)$ as the (approximate) *M_t-optimal mode sequence*, and define $M^*_{a:b}(M_t)$ as the restriction of this sequence to frames a through b, and adopt the shorthand $M^*_a \triangleq M^*_{a:a}$. This approximation, similar to that used by Pavlovic et al. [87] for the learning of switching linear models of human motion, treats the history of the mode sequence up to time $t-1$ collectively as a nuisance parameter, replacing its value with the corresponding maximum *a posteriori* estimate given M_t and $Y_{1:t}$.

The inference proceeds in two passes, a forward, *filtering* pass, followed by a backward, *smoothing* pass. Table 3.4 summarizes the quantities propagated in these passes, as well as the necessary input distributions (the conditional dependences on the r.h.s. of the factorization (3.42)). The designation ($\approx$) means the referenced quantity is approximate.

Symbol	Quantity	Description
$\tau^*(M_t, S_t)$	$P(S_t\|M^*_{1:t-1}(M_t), M_t, Y_{1:t-1})$	Predicted posterior given M_t-optimal mode sequence
$\mu^*(M_t, S_t)$	$P(S_t\|M^*_{1:t-1}(M_t), M_t, Y_{1:t})$	Smoothed posterior given M_t-optimal mode sequence
$J(M_t)$	$\max_{M_{1:t-1}} P(M_{1:t}\|Y_{1:t})$ ($\approx$)	Objective at time t
$M^*_{t-1}(M_t)$	$\operatorname{argmax}_{M_{t-1}} \max_{M_{1:t-2}} P(M_{1:t}\|Y_{1:t})$ ($\approx$)	Backpointer
M^*_t	$\operatorname{argmax}_{M_t} \max_{M_{1:t-1}, M_{t+1:N}} P(M_{1:N}\|Y_{1:N})$ ($\approx$)	MAP mode at time t
$\sigma^*_t(S_t)$	$P(S_t\|M^*_{1:N}, Y_{1:N})$	Smoothed posterior
$\mu_0(M_t, S_t, M_{t+1})$	$P(S_{t+1}, Y_{t+1}\|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t})$	Intermediate
$\tau(M_t, S_t, M_{t+1})$	$P(S_{t+1}\|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t})$	Intermediate
$\mu(M_t, S_t, M_{t+1})$	$P(S_{t+1}\|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t+1})$	Intermediate
$\Sigma_0(M_t, M_{t+1})$	$P(Y_{t+1}\|M^*_{1:t-1}(M_t), M_{t+1}, Y_{1:t+1})$	Intermediate
$J_0(M_t, M_{t+1})$	$\max_{M_{1:t-1}} P(M_{1:t+1}\|Y_{1:t+1})$ ($\approx$)	Intermediate

Table 3.4: *Approximate Viterbi inference inputs and propagated quantities*

The computation of $M^*_{1:N}$ and $\{\sigma^*(S_t)\}_{t=1}^N$ via Table 3.4 satisfies (3.72) and (3.73), as desired.

To begin, the filtering pass is initialized as follows.

$$\begin{aligned}
\mu^*(S_1, M_1) &= P(M_1, S_1|Y_1) \\
&= \frac{P(S_1|M_1)P(Y_1|S_1)}{\sum_{S_1} P(S_1|M_1)P(Y_1|S_1)} \\
J(M_1) &= P(M_1|Y_1) \\
&= \frac{P(M_1)\sum_{S_1} P(S_1|M_1)P(Y_1|S_1)}{\sum_{M_1} P(M_1)\sum_{S_1} P(S_1|M_1)P(Y_1|S_1)}
\end{aligned} \tag{3.76}$$

Then, for $t \in 1 : N-1$, the filtering recursions proceed:

$$\begin{aligned}
\tau\left(M_t, S_t, M_{t+1}\right) &= \sum_{S_t} \mu^*\left(M_t, S_t\right) P\left(S_{t+1}|M_t, M_{t+1}, S_t\right) \\
\mu_0\left(M_t, S_t, M_{t+1}\right) &= P\left(Y_{t+1}|S_{t+1}\right) \tau\left(M_t, S_t, M_{t+1}\right) \\
\Sigma_0\left(M_t, M_{t+1}\right) &= \sum_{S_{t+1}} \mu_0\left(M_t, S_t, M_{t+1}\right) \\
J_0\left(M_t, M_{t+1}\right) &= J\left(M_t\right) P\left(M_{t+1}|M_t\right) \Sigma_0\left(M_t, M_{t+1}\right) \\
\mu\left(M_t, S_t, M_{t+1}\right) &= \frac{\mu_0\left(M_t, S_t, M_{t+1}\right)}{\Sigma_0\left(M_t, M_{t+1}\right)} \\
M_t^*\left(M_{t+1}\right) &= \underset{M_t}{\operatorname{argmax}}\ J_0\left(M_t, M_{t+1}\right) \\
J\left(M_{t+1}\right) &= \frac{J_0\left(M_t, M_{t+1}\right)}{P\left(Y_{t+1}|Y_t\right)} \\
\mu^*\left(S_t, M_t\right) &= \mu\left(S_{t+1}, M_t^*\left(M_{t+1}\right), M_{t+1}\right) \\
\tau^*\left(S_t, M_t\right) &= \tau\left(S_{t+1}, M_t^*\left(M_{t+1}\right), M_{t+1}\right)
\end{aligned} \tag{3.77}$$

For $t \geq 1$, $\mu^*(M_t, S_t)$ and $M_t^*(M_{t+1})$ are stored as well as $\tau^*(M_t, S_t)$ for $t \geq 2$, and $J(M_N)$. These quantities are necessary for efficient computation of the smoothing pass. The latter is initialized as follows.

$$\begin{aligned}
M_N^* &= \underset{M_N}{\operatorname{argmax}}\ J\left(M_N\right) \\
\sigma^*(S_N) &= \mu^*(S_N, M_N^*)
\end{aligned} \tag{3.78}$$

Smoothing recursions proceed as t decreases from $N-1$ down to 1, as follows:

$$\begin{aligned} M_t^* &= M_t^*(M_{t+1}^*) \\ \sigma^*(S_t) &= \mu^*(S_t, M_t^*) \sum_{S_{t+1}} \frac{\sigma^*(S_{t+1}) P(S_{t+1}|S_t, M_t^*, M_{t+1}^*)}{\tau^*(S_{t+1}, M_{t+1}^*)} \end{aligned} \tag{3.79}$$

Recursions (3.76 - 3.78) are derived in Appendix A.

Finally, we still need to compute the smoothed posteriors for the individual components of S_t as required by (3.44). These are given by marginalizing out the other components of S_t according to the definitions (3.40). There are two cases: $M_t^* \in \mathcal{P}$, or $M_t^* \in \mathcal{Q}$:

$$\begin{aligned} P(N_t|M_{1:N}^*, Y_{1:N}) &= \sum_{T_t, A_t} P(S_t|M_{1:N}^*, Y_{1:N}),\ M_t^* \in \mathcal{P} \\ P(N_t|M_{1:N}^*, Y_{1:N}) &= \sum_{T_t, A_t^Q} P(S_t|M_{1:N}^*, Y_{1:N}),\ M_t^* \in \mathcal{Q} \end{aligned} \tag{3.80}$$

$$\begin{aligned} P(T_t|M_{1:N}^*, Y_{1:N}) &= \sum_{N_t, A_t} P(S_t|M_{1:N}^*, Y_{1:N}),\ M_t^* \in \mathcal{P} \\ P(T_t|M_{1:N}^*, Y_{1:N}) &= \sum_{N_t, A_t^Q} P(S_t|M_{1:N}^*, Y_{1:N}),\ M_t^* \in \mathcal{Q} \end{aligned} \tag{3.81}$$

$$\begin{aligned} P(A_t|M_{1:N}^*, Y_{1:N}) &= \sum_{T_t, N_t} P(S_t|M_{1:N}^*, Y_{1:N}),\ M_t^* \in \mathcal{P} \\ P(A_t^Q|M_{1:N}^*, Y_{1:N}) &= \sum_{T_t, N_t} P(S_t|M_{1:N}^*, Y_{1:N}),\ M_t^* \in \mathcal{Q} \end{aligned} \tag{3.82}$$

3.7.2 Estimation of free parameters in the mode transition dependence

Recall from Section 3.6.2 that the free parameters of the mode transition dependence $P(M_{t+1}|M_t)$, may be encoded in the vector θ_M:

$$\theta_M \triangleq \text{Vec}\left(\bigcup_{j\in M}\bigcup_{k\in\mathcal{S}_j}\{p_{k|j}\}\right) \tag{3.83}$$

where $\mathcal{S}_j \subset \mathcal{M}$ denotes the set of possibilities for k for which $p_{k|j} \triangleq P\left(M_{t+1}=k|M_t=j\right)$ represents a transition probability in the standard note evolution grammar (3.47).

The EM algorithm for estimating θ_M, introduced in Section 3.6.2, begins with an initial guess, i.e., $\theta_M^{(0)}$, and proceeds over iterations i, updating the estimate $\theta_M^{(i)}$. Iterations repeat until convergence. Each iteration updating $\theta_M^{(i)} \rightarrow \theta_M^{(i+1)}$ consists of two steps:

- **E-step**: Compute as follows:

$$\sigma^{(2)}(M_t, M_{t+1}) = P(M_t, M_{t+1}|Y_{1:N}, \theta_M^{(i)}) \tag{3.84}$$

  for all $t \in 1:N-1$ and $M_t, M_{t+1} \in \mathcal{M}$.

- **M-step**: Update for each $j \in \mathcal{M}, k \in \mathcal{S}_j$:

$$p_{k|j}^{(i+1)} = \frac{\sum_{t=1}^{N-1}\sigma^{(2)}(M_t=j, M_{t+1}=k)}{\sum_{k\in\mathcal{M}}\sum_{t=1}^{N-1}\sigma^{(2)}(M_t=j, M_{t+1}=k)} \tag{3.85}$$

A complete derivation of the EM algorithm steps is provided in Appendix B, Section B.1 while the computation of the pairwise smoothed posterior $P(M_t, M_{t+1}|Y_{1:N}, \theta_M^{(i)})$ is addressed in Section B.2.

3.8 Postprocessing

The goal of postprocessing is to take the maximum *a posteriori* mode sequence, $M^*_{1:N}$, (3.43) and the smoothed note posterior $P(N_t|M^*_{1:N}, Y_{1:N})$, and produce a string of distinct note events. These events can be stored in a MIDI file. With additional metrical information, one may further process the note event stream to produce a score-based representation. However, doing so properly depends on the ability to model uncertainties in metrical structure, and to integrate such models with signal information, which lies beyond the scope of the current work. By augmenting the state N_t with past note values and intervals, as well as higher-level information such as key, harmony, meter, harmonic rhythm, and beat position, the transition distribution $P_{note\_trans}(N_1|N_0)$ may represent such uncertainties. Since this distribution is activated upon transition into the first pitched frame of a new note event, following (3.55), it serves thus to integrate both signal-level and symbolic-level dependences. Forthcoming work by Leistikow [71] makes explicit, among other things, the use of the augmented N_t representation in the modeling of metrical structure (i.e., determination of bar lines) and hence the production of scores from MIDI files. The bar line determination problem is well studied; see [3, 21, 30, 113] among others. Other issues which aid the MIDI $\rightarrow$ score conversion, such as pitch spelling determination, are addressed in [22, 23] and the listed references therein, among other sources.

In the present (MIDI file) output, each event consists of an onset time, note value, and duration. Additionally, we provide a sub-segmentation into transient and pitched regions. Since the nominally monophonic model (Figure 3.10) is restricted to having for each note event, at most one transient region followed by a pitched region, it suffices for the sub-segmentation to specify the duration of the transient region. Table 3.5 summarizes the symbols defined to represent these quantities; here, all symbols refer to the k^{th} note event.

Now let $\mathcal{Z}$ be a collection of distinct integers, and let $\min(\mathcal{Z})$ be the minimum integer in the collection if $\mathcal{Z}$ is nonempty. Define:

$$\overset{+}{\min}(\mathcal{Z}) \triangleq \begin{cases} \infty, & \mathcal{Z} = \emptyset \\ \min(\mathcal{Z}) & \text{otherwise} \end{cases} \tag{3.86}$$

Symbol	Description
$o^{(k)}$	Onset frame for note event
$p^{(k)}$	First pitched frame in note event
$d^{(k)}$	Note event duration
$e_+^{(k)}$	One frame beyond end of note event
$N^{*(k)}$	MIDI note value

Table 3.5: *Transcription output quantities*

The postprocessing algorithm iterates over note events k, stopping only when either the onset frame, pitch boundary, or the advanced end point ($o^{(k)}$, $p^{(k)}$, or $e_+^{(k)}$), are infinite. This stopping condition indicates that there is not enough signal to determine information about the current or subsequent note events.

The onset frame for the first event is initialized as follows.

$$o^{(1)} = \overset{+}{\min}\{t \geq 1 : M_t^* \in \mathcal{O}\} \tag{3.87}$$

This search for an explicit onset automatically discards tail portions of note events which are truncated by the beginning of the signal.

In general, the recursions used to extract note events information are as follows:

$$\begin{aligned} o^{(k)} &= \overset{+}{\min}\left\{t \geq e_+^{(k-1)} : M_t^* \in \mathcal{O}\right\} \\ p^{(k)} &= \overset{+}{\min}\left\{t \geq o^{(k)} : M_t^* \in \mathcal{P}\right\} \\ e_+^{(k)} &= \overset{+}{\min}\left\{t \geq p^{(k)} : M_t^* \notin \mathcal{C}\right\} \end{aligned} \tag{3.88}$$

If $k = 1$, the initialization (3.87) is used in place of (3.88) in case of $o^{(k)}$. As indicated, $e_+^{(k)}$ lies one frame beyond the last frame of the note event. The duration of note k is just the simple difference between $e_+^{(k)}$ and $o^{(k)}$, unless $e_+^{(k)}$ has been truncated by the end of the signal. In the latter case, the duration is that of the truncated part: $N - o^{(k)} + 1$.

$$d^{(k)} = \min(e_+^{(k)}, N+1) - o^{(k)} \tag{3.89}$$

To obtain the MIDI note value, we extract:

$$N^{*(k)} \quad = \quad \underset{n}{\text{argmax}} \;\; P(N_{p^{(1)}+c} = n | Y_{1:N}) \tag{3.90}$$

where

$$c \quad = \quad \min(c_0, e_+^{(k)} - p^{(1)} - 1). \tag{3.91}$$

Here c is a margin variable ensuring that the maximum *a posteriori* pitch value assigned to the entire event is sampled from a frame which is some distance away from the end of the transient region. The canonical value of c, ignoring truncation effects, is c_0; $c_0 = 3$ is used to generate the examples of Section 3.9.

We note that the the algorithm seems relatively insensitive to c_0 due to consistency of N_t during pitch regions.. Clearly, the greater the *a priori* consistency, the more consistent the maximum *a posteriori* estimates: signal information is weighted more uniformly during pitched regions to estimate N_t along any point in the region; as apparent in the observation weightings for the legato model example of Figure 3.5.

Recalling (3.53), the consistency of N_t during pitched regions is captured by the following distribution:

$$P\left(N_{t+1} | N_t, M_{t+1} = \text{'CP'}, M_t \in \mathcal{P}\right) \quad \sim \quad \text{E2}\left(N_{t+1} | N_t, \alpha_N, \alpha_N\right) \tag{3.92}$$

where the double-sided exponential, E2, is defined by (3.52). The amount of consistency is governed by α_N in the sense that $\alpha_N \downarrow 0$ indicates that all note values must be identical during the entire pitched region.

In practice, it is virtually unheard of for the maxmized note posterior to fail to be identical during the pitched portions of note events, as even extreme vibrato effects may be absorbed by tuning offset variations. Hence, there is no loss of generality in setting $\alpha_N = 0$, implying that the actual value of c_0 specified in (3.90) is immaterial, as long as $c_0 > 0$.

3.9 Results

The system for joint onset detection, transient region identification, and melody transcription developed in the preceding sections has been applied to a variety of piano and violin recordings. While each recording arises from a monophonic score, in actuality they contain instances of polyphony due to reverberation, note overlaps from legato playing, and background instrumentation. Furthermore, expressive pitch variations such as portamento and vibrato occur in the violin passages. The goal of this section is to provide a detailed analysis of the system's performance for one representative example of each type: a piano recording in which the performer also sings in the background, and a violin recording exhibiting significant reverberation, legato playing and expressive pitch variation.

The piano recording consists of the introductory motive of Bach's Invention II in C minor (BWV 773), performed by Glenn Gould. The top section of Figure 3.17 displays the time domain waveform. While this recording lacks significant reverberation, the legato playing style nevertheless causes many notes to overlap. A further complication is that the performer (Gould) accompanies himself with low-amplitude vocalizations, with fundamentals approximately two octaves below those of the piano[10]. Nevertheless, onsets are clearly visible and may be determined by standard heuristic time domain approaches which monitor jumps in the amplitude envelope; see, for instance, the algorithm of Schloss [102] used in the automatic transcription of percussive music. The purpose of this example is mainly to show robustness to low levels of background instrumentation.

The violin recording is an excerpt from the third movement of Bach's solo violin Sonata No. 1 in G minor (BWV 1001), performed by Nathan Milstein. The top section of Figure 3.18 displays the time domain waveform. This recording is awash in reverberation, which makes it difficult to detect onsets visually or by monitoring the amplitude envelope. Furthermore, excessive reverberation combined with legato playing induces significant polyphony due to overlapping notes. Expressive pitch variations, particularly portamento, create ambiguities concerning onset locations. A

[10]This idiosyncrasy of Gould has been well documented; see Malone [80] for an interesting study of counterpoint with regards to the piano material as appearing in Gould's vocalizations.

significant "slide" (legato/portamento transition), for instance, exists between the third and fourth notes.

3.9.1 Primary inference

Primary inference concerns the determination of the maximum *a posteriori* mode sequence $M^*_{1:N}$ and the and the computation of the smoothed posteriors for note, tuning offset, and pitched/non-pitched reference amplitude (3.80 - 3.82). Figure 3.17 displays results for the piano example. The time domain waveform is displayed at the top of Figure 3.17. Vertical lines designate frame boundaries. In the section labeled "Onset", a black rectangle is displayed for each frame t for which $M^*_t \in \mathcal{O}$. The section labeled "Modes" displays for each $M \in \mathcal{M}$, a black rectangle for each frame in which $M^*_t = M$. The remaining sections, entitled "Notes", "Tunings", "Amps", and "NullAmps", display rectangles with sizes depending on the value of the posterior for each quantity and frame. (e.g., the "Tunings" section displays $P(T_t|M^*_{1:N}, Y_{1:N})$). Rectangle sizes vary logarithmically according to posterior probability, with the smallest visible rectangle corresponding to a probability of 0.03, and the largest, 1.0.

In this example, since visual inspection of the amplitude envelope essentially indicates the proper segmentation, it is easy to check that the sequence $M^*_{1:N}$ and hence the onset determination are valid. The note posteriors, $P(N_t|M^*_{1:N}, Y_{1:N})$ concentrate almost perfectly about the correct note values during pitched portion of note events, despite the occasional presence of the performer's voice, and the overlapping decays (significant overlaps are observed between the second and third notes, as well as between the third and fourth). The concentration of the note posterior extends beyond the ends of these regions, encompassing null and transient regions until the beginning of the pitched region for the following event. The latter observation is consistent with the reuse of N_t during non-pitched regions to "memorize" the previous note value, as discussed in Section 3.6.2. The tuning posterior, $P(T_t|M^*_{1:N}, Y_{1:N})$, seems remarkably consistent and only slightly influenced by note boundaries (the maximum of this posterior deviates ± 10 cents for the third and fifth notes). The (pitched) amplitude posterior, $P(A_t|M^*_{1:N}, Y_{1:N})$, indicates decaying envelopes as expected for piano.

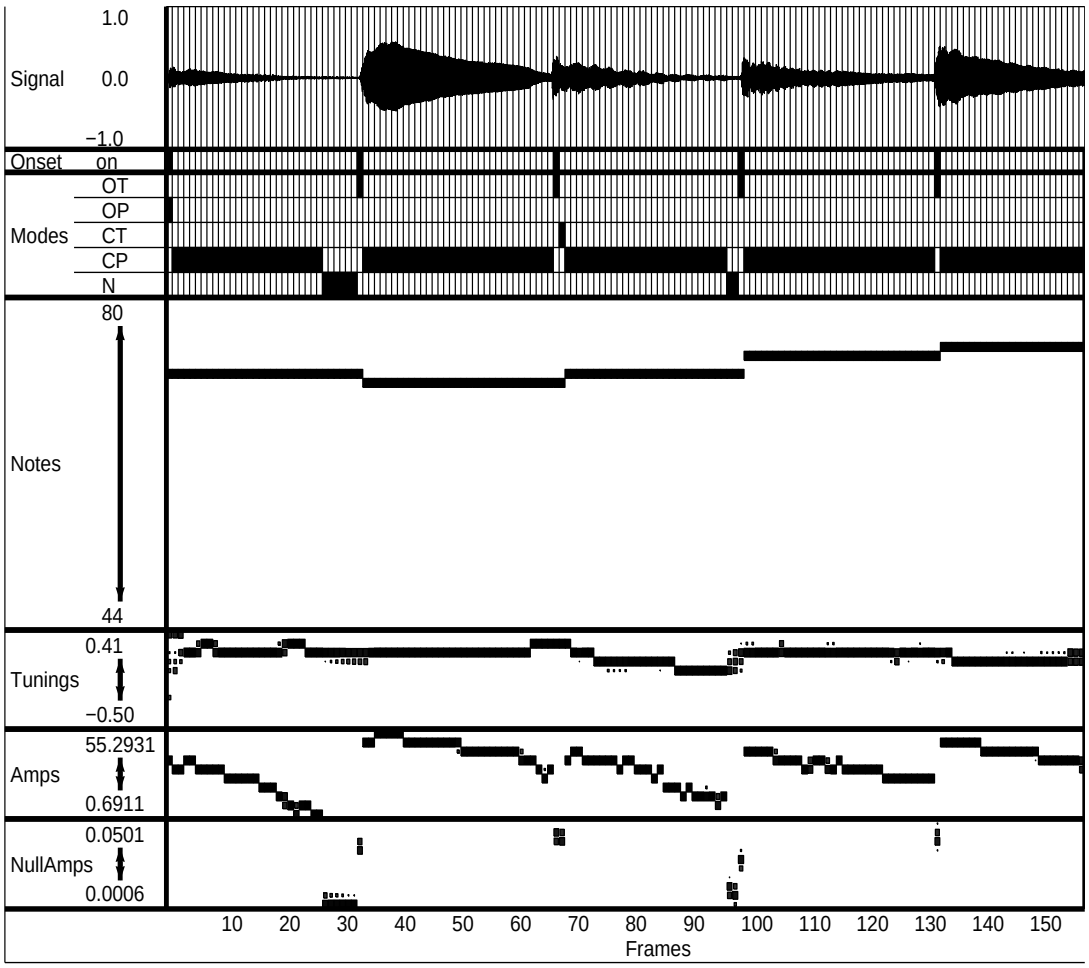


Figure 3.17: *Piano example: Introductory motive of Bach's Invention 2 in C minor (BWV 773), performed by Glenn Gould*

Corresponding results for the violin example are displayed in Figure 3.18. Since onsets are not directly visible by eye, checking the validity of the segmentation in the violin example becomes a painstaking process in which one must repeatedly splice and listen to various extracts, each time shifting the possible note boundaries. This process may take several minutes for each second of music. Further complications arise thanks the to shortness of the segments and the significant polyphony induced by reverberation and legato playing. Nevertheless, it seems difficult to improve on the resultant segmentation as indicated by $M^*_{1:N}$.

Through maximizing the note posterior, $P(N_t|M^*_{1:N}, Y_{1:N})$, all notes are correctly detected except for the initial "grace note" comprising the first three frames[11]. For these frames, the detected pitch is one octave below the notated pitch. The tuning posterior is less consistent than in the piano example, primarily due to portamento. We see that the overt slide between the third and fourth notes manifests in the drift of the maximized tuning posterior about the onset of the fourth note.

3.9.2 Estimation of mode transition dependence

The convergence of the EM iterations for estimating $P(M_{t+1}|M_t)$ is displayed in Figure 3.19 for the Poisson initialization and in Figure 3.20 for a uniform initialization. The latter is provided for purposes of comparison. These figures refer to the piano example generating the primary inference results of Figure 3.17. States labeled on the horizontal axis correspond to M_{t+1}; the vertical axis, M_t. Black rectangles are used to display the transition probabilities $P(M_{t+1}|M_t)$; the size of each rectangle varies logarithmically with the probability value according to the schema of the previous section (Figures 3.17 and 3.18). That is, the smallest visible rectangle corresponds to a transition probability of 0.03 while the largest corresponds to a probability of 1.0.

Recall that the Poisson initialization encodes knowledge of the "cyclic succession" schema (Figure 3.10), where a transient region of expected length N_T is followed by a pitched region of expected length N_P, followed by a null region of expected length N_N, and repeating until the end of the signal. Individual region lengths are modeled

[11]The note in question is not an actual grace note, but an artifact due to truncation of all but three frames from the previous note event.

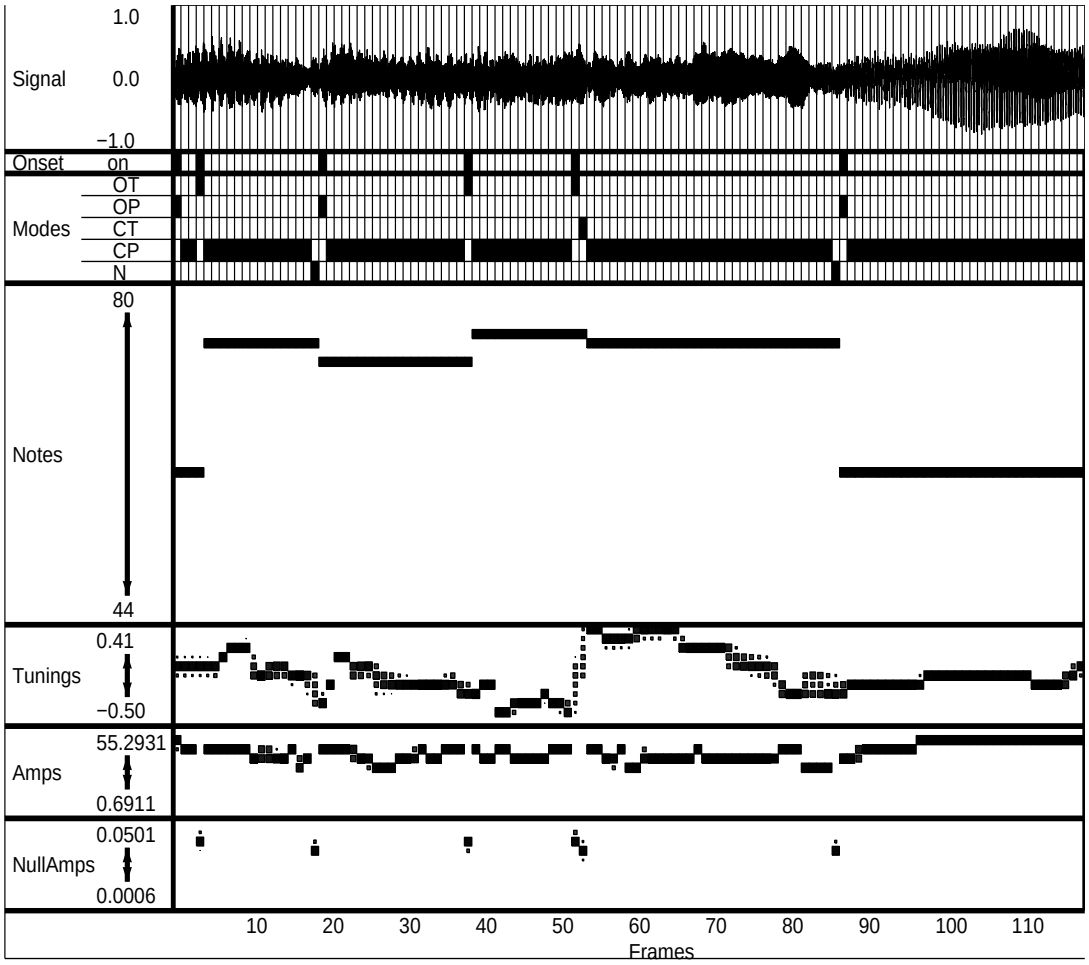
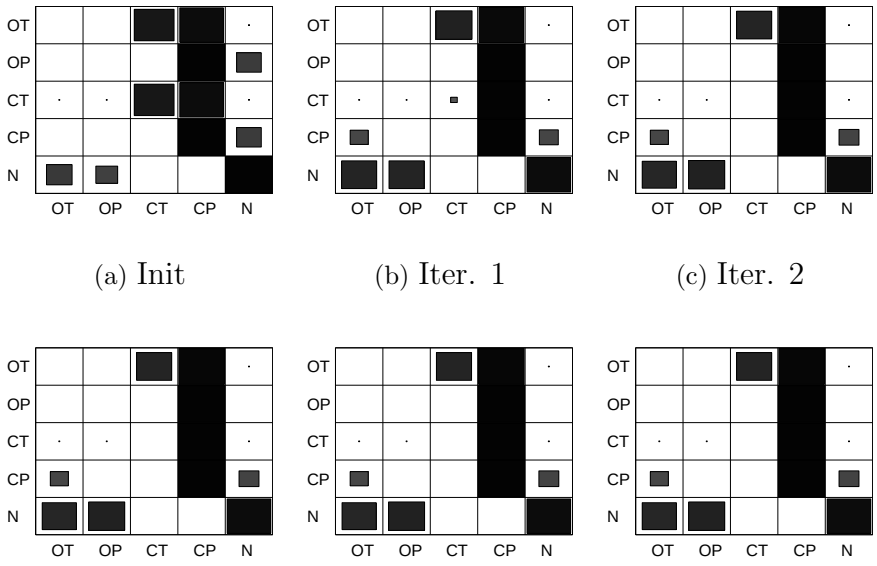


Figure 3.18: *Primary inference results on an excerpt from the third movement of Bach's solo violin Sonata No. 1 in G minor (BWV 1001), performed by Nathan Milstein*



(a) Init (b) Iter. 1 (c) Iter. 2

(d) Iter. 3 (e) Iter. 4 (f) Iter. 5

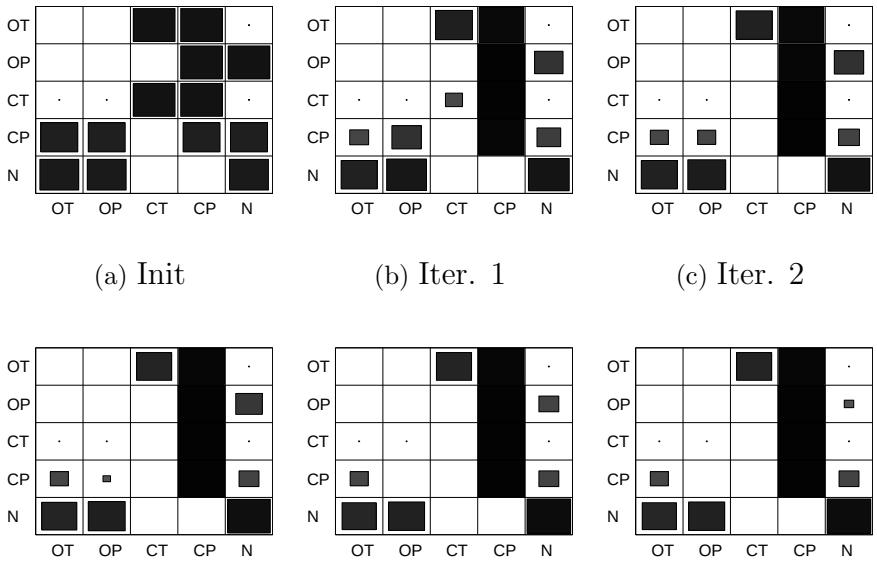
Figure 3.19: *EM convergence results beginning from Poisson initialization*

as independent, exponentially distributed random variables. We choose mean lengths of $N_T = 23$ ms, $N_P = 580$ ms, and $N_N = 348$ ms [12]. The EM convergence, however, appears relatively insensitive to modest variations in mean lengths.

The uniform initialization attempts to remove all prior knowledge from the specification of $\theta_M^{(0)}$, distributing transition probabilities in a conditionally uniform manner given $\mathcal{S}_j \in \mathcal{M}$ for each $j \in \mathcal{M}$ after accounting for spurious transitions. To represent the latter, let $\mathcal{S}_j^{spur}$ denote for each $j \in \mathcal{M}$, the collection of $k \in \mathcal{M}$ for which the $j \to k$ transition is spurious, meaning that it appears as a dotted line in Figure 3.15. Then, according to (3.83), we may represent $\theta_M^{(0)}$ as follows:

$$\theta_M^{(0)} \triangleq \text{Vec}\left(\bigcup_{j \in M} \bigcup_{k \in \mathcal{S}_j} \left\{p_{k|j}^{(0)}\right\}\right) \tag{3.93}$$

[12] With 44100 Hz sampling rate and 1024-sample frames, these settings are achieved with $N_T = 1$ frame, $N_P = 25$ frames, and $N_N = 15$ frames.



(a) Init (b) Iter. 1 (c) Iter. 2

(d) Iter. 3 (e) Iter. 4 (f) Iter. 5

Figure 3.20: *EM convergence results beginning from uniform initialization*

where $p^{(0)}_{k|j}$ is the initial value of $P(M_{t+1} = k|M_t = j)$ and $\mathcal{S}_j$ represents for each $j \in \mathcal{M}$, the collection of $k \in \mathcal{M}$ for which $p^{(0)}_{k|j}$ corresponds to the standard note evolution grammar as represented by solid lines in Figure 3.15. Clearly $\mathcal{S}_j \cap \mathcal{S}_j^{spur} = \emptyset \; \forall j \in \mathcal{M}$. Hence, for each $j \in \mathcal{M}$, the probability accorded to $\mathcal{S}_j$ is one minus the probability accorded to $\mathcal{S}_j^{spur}$. The latter is distributed uniformly among the elements of $\mathcal{S}_j$:

$$p^{(0)}_{k|j} \;=\; \frac{1 - \#(\mathcal{S}_j^{spur}) \cdot \epsilon}{\#(\mathcal{S}_j)} \quad \forall j \in \mathcal{M}, k \in \mathcal{S}_j \tag{3.94}$$

As Figures (3.19 - 3.20) indicate, the EM under uniform initialization takes at least five iterations to converge, while the Poisson initialization converges after only two iterations. This underscores the usefulness of the Poisson specification of $\theta^{(0)}_M$ even though the latter differs substantially from the maximum likelihood fit to the data as approximated by $\theta^{(i)}_M$ after convergence.

3.10 Conclusions and future work

As demonstrated in the previous section, the proposed method is able to accurately segment and extract melodies from framewise STFT peak data obtained from nominally monophonic recordings. The method not only identifies note boundaries, it yields a sub-segmentation of each note into transient and steady-state regions. The latter proves useful for the analysis-synthesis tasks of Chapter 2, for instance the time and pitch scaling discussed in Section 2.2, and the transient-specific processing introduced in Section 2.3. Additionally, the sub-segmentation finds application in adaptive window switching for audio transform coding [36]. Since transient regions generally contain broadband, rapidly time-varying spectral content, it is advantageous to use short windows to analyze these regions, because short windows sacrifice frequency resolution for increased time resolution. Shorter windows are used for transients and longer windows for the steady-state parts.

The proposed method proves robust to noise, interference, expressive pitch variations resulting from portamento and vibrato, and instances of polyphony resulting from background instrumentation, note overlaps due to legato playing, and reverberation. Robustness is aided by the *temporal integration* caused by modeling prior structural information at the signal level as captured by the "standard note evolution" grammar (3.47) as well as the expectation that pitch and amplitude characteristics will be consistent throughout pitched regions of note events. As a result, all frames within each pitched region are used to estimate signal characteristics (pitches and amplitudes) during any point in the region.

Furthermore, the proposed method facilitates the modeling of higher-level structural attributes via melodic expectations, and the *integration* of such attributes with the signal-level information. Melodic expectations are presently modeled via $P_{note\_trans}(N_1|N_0)$, the note-to-note transition distribution introduced in Section 3.6.2. Unfortunately, this first-order Markov dependence may not capture the majority of expectations which arise in practice, particularly in the context of Western tonal music [69]. Forthcoming work by Leistikow [71] considers the augmentation of the N_t-encoding to include past notes and intervals, as well as higher-level context (e.g.,

key, harmony, meter, harmonic rhythm, and beat position) so that these expectations may be represented using a first-order Markov dependence, which may be incorporated into the present method. Section 3.10.1 summarizes key features of this approach.

In addition to melody, another important high-level structural attribute is *rhythm*. Rhythm manifests in quasi-regular patterns of onset times (and other region boundary times) about which listeners develop *temporal expectations*. The latter take the form of a *hazard rate* influencing the distribution $P(M_{t+1}|M_t)$. Temporal expectations involving rhythmic structure may be modeled via *probabilistic phase locking networks*, as introduced in Section 3.10.1.

Subsequent work should also address several limitations of the proposed method. First, the method currently lacks the ability to encode actual instances of polyphony, so it cannot be used to transcribe recordings generated from polyphonic scores. In Section 3.10.3, a straightforward polyphonic extension is proposed. Second, the method's temporal resolution for determining segment boundaries is restricted by the frame rate (equivalent to the STFT hop size). For instance, the results of Figures 3.17 and 3.18 use a frame rate of 512 samples at 44.1 kHz (11.6 ms), which may not suffice for either analysis-synthesis or transform coding applications. Nevertheless, the frame-resolution output yields significant information about the location of onsets and other transient boundaries, as well as signal models before and after change. As discussed in Section 3.10.4, this information may be useful in subsequent sample accurate processing. In this section, we propose additional applications in interactive audio editing. For instance, given a nominally monophonic recording with overlapping notes, one may select all note events of a given type (such as the stable notes with respect to the current harmony), demix them from the recording, process them individually, and mix the processed versions in with the rest of the recording.

3.10.1 Modeling melodic expectations

Currently, one may integrate melodic expectations into the present system in the form of a stationary, first-order Markov dependence, $P_{note\_trans}(N_1|N_0)$, where N_1

represents the current note and N_0 the previous. However, this dependence proves insufficient to model even the simplest melodic expectations which arise in most musics, for instance Western tonal music.

Much recent work in the music cognition literature (cf. Krumhansl [64], Narmour [85], Schellenberg [101], and Larson and McAdams [69]) addresses the fundamental melodic expectations which arise in Western tonal music. Unfortunately, these expectations are given in *rule based*, instead of probabilistic, form. To be of use in the present system, these models must admit a stationary first-order Markov probabilistic representation with respect to some encoding which advances on a note-by-note basis. In forthcoming work, Leistikow [71] solves this issue by expanding the note representation to include additional state information which makes the Markov representation possible. Leistikow's representation augments the current note value with past notes and intervals, as well as higher-level contextual attirbutes, for instance, key, harmony, meter, harmonic rhythm, and beat position.

We may summarize Leistikow's representation as follows. Let X_k denote the augmented note representation, where k is the note index. The goal is to encode the melodic expectations as described in the aforementioned music cognition work in the form of $P(X_1|X_0)$ (which by assumed stationarity equals $P(X_k|X_{k-1})$ for any k). A common feature of the aforementioned music cognition work is that, given particular values of X_{k-1}, certain values of X_k (or subsets of such values) are expected to occur more frequently than others. This expectation does not *determine* $P(X_k|X_{k-1})$, inasmuch as it *constrains* it; the form of such constraints are clearly convex over the product space of simplices representing possibilities for $P(X_k|X_{k-1})$. For example, consider the "musical force" expectations introduced by Larson and McAdams [69], namely inertia, magnetism, and gravity.

- *Inertia* says that if a melody makes at least two stepwise transitions in a given direction (up or down), the following note will be more likely to continue that direction than reverse course.

- *Magnetism* involves the current key, which generates a collection of *stable* note values corresponding to the tonic major triad (i.e., if the key is $'$C Major$'$, the

'C', 'E', and 'G' notes constitute the stable notes for each octave). Magnetism says that if the choice is between a stepwise transition towards or away from the nearest stable note, the transition towards that note will be more likely; i.e., in the 'C Major' example, the transition 'F' $\rightarrow$ 'E' is more likely than 'F' $\rightarrow$ 'G'.

- Similarly, *gravity* says that unstable notes are more likely to descend then ascend along stepwise transitions.

Now, consider

$$X_k \triangleq \left\{ N_k^{(0)}, I_k^{(1)}, I_k^{(2)}, K_k \right\} \tag{3.95}$$

where $N_k^{(0)}$ is the current (k^{th}) note value, $I_k^{(1)} \in \mathcal{I}$ is a type designation for the interval leading up to $N_k^{(0)}$, $I_k^{(2)}$ is the preceding interval designation, and K_k is the key. The set $\mathcal{I}$ consists of five interval types:

$$\mathcal{I} = \{'\mathrm{R}', '\mathrm{SU}', '\mathrm{SD}', '\mathrm{LU}', '\mathrm{LD}'\} \tag{3.96}$$

where

- 'R' means *repeat*; the current and previous notes are identical
- 'SU' means *step up*; the current note is one or two semitones above the previous
- 'SD' means *step down*; the current note is one or two semitones below the previous
- 'LU' means *leap up*; the current note is at least three semitones above the previous
- 'LD' means *leap down*; the current note is at least three semitones below the previous

Since many elements of the pair $\{X_{k-1}, X_k\}$ are redundant (e.g., $I_k^{(2)} = I_{k-1}^{(1)}$) $P(X_k|X_{k-1})$ factors into $P(K_k|K_{k-1})$, $P(N_k^{(0)}|N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k)$, and a number of degenerate

(deterministic) distributions:

$$\begin{aligned} P(X_k|X_{k-1}) &= P(K_k|K_{k-1})P(N_k^{(0)}|N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k) \\ &\quad \times P(I_k^{(1)}|N_{k-1}^{(0)})P(I_k^{(2)}|I_{k-1}^{(1)}) \end{aligned} \tag{3.97}$$

Assuming $P(K_k|K_{k-1})$ is predetermined (the key may be considered constant for this example), the specification of $P(X_k|X_{k-1})$ reduces to the specification of $P(N_k^{(0)}|N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k)$. The latter models the expectation for the current note given the previous note, key, and preceding interval types.

Each of Larson and McAdams' tendencies (inertia, gravity, magnetism) may be represented as a linear constraint on a simplex representing possible values of $N_k^{(0)}$, for some fixed value of $N_{k-1}^{(0)}$, $I_{k-1}^{(1)}$, $I_{k-1}^{(2)}$, and K_k. Generally there exist many possible distributions satisfying these constraints. The distribution proposed by Leistikow [71] is that which effectively maximizes the number of pieces which are in some sense "close" to those generated by $P(X_k|X_{k-1})$, namely the distribution which maximizes the entropy rate of the note process $\{N_1^{(0)}, N_2^{(0)}, \ldots\}$ (considering $P(K_k|K_{k-1})$ as fixed). Let $\pi(N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k)$ represent a certain prior distribution; define the functional $J(P, \pi)$ as follows:

$$J(P, \pi) = E_{\pi(N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k)} \left[\log P(N_k^{(0)}|N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k) \right] \tag{3.98}$$

For fixed $\pi(N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k)$, it is easily shown that $J(P, \pi)$ is concave on the product of simplices representing free parameters in $P(N_k^{(0)}|N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k)$. Since the Larson-McAdams constraints are convex on this space, the determination of P maximizing $J(P, \pi)$ is a convex problem, and can be solved by one of the many available software packages, for instance PDCO (primal-dual method for convex objectives) [100].

The entropy rate of the note process equals constant terms plus $J(P, \mu)$, where $\mu(N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k)$ is the corresponding component of the stationary distribution for $P(X_k|X_{k-1})$, this of course assuming the constraints allow this stationary distribution to exist. Unfortunately, $P(X_k|X_{k-1})$ depends on $P(N_k^{(0)}|N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k)$;

this implicit circularity fails to guarantee convexity for the entropy rate maximization. Preliminary studies, however, show excellent results in practice using an iterative approach: First $\pi(N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k)$ is initialized as uniform. Then $P(N_k^{(0)}|N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k)$ is chosen to maximize $J(P, \pi)$ as in (3.98). Subsequently a new π is chosen by solving for the stationary distribution of $P(X_k|X_{k-1})$ given $P(N_k^{(0)}|N_{k-1}^{(0)}, I_{k-1}^{(1)}, I_{k-1}^{(2)}, K_k)$. The latter two steps repeat until convergence.

So far, stationary first-order Markov models of musical expectations have been developed in [71] by translating rule-based constructs from the music cognition literature into probability distributions via entropy-rate maximization. The latter is solvable as a sequence of convex optimization problems. A remaining task is to incorporate higher-level constructs involving meter and beat position, as metrical accents greatly influence melodic expectations [73, 45, 12] , Another primary task is to integrate these melodic expectation models with the present signal-level models for melody extraction and segmentation, and test the result on a representative corpus.

3.10.2 Modeling temporal expectations from rhythm via probabilistic phase locking networks

The presence of *rhythm* in most musics guarantees some regularity as to the locations of segment boundaries, especially note onsets. Such regularity allows one to accurately predict where the next boundary will occur. We call the associated predictive distributions *temporal expectations*, analogously to the melodic expectations discussed above. Temporal expectations may be combined with subsequent signal observations to improve the segment boundary detection. The end result is a fully integrated Bayesian framework for joint tempo tracking and onset detection which operates directly on signal observations. Presently, recent literature on audio-based tempo tracking [49, 51, 18, 65] seems to consider onset detection separately from the tempo induction, as the latter uses already detected onsets as observations. Such decoupled approaches make it difficult for temporal expectations associated with the tempo hypothesis to inform the onset detection, as is possible with an integrated Bayesian approach.

The simplest possible scenario concerns an unknown quasi-periodic structure, which can be thought of as a jittery metronome with additional, smooth tempo drift. To illustrate how temporal expectations are encoded by this structure, we may consider once again the legato model of Section 3.3.3, where $M_t \in \{'\mathrm{O}', '\mathrm{C}'\}$; $'\mathrm{O}'$ indicating onset (abrupt change), and $'\mathrm{C}'$ indicating continuation of the underlying state quantity S_t, for which Y_t constitutes a noisy observation. Two additional hidden variables interact with the M_t-layer: T_t, the unknown tempo (representing the inverse of the number of frames between onsets), and τ_t, the elapsed duration since the last onset. To allow for quasi-perodicity rather than strict periodicity, T_t may vary smoothly across frames; we expect the actual onset interarrival times to have additional jitter.

While the elapsed duration is a function of the onset incidence, and the onset incidence is influenced by the elapsed duration, this circularity does not cause problems in practice. As such, the dependences among the aforementioned variables may be encoded in a directed acyclic graph, as shown in Figure 3.21. We call the proposed structure a *probabilistic phase locking network*.

We address each dependence not already discussed in Section 3.3.3 as follows.

- $P(T_{t+1}|T_t)$ models tempo drift. Following [21], $\log T_{t+1}$ equals $\log T_t$ plus Gaussian noise.

- $P(\tau_{t+1}|\tau_t, M_{t+1})$ models the elapsed duration since the previous onset. This distribution is deterministic, modeling a counter which resets upon $M_{t+1} = '\mathrm{O}'$. Onset locations are considered quantized to the beginning of the frame; more precisely, the event $M_t = '\mathrm{O}'$ corresponds to the event that an onset occurs in $[t - 1/2, t + 1/2)$ where t is measured in frames. Elapsed durations (as a matter of definition) are measured from the end of the frame. Hence, $P(\tau_{t+1}|\tau_t, M_{t+1} = '\mathrm{O}')$ concentrates on $\tau_{t+1} = 1$ (reset); $P(\tau_{t+1}|\tau_t, M_{t+1} = '\mathrm{C}')$ concentrates on $\tau_{t+1} = \tau_t + 1$ (increment).

- The *temporal expectation* $P(M_{t+1}|\tau_t, T_{t+1}, M_t)$ models the probability that an onset is assigned to frame $t+1$ given elapsed duration and tempo; for this simple example M_t may be dropped from the conditioning if we allow onsets to occur

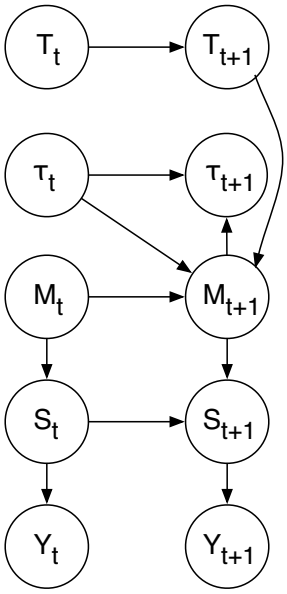


Figure 3.21: *Probabilistic phase locking network for modeling quasi-periodic stream of abrupt-change events*

in adjacent frames. Now, let Z represent the random interarrival time between successive onsets. Given T_{t+1}, we expect $\log Z$ to equal the logarithm of the tempo period, which is $\log 1/T_{t+1}$, plus Gaussian noise. The pdf of O is hence completely specified given T_{t+1} and we represent it as such: $p_Z(z|T_{t+1})$.

Given additionally τ_t, the probability that $M_{t+1} = \text{'O'}$ is equal to the event that $Z < \tau_t + 1/2$. Since τ_t is observed we know that $Z \geq \tau_t - 1/2$, so $P(M_{t+1} = \text{'O'}|\tau_t, T_{t+1}, M_t)$ is equal to the *hazard rate*, or $P(Z < \tau_t + 1/2|O \geq \tau_t - 1/2)$:

$$P(M_{t+1} = \text{'O'}|\tau_t, T_{t+1}, M_t) \quad = \quad \frac{\int_{\tau_t - 1/2}^{\tau_t + 1/2} p_Z(z|T_{t+1})dz}{1 - \int_0^{\tau_t - 1/2} p_Z(z|T_{t+1})dz} \tag{3.99}$$

In standard, nominally monophonic musical examples, the expected note durations are not constant. They still relate to each other through the tempo period via *notated* durations; e.g., quarter notes, eighth notes, triplets etc. We represent this situation by introducing two additional variables. The anticipated duration between onsets

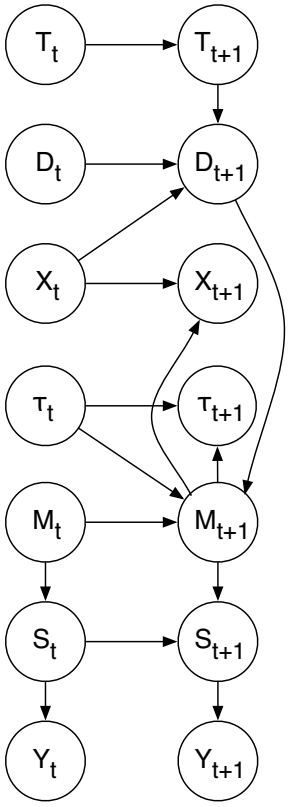


Figure 3.22: *Probabilistic phase-locking network for nominally monophonic temporal expectation model*

is measured by $1/D_t$ in units of frames, while X_t represents the current notated duration; X_t is permitted to change only upon the onset of a new note, $M_t = \text{'O'}$. One may compute D_t by knowing the frame rate, T_t, and X_{t-1}.

Figure 3.22 represents the nominally monophonic temporal expectation model. Additional dependences represented in this figure which have not been previously introduced are as follows.

- $P(M_{t+1} = \text{'O'}|\tau_t, D_{t+1}, M_t)$ is evaluated using (3.99) with D_{t+1} in place of T_{t+1}.
- $P(D_t|T_t, X_t)$ is deterministic. If the tempo is such that $1/T_t$ represents the expected onset interarrival time in frames between quarter notes, the period $1/D_t$ may be adjusted according to the notated duration represented by X_t.

- $P(X_{t+1}|X_t, M_{t+1} = \text{'C'})$ concentrates on X_t because the notated duration cannot change until a new onset appears and we move on to the next note. Then $P(X_{t+1}|X_t, M_{t+1} = \text{'O'})$ describes the anticipated distribution of notated durations for the next note. Of course, it is difficult to model this distribution without augmenting X_t to include metrical information (meter and beat position); in the meantime we choose a uniform distribution among the available possibilities for X_{t+1} until an improved solution may be found. Alternatively, $P(X_{t+1}|X_t, M_{t+1} = \text{'O'})$ may be adapted to a representative corpus using EM.

The proposed modeling of temporal expectations seems promising for two reasons: first, it enables the *joint* tracking of tempo and note onsets directly from audio signal observations, rather than tracking tempo from onset information alone, as is done in recent literature [49, 51, 18, 65]; moreover, the probabilistic modeling of temporal expectations may be of interest in music cognition research, because these expectations explicitly encode the anticipation that an event is about to occur. For instance, we may investigate how to create expectations which are continually deferred, to build up tension.

A major challenge which has not been so far addressed is the adaptation of the temporal expectation models in Figures 3.21 and 3.22 to account for transient information. By so doing, M_t can take on the full set of possibilities in $\mathcal{M}$, which enables the temporal expectation models to be integrated into the current melody extraction and segmentation method to further improve our abilities to detect all types of region boundaries, not just onsets. A further task is to expand the encoding of X_t to include metrical information as the latter also influences melodic expectations (i.e., given a particular harmony, it is more likely for stable notes with respect to this harmony to occur on downbeats than other beat positions [73]). Hence, we may assess the *interaction* between melodic and rhythmic tendencies through different types of accentual patterns involving the meter. This in turn will allow the tracking of melodic patterns to improve our abilities to track rhythmic patterns and vice versa, as both are influenced by metrical information.

3.10.3 Polyphonic extensions

The extension to polyphony is conceptually straightforward. Unfortunately, the inference methods of Section 3.7 encounter computational difficulties due to the combinatorial explosion in the number of aggregate mode possibilities. To see this, let the range of note values $\mathcal{N} = N_{min} : N_{max}$, and for each $N \in \mathcal{N}$, assign hidden variables $M_t^{(N)}$ and $S_t^{(N)}$ to model the mode and state information for that particular note[13]. The number of possibilities for the aggregate mode variable, $M_t^{(N_{min}:N_{max})}$, grows exponentially with N. Both the primary inference, discussed in Section 3.7.1, and the EM algorithm for estimating free parameters in the mode transition dependence (Section 3.7.2) yield computational costs which are quadratic in the number of possibilities for M_t. This quadratic cost arises from various quantities propagated in primary inference and EM recursions which involve both M_t and M_{t+1}; i.e., $\tau\left(M_t, S_t, M_{t+1}\right)$ and $\mu\left(M_t, S_t, M_{t+1}\right)$ in (3.77), and $\sigma^{(2)}(M_t, M_{t+1})$ in (3.84) and (3.85).

Further difficulties arise in the modeling of $P(Y_t|S_t^{(N_{min}:N_{max})})$, as the methods discussed in Chapter 4 concern only single-pitch hypotheses. These methods must be extended to the case of multiple pitches. Such an extension has already been developed by Leistikow *et al.* and successfully applied in the context of Bayesian chord recognition from single-frame STFT data [72]. Hence, the primary difficulty in the polyphonic extension remains the computability of the inference. We believe this difficulty may be resolved by sequential Monte Carlo (SMC) techniques, in particular adaptations of the *particle filter* [38, 32].

To assess the applicability of SMC, we consider the expected concentration of the posterior quantities which are actually propagated in the inference. These quantities consist of various marginalizations of filtered or smoothed posteriors evaluated over adjacent pairs of states; i.e.:

$$P\left(M_t^{(N_{min}:N_{max})}, S_t^{(N_{min}:N_{max})}, M_{t+1}^{(N_{min}:N_{max})}, S_{t+1}^{(N_{min}:N_{max})}|Y_{1:t}\right)$$

$$P\left(M_t^{(N_{min}:N_{max})}, S_t^{(N_{min}:N_{max})}, M_{t+1}^{(N_{min}:N_{max})}, S_{t+1}^{(N_{min}:N_{max})}|Y_{1:N}\right)$$

[13]Here the state information can be reduced because the N_t-component of $S_t^{(N)}$ equals N.

While the nominal space of joint possibilities for $M_t^{(N_{min}:N_{max})}$, $S_t^{(N_{min}:N_{max})}$, $M_{t+1}^{(N_{min}:N_{max})}$, and $S_{t+1}^{(N_{min}:N_{max})}$ remains exponentially large in the size of the note range, the vast majority of such possibilities, *given* adequate signal observations, are expected to have negligible probability. For instance, usually we can expect only a limited number of notes to be sounding at any given time, which means for most $N \in \mathcal{N}$, $M_t^{(N)}$ concentrates on $'\mathrm{N}'$. Furthermore, rhythmic structure indicates that onsets and the locations of transient regions will be highly synchronized, and harmonic structure indicates that all but a few note combinations are likely to occur simultaneously. As such, it is plausible that these posterior distributions may be well-represented[14] by a reasonably-sized collection of weighted particles, each particle corresponding one of the joint possibilities with non-negligible posterior probability. A byproduct is that since there is effectively no limit on the *nominal* size of the space for $S_t^{(N_{min}:N_{max})}$, we may forego the discretization of the remaining state quantities (amplitudes, tuning offsets) altogether, since these quantities are naturally continuous-valued. Rather than treating amplitudes and tuning offsets as nuisance parameters, as is done formally in the postprocessing stage, we can extract more meaningful information from the posteriors $P(A_t^{(N_{min}:N_{max})}|M_{1:N}^{(N_{min}:N_{max})}, Y_{1:N})$ and $P(T_t^{(N_{min}:N_{max})}|M_{1:N}^{(N_{min}:N_{max})}, Y_{1:N})$ in order to track expressive pitch and amplitude variations, following the interpretation of Figure 3.18 given in Section 3.9.1.

The proposed approximate inference strategies using SMC are presently under development. This development proceeds in three stages: first, we replicate the present nominally monophonic model in order to test the SMC approximation; second, we eliminate the discretization of tuning offsets and amplitudes in the monophonic case; third, we complete the polyphonic extension.

3.10.4 Interactive audio editing

In recording applications, it is common that individual instruments or small groups of instruments are recorded on separate tracks. The majority of tracks hence satisfy

[14]The idea of "well-representation" means at the very least that, as the number of particles becomes sufficiently large, the weighted sample average converges to the posterior mean; see, e.g. [27] for more rigorous definitions and convergence properties.

the nominally monophonic assumption. Often, prior to mixing, individual tracks are edited to correct timing or intonation[15] or to generate other, creative transformations of the sound material. Using the standard "waveform only" visual representation, the editing process may become quite time consuming, as sections of the recording must be spliced by hand and repeatedly listened to in order to discern note boundaries and regions containing pitch information.

To this end, the present method generates a map of detected onsets, transient and pitched regions, and note values, as well as approximate trajectories for tuning offset and amplitude envelope information for each individual track[16], as long as the latter satisfies the nominally monophonic assumption. This map may be displayed in conjunction with, or overlaid on top of, the waveform representation. We expect that the combined representation will facilitate the editing process, as the time consuming detection problems become automated. Furthermore, making the map *editable* opens up new creative possibilities: one can slide note regions around with the mouse, modifying time and pitch information; one can also select certain types of notes or note regions and apply specific processing to just these regions. For instance, dynamic range modifications (e.g., compression or expansion) may be applied to just transient regions in order to sharpen attacks and increase the track's presence in the mix without changing its volume. Another example concerns the "correction" of a violinist's intonation, taking care that the end result does not destroy expressive qualities. If the current harmony is known, one may correct the intonation of just the stable pitches while leaving other notes unprocessed. This might make the performer sound more "in tune" with the rest of the ensemble while preserving more nuanced performance characteristics which prove otherwise difficult to model [4].

Of course, to implement such region-dependent changes, due to the possibility of overlapping notes it becomes necessary to demix these notes, extract them individually, apply transformations as desired, then reconstitute the results. Unfortunately, the segmentation's temporal resolution is limited to the frame rate; subsequent sample-accurate segmentation may be required. Fortunately, the present, frame-based

[15]Time and pitch corrections are especially common in vocal recordings.

[16]Imagine the posterior plot shown in Figure 3.18, but with segment boundaries and note regions clearly delineated according to the postprocessing discussed in Section 3.8.

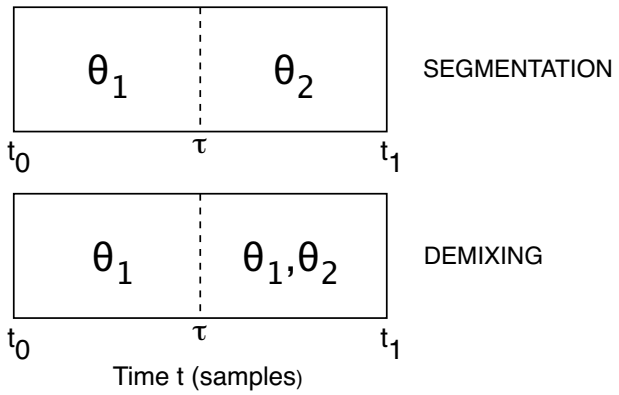


Figure 3.23: *Schematics for sample accurate segmentation and demixing of overlapping audio sources*

method greatly facilitates subsequent sample-accurate processing: it restricts boundaries to frame neighborhoods as well as yields approximate information about possible signal models immediately before and after these boundaries.

Given appropriate signal models, both sample-accurate segmentation and demixing may be performed by maximum-likelihood estimation as described in [61]. Figure 3.23 displays schematics for both the standard segmentation problem and the demixing problem. This figure represents the situation where it is known that exactly one segment boundary occurs at $t = \tau$, where $\tau \in t_0 : t_1$ and t, τ are measured in *samples*. The joint distribution of these samples, $y_{t_0:t_1}$, may be modeled:

$$P(y_{t_0:t_1}|\theta_1, \theta_2, \tau) = \prod_{t=t_0}^{\tau-1} P(y_t|y_{1:t-1}, \theta_1) \prod_{t=\tau}^{t_1} P(y_t|y_{1:t-1}, \theta_1, \theta_2), \qquad (3.100)$$

Here $\prod_{t=t_0}^{\tau-1} P(y_t|y_{1:t-1}, \theta_1)$ represents the *signal model* before the segment boundary parameterized by θ_1. This parameter may encode pitch and amplitude characteristics, as well as the type of model (transient, pitched, and so forth). The signal model for the region after the segment boundary is given by $P(y_t|y_{1:t-1}, \theta_1, \theta_2)$; where θ_2 encodes the new information present at time τ. The estimated segment boundary location, $\hat{\tau}$,

is the maximum-likelihood estimate:

$$\hat{\tau} \;=\; \underset{\tau \in t_0:t_1}{\operatorname{argmax}} \; P(y_{t_0:t_1} | \theta_1, \theta_2, \tau) \tag{3.101}$$

Usually, these signal models may be represented as stationary Markov processes; in many cases (e.g., autoregressive models, or the linear Gaussian state space models for sinusoidal parameter estimation used by Cemgil *et. al.* [20, 18] for sample accurate processing), each likelihood update as τ increments can be performed in constant time, making the overall computational complexity of the likelihood maximization linear in the region length. It is hoped that the identified frame neighborhood regions are short enough to make such complexity issues irrelevant.

Chapter 4

Evaluating pitch content hypotheses

4.1 Introduction

We begin by summarizing briefly several goals stated in Chapter 3 concerning the evaluation of pitch content hypotheses with respect to STFT peak observations. Let Y_t denote the collection of peaks for the t^{th} frame; we may represent $Y_t = \{F, A\}$, where:

$$\begin{aligned} F &\triangleq \{F(1), F(2), \ldots, F(N_o)\} \\ A &\triangleq \{A(1), A(2), \ldots, A(N_o)\} \end{aligned} \tag{4.1}$$

where $F(k)$ denotes the frequency of the k^{th} lowest-frequency STFT peak, $A(k)$ the corresponding amplitude, and N_o the number of observed peaks. The primary goal consists of evaluating the likelihood of Y_t with respect to both pitch and non-pitch hypotheses, as there is no guarantee that the underlying signal contains significant pitch content, for instance during transient regions. The pitch hypothesis likelihood is denoted as $P(Y_t|N_t, T_t, A_t)$, where

- N_t is an integer representing the note semitone value; i.e., $N_t = 60$ corresponds to the note $'\mathrm{C4}'$.

- $T_t \in [0.5, 0.5)$ is a fractional tuning offset, representing the deviation from N_t in semitones.

- A_t is a *reference amplitude*, in the sense that amplification of the input signal by some constant causes a proportionate change in A_t.

The non-pitch hypothesis likelihood is denoted as $P(Y_t|A_t^Q)$, where A_t^Q is a reference amplitude for the overall signal level[1].

The proposed model for pitch hypotheses actually subsumes the model for non-pitch hypotheses, because the former explicitly accounts for *spurious* peaks which arise from signal content unrelated to the pitch hypothesis, for instance noise, interference, and other non-pitched signals. Hence, the evaluation of $P(Y_t|A_t^Q)$ may proceed using the evaluation for pitch hypotheses under the constraint that all peaks are spurious.

4.2 The proposed model

The proposed model makes use of a harmonic template to govern the distribution of spectral peak frequencies, inspired by the approach of Goldstein [47]. However, many cases exist where there is prior information concerning timbre, resulting from full or partial knowledge of the instruments used in the recording. Consequently, the proposed template involves spectral peak amplitudes as well as frequencies to exploit knowledge of timbre in the disambiguation of pitch determinations. For instance, if it is known *a priori* that a certain instrument's timbre emphasizes even harmonics, it will be considerably less likely that the second harmonic is mistaken for the fundamental in assigning pitch values to recordings using that instrument.

Another deviation from Goldstein's template-based model is the explicit accounting for *spurious* peaks. The latter are peaks observed in the STFT which do not arise

[1]Currently, there are no efforts to model signal characteristics for non-pitch hypotheses beyond the reference amplitude. Subsequent revisions may focus on characterizing the spectral envelope in terms of psychoacoustically relevant features, for instance mel frequency cepstral coefficients (MFCC's). The latter have been demonstrated quite useful in the perceptual discrimination of timbre [114], as well as a variety of musical information retrieval tasks which exploit timbral characteristics of non-pitched sounds [24, 43].

from sinusoidal components indicated in the template. Spurious peaks arise primarily from noise or background instrumentation. Under low noise conditions, sidelobes may cause spurious detections, although the latter behavior is rare due to the thresholding used in preprocessing stages.

Furthermore, the proposed model accounts for *missing* or undetected peaks. These peaks exist in the template, but are not observed in the STFT. Three common causes of missing peaks are as follows: the designated sinusoidal component may fall below the preprocessing threshold; it may be of such low amplitude as to dissappear below the noise floor, or be absent entirely from the input signal (e.g., clarinet sounds are generally missing even harmonics); it may collide with neighboring peaks and hence fail to be resolved.

4.2.1 Preprocessing

Issues surrounding spurious and missing peaks are clarified by considering the algorithm's preprocessing stages. The goal of preprocessing is to take a signal frame and extract from it a peaklist Y_t. Figure 4.1 shows the preprocessing stage for a frame hopped every $T/2$ samples.

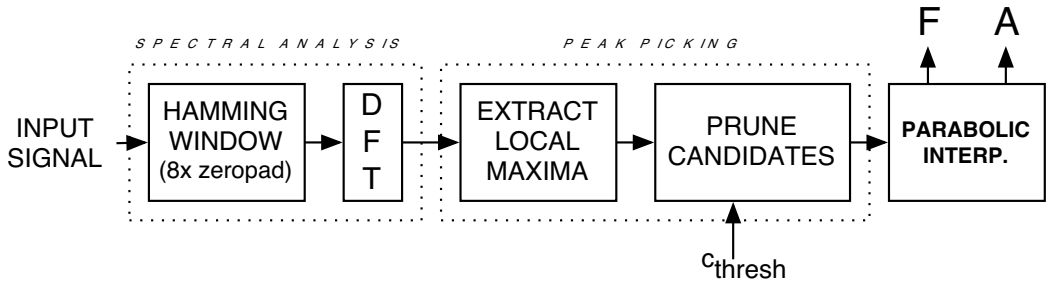


Figure 4.1: *Preprocessing steps for pitch likelihood evaluation*

In the spectral analysis stage, a section of length T is extracted and multiplied by a Hamming window of the same length, then zeropadded by a factor of eight. A discrete Fourier transform (DFT) of length $8T$ is taken. All local maxima of the DFT magnitude are first considered as candidate peaks; accepted candidates must satisfy the following:

- The local maximum's magnitude must exceed all DFT magnitudes up to eight bins to the right and left of the maximum's bin position. The eight-bin distance represents half of the Hamming window's mainlobe width under the specified zeropadding factor.

- The maximum's magnitude must be no greater than c_{thresh} times that of the global maximum. Usually $c_{thresh} \in [0.01, 0.1]$.

Let $X_t[k]$ denote the DFT evaluated at bin k. Suppose at bin k^*, an accepted candidate is found. The log magnitudes of DFT bins $X_t[k^* - 1]$, $X_t[k^*]$, and $X_t[k^* + 1]$ are presented to a parabolic interpolation algorithm following the PARSHL approach of Smith and Serra [110].

The fitted parabola approximates the log magnitude discrete time Fourier transform (DTFT) of the input frame about the peak position; i.e., if $X_t(\omega)$ represents the DTFT of the windowed frame:

$$\log |X_t(\omega)| \approx A - B(\omega - C)^2 \tag{4.2}$$

where the approximation is valid for $\omega \approx 2\pi k^*/T$ (the latter is the radian frequency corresponding to bin k^*). From (4.2), the peak frequency is estimated as C, and the amplitude is estimated as e^A.

4.2.2 The harmonic template

In the ideal case, in the absence of noise and other uncertainties, the harmonic template describes peaks with frequencies in a harmonic series out to the Nyquist limit. If N_i denotes the number of template peaks, and f_0 the hypothesized fundamental pitch, ideally $N_i = \lceil \pi/f_0 \rceil$. Since most acoustic instruments have decreasing spectral energies of their harmonic portions beyond some critical frequency, in practice the number of template peaks is chosen between three and seven, unless doing so generates template peak frequencies beyond the Nyquist limit.

The ideal frequency of the k^{th} template peak is kf_0; the ideal amplitude follows the *timbral hypothesis* $A_0 \cdot T(k)$, the latter arising from knowledge of the instrument.

Here the role of A_0 as a *reference amplitude* becomes apparent: if the input signal is multiplied by a fixed constant, A_0 will be multiplied by this constant. In practice, A_0 is chosen such that $\text{argmax}_{k \in 1:N_i} \; T(k) = 1$. One may interpret A_0 as the maximum template peak amplitude in the absence of noise.

Of course, the ideal template as stated represents only *deterministic* effects. To be robust to the variety of real-world signals perceived as pitched, we must account for variations in the anticipated signal characteristics due to different source instruments, recording conditions, interference, and the suboptimality of preprocessing stages. Interference from spurious events may perturb both peak frequencies and amplitudes; additional deviations may result from the imperfect nature of the finite zeropadding and quadratic DFT interpolation used in preprocessing, though we expect the latter to be insignificant[2]. There may also exist uncertainty concerning the harmonic structure. Many sounds with perceptible pitch content contain significant amounts of inharmonicity, such as piano and marimba. As the instrument may be unknown *a priori*, it becomes important to account for some acceptable range of variation in the harmonic structure. Likewise, the amplitudes of various harmonics may deviate from the timbral hypothesis $T(k)$. Even if the instrument is known, recording circumstances (equalization, distortion, etc.) may cause significant deviations from the anticipated spectral envelope.

Hence, the harmonic template is represented *probabilistically*, as a set of joint distributions over frequencies and amplitudes, one joint distribution for each template peak. That is, if $F(j)$ and $A(j)$ constitute, respectively, the frequency and amplitude of the j^{th} observed STFT peak, the latter corresponding to the k^{th} template peak, the variation of $F(j)$ and $A(j)$ is encoded by the conditional distribution $P_k(F(j), A(j)|f_0, A_0)$.

[2]We define the "perfect" preprocessing stage as that which for a single sinusoid with frequency ω embedded in additive white Gaussian noise, estimates ω with zero bias and minimum variance. It is well known [97] that the maximum-likelihood frequency estimate is the frequency of the DTFT magnitude peak. The latter becomes asymptotically unbiased and minimum variance as the number of samples tends to infinity, achieving the Cramer-Rao lower variance bound. The preprocessing and peak extraction (Figure 4.1) closely approximates the maximum-likelihood estimate, as discussed in Section 4.3.3.

The *harmonic template*, then, represents the *collection* of such distributions:

$$\text{TEMPLATE} \quad = \quad \{P_k\left(F(j), A(j)|f_0, A_0\right)\}_{k=1}^{N_i} \tag{4.3}$$

The precise formulation of template distributions is discussed in Section 4.3.3. Template indices are sorted in terms of increasing frequency mean; i.e.:

$$E_{k+1}\left(F(i)|f_0, A_0\right) \quad \geq \quad E_k\left(F(j)|f_0, A_0\right) \quad \forall k \in 1\!:\!N_i \tag{4.4}$$

where E_k denotes the expectation under P_k, $F(i)$ is the frequency of the observed peak originating from template peak $k+1$, and $F(j)$ is the observed frequency originating from template peak k.

An additional consideration is the *peak non-interaction hypothesis*. In the absence of spurious or missing peaks; i.e., if every observed peak corresponds to exactly one template peak, non-interaction stipulates that the observed peak likelihood, $P(F, A|f_0, A_0)$, factors as a product distribution over individual template distributions:

$$P(F, A|f_0, A_0) \quad = \quad \prod_{k=1}^{N_o} P_k(F(k), A(k)|f_0, A_0) \tag{4.5}$$

The non-interaction hypothesis says, effectively, that neighboring template peaks exert no influence on an observed peak, *given* its correspondent. This hypothesis merits criticism in the following sense: if template peaks are sufficiently close in frequency that they approach the STFT's resolution limit, neighboring components will clearly bias observed frequencies and amplitudes corresponding to the given component. Nevertheless, as tolerance for such interferences is already encoded in the distribution $P_k(F(k), A(k)|f_0, A_0)$, the avoidance of an explicit encoding of peak interactions seems not to cause problems in practice.

4.2.3 Representing the linkage between template and observed peaks

Real-world signals generally lead to spurious detections (observed peaks which have no correspondent in the template) and missing peaks (template peaks which are unobserved). Without knowing which observed peaks map to which template peaks, it becomes difficult to evaluate the overall peak likelihood via template distributions. Additionally we encode the distribution for the possibility that the observed peak with frequency $F(k)$ and amplitude $A(k)$ is spurious, as $P_{\prime \mathrm{S}\prime}(F(j), A(j)|f_0, A_0)$.

The correspondence between observed and template peaks (plus the spurious possibility) is encoded via the *linkmap* $L : \mathcal{J}_o \rightarrow \mathcal{J}_i$ where $\mathcal{J}_o \stackrel{\Delta}{=} 1 : N_o$ denotes the set of observed peak indices, $\mathcal{J}_i \stackrel{\Delta}{=} 1 : N_i \cup \prime\mathrm{S}\prime$ denotes the set of template peak indices plus $\prime\mathrm{S}\prime$, which is the spurious possibility. In other words, if j is the index of an observed peak; $L(j)$ returns the index of the corresponding input peak, except when $L(j) = \prime\mathrm{S}\prime$, which means the j^{th} observed peak is spurious. Figure 4.2 illustrates an example linkmap where $L(1) = 1$, $L(2) = 2$, $L(3) = \prime\mathrm{S}\prime$, $L(4) = \prime\mathrm{S}\prime$, and $L(5) = 4$. In the figure, template peaks are shown as circles and observed peaks as "X's". Frequencies (or mean frequencies in the case of template peaks) are represented by the horizontal position of each peak symbol; amplitudes (mean amplitudes in the case of template peaks) by the peak symbol's relative size.

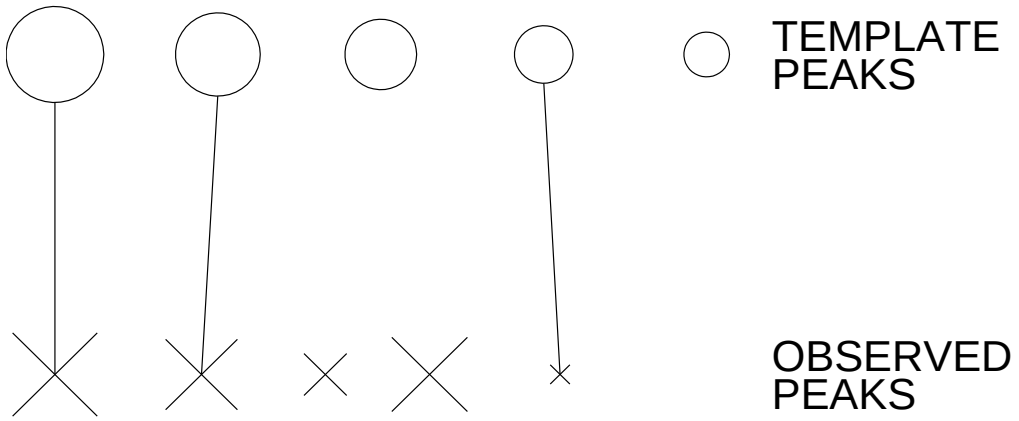


Figure 4.2: *Example linkmap*

Given linkmap L, the STFT peaks' likelihood factors as a product distribution over individual template distributions:

$$P(F, A|L, f_0, A_0) = \prod_{j=1}^{N_o} P_{L(j)}(F(j), A(j)|f_0, A_0) \tag{4.6}$$

Since L is unknown, we marginalize with respect to a *prior* $P(L)$:

$$\begin{aligned} P(F, A|f_0, A_0) &= \sum_{L\in\mathcal{L}} P(L)P(F, A|L, f_0, A_0) \\ &= \sum_{L\in\mathcal{L}} P(L)\prod_{j=1}^{N_o} P_{L(j)}(F(j), A(j)|f_0, A_0) \end{aligned} \tag{4.7}$$

where $\mathcal{L}$ denotes the set of *valid* linkmaps.

A linkmap is *valid* if and only if the map is one-to-one and the links as represented by segments conjoining peak symbol centers do not intersect in the graphical representation (Figure 4.2), Formally, this means for any $j^{(0)}$ and $j^{(1)}$ in $\mathcal{J}_o$, *any* of the following statements are true:

- **V1** if $L(j^{(0)}) \in 1\!:\!N_i$ and $L(j^{(1)}) \in 1\!:\!N_i$, then $j^{(1)} > j^{(0)} \leftrightarrow L(j^{(1)}) > L(j^{(0)})$.
- **V2** $L(j^{(0)}) = \text{'S'}$
- **V3** $L(j^{(1)}) = \text{'S'}$

4.3 Distributional specifications

To evaluate the likelihood of the pitch hypothesis, $P(F, A|f_0, A_0)$ using (4.7), we must specify the following distributions:

- The *linkmap prior* $P(L)$
- The collection of *template distributions* $\{P_k\left(F(j), A(j)|f_0, A_0\right)\}_{k=1}^{N_i}$
- The *spurious distribution* $P_{\text{S}'}\left(F(j), A(j)|f_0, A_0\right)$

Recall that the symbolic linkmap representation, L, is asymmetric in that it describes the assignment *from* observed *to* template peaks. With such a representation, it becomes easy to evaluate both template and spurious distributions. However, the evaluation of the prior concerns matters such as which template peaks are unobserved in the STFT. In this case, it becomes convenient to access a *dual* representation of the linkmap, which describes the reverse assignment, *from* template *to* observed peaks. The following section gives an algorithm for obtaining the dual linkmap representation for any valid linkmap (and vice versa), establishing the inherent equivalency of both representations.

4.3.1 Dual linkmap representation

Given a linkmap $L \in \mathcal{L}$, define the *dual linkmap* $M : \mathcal{K}_i \rightarrow \mathcal{K}_o$, where $\mathcal{K}_i \triangleq 1 : N_i$, $\mathcal{K}_o \triangleq 1 : N_o \cup \text{'M'}$, where 'M' designates the possibility of a *missing* peak. In other words, for the k^{th} template peak; $M(k)$ returns the index of the corresponding observed peak, except when $M(j) = \text{'M'}$, meaning the k^{th} template peak is unobserved in the STFT. We derive M as follows.

$$M(k) \triangleq \begin{cases} j \in 1\!:\!N_o, & L(j) = k \\ \text{'M'}, & L(j) \neq k \;\; \forall j \in 1\!:\!N_o \end{cases} \tag{4.8}$$

For instance, in the example of Figure 4.2, $M(1) = 1$, $M(2) = 2$, $M(3) = \text{'M'}$, $M(4) = \text{'M'}$, and $M(5) = 4$.

For $M : \mathcal{K}_i \rightarrow \mathcal{K}_o$ to be a valid functional mapping, we must show additionally that each $j \in 1 : N_o$ for which $M(k) = j$ is unique. Assume to the contrary there exists $j^{(0)}, j^{(1)} \in 1\!:\!N_o$ and $k \in 1\!:\!N_i$ for which $L(j^{(0)}) = k$, $L(j^{(1)}) = k$, but $j^{(0)} \neq j^{(1)}$. Either $j^{(0)} > j^{(1)}$ or $j^{(0)} < j^{(1)}$. If $j^{(0)} > j^{(1)}$, by $L \in \mathcal{L}$ and validity condition **V1**, $L(j^{(0)}) > L(j^{(1)})$ implies $k > k$ which is a contradiction. Similarly, $j^{(0)} < j^{(1)}$ implies $k < k$. Hence M belongs to the set of functional mappings $M^* : \mathcal{K}_i \rightarrow \mathcal{K}_o$.

By the same reasoning, we may show that M is unique given $L \in \mathcal{L}$. Assume to the contrary there exists $M, M' \in \mathcal{M}^*$, both satisfying (4.8), for which $M \neq M'$. Then there must exist some $k \in \mathcal{K}_i$ for which either:

- (a) $M(k) = j^{(0)} \in 1\!:\!N_o$, $M'(k) = j^{(1)} \in 1\!:\!N_o$, and $j^{(0)} \neq j^{(1)}$
- (b) $M(k) = j^{(0)} \in 1\!:\!N_o$ and $M^{'}(k) = {}'\mathrm{M}'$
- (c) $M(k) = {}'\mathrm{M}'$ and $M^{'}(k) = j^{(1)} \in 1\!:\!N_o$

The latter two cases are similar, so only (b) will be addressed. If (a) holds, then $L(j^{(0)}) = L(j^{(1)})$ with $j^{(0)} \neq j^{(1)}$, but by preceding arguments, $L \neq \mathcal{L}$, which is a contradiction. If (b) holds, we have simultaneously $L(j^{(0)}) = k$, $L(j^{(1)}) \neq k$. Hence $M = M^{'}$, so that M is unique.

By uniqueness of the correspondence $L \rightarrow M$, we may query the range of this correspondence, if indeed it is as large as $\mathcal{M}^*$. However, if $L \in \mathcal{L}$, it is readily shown $M \in \mathcal{M}$, where $\mathcal{M}$ constitutes the set of all mappings $\mathcal{K}_i \rightarrow \mathcal{K}_o$ for which any of the following validity conditions apply:

- **V1M** if $M(k^{(0)}) \in 1\!:\!N_o$, $M(k^{(1)}) \in 1\!:\!N_o$, then $k^{(1)} > k^{(0)} \leftrightarrow M(k^{(1)}) > M(k^{(0)})$.
- **V2M** $M(j^{(0)}) = {}'\mathrm{M}'$
- **V3M** $M(j^{(1)}) = {}'\mathrm{M}'$.

To show, suppose $L \in \mathcal{L}$, yet none of the conditions **V1M**–**V3M** hold. Defining $j^{(0)} = M(k^{(0)})$, $j^{(1)} = M(k^{(1)})$, then there exists $j^{(0)}, j^{(1)} \in \mathcal{J}_o$, $j^{(0)} \geq j^{(1)}$ but $L(j^{(0)}) < L(j^{(1)})$, which contradicts $L \in \mathcal{L}$. Clearly **V1M**–**V3M** are symmetric to **V1**–**V3**.

Finally, for all $M \in \mathcal{M}$, we may define the reverse correspondence $M \rightarrow L$, where

$$L(j) \triangleq \begin{cases} k \in 1\!:\!N_o, & M(k) = j \\ {}'\mathrm{S}', & M(k) \neq j \quad \forall k \in 1\!:\!N_i \end{cases} \tag{4.9}$$

By symmetry of the definitions (4.8) and (4.9), it follows that L defined as such is a unique member of $\mathcal{L}$. Hence, the valid representation spaces $\mathcal{L}$ and $\mathcal{M}$ exist in one-to-one correspondence, with each $L \in \mathcal{L}$ mapping to a unique $M \in \mathcal{M}$ and vice versa. We conclude that L, M are *equivalent* (i.e., dual) representations for the same underlying structure.

4.3.2 Prior specification

When specifying the prior, $P(L)$, it becomes convenient to consider L paired with its dual representation: $\{L, M\}$, where M is defined by (4.8) in the preceding section.

We first model the process which accounts for missing peaks in the STFT, then we account for the incidence of spurious peaks. The information concerning missing peaks may be encoded in the vector 1_M, defined as follows.

$$1_M(k) \quad \stackrel{\Delta}{=} \quad 1_{\{M(k)='\mathrm{M}'\}}, \quad \forall k = 1 : N_i \tag{4.10}$$

In other words, $1_M(k) = 0$ means that the k^{th} template peak is observed; $1_M(k) = 1$ means that it is missing.

Similarly, the generation of spurious peaks may be encoded in the vector 1_S:

$$1_S(j) \quad \stackrel{\Delta}{=} \quad 1_{\{L(k)='\mathrm{S}'\}}, \quad \forall l = 1 : N_o \tag{4.11}$$

The spurious peaks' generation is modeled as statistically independent of the process responsible for missing peaks. In reality, these processes are interdependent. For instance, a prominent sinusoidal component from an interference event for which the frequency matches the mean frequency of one of the template peaks may appear in the STFT as a spurious peak, also annihilating the corresponding template peak. In practice, acceptable results are achieved using the independence assumption, especially since the majority of spurious peaks' frequencies are seen to differ substantially from any template peak's mean frequency. Hence:

$$P(L) \quad \propto \quad P(1_M)P(1_S) \tag{4.12}$$

where 1_M and 1_S are derived from L and 1_M from M. The proportionality in (4.12) ensures that the resultant distribution sums to unity. The process modeled by $P(1_S)$ produces in general a variable number of spurious peaks, but the number of missing peaks (via 1_M) and the given number of observed peaks necessarily *fixes* the number of spurious peaks. Hence, the model (4.12) actually describes the restriction of a more general process to a fixed number of observed peaks. The proportionality effectively

enforces the conditioning implied by this restriction.

The template peak survival is modeled according to the assumption that higher-frequency peaks are less likely to survive, due to their lower expected amplitude. We further assume the *peak non-interaction hypothesis*: the survival of any given peak is not predicated upon the survival of other peaks. As such the distribution of 1_M factors as a product distribution over each $1_M(k)$; the latter is modeled with geometrically decaying survival probability:

$$P(1_M) \;=\; \prod_{k=1}^{N_i} \phi_{surv}^{k(1-1_M(k))}(1-\phi_{surv}^{k})^{1_M(k)} \tag{4.13}$$

Then the spurious peak generation is modeled as a Poisson process, which indicates a Poisson-distributed number of spurious peaks:

$$P(1_S) \;=\; e^{-\lambda_{spur}} \frac{\lambda_{spur}^{N_{\prime S\prime}}}{(N_{\prime S\prime})!} \tag{4.14}$$

where λ_{spur} denotes the expected number of spurious peaks in the frequency interval $[0, \pi)$ and $N_{\prime S\prime}$ denotes the actual number of spurious peaks:

$$N_{\prime S\prime} \;=\; \sum_{j=1}^{N_o} 1_S(k) \tag{4.15}$$

4.3.3 Template distribution specification

We now consider the modeling of the template distributions in (4.7). Frequency and amplitude observations are modeled as statistically independent, each depending only on their respective (fundamental frequency; reference amplitude) hypotheses:

$$P_k(F(j), A(j)|f_0, A_0) \;=\; P_k(F(j)|f_0)P_k(A(j)|A_0) \tag{4.16}$$

The frequency observation is modeled as Gaussian, following the model of Goldstein [47]:

$$P_k(F(j)|f_0) \sim \mathcal{N}(F(j)|\mu_{f,k}, \sigma_{f,k}) \tag{4.17}$$

Absent knowledge of harmonic structure, mean frequencies are modeled as integer multiples of f_0. Uncertainties due to additive noise and inharmonicity are absorbed in the variance term $\sigma_{f,k}$. Hence:

$$\mu_{f,k} = kf_0 \tag{4.18}$$

Regarding the specification of $\sigma_{f,k}$, we find that variances due to the most common sources of uncertainty admit the form of a multivariate polynomial relation with respect to fundamental and harmonic number; i.e.:

$$\sigma^2_{f,k} = \sum_{n=0}^{n_{max}} \sum_{m=0}^{m_{max}} C_{m,n} f_0^n k^m \tag{4.19}$$

We consider in turn, uncertainties due to additive white Gaussian noise, fourth-order stiffness behavior (a common form of inharmonicity found in acoustic instruments such as piano and marimba), and the psychoacoustic considerations addressed by Goldstein's model.

The case of additive noise in light of our peak extraction method (Figure 4.1) is discussed briefly in Section 4.2.2. To review the argument, our peak extraction approaches the DTFT magnitude estimator of Rife and Boorstyn [97], which the authors derive as a maximum-likelihood estimator (MLE). Suppose $y_{1:T}$ is a single complex sinusoid with true amplitude A, frequency ω, and phase ϕ. Suppose further that A, ω, and ϕ are unknown, and that the signal is embedded in additive Gaussian white noise with variance σ_n^2. Let $\hat{\omega}_{MLE}$ be the estimate of ω corresponding to the

joint MLE. Then:

$$\hat{\omega} \;=\; \underset{\omega}{\operatorname{argmax}} \left| \sum_{t=1}^{T} e^{-j\omega t} y_t \right| \tag{4.20}$$

where $j \triangleq \sqrt{-1}$.

It follows that $\hat{\omega}$, being a MLE, is asymptotically unbiased, achieving the Cramer-Rao lower variance bound[97]. In other words, for sufficiently large frame length T:

$$Var(\hat{\omega}) \;\approx\; \frac{12\sigma_n^2}{A^2 T(T^2-1)} \tag{4.21}$$

The key differences between the proposed peak extraction method, discussed in Section 4.2.1, and the MLE approach of Rife and Boorstyn are as follows. First, the proposed method is encumbered by the finite resolution and imperfect interpolation in the frequency domain. The eight-times-zeropadded DFT produces a sampling of the DTFT at frequencies which are integer multiples of $\pi/(4T)$. The exact frequency value maximizing the DTFT magnitude usually occurs between these values. Quadratic interpolation of the log magnitude about zeropadded-DFT maxima recovers substantial accuracy, but is only exact in the case of an infinite Gaussian window: it cannot be exact for all analysis window shapes. Second (4.20) implies a rectangular window, while the proposed method uses a tapered (Hamming) window. The reason, which constitutes yet another primary difference, is that most signals of interest contain multiple component frequencies. The DTFT peak caused by one component frequency may interfere with either sidelobes or mainlobes caused by the other components.

Figure 4.3 shows an example DTFT containing sidelobe interference. To generate this example, the input consists of a target sinusoid corrupted by an interfering sinusoid at higher frequency and three times the magnitude of the target. The upper graph displays DTFT magnitudes individually for each sinusoid (target = solid line; interference = dotted line); the lower displays the DTFT magnitude of the target

(dotted line) vs. the DTFT magnitude of the actually observed mixture (solid line). Use of a tapered window suppresses sidelobe interference at the expense of widening the mainlobe (see Figure 4.4); if two mainlobes interact (guaranteed for the Hamming window if the distance between any two component frequencies is less than $4\pi/T$: see Figure 4.5), the estimated frequency may shift or the peak may disappear altogether.

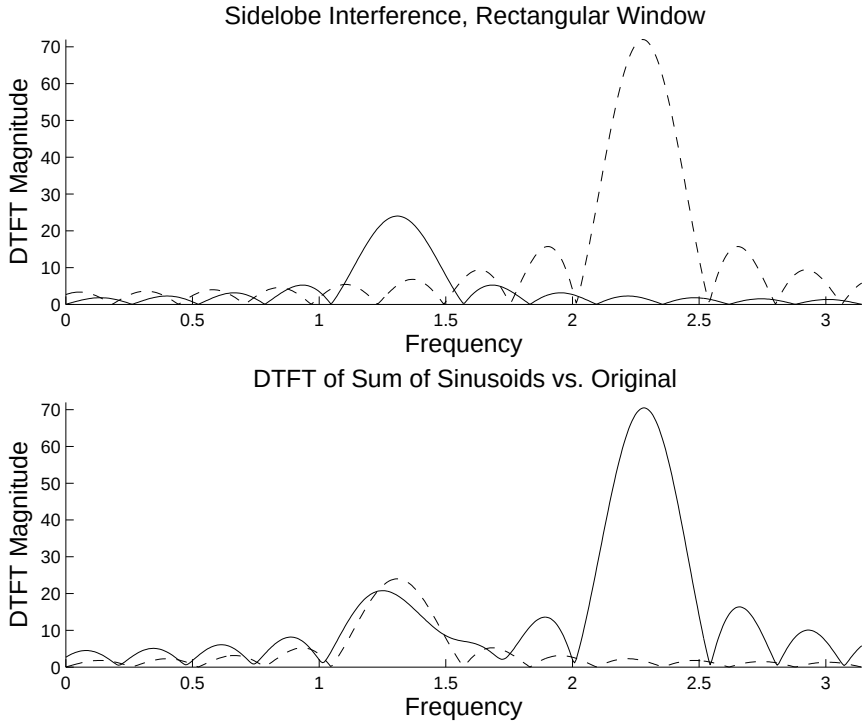


Figure 4.3: *Sidelobe interference for rectangular window*

Under an eight-fold zeropadding factor, quadratic interpolation frequency biases appear negligible in light of high-noise operating conditions (0 to -20 dB noise with frame lengths from 256 to 16384 samples). A zeropadding factor of two or four usually suffices; i.e., the factor of eight is chosen to exist comfortably beyond the point of diminishing returns[3]. Mainlobe interference remains a problem; however, if template peaks' mean frequencies are spaced far enough apart that the underlying components are resolved in the DTFT, it becomes more likely that a linked peak is replaced by a spurious peak caused by the interference. The latter is already handled

[3]See [1] for a recent study on frequency biases due to quadratic interpolation.

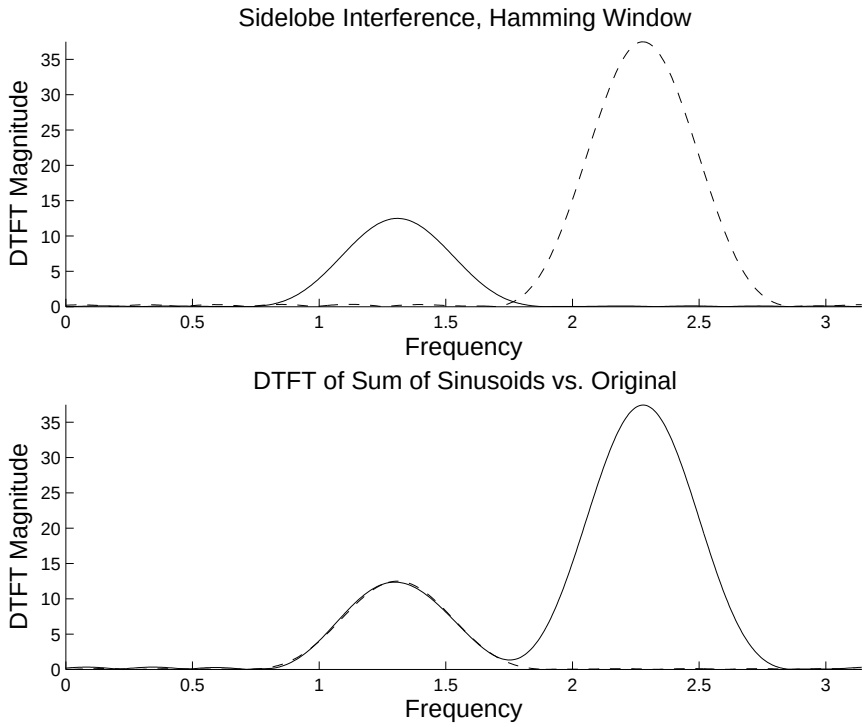


Figure 4.4: *Sidelobe interference for Hamming window*

by the linkmap encoding. As such, gross frequency estimation errors are usually not observed.

Hence, the form of the uncertainty due to additive Gaussian white noise via the proposed preprocessing and peak extraction method seems adequately represented by the Gaussian uncertainty of the MLE (4.21) under similar conditions. In practice, the realized variance of the frequency estimate is two to three times greater than that predicted by the Cramer-Rao bound, depending on the number of data samples. This "Cramer-Rao" uncertainty appears independent of fundamental frequency and harmonic number, accounting for the $m = n = 0$ term in (4.19).

Next, we consider a common source of uncertainty in harmonic structure, which arises from an unknown fourth-order dispersion coefficient representing the inherent stiffness of the wave propagation medium. Bar instruments such as marimba, vibraphone, and xylophone contain significant dispersion as well as instruments with thick metallic strings such as piano.

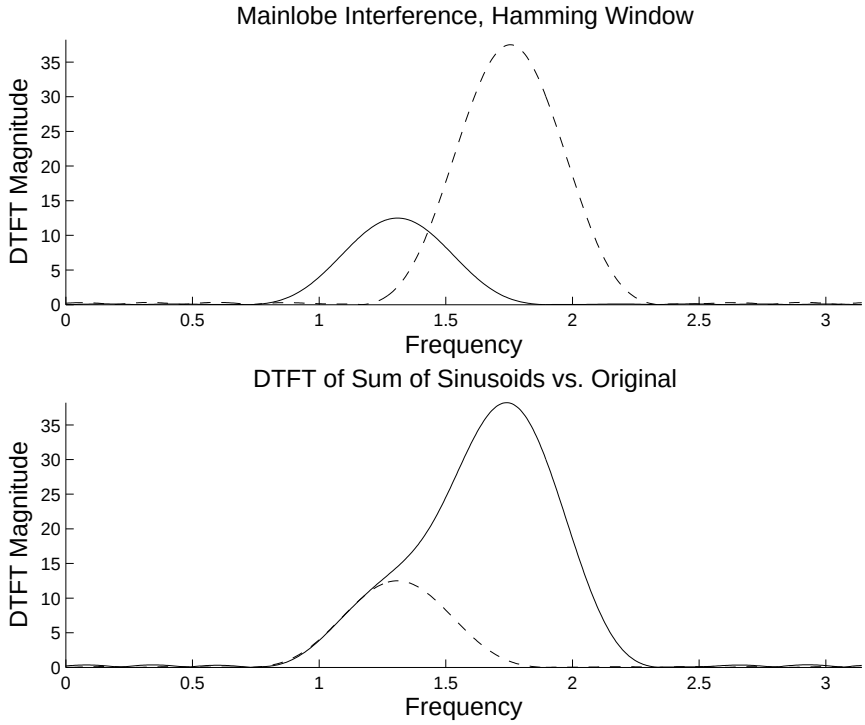


Figure 4.5: *Mainlobe interference for Hamming window*

Consider the general linear second-order-time wave equation [11]:

$$\frac{\partial^2 y}{\partial t^2} + 2\sum_{k=0}^{M} q_k \frac{\partial^{2k+1} y}{\partial x^{2k} \partial t} + \sum_{k=1}^{N} r_k \frac{\partial^{2k} y}{\partial x^{2k}} \quad = \quad 0 \tag{4.22}$$

where t denotes (continuous) time, y displacement, and x the spatial position. Odd-order (spatial) terms (the q_k-terms) contribute primarily frequency-dependent losses; even-order terms influence primarily harmonic structure. A simplification of (4.22) up to fourth-order spatial terms is as follows [11]:

$$\frac{\partial^2 y}{\partial t^2} \quad = \quad c^2 \frac{\partial^2 y}{\partial x^2} - \kappa^2 \frac{\partial^4 y}{\partial x^4} - 2b_1 \frac{\partial y}{\partial t} + 2b_2 \frac{\partial^3 y}{\partial x^2 \partial t} \tag{4.23}$$

Let us consider fixed boundary conditions; i.e., the displacements and second spatial

derivatives are zero at $x = 0$ and $x = L$. Then, with small loss terms $b_1, b_2 \ll 1$ [42]:

$$f_k \approx k f_0 \left|1 + k^2 B\right|^{1/2} \tag{4.24}$$

where f_k is the frequency of the k^{th} partial, and [11]:

$$\begin{aligned} f_0 &= \frac{c\pi}{L} \\ B &= \frac{k^2 f_0^2}{c^4} \end{aligned} \tag{4.25}$$

Physically $B > 0$; we consider B to be exponentially distributed with mean λ_B. The latter reflects the desired level of inharmonicity to which we expect to be robust. As the actual level is unknown, we absorb the expected total squared error in the variance term; i.e., $\sigma_{f,k} = E(f_k - k f_0)^2$. From (4.24) and (4.25), we obtain:

$$\sigma_{f,k}^2 = k^4 f_0^2 \lambda_B \tag{4.26}$$

Hence, variance scaling due to uncertainty about harmonic structure accounting for fourth-order dispersive effects corresponds to the $n = 2, m = 4$ term in (4.19).

Finally, we recall the variance scaling used in Goldstein's harmonic template method [47], which is motivated by psychoacoustic considerations:

$$\sigma_{f,k}^2 = K^2 k^2 f_0^2 \tag{4.27}$$

This scaling may be derived from a supposed logarithmic tolerance for frequency deviations. In other words, let:

$$\begin{aligned} f_k &= \exp(X_k) \\ X_k &\triangleq \log(k f_0) + \epsilon \\ \epsilon &\sim \mathcal{N}\left(0, \sigma_\epsilon^2\right) \end{aligned} \tag{4.28}$$

Then the moment generating function of X_k, $M(\alpha) \stackrel{\Delta}{=} E(e^{\alpha X_k})$, is as follows:

$$E(e^{\alpha X_k}) = (kf_0)^\alpha e^{\alpha^2 \lambda/2} \tag{4.29}$$

From (4.29) the mean and variance terms for f_k may be derived.

$$\begin{aligned} E(f_k) &= kf_0 e^{\lambda/2} \\ Var(f_k) &= k^2 f_0^2 \left(e^{2\lambda} - e^{\lambda}\right) \end{aligned} \tag{4.30}$$

For λ small, $E(f_k) \sim kf_0$, and the variance remains proportional to $k^2 f_0^2$ as in Goldstein's variance scaling (4.27).

We note that the latter variance expressions, (4.26, 4.27), conform to a simplified version of the general term in (4.19):

$$\sigma_{f,k}^2 \propto (kf_0)^2 k^p \tag{4.31}$$

where $p = 0$ for the Goldstein model and $p = 2$ for the inharmonicity model. Only the additive noise term fails to conform to (4.31). Thus, in practice, we consider only two terms: a constant term accounting for additive noise, and a term accounting for timbral variation via (4.31).

$$\sigma_{f,k}^2 = \sigma_F^2 + C_{harm}(kf_0)^2 k^p \tag{4.32}$$

Empirical testing on a limited number of examples (mostly piano and violin) favors p between 0 and 1. This lies between the variance scaling proposed by Goldstein, ($p = 0$), and our proposed scaling due to unknown stiffness ($p = 2$), although somewhat closer to the Goldstein model.

We now consider the amplitude distribution $P_k(A(j)|A_0)$. Ideally, as previously discussed in (Section 4.2.2) $A(j) \sim T(k)$, where $T(k)$ is a *timbral hypothesis* describing the spectral envelope[4] as a function of harmonic number k. $T(k)$ models a very

[4]For most acoustic instruments, timbre varies with fundamental frequency. The timbre of the lowest note on the piano is much brighter, for instance, than that of the highest note. However, since the exact manner of variation is difficult to generalize across different instruments, we do not

coarse envelope, similar to the spectral envelopes derived from linear predictive coding (LPC) or cepstral methods using very few coefficients. This modeling disregards the fact that some template harmonics are missing or undetectable in the source. For instance, clarinet timbres generally lack even harmonics, yet the proposed $T(k)$ decaying smoothly as a function of k. No problems arise, however, as the linkmap already encodes the instance of missing harmonics. Any prior expectation concerning missing harmonics may just as easily be addressed by the linkmap prior, rather than explicitly via $T(k)$.

In practice $T(k)$ is unknown: we lack prior knowledge of the instrument(s) used in the recording. Another source of uncertainty comes from additive noise. Suppose that the input consists of a sinusoid corrupted by additive Gaussian white noise, and the STFT peak amplitude in the absence of noise is $T(k)$. If a rectangular window is used in preprocessing, the distribution of an appropriately normalized version of the squared peak amplitude can be modeled by a χ^2 distribution with two degrees of freedom [63, 54]:

$$P_k\left(A^2(j)/\sigma_A^2 \middle| A_0, T(k)\right) \sim \chi^2_{2,A_0^2T^2(k)/\sigma_A^2} \tag{4.33}$$

where σ_A^2 is the variance of the additive noise in the frequency domain, and $\chi^2_{p,q}$ denotes the χ^2 distribution with p degrees of freedom and noncentrality parameter q. To address the unknown $T(k)$, we assume a geometrically decaying envelope for the noncentrality, absorbing the remaining uncertainty as excess variance in (4.33). As a result:

$$P_k\left(A^2(j)/\sigma_A^2\right) \sim \chi^2_{2,A_0^2c_A^{2(k-1)}/\sigma_A^2} \tag{4.34}$$

Here c_A represents the rolloff of upper harmonic amplitudes. This rolloff remains a nuisance parameter; ideally, c_A would be marginalized respect to some noninformative prior; however, this step remains unexplored due to tractability issues. However, the implied relation, $T(k) = c_A^{k-1}$, does reduce the number of unknown parameters from N_i (the number of template peaks) down to one.

model it at present.

4.3.4 Spurious distribution specification

Finally, we consider the distributional modeling for spurious peaks. Paralleling the situation for template peaks (4.16), frequency and amplitude observations for spurious peaks are modeled as statistically independent:

$$P_{\mathrm{S}'}(F(j), A(j)|f_0, A_0) \;=\; P_{\mathrm{S}'}(F(j))P_{\mathrm{S}'}(A(j)) \tag{4.35}$$

We note that the spurious distribution does not actually depend on f_0 or A_0.

The frequency distribution follows a generative Poisson model, discussed in Section 4.3.2. Since the linkmap fixes the number of spurious peaks, each frequency is modeled as conditionally uniform:

$$P_{\mathrm{S}'}(F(j)) \;\sim\; U(0,\pi) \tag{4.36}$$

The amplitude distribution is modeled as the result of pure Gaussian noise in the time domain. From (4.33) and the preceding section's discussion, the squared amplitude normalized by the noise variance admits a *central* χ^2 distribution with two degrees of freedom, the deterministic component in (4.33) being absent. As such,

$$P_{\mathrm{S}'}\left(A^2(j)/\sigma^2_{A,spur}\right) \;\sim\; \chi^2_{2,0} \tag{4.37}$$

In the event we are not modeling a pitch hypothesis, but a transient or noise hypothesis for which all peaks are spurious, we take $\sigma^2_{A,spur} = \left(A^Q\right)^2$, where A^Q represents the overall signal level as defined in Section 4.1.

4.4 Results for exact enumeration

Recall (4.7) that the exact likelihood evaluation proceeds as a summation over all valid linkmaps $L \in \mathcal{L}$, where the latter is described according to the validity conditions **V1-V3** introduced in Section 4.2.3.

To describe an exact enumeration of $\mathcal{L}$, we partition $\mathcal{L}$ according to the number

of template linkages; i.e., the number of output peaks which map to template peaks. The minimum such number is zero, and the maximum is $\min(N_o, N_i)$. Now let m represent this number of template linkages, and let $\mathcal{L}_m$ denote the corresponding partition of $\mathcal{L}$. For each $m \in \{0 : \min(N_o, N_i\}$, we form a linkmap by choosing a subset of $\mathcal{S}_{o,m} \subset 1 : N_o$ containing m output peaks and mapping it to a subset $\mathcal{S}_{i,m} \subset 1 : N_i$ containing m template peaks. That is, if $\mathcal{S}_{o,m}$ and $\mathcal{S}_{i,m}$ consist of the index sets:

$$\begin{aligned} \mathcal{S}_{o,m} &= \{s_{o,m}(1), s_{o,m}(2), \ \ldots \ , s_{o,m}(m)\} \\ \mathcal{S}_{i,m} &= \{s_{i,m}(1), s_{i,m}(2), \ \ldots \ , s_{i,m}(m)\} \end{aligned} \tag{4.38}$$

the linkmap is defined by

$$\begin{aligned} L(s_{o,m}(k)) &= s_{i,m}(k), \ \forall k = 1 : m \\ L(j) &= \text{'S'}, j \notin \mathcal{S}_{o,m} \end{aligned} \tag{4.39}$$

Now, there is no loss of generality if we fix the ordering of $\mathcal{S}_{o,m}$; e.g., such that the corresponding output peaks are sorted by increasing frequency:

$$F(s_{o,m}(k)) \ < \ F(s_{o,m}(l)), \ \forall 1 \leq k < l \leq m \tag{4.40}$$

But (4.39) and validity condition **V1** of Section 4.2.3 require that $\mathcal{S}_{i,m}$ be sorted in the same way; i.e,

$$F(s_{i,m}(k)) \ < \ F(s_{i,m}(l)), \ \forall 1 \leq k < l \leq m \tag{4.41}$$

Hence, exactly one valid linkmap $L \in \mathcal{L}_m$ exists for each pair of subsets $\mathcal{S}_{o,m}, \mathcal{S}_{i,m}$. It follows that the enumeration of each $\mathcal{L}_m$ consists of an inner loop enumerating the $\binom{N_o}{m}$ distinct subsets of $1 : N_o$ with m elements enclosed in an outer loop enumerating the $\binom{N_o}{m}$ distinct subsets of $1 : N_i$ with m elements. As such, the total number of valid

linkmaps may be expressed:

$$\#\{\mathcal{L}\} = \sum_{m=0}^{\min(N_o,N_i)} \binom{N_o}{m}\binom{N_i}{m} \tag{4.42}$$

If $N_o = N_i = N$, (4.42) simplifies accordingly:

$$\begin{aligned}\#\{\mathcal{L}\} &= \sum_{m=0}^{N} \binom{N}{m}^2 \\ &= \sum_{m=0}^{N} \binom{N}{m}\binom{N}{N-m} \\ &= \binom{2N}{N}\end{aligned} \tag{4.43}$$

The final step of (4.43) is justified by the following argument. Consider a collection of $2N$ objects partitioned into two groups of N objects each. Choosing N from these $2N$ objects is the same as choosing m from the first group and $N-m$ from the second group. How the objects are chosen within each group is arbitrary, so there are $\binom{N}{m}$ times $\binom{N}{N-m}$ possibilities for each m. Finally, we must sum over m: between 0 and N objects may be chosen from the first group.

From Stirling's approximation, the following asymptotic behavior is derived [62]:

$$\binom{2N}{N} = \frac{4^N}{\sqrt{\pi N}}\left[1 - \mathcal{O}\left(\frac{1}{N}\right)\right] \tag{4.44}$$

Hence for the exact enumeration, the number of valid linkmaps (hence terms in the likelihood summation) grows exponentially with the problem size as measured by $N = \max(N_o, N_i)$. For large problems, computations may be reduced by pre-computing the $N_o(N_i + 1) = \mathcal{O}(N^2)$ individual peak likelihood terms of the form $P_{L(j)}(F(j), A(j)|f_0, A_0)$ in (4.7). Nevertheless, one still must form an exponential number of products and sum over an exponential number of terms. Although this complexity may seem distressing, we find that in all of the examples investigated, most of the likelihood concentrates in very few linkmaps, as long as the input signal

contains salient pitch content. That observation motivates the stochastic approximation pursued in Section 4.5. The latter adaptively pursues just those linkmaps which, collectively, contain virtually all of the likelihood. The stochastic approximation sums over these, neglecting the remainder of the summation.

We now investigate results for a typical case. Here the input consists of a single 227 ms frame of an $'A4'$ piano tone (nominally 440 Hz). The piano tone is recorded at 44.1 kHz with -14 dB additive Gaussian white noise. The analysis is artificially truncated to the first seven observed and template peaks to facilitate a tractable computation. Of the seven observed peaks, at least two appear spurious, and two of the seven template peaks appear missing. Here, $\mathcal{L}$ contains 3432 linkmaps.

Table 4.1 summarizes the model parameter settings used to generate this example.

Parameter	Type	Description	In Equation	Value
ϕ_{surv}	Linkmap prior	Survival exponent	(4.13)	0.55
λ_{spur}	Linkmap prior	Spurious peak rate per $[0, \pi)$	(4.14)	10.0
σ_F^2	Template frequency	Frequency variance (additive noise)	(4.32)	0 (not used)
C_{harm}	Template frequency	Degree of harmonic uncertainty	(4.32)	0.05
p	Template frequency	Frequency variance scaling exponent	(4.32)	0
c_A	Template amplitude	Expected timbral decay	(4.34)	0.35
σ_A^2	Template amplitude	Timbral uncertainty/amplitude noise level	(4.34)	$(0.5A_0)^2$
$\sigma_{A,spur}^2$	Spurious amplitude	Spurious level (synonymous w/$(A^Q)^2$)	(4.37)	$(0.05A_0)^2$

Table 4.1: *Model parameter settings for exact enumeration example*

Figure 4.6 displays the resultant likelihood $P(F, A|f_0, A_0)$ raised to the 0.05 power versus candidate frequency f_0. Here the reference amplitude A_0 is treated as an unknown nuisance parameter. We estimate A_0 as the maximum peak amplitude:

$$A_0 \quad = \quad \max_{k=1:N_o} A(k) \tag{4.45}$$

The reason that the 0.05 likelihood power is taken in Figure 4.6, is that interesting secondary features, such as the local maxima of the likelihood surface near subharmonics of f_0, may not be visible otherwise. We observe that the global likelihood maximum occurs at $f_0 = 0.0628$ radians per sample. At a sampling rate of 44.1 kHz, this corresponds to a 441 Hz fundamental, which is virtually indistinguishable from the nominal frequency of 440 Hz. Other local maxima correspond to subharmonics.

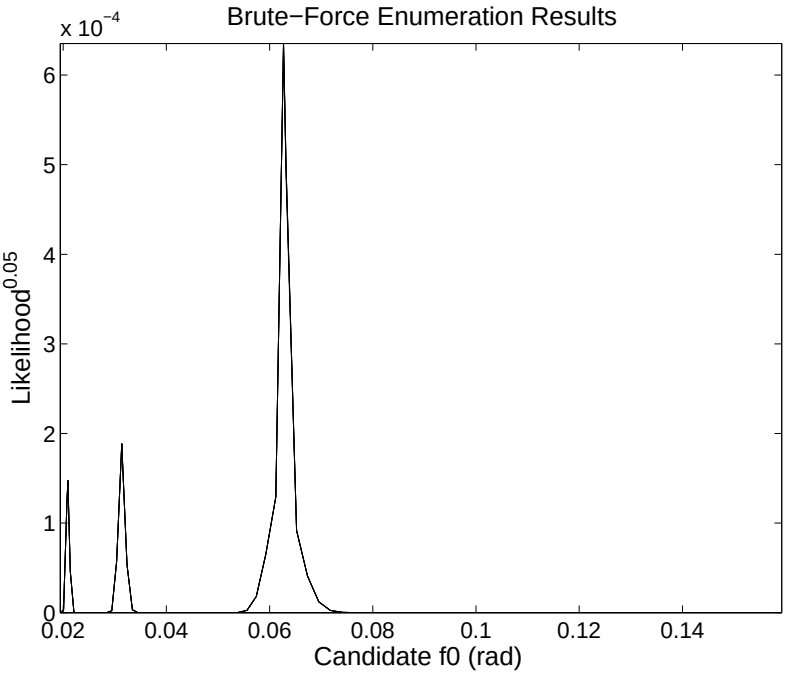


Figure 4.6: *Likelihood evaluation results for exact enumeration, piano example*

The likelihood ratio between the global maximum and any other local maximum is at least $1.93 \cdot 10^{10}$, indicating marked suppression of subharmonic ambiguity.

Now we consider the necessity of enumerating all linkmaps in $\mathcal{L}$, as opposed to a few linkmaps which contribute most to the likelihood evaluation. Via (4.7), the contribution of each individual linkmap as a function of hypotheses f_0 and A_0, which we designate as $\pi_0(L|f_0, A_0)$, may be expressed:

$$\pi_0(L|f_0, A_0) \;=\; P(L)P(F, A|L, f_0, A_0) \tag{4.46}$$

This means that via (4.7, 4.46),

$$P(F, A|f_0, A_0) \;=\; \sum_{L \in \mathcal{L}} \pi_0(L|f_0, A_0) \tag{4.47}$$

Now, define the *likelihood concentration*, $\rho_{conc}\,(l|f_0, A_0)$ as follows:

$$\rho_{conc}\,(l|f_0, A_0) \;\triangleq\; \sup_{\mathcal{S}\subset\mathcal{L}:\#(\mathcal{S})=l} \frac{\sum_{L\in\mathcal{S}} \pi_0(L|f_0, A_0)}{\sum_{L\in\mathcal{L}} \pi_0(L|f_0, A_0)} \tag{4.48}$$

In other words, $\rho_{conc}\,(l|f_0, A_0)$ represents the fraction of the overall likelihood contributed by the l linkmaps with the greatest contributions $\pi_0(L|f_0, A_0)$. Figure 4.7 displays $\rho_{conc}\,(l|f_0, A_0)$ vs. f_0 for the piano example for $l \in 1:3$; Table 4.2 displays the concentration averaged over f_0 and the percentage of f_0 for which the concentration exceeds 99%.

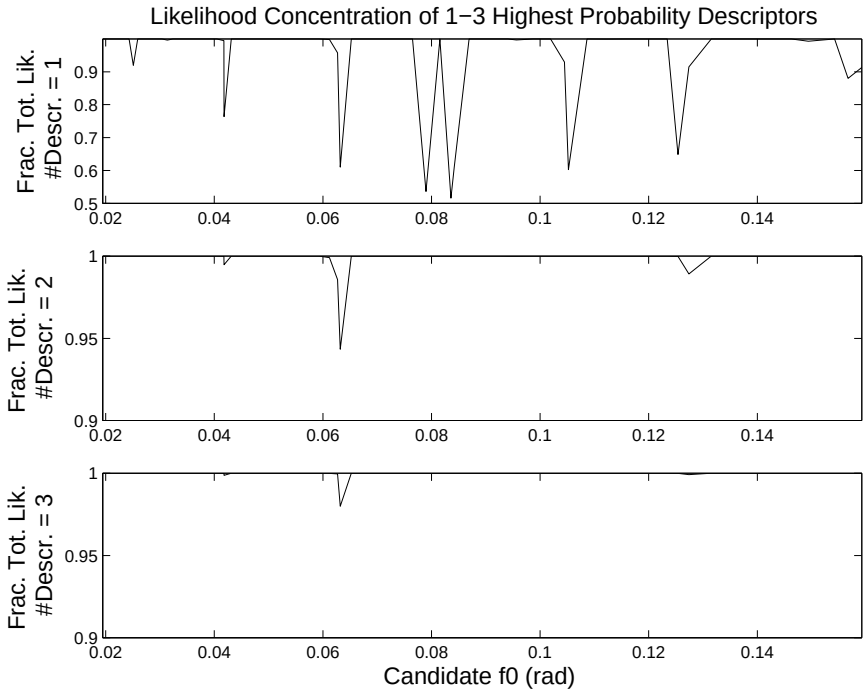


Figure 4.7: *Likelihood concentration for 1-3 top descriptors*

For the typical piano example, virtually all of the likelihood concentrates in just three linkmaps. If we knew in advance which linkmaps these were, we could just evaluate $\pi_0(L|f_0, A_0)$ with respect to these and neglect the rest of the summation (4.46).

#Linkmaps	Average Likelihood Fraction	Exceed 99% Fraction
1	.9581	.8312
2	.9970	.9610
3	.9997	.9870

Table 4.2: *Likelihood concentration for 1-3 top descriptors*

4.5 MCMC approximate likelihood evaluation

As discussed in the previous section, we are interested in identifying a small subset of linkmaps which contribute virtually all of the likelihood to the summation (4.46), so that we can sum over this subset and neglect the rest of the terms, resulting in tremendous computational savings. To this end, we construct a Markovian random walk on $\mathcal{L}$, described by initial value L_0 and transition distribution $P(L_{i+1}|L_i)$, for which the stationary distribution, $\pi(L)$, is proportional to the likelihood contribution $\pi_0(L|f_0, A_0)$ raised to some power $\kappa > 1$:

$$\pi(L) \quad \propto \quad \pi^{\kappa}(L|f_0, A_0) \tag{4.49}$$

As $\kappa \to \infty$, the stationary distribution concentrates on the set of linkmaps achieving the maximum contribution. (Under normal operating conditions, we expect with probability one that there is just one linkmap in this set.) For the piano example, Table 4.2 shows that the linkmap with the maximum likelihood contribution contributes on average 95.8% of the likelihood, which is inadequate for most purposes. Selecting $\kappa \in [1.5, 5.0]$ seems to achieve desirable concentration levels when all the linkmaps visited by the random walk are taken into account. As the random walk is likely to revisit linkmaps, we *hash* likelihood computations for each linkmap.

Given $\pi(L)$, we construct $P(L_{i+1}|L_i)$ via the Metropolis-Hastings algorithm [40] as follows. First, given L_i, a candidate $L_i^{'}$ is chosen according to the *sampling distribution* $q(L_i^{'}|L_i)$. Second, we decide either to accept $L_i^{'}$, upon which $L_{i+1} = L_i^{'}$, or we reject

it, taking $L_{i+1} = L_i$. Acceptance occurs with probability $\min(1, r(L_i, L_i^{'}))$, where

$$r(L_i, L_i^{'}) = \frac{\pi(L_i^{'})q(L_i|L_i^{'})}{\pi(L_i)q(L_i^{'}|L_i)} \tag{4.50}$$

If $P(L_i^{'}|L_i)$ is *irreducible*, meaning that starting from any initial L_0, any $L \in \mathcal{L}$ can be reached in a finite number of steps with positive probability, and *aperiodic*, meaning that for each $L \in \mathcal{L}$, the greatest common divisor of the set $\{n : p_L^n > 0\}$ is unity where p_L^n denotes the probability that the chain beginning in state L will return to L in n steps, the convergence of the chain to $\pi(L)$ is guaranteed [40]. If $\pi(L) > 0$ for all $L \in \mathcal{L}$, the irreducibility and aperiodicity of $P(L_i^{'}|L_i)$ follows from the irreducibility of $q(L_i^{'}|L_i)$. To ensure rapid convergence, we adhere to the following principles concerning initialization and sampling strategies:

- *Favorable initialization* The initial linkmap, L_0, should be chosen such that $\pi(L_0)$ is as large as possible.
- *Sampling via adjacency* The sampling distribution, $q(L_i^{'}|L_i)$, should concentrate on those $L_i^{'}$ which are close to L_i under $\pi(\cdot)$, meaning that the difference $\left|\pi(L_i^{'}) - \pi(L_i)\right|$ is minimized.

Favorable initialization is approached by taking L_0 as the output of some heuristic peak matching algorithm. Here we adopt a method of McAulay and Quatieri [81], termed *MQ-initialization*[5]. For the piano example, the average likelihood concentration of the linkmap derived from MQ-initialization is 0.1149 (Table 4.3), while the maximum achievable concentration for a single linkmap is 0.9581 (Table 4.2). Hence, it seems there is significant room for improvement in the initialization strategy; indeed, alternative peak matching strategies such as [110] merit further investigation. Nonetheless, the MQ-initialization followed by MCMC iterations adhering to the proposed sampling strategy achieves excellent results (Table 4.3).

We now discuss our proposed sampling strategy which is irreducible and which exploits some notion of adjacency in $\mathcal{L}$. Candidate $L_i^{'}$, is derived from L_i via one of

[5]The peak matching strategy in [81] was originally designed to connect sinusoidal peak trajectories across frames, rather than match peaks to a template. Nevertheless, the aims are similar.

the following categories of moves:

- **Q1** *Remove a link* We choose an index $j \in 1:N_o$ for which $L(j) \in 1:N_i$ and set $L(j) = 'S'$.

- **Q2** *Add a non-intersecting link* We choose $j \in 1:N_i$ for which $L(j) = 'S'$ and $k \in 1:N_o$ for which no pair $\{l \in 1:N_o, m \in 1:N_i\}$ exists with $L(l) = m$ and either of the following intersection conditions:

  - $l < k$, $m \geq j$
  - $l > k$, $m \leq j$

- **Q3** *Switch a link to adjacent template peak* We choose $j \in 1:N_o$ for which $L(j) = k \in 1:N_i$ and specify either $L(j) = k-1$ or $L(j) = k+1$. The target value must remain in the range $1:N_i$ and the resultant link must not intersect any other. For instance, if $L(j) = k+1$, we must have $k \in 1:N_i - 1$ and no pair $\{l \in 1:N_o, m \in 1:N_i\}$ exists with $L(l) = m$, and either

  - $l < k+1$, $m \geq j$
  - $l > k+1$, $m \leq j$

  The case $L(j) = k-1$ is symmetric.

- **Q4** *Switch a link to adjacent observed peak* We choose $j \in 1:N_o$ for which $L(j) = k \in 1:N_i$ and either $j-1$ for which $L(j-1) = 'S'$ and $j \in 2:N_o$, or $j+1$ for which $L(j+1) = 'S'$ and $j \in 1:N_{o-1}$, and assign $L(j-1) = k$ (or $L(j+1) = k$). The resultant link must not intersect any other. That means additionally, for the $j+1$ case, no pair $\{l \in 1:N_o, m \in 1:N_i\}$ exists with $L(l) = m$, and either

  - $l < k$, $m \geq j$
  - $l > k$, $m \leq j$

  The $j-1$ case is symmetric.

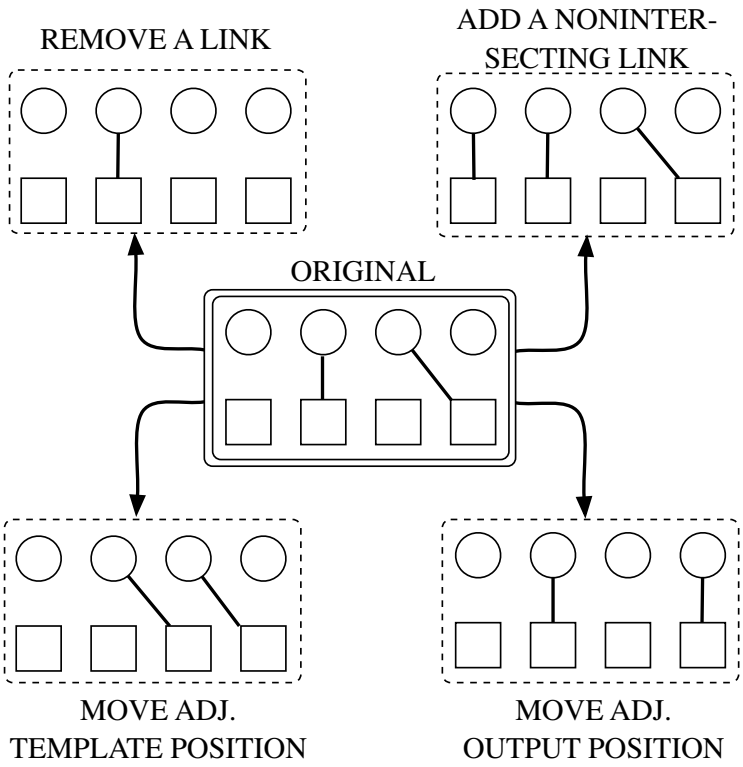


Figure 4.8: *Move possibilities for MCMC sampling strategy*

Example moves are displayed in Figure 4.8.

Given L_i, the set of move possibilities for each category is computed. A category is selected equiprobably over the categories with at least one possibility, then a move is selected equiprobably among the possibilities for that category.

Note that we may reach any linkmap from any other by removing then adding links one by one. This guarantees the irreducibility of $q(L_i'|L_i)$ because each remove/add possibility has positive probability, and the maximum number of links is finite. Because $q(L_i'|L_i)$ is irreducible the entire chain is irreducible and aperiodic, thus guaranteeing convergence to $\pi(L)$. The role of the latter "switching" categories is to speed convergence. A common source of ambiguity arises when two observed peaks are closely spaced in frequency about the mean frequency of a template peak: either observation may link to the template peak. Without the ability to switch links among adjacent observed peaks, we are forced to traverse the unlikely possibility for

which both observed peaks are considered spurious. The switching categories thereby provide valuable "shortcuts" towards convergence.

Under identical conditions generating Figures 4.6 and 4.7, Figure 4.9 compares the MCMC likelihood approximation averaged over 1000 trials and the likelihood from MQ-initialization alone with the exact likelihood evaluation. Each trial involves 200 MCMC iterations. We vary parameter κ, defined via (4.49), according to the annealing schedule

$$\kappa_0 = 0.05 \tag{4.51}$$

$$\kappa_i = \min(1.03\kappa_{i-1}, 5.0) \tag{4.52}$$

Figure 4.9 displays likelihood surfaces for exact evaluation, MCMC approximation, and the MQ-initialization alone. Here the exact and MCMC-approximate results

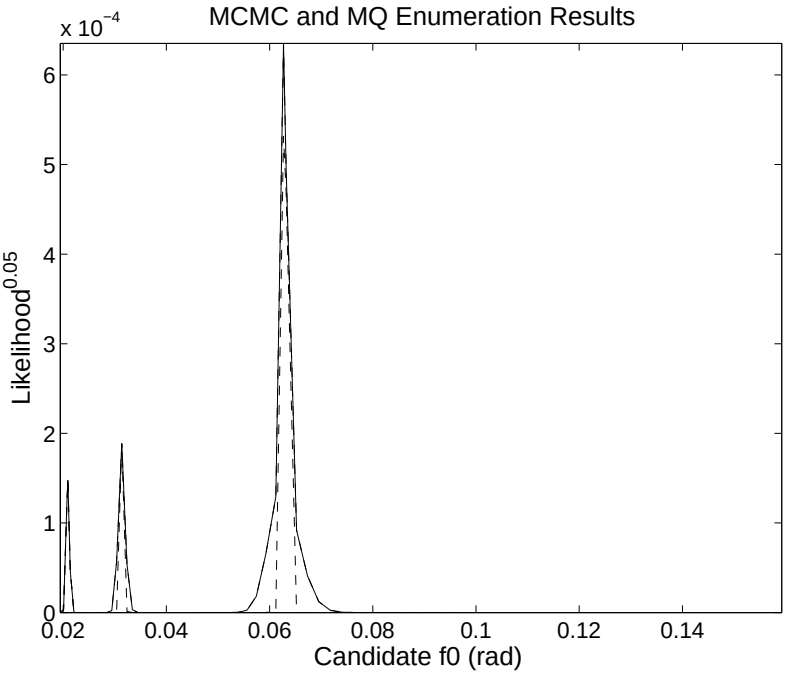


Figure 4.9: *Likelihood evaluation results for exact enumeration, MCMC approximation, and MQ-initialization for piano example*

are plotted via solid lines; the MQ-initialization result appears via dotted line. Exact

and MCMC-approximate results appear indistinguishable, while the MQ-initialization likelihood approaches the exact likelihood only about the correct frequency and subharmonics; elsewhere, almost none of the likelihood is captured via initialization.

Likelihood concentration results are summarized in Table 4.3. On average (for

Method	**Average Likelihood Fraction**	**Exceed 99% Fraction**
MQ-Initialization Only	.1991	.1948
MCMC	$1 - 3.1819 \cdot 10^{-13}$	1

Table 4.3: *Likelihood concentrations of MCMC vs. MQ-initialization*

1000 trials covering all f_0-candidates), all but $3.1819 \cdot 10^{-13}$ of the exact likelihood is captured by the MCMC evaluation. Despite virtually identical results, the latter obtains significant computational savings. Over 200 iterations the MCMC evaluation visits on average 22.38 unique linkmaps per candidate f_0, while the exact method requires 3432 linkmaps. Hence the MCMC approximation yields over a hundredfold reduction in computational effort. Due to the exponential complexity of the exact evaluation (4.44), the computational savings are expected to be even greater as the number of template or observed peaks increases.

Situations arise, however, where the accuracy of the MCMC approximation may be unnecessary, for instance when pitch content is salient. If in addition we must evaluate a large number of pitch hypotheses, it is important to have an alternative approximation strategy, which may sacrifice some of the accuracy of the MCMC approximation in favor of increased computational savings. For instance, the Bayesian segmentation framework of Chapter 3 requires one evaluation per hypothesized note value, tuning offset, and reference amplitude for each signal frame. There the total number of hypotheses per frame may be in the hundreds of thousands or greater.

4.6 Deterministic approximate likelihood evaluation

In this section, we develop a deterministic approximate likelihood evaluation which may save computations at the expense of accuracy when compared to the MCMC method. While the resultant likelihood surface may not match the exact result, primary salient features are nevertheless retained. Moreover, the approximation has been successfully incorporated into Bayesian contexts, for instance the joint segmentation and melody retrieval engine discussed in Chapter 3. Here the method is seen to yield acceptable results even though the input signal contains significant noise and reverberation.

The deterministic approximation is motivated by the form of the exact evaluation, recalling (4.7):

$$P(F, A|f_0, A_0) \;=\; \sum_{L\in\mathcal{L}} P(L) \prod_{j=1}^{N_o} P_{L(j)}(F(j), A(j)|f_0, A_0) \tag{4.53}$$

If $P(L)$ is uniform and $\mathcal{L}$ forms a Cartesian product space over the individual elements $L(j)$, we may exchange sums and products in (4.53) to obtain an expression requiring only $\mathcal{O}(N_o N_i)$ template distribution evaluations of the form $P_{L(j)}(F(j), A(j)|f_0, A_0)$. With $N_o = N_i = N$, the proposed approximation is quadratic in N, as opposed to the exact method which is $\mathcal{O}(4^N/\sqrt{N})$ (4.44).

4.6.1 Uniform linkmap prior approximation

Unfortunately, it becomes difficult to approximate $P(L)$ as uniform. In theory, the variation of $P(L)$ over $\mathcal{L}$ may be quite significant. Via (4.12 - 4.15), we determine the range of $P(L)$ as a function of the survival exponent ϕ_{surv}, the spurious rate λ_{spur},

and the number of template and observed peaks (N_i, N_o):

$$\begin{aligned} P_{min}(L) &\triangleq \min_{L\in\mathcal{L}} P(L) \\ &= \left[\min_{j=1:N_o} e^{-\lambda_{spur}} \frac{\lambda_{spur}^j}{j!}\right] \prod_{k=1}^{N} \min\left(\phi_{surv}^k, 1-\phi_{surv}^k\right) \\ P_{max}(L) &\triangleq \max_{L\in\mathcal{L}} P(L) \\ &= \left[\max_{j=1:N_o} e^{-\lambda_{spur}} \frac{\lambda_{spur}^j}{j!}\right] \prod_{k=1}^{N} \max\left(\phi_{surv}^k, 1-\phi_{surv}^k\right) \end{aligned} \tag{4.54}$$

Under typical conditions ($\phi_{surv} = 0.95, \lambda_{spur} = 3.0$), Figure 4.10 tracks the evolution of $P_{min}(L)$ and $P_{max}(L)$ for $N_i = N_o = N$, $N \in 1{:}10$.

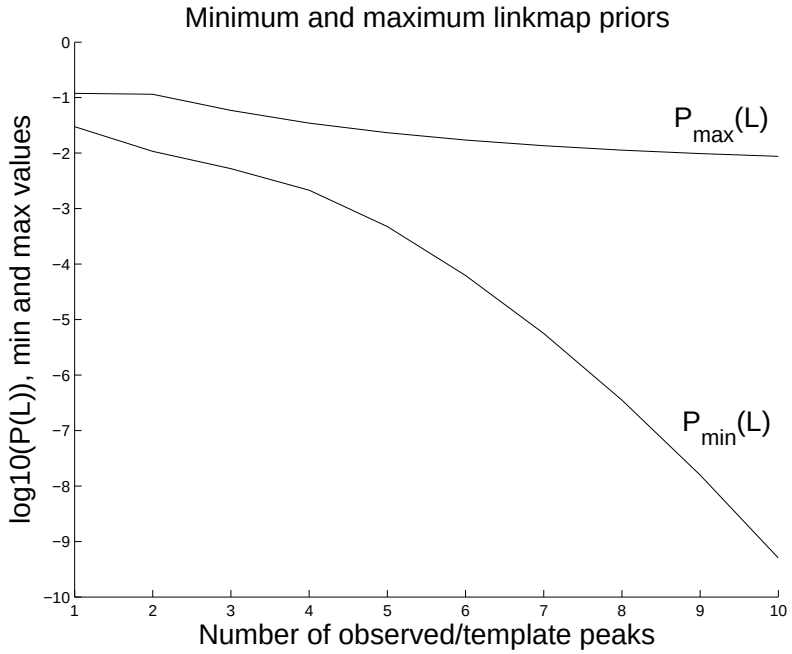


Figure 4.10: *Range of* $P(L)$ *given* $\phi_{surv} = 0.95$, $\lambda_{spur} = 3.0$ *for* $N_o = N_i \in 1{:}10$

Of course, this worst-case analysis somewhat exaggerates the effect of the $P(L)$-variation on the overall likelihood evaluation. Recalling the primary motivation behind the MCMC approximation, virtually all of the likelihood concentrates in just a

few linkmaps: i.e., given $\epsilon > 0$, there exists $\mathcal{S} \subset \mathcal{L}$, $\#(\mathcal{S}) \ll \#\mathcal{L}$, for which:

$$\sum_{L \in \mathcal{L}} P(L)P(F, A|L, f_0, A_0) - \sum_{L \in \mathcal{S}} P(L)P(F, A|L, f_0, A_0) \quad < \quad \epsilon \qquad (4.55)$$

That is, we may discard the summation over $\mathcal{L} \backslash \mathcal{S}$: we are only concerned with the variation of $P(L)$ over $\mathcal{S}$. The latter is connected by the adjacency moves **Q1**–**Q4** as shown in Figure 4.8: each move modifies at most one link in the linkmap L. If $\mathcal{S}$ is sufficiently small, each $L \in \mathcal{S}$ may be reached from some other $L^{'} \in \mathcal{S}$ by modifying a small number of links. We expect, therefore, the *effective* variation of $P(L)$, which is the variation inside $\mathcal{S}$, to be significantly smaller than the variation over the entire space.

In the Bayesian segmentation context, the overall deterministic approximation, which may be considered a further reduction of the uniform-$P(L)$ approximation, seems to yield acceptable results as presented in Section 3.9. For instance, the violin example of Figure (3.18) contains significant regions of overlapping pitch content due to reverberation and legato playing. Nevertheless, the values of all notes of significant length are correctly determined, the initial grace note notwithstanding. Of course, these results are aided by the integration of contextual information across frames. Nonetheless, it is significant that the system *as a whole* is able to glean enough pitch content to detect note events, correctly identify their values, and determine expressive pitch fluctuations surrounding these events, which in some sense justifies the applicability of the uniform linkmap prior approximation.

Perhaps a deeper explanation for the success of the uniform approximation in Bayesian contexts comes via maximum entropy arguments [55]. The uniform linkmap prior maximizes entropy over all choices of this prior, absent constraints [25]. In other words, the uniform prior models probabilistically the largest class of linkmaps, hence retaining the ability to generalize to the greatest variety of situations where nothing else is known.

4.6.2 Product linkmap space

The remainder of the deterministic approximation begins aacording to the uniform-$P(L)$ approximation:

$$P(F, A|f_0, A_0) \approx \frac{1}{\#\mathcal{L}} \sum_{L \in \mathcal{L}} \prod_{j=1}^{N_o} P_{L(j)}(F(j), A(j)|f_0, A_0) \tag{4.56}$$

Now, suppose $\mathcal{L}$, the set of valid linkmaps, is replaced by $\mathcal{L}^*$, defined as the set of *all* such maps $\mathcal{J}_o \rightarrow \mathcal{J}_i$, where (recalling the definitions in Section 4.2.3) $\mathcal{J}_o \stackrel{\Delta}{=} 1\!:\!N_o$, $\mathcal{J}_i \stackrel{\Delta}{=} 1\!:\!N_i \cup \mathcal{S}$, regardless of validity. We may write $\mathcal{L}^*$ as a Cartesian product space:

$$\mathcal{L}^* = l_1^* \otimes l_2^* \otimes \ldots \otimes l_{N_o}^* \tag{4.57}$$

where l_j^* denotes the set of possible maps from the index j to $\mathcal{J}_i$. Each map $j \rightarrow \mathcal{J}_i$ corresponds to a possibility for $L(j)$ in (4.56). Extending the summation over $\mathcal{L}^*$ recasts (4.56) as

$$\begin{aligned} P(F, A|f_0, A_0) &\approx \frac{1}{\#\mathcal{L}} \sum_{L \in \mathcal{L}^*} \prod_{j=1}^{N_o} P_{L(j)}(F(j), A(j)|f_0, A_0) \\ &= \frac{1}{\#\mathcal{L}} \sum_{L(1) \in \mathcal{J}_i} \sum_{L(2) \in \mathcal{J}_i} \cdots \sum_{L(N_o) \in \mathcal{J}_i} \prod_{j=1}^{N_o} P_{L(j)}(F(j), A(j)|f_0, A_0) \end{aligned} \tag{4.58}$$

Interchanging sums and products yields the final form of the approximation:

$$P(F, A|f_0, A_0) \approx \frac{1}{\#\mathcal{L}} \prod_{j=1}^{N_o} \sum_{L(j) \in \mathcal{J}_i} P_{L(j)}(F(j), A(j)|f_0, A_0) \tag{4.59}$$

The challenge, of course, is to assess the degree by which summation over *invalid* linkmaps, meaning linkmaps in $\mathcal{L}^*$ which are not in $\mathcal{L}$, affects the quality of the

approximation. Let $\bar{\mathcal{L}} \triangleq \mathcal{L}^* \backslash \mathcal{L}$ and define $\eta(F, A|f_0, A_0)$ as the latter-stage approximation error:

$$\eta(F, A|f_0, A_0) \approx \frac{1}{\#\mathcal{L}} \sum_{L \in \bar{\mathcal{L}}} \prod_{j=1}^{N_o} P_{L(j)}(F(j), A(j)|f_0, A_0) \tag{4.60}$$

We obtain that each product term on the r.h.s. of (4.60), and hence $\eta(F, A|f_0, A_0)$ becomes negligible under the following conditions:

- The pitch content is salient, meaning that for $L(j) \in 1 : N_i$, the standard deviation of each template distribution $P_{L(j)}(F(j), A(j)|f_0, A_0)$ with respect to $F(j)$ is negligible with respect to the difference in means among neighboring distributions[6].

- The observed peaks are well separated in frequency. In particular:

$$\min_{j \in 2:N_o} (F(j) - F(j-1)) \gg \max_{k \in 1:N_i} \sigma_k^{(F)} \tag{4.61}$$

  where $\sigma_k^{(F)}$ denotes the standard deviation of $F(j)$ under $P_k(F(j), A(j)|f_0, A_0)$.

According to the validity conditions **V1** – **V3** defined in Section 4.2.3, each $L \in \mathcal{L}$ has the property that there exists $j^{(0)}, j^{(1)} \in 1 : N_i$, for which one of the following "invalidity" conditions hold[7]:

- **IV1** *Links intersect* $j^{(1)} > j^{(0)}$; $L(j^{(1)}) < L(j^{(0)})$

- **IV2** *Multiple links per template peak* $j^{(1)} > j^{(0)}$; $L(j^{(1)}) = L(j^{(0)})$

- **IV3** *Multiple links per observed peak* $j^{(1)} = j^{(0)}$; $L(j^{(1)}) < L(j^{(0)})$

Suppose that $P_{L(j^{(0)})}(F(j^{(0)}), A(j^{(0)})|f_0, A_0)$ is negligibly small. Then, since this term is one of the product terms (4.60), the entire product corresponding to L is annihilated. Otherwise, by the pitch salience hypothesis, $F(j^{(0)})$ must be close

[6] By "neighboring distributions" we mean $P_{L(j)-1}(F(j), A(j)|f_0, A_0)$ and $P_{L(j)+1}(F(j), A(j)|f_0, A_0)$, where applicable.

[7] The implicit assumption that $j^{(1)} \geq j^{(0)}$ is without loss of generality.

to the frequency mean of the template distribution corresponding to $L(j^{(0)})$. Let us now consider each condition, **IV1**–**IV3**. By the peak separation hypothesis, $j^{(1)} > j^{(0)}$ implies that $F(j^{(1)})$ exceeds $F(j^{(0)})$ by a non-negligible amount. Either condition (**IV1** or **IV2**) requires $L(j^{(1)}) \leq L(j^{(0)})$. By implication, $F(j^{(1)})$ significantly exceeds the frequency mean of the template distribution under $L(j^{(1)})$; consequently, $P_{L(j^{(1)})}(F(j^{(1)}), A(j^{(1)})|f_0, A_0)$ becomes negligible, annihilating the product corresponding to L in (4.60). For the remaining condition, **IV3**, $j^{(1)} = j^{(0)}$, but $L(j^{(1)}) < L(j^{(0)})$. By the pitch salience hypothesis, the frequency mean of the template distribution under $L(j^{(1)})$ will be significantly less than the mean under $L(j^{(0)})$ when compared with the frequency standard deviation under $L(j^{(1)})$. Hence $P_{L(j^{(1)})}(F(j^{(1)}), A(j^{(1)}|f_0, A_0) = P_{L(j^{(1)})}(F(j^{(0)}), A(j^{(1)}|f_0, A_0)$ becomes negligible, annihilating the product corresponding to L in (4.60). Since the error contribution for each term $L \in \bar{\mathcal{L}}$ becomes negligible, and there are a finite number of such terms, $\eta(F, A|f_0, A_0)$ hence becomes negligible.

4.6.3 Computational considerations

The computational cost of the deterministic approxmation, via (4.59), is $\mathcal{O}(N_i N_o)$. Under $N_i = N_o = N$ this becomes $\mathcal{O}(N^2)$, as opposed to $\mathcal{O}(4^N/\sqrt{N})$ (4.44) for the exact method.

In theory, either the MCMC or the deterministic approximation may be faster for a given application; in practice, the deterministic method seems to take 10-50% of the time of the MCMC method for the Bayesian segmentation examples reviewed in Section 3.9. Unfortunately, it becomes difficult to draw more general conclusions. First, it becomes uncertain how these results generalize to the almost limitless variety of instruments, recording conditions, and background noises manifest in typical sound examples. Second, thanks to the algorithmic complexity of both approaches, it is difficult to verify that both algorithms have been implemented in an equally efficient (let alone optimally efficient) manner. The reason the MCMC method may theoretically require less computations is that the set of linkmaps spanned by the traversal may not involve the exhaustive set of template distribution evaluations computed by the

deterministic approach[8]. However, the latter lacks many sources of overhead inherent to the MCMC approach; e.g., the linkmap prior evaluation, the Metropolis-Hastings acceptance-rejection strategy, and the maintenance of numerous hashtables. Both Metropolis-Hastings and hashtable maintenance incur costs once per MCMC iteration as opposed to once per unique linkmap visited, or once per template distribution evaluation.

Ultimately, the user is encouraged to implement both deterministic and MCMC approximations, assessing computational costs in terms of how well each method achieves the desired performance goals. However, the results of Section 3.9 seem quite encouraging as regards the deterministic approximation.

[8]One would expect this to be the case for "clean" data, meaning signals for which most of the likelihood concentrates in one or two linkmaps.

Appendix A

Approximate Viterbi inference recursions

This appendix derives the filtering and smoothing recursions given in Section 3.7.1. Recall that the goals are to compute:

$$M^*_{1:N} = \underset{M_{1:N}}{\operatorname{argmax}}\ P\left(M_{1:N}|Y_{1:N}\right) \tag{A.1}$$

$$\sigma^*(S_t) = P(S_t|M^*_{1:N}, Y_{1:N}), \quad \forall t \in 1\!:\!N \tag{A.2}$$

from the distributions given in the factorization of $P(M_{1:N}, S_{1:N}, Y_{1:N})$ (3.42):

$$\begin{aligned} P(M_{1:N}, S_{1:N}, Y_{1:N}) &= P(M_1)P(S_1|M_1)P(Y_1|S_1) \\ &\times \prod_{t=2}^{N} P(M_t|M_{t-1})P(S_t|S_{t-1}, M_{t-1}, M_t)P(Y_t|S_t) \end{aligned} \tag{A.3}$$

The factorization (A.3) is represented by the directed acyclic graph of Figure A.1.

Quantities propagated in filtering and smoothing recursions as well as the necessary input distributions given on the r.h.s. of (A.3) are summarized in Table A.1,

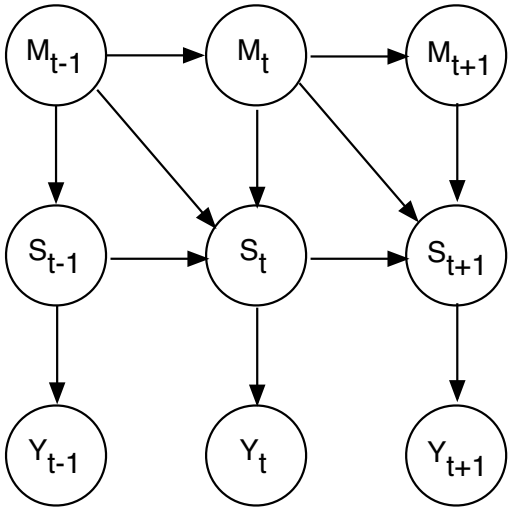


Figure A.1: *Directed acyclic graph for the factorization of* $P(M_{1:N}, S_{1:N}, Y_{1:N})$

where the following notation is used:

$$M^*_{1:t-1}(M_t) \approx \underset{M_{1:t-1}}{\operatorname{argmax}} P(M_{1:t-1}|M_t, Y_{1:t}) \tag{A.4}$$

In general, we refer to $M^*_{1:t-1}(M_t)$ as the M_t-*optimal mode sequence*; we define $M^*_{a:b}(M_t)$ as the corresponding subsequence for frames between a and b, $a \leq b$ assumed, and adopt the shorthand $M^*_a(M_t) \triangleq M^*_{a:a}(M_t)$.

These recursions depend on the approximation:

$$P(Y_{t+1}|M_{1:t+1}, Y_t) \approx P(Y_{t+1}|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t}) \tag{A.5}$$

The meaning and applicability of (A.5) are described in Section 3.7.1. With the distributional terms on the r.h.s. of (A.3) serving as inputs, the outputs of the recursions are taken as $M^*_{1:N}$ and $\sigma^*(S_{1:N})$, which satisfy (A.1) and (A.2) as desired.

Symbol	Quantity	Description
$\pi(M_1, S_1)$	$P(M_1, S_1)$	Prior
	$P(M_{t+1}\|M_t)$	Mode transition dependence
	$P(S_{t+1}\|S_t, M_t, M_{t+1})$	State transition dependence
	$P(Y_t\|S_t)$	Observation likelihood
$\tau^*(M_t, S_t)$	$P(S_t\|M^*_{1:t-1}(M_t), M_t, Y_{1:t-1})$	Predicted posterior given M_t-optimal mode sequence
$\mu^*(M_t, S_t)$	$P(S_t\|M^*_{1:t-1}(M_t), M_t, Y_{1:t})$	Smoothed posterior given M_t-optimal mode sequence
$J(M_t)$	$\max_{M_{1:t-1}} P(M_{1:t}\|Y_{1:t})$ $(\approx)$	Objective at time t
$M^*_{t-1}(M_t)$	$\operatorname{argmax}_{M_{t-1}} \max_{M_{1:t-2}} P(M_{1:t}\|Y_{1:t})$ $(\approx)$	Backpointer
M^*_t	$\operatorname{argmax}_{M_t} \max_{M_{1:t-1}, M_{t+1:N}} P(M_{1:N}\|Y_{1:N})$ $(\approx)$	Maximum *a posteriori* mode at time t
σ^*_t	$P(S_t\|M^*_{1:N}, Y_{1:N})$	Smoothed posterior
$\mu_0(M_t, S_{t+1}, M_{t+1})$	$P(S_{t+1}, Y_{t+1}\|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t})$	Intermediate
$\tau(M_t, S_{t+1}, M_{t+1})$	$P(S_{t+1}\|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t})$	Intermediate
$\mu(M_t, S_{t+1}, M_{t+1})$	$P(S_{t+1}\|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t+1})$	Intermediate
$\Sigma_0(M_t, M_{t+1})$	$P(Y_{t+1}\|M^*_{1:t-1}(M_t), M_{t+1}, Y_{1:t+1})$	Intermediate
$J_0(M_t, M_{t+1})$	$\max_{M_{1:t-1}} P(M_{1:t+1}\|Y_{1:t+1})$ $(\approx)$	Intermediate

Table A.1: *Quantities propagated in approximate Viterbi inference*

The filtering recursions update the following quantities:

$$\begin{aligned} J(M_t) &\approx \max_{M_{1:t-1}} P(M_{1:t}|Y_{1:t}) \\ M^*_{1:t-1}(M_t) &\approx \operatorname*{argmax}_{M_{1:t-1}} P(M_{1:t}|Y_{1:t}) \\ \tau^*(M_t, S_t) &\approx P(S_t|M^*_{1:t-1}(M_t), M_t, Y_{1:t-1}) \\ \mu^*(M_t, S_t) &\approx P(S_t|M^*_{1:t-1}(M_t), M_t, Y_{1:t}) \end{aligned} \tag{A.6}$$

For purposes of interpretation, we assume that the approximations in (A.6) are exact. In this case, the value of M_t maximizing $J(M_t)$ retrieves the maximum *a posteriori* mode trajectory given $Y_{1:t}$; i.e., $M^*_{1:t}$. Thanks to the nesting property:

$$M^*_{1:t-2}(M_t) = M^*_{1:t-2}\left(M^*_{t-1}(M_t)\right) \tag{A.7}$$

it is necessary only to store $M^*_{t-1}(M_t)$, as the remainder of the past trajectory can be unraveled by recursive application of (A.7), i.e.

$$M^*_s = M^*_s(M^*_{s+1}) \;\; \forall s \in 1 : t-1 \tag{A.8}$$

Assuming that the quantities of (A.6) have already been computed for frame t

over all M_t and S_t, we update $J(M_{t+1})$:

$$\begin{aligned} J\left(M_{t+1}\right) &= \max_{M_{1:t}} P\left(M_{1:t}, M_{t+1} | Y_{1:t+1}\right) \\ &= \max_{M_{1:t}} \frac{P\left(M_{1:t}, M_{t+1}, Y_{t+1} | Y_t\right)}{P\left(Y_{t+1} | Y_t\right)} \\ &= \frac{1}{P\left(Y_{t+1} | Y_t\right)} \max_{M_t} \max_{M_{1:t-1}} \big[P\left(M_{1:t} | Y_{1:t}\right) P\left(M_{t+1} | M_{1:t}, Y_{1:t}\right) \\ &\qquad \times P\left(Y_{t+1} | M_{1:t+1}, Y_{1:t}\right) \big] \end{aligned} \tag{A.9}$$

The conditional independence relations of (A.3) yield the simplification:

$$P(M_{t+1} | M_{1:t}, Y_{1:t}) = P(M_{t+1} | M_t) \tag{A.10}$$

Unfortunately, there lacks a corresponding simplification for $P(Y_{t+1} | M_{1:t+1}, Y_{1:t})$; this is addressed by the approximation (A.5). As a result, (A.5) may be expanded by marginalizing over S_{t+1}:

$$\begin{aligned} P(Y_{t+1} | M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t}) &= \sum_{S_{t+1}} P(Y_{t+1}, S_{t+1} | M^*_{1:t-1}(M_t), M_{t+1}, M_{t+1}, Y_{1:t}) \\ &= \sum_{S_{t+1}} P(Y_{t+1} | S_{t+1}) \tau(M_t, S_{t+1}, M_{t+1}) \end{aligned} \tag{A.11}$$

where $\tau(M_t, S_{t+1}, M_{t+1})$, the precursor to the $t+1$-frame predicted posterior, is defined as follows.

$$\tau(M_t, S_{t+1}, M_{t+1}) \triangleq P(S_{t+1} | M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t}) \tag{A.12}$$

This precursor is updated from the previously stored posterior, $\mu^*(M_t, S_t)$, accordingly:

$$\begin{aligned}
\tau(M_t, S_{t+1}, M_{t+1}) &= P\left(S_{t+1}|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t}\right) \\
&= \sum_{S_t} P\left(S_t, S_{t+1}|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t}\right) \\
&= \sum_{S_t} P\left(S_t|M^*_{1:t-1}\left(M_t\right), M_t, Y_{1:t}\right) P\left(S_{t+1}|M_t, M_{t+1}, S_t\right) \\
&= \sum_{S_t} \mu^*\left(M_t, S_t\right) P\left(S_{t+1}|M_t, M_{t+1}, S_t\right) \qquad \text{(A.13)}
\end{aligned}$$

The third step follows from the conditional independence relations indicated by the factorization (A.3).

Now, defining the precursor to the $t+1$-frame filtered posterior:

$$\mu(M_t, S_{t+1}, M_{t+1}) \triangleq P(S_{t+1}|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t+1}) \qquad \text{(A.14)}$$

it is easily verified:

$$\begin{aligned}
\mu(M_t, S_{t+1}, M_{t+1}) &= \frac{\mu_0(M_t, S_{t+1}, M_{t+1})}{P(Y_{t+1}|M^*_{1:t+1}(M_t), M_t, M_{t+1}, Y_{1:t})} \\
P(Y_{t+1}|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t}) &= \sum_{S_{t+1}} \mu_0(M_t, S_{t+1}, M_{t+1}) \qquad \text{(A.15)}
\end{aligned}$$

where

$$\begin{aligned}
\mu_0(M_t, S_{t+1}, M_{t+1}) &\triangleq P(S_{t+1}, Y_{t+1}|M^*_{1:t-1}(M_t), M_t, M_{t+1}, Y_{1:t}) \\
&= P(Y_{t+1}|S_{t+1})\tau(M_t, S_{t+1}, M_{t+1}) \qquad \text{(A.16)}
\end{aligned}$$

By substituting (A.14) into the approximation (A.5), and then the result into

(A.9), the update of $J(M_{t+1})$ simplifies as follows.

$$\begin{aligned} J(M_{t+1}) &\approx \frac{1}{P(Y_{t+1}|Y_t)} \max_{M_t} \max_{M_{1:t-1}} P(M_{1:t}|Y_{1:t}) P(M_{t+1}|M_t) \sum_{S_{t+1}} \mu_0(M_t, S_{t+1}, M_{t+1}) \\ &= \frac{1}{P(Y_{t+1}|Y_t)} \max_{M_t} J(M_t) P(M_{t+1}|M_t) \sum_{S_{t+1}} \mu_0(M_t, S_{t+1}, M_{t+1}) \end{aligned} \tag{A.17}$$

Then for each M_{t+1}, the value of M_t achieving the maximum on the r.h.s. of (A.17) is stored as $M_t^*(M_{t+1})$. Finally, the filtered and smoothed posteriors may be updated from the respective precursors:

$$\begin{aligned} \tau^*(M_{t+1}, S_{t+1}) &= \tau(M_t^*(M_{t+1}), S_{t+1}, M_{t+1}) \\ \mu^*(M_{t+1}, S_{t+1}) &= \mu(M_t^*(M_{t+1}), S_{t+1}, M_{t+1}) \end{aligned} \tag{A.18}$$

Hence, the filtering updates for $J(M_{t+1})$, $M_t^*(M_{t+1})$, $\tau^*(M_{t+1}, S_{t+1})$, and $\mu^*(M_{t+1}, S_{t+1})$ are now expressed in terms of the component distributions on the r.h.s. of (A.3), as desired. Strictly speaking, it is not necessary to propagate $\tau^*(M_t, S_t)$; however, storing the latter for $t \geq 2$ may reduce computations in the smoothing pass.

To initialize the filtering pass, it becomes necessary to supply $\mu^*(S_1, M_1)$ and $J(M_1)$, as follows.

$$\begin{aligned} \mu^*(S_1, M_1) &= P(M_1, S_1|Y_1) \\ &= \frac{P(S_1|M_1)P(Y_1|S_1)}{\sum_{S_1} P(S_1|M_1)P(Y_1|S_1)} \\ J(M_1) &= P(M_1|Y_1) \\ &= \frac{P(M_1)\sum_{S_1} P(S_1|M_1)P(Y_1|S_1)}{\sum_{M_1} P(M_1)\sum_{S_1} P(S_1|M_1)P(Y_1|S_1)} \end{aligned} \tag{A.19}$$

The filtering recursions, as derived via (A.9 – A.18) may be summarized:

$$\begin{aligned}
\tau\left(M_t, S_{t+1}, M_{t+1}\right) &= \sum_{S_t} \mu^*\left(M_t, S_t\right) P\left(S_{t+1} | M_t, M_{t+1}, S_t\right) \\
\mu_0\left(M_t, S_{t+1}, M_{t+1}\right) &= P\left(Y_{t+1} | S_{t+1}\right) \tau\left(M_t, S_{t+1}, M_{t+1}\right) \\
\Sigma_0\left(M_t, M_{t+1}\right) &= \sum_{S_{t+1}} \mu_0\left(M_t, S_{t+1}, M_{t+1}\right) \\
J_0\left(M_t, M_{t+1}\right) &= J\left(M_t\right) P\left(M_{t+1} | M_t\right) \Sigma_0\left(M_t, M_{t+1}\right) \\
\mu\left(M_t, S_{t+1}, M_{t+1}\right) &= \frac{\mu_0\left(M_t, S_{t+1}, M_{t+1}\right)}{\Sigma_0\left(M_t, M_{t+1}\right)} \\
M_t^*\left(M_{t+1}\right) &= \underset{M_t}{\operatorname{argmax}}\ J_0\left(M_t, M_{t+1}\right) \\
J\left(M_{t+1}\right) &= \frac{J_0\left(M_t, M_{t+1}\right)}{P\left(Y_{t+1} | Y_t\right)} \\
\mu^*\left(M_{t+1}, S_{t+1}\right) &= \mu\left(M_t^*(M_{t+1}), S_{t+1}, M_{t+1}\right) \\
\tau^*\left(M_{t+1}, S_{t+1}\right) &= \tau\left(M_t^*(M_{t+1}), S_{t+1}, M_{t+1}\right)
\end{aligned} \tag{A.20}$$

The initialization (A.19) and filtering recursions (A.20) verify the corresponding relations in Section 3.7.1 (3.76, 3.77), as was to be shown.

The goal of the smoothing pass is to supply the optimal mode trajectory $M_{1:N}^*$ and the smoothed posterior $\sigma^*(S_{1:N})$ according to (A.1) and (A.2). As such, we initialize this pass by taking M_N^* as the maximum *a posteriori* choice, from the definitions in Table A.1:

$$M_N^* = \underset{M_N}{\operatorname{argmax}}\ J(M_N) \tag{A.21}$$

Then, via the nesting relation (A.8), past values obey the recursion:

$$M_t^* = M_t^*(M_{t+1}^*) \quad \forall t \in 1 : N-1 \tag{A.22}$$

At the time that M_t^* is known, $\sigma^*(S_t)$ may be updated from $\sigma^*(S_{t+1})$ and the posteriors computed in the filtering pass, $\mu^*(M_t, S_t)$ and $\tau^*(M_t, S_t)$:

$$\begin{aligned}
\sigma^*(S_t) &= P(S_t|M_{1:N}^*, Y_{1:N}) \\
&= \sum_{S_{t+1}} P(S_{t+1}|M_{1:N}^*, Y_{1:N}) P(S_t|S_{t+1}, M_{1:N}^*, Y_{1:N}) \\
&= \sum_{S_{t+1}} P(S_{t+1}|M_{1:N}^*, Y_{1:N}) P(S_t|S_{t+1}, M_{1:t+1}^*, Y_{1:t}) \\
&= P(S_t|M_{1:t}^*, Y_{1:t}) \sum_{S_{t+1}} \frac{P(S_{t+1}|M_{1:N}^*, Y_{1:N})}{P(S_{t+1}|M_{1:t+1}^*, Y_{1:t})} P(S_{t+1}|S_t, M_t^*, M_{t+1}^*) \\
&= \mu^*(S_t, M_t^*) \sum_{S_{t+1}} \frac{\sigma^*(S_{t+1}) P(S_{t+1}|S_t, M_t^*, M_{t+1}^*)}{\tau^*(S_{t+1}, M_{t+1}^*)} \qquad \text{(A.23)}
\end{aligned}$$

Finally, the smoothed posterior is initialized:

$$\sigma^*(S_N) = \mu^*(M_N^*, S_N) \qquad \text{(A.24)}$$

To conclude the derivation, we note that the recursion (A.23) and associated initialization (A.24) verify the corresponding relations in Section 3.7.1 (3.78, 3.79), as was to be shown.

Appendix B

Learning the mode transition dependence

The purpose of this appendix is to derive the expectation-maximization (EM) algorithm steps discussed in Section 3.7.2. The appendix consists of two parts. First, Section B.1 derives the overall approach as summarized by (3.84) and (3.85). However, this approach depends on the smoothed pairwise mode posterior, $P(M_t, M_{t+1}|Y_{1:N})$, for all $t \in 1\!:\!N-1$. To this end, Section B.2 derives the Bayesian inference methodology responsible for computing this posterior in an efficient manner.

B.1 Derivation of EM approach

To begin, define:

$$\begin{aligned} p_{k|j} &\triangleq P\left(M_{t+1} = k | M_t = j\right) \quad \forall j, k \in \mathcal{M} \\ \theta_M &\triangleq \text{Vec}\left(\bigcup_{j \in M} \bigcup_{k \in \mathcal{S}_j} \left\{p_{k|j}\right\}\right) \end{aligned} \tag{B.1}$$

where $\mathcal{S}_j \subset \mathcal{M}$ denotes the set of possibilities for k for which $p_{k|j}$ represents a transition probability in the standard note evolution grammar (3.47), which we recall as

follows:

$$\begin{aligned} \text{'OT'} &\rightarrow \text{'CT','CP'} \\ \text{'OP'} &\rightarrow \text{'CP','N'} \\ \text{'CT'} &\rightarrow \text{'CT','CP'} \\ \text{'CP'} &\rightarrow \text{'CP','N','OT','OP'} \\ \text{'N'} &\rightarrow \text{'OT','OP'} \end{aligned} \tag{B.2}$$

The generic EM algorithm, following [28], begins with an initial guess for θ_M; i.e. $\theta_M^{(0)}$, and proceeds over iterations i, updating $\theta_M = \theta_M^{(i)}$. Each iteration comprises two steps. The *expectation* step computes the expected log likelihood of $M_{1:N}$, $S_{1:N}$, and $Y_{1:N}$ given θ_M where $M_{1:N}$ and $S_{1:N}$ are generated according to $P\left(M_{1:N}, S_{1:N}|Y_{1:N}\right)$. That is, we form

$$Q(\theta_M|\theta_M^{(i)}) = E_{P\left(M_{1:N},S_{1:N}|Y_{1:N},\theta_M^{(i)}\right)}\left[\log P\left(M_{1:N}, S_{1:N}, Y_{1:N}|\theta_M\right)\right] \tag{B.3}$$

The *maximization* step chooses $\theta_M^{(i+1)}$ as a value of θ_M maximizing $Q(\theta_M|\theta_M^{(i)})$.

First evaluating the expectation step, the log likelihood decomposes via the factorization (3.42):

$$\begin{aligned} P(M_{1:N}, S_{1:N}, Y_{1:N}) &= P(M_1)P(S_1|M_1)P(Y_1|S_1) \\ &\times \prod_{t=2}^{N} P(M_t|M_{t-1})P(S_t|S_{t-1}, M_{t-1}, M_t)P(Y_t|S_t) \end{aligned} \tag{B.4}$$

Using (B.4), (B.3) may be written:

$$\log P\left(M_{1:N}, S_{1:N}, Y_{1:N}|\theta_M\right) = \text{const} + \sum_{t=1}^{N-1} \log P\left(M_{t+1}|M_t, \theta_M\right) \tag{B.5}$$

where the "const" term absorbs terms which do not depend on θ_M. Hence, in place

of $Q(\theta_M|\theta_M^{(i)})$, it becomes equivalent to maximize $Q'(\theta_M|\theta_M^{(i)})$:

$$\begin{aligned} Q'(\theta_M|\theta_M^{(i)}) &\triangleq E_{P\left(M_{1:N},S_{1:N}|Y_{1:N},\theta_M^{(i)}\right)} \sum_{t=1}^{N-1} \log P\left(M_{t+1}|M_t,\theta_M\right) \\ &= \sum_{t=1}^{N-1} E_{P\left(M_t,M_{t+1}|Y_{1:N},\theta_M^{(i)}\right)} \log P\left(M_{t+1}|M_t,\theta_M\right) \\ &= \sum_{t=1}^{N-1} \sum_{j\in\mathcal{M}} \sum_{k\in\mathcal{M}} \log p_{k|j} P(M_t=j, M_{t+1}=k|Y_{1:N},\theta_M^{(i)}) \end{aligned} \tag{B.6}$$

Since the terms in the inner summation where $k \in \mathcal{M}\backslash\mathcal{S}_j$ do not depend on θ_M, as evident from (B.1), they may be neglected. Hence, it becomes equivalent to maximize:

$$Q''(\theta_M|\theta_M^{(i)}) \triangleq \sum_{t=1}^{N-1} \sum_{j\in\mathcal{M}} \sum_{k\in\mathcal{S}_j} \log p_{k|j} P(M_t=j, M_{t+1}=k|Y_{1:N},\theta_M^{(i)}) \tag{B.7}$$

This maximization is constrained by the fact for each $j \in \mathcal{M}$, $\left\{p_{k|j}\right\}_{k\in\mathcal{M}}$ forms a probability distribution, i.e.:

$$\begin{aligned} p_{k|j} &\geq 0 \quad \forall j,k \in \mathcal{M} \\ \sum_{k\in\mathcal{M}} p_{k|j} &= 1, \quad \forall j \in \mathcal{M} \end{aligned} \tag{B.8}$$

To accomplish the constrained minimization of (B.7), we form the Lagrangian:

$$J(\theta_M) = Q''(\theta_M|\theta_M^{(i)}) + \sum_{j\in\mathcal{M}} \lambda_j \left(\sum_{k\in\mathcal{M}} p_{k|j} - 1 \right) \tag{B.9}$$

Differentiating $J(\theta_M)$ with respect to each free parameter, $p_{k|j} \in \mathcal{S}_j$ (for all $j \in \mathcal{M}$) obtains as follows.

$$p_{k|j}, k\in\mathcal{S}_j = \frac{-1}{\lambda_j} \sum_{t=1}^{N-1} P\left(M_t=j, M_{t+1}=k|Y_{1:N},\theta_M^{(i)}\right) \tag{B.10}$$

If the $\{\lambda_j\}_{j\in\mathcal{M}}$ are chosen to satisfy the constraints (B.8), we obtain for the maximization step:

$$\theta_M^{(i+1)} = \bigcup_{j\in\mathcal{M}} \bigcup_{k\in\mathcal{S}_j} \left\{ p_{k|j}^{(i+1)} \right\} \tag{B.11}$$

with, for each $j \in \mathcal{M}$, $k \in \mathcal{S}_j$:

$$p_{k|j}^{(i+1)} = \frac{\sum_{t=1}^{N-1} P\left(M_t = j, M_{t+1} = k | Y_{1:N}, \theta_M^{(i)}\right)}{\sum_{k\in\mathcal{M}} \sum_{t=1}^{N-1} P\left(M_t = j, M_{t+1} = k | Y_{1:N}, \theta_M^{(i)}\right)} \tag{B.12}$$

which verifies (3.84, 3.85), as was to be shown.

B.2 Computation of smoothed pairwise mode posteriors

We now address the computation of the unknown terms in (B.12). In other words, we need to compute, for all $t \in 1:N-1$ and $M_t, M_{t+1} \in \mathcal{M}$:

$$\sigma^{(2)}(M_t, M_{t+1}) \triangleq P(M_t, M_{t+1} | Y_{1:N}, \theta_M^{(i)}) \tag{B.13}$$

The inference of $\sigma^{(2)}(M_t, M_{t+1})$ proceeds as a result of the standard Bayesian posterior inference of the hidden variables, $P(M_t, S_t|Y_{1:N})$, for all $t \in 1:N$, with a few modifications. This inference proceeds in two stages, taking as input the conditional distributions on the r.h.s. of the factorization (B.4). In the *filtering* pass, we compute the *filtered posteriors* $P(M_t, S_t|Y_{1:N})$ recursively for all $t \in 1:N$. In the *smoothing* pass, we compute the *smoothed posteriors* $P\left(M_t, S_t|Y_{1:N}\right)$ recursively for $t = N$ down to 1, in conjunction with the *pairwise mode posteriors* $P\left(M_t, M_{t+1}|Y_{1:N}\right)$, to satisfy (B.13). Quantities propagated in filtering and smoothing passes as well as necessary inputs are summarized in Table B.1.

Symbol	Quantity	Description
$\pi(M_1, S_1)$	$P(M_1, S_1)$	Prior
	$P(M_{t+1}\|M_t)$	Mode transition dependence
	$P(S_{t+1}\|S_t, M_t, M_{t+1})$	State transition dependence
	$P(Y_t\|S_t)$	Observation likelihood
$\tau(M_t, S_t)$	$P(M_t, S_t\|Y_{1:t-1})$	Predicted posterior
$\mu(M_t, S_t)$	$P(M_t, S_t\|Y_{1:t})$	Filtered posterior
$\sigma(M_t, S_t)$	$P(M_t, S_t\|Y_{1:T})$	Smoothed posterior
$\sigma^{(2)}(M_t, M_{t+1})$	$P(M_t, M_{t+1}\|Y_{1:T})$	Pairwise mode posterior
$\Psi(M_t, S_{t+1}, M_{t+1})$	$\frac{P(M_t, S_{t+1}, M_{t+1})}{P(M_{t+1}\|M_t)}$	Intermediate
$\phi(M_t, S_t)$	$\frac{P(M_t, S_t\|Y_{1:N})}{P(M_t, S_t\|Y_{1:t-1})}$	Intermediate

Table B.1: *Quantities propagated in standard Bayesian posterior inference*

The filtering pass is initialized accordingly:

$$\begin{aligned} \mu(M_1, S_1) &= P(M_1, S_1|Y_1) \\ &= \frac{P(S_1|M_1)P(Y_1|S_1)}{\sum_{S_1} P(S_1|M_1)P(Y_1|S_1)} \end{aligned} \tag{B.14}$$

The updating of $\mu(M_{t+1}, S_{t+1})$ proceeds in two stages; first, the *time update* computes the predicted posterior $\tau(M_{t+1}, S_{t+1})$, and by so doing computes also the precursor $\Psi(M_t, S_{t+1}, M_{t+1})$, which is not a distribution itself, but actually the ratio of two distributions; second, the *measurement update* computes $\mu(M_{t+1}, S_{t+1})$ from $\tau(M_{t+1}, S_{t+1})$. Now, for the sake of filtering alone it is not strictly necessary to compute the precursor, as the time update computations can just as easily be rearranged to compute only $\tau(M_{t+1}, S_{t+1})$. The main additional cost of computing $\Psi(M_t, S_{t+1}, M_{t+1})$ is storage. However, caching the latter facilitates computation of the pairwise mode posteriors in the smoothing pass. The time update is

$$\begin{aligned} \tau(M_{t+1}, S_{t+1}) &= P(M_{t+1}, S_{t+1}|Y_{1:t}) \\ &= \sum_{M_t} P(M_t, M_{t+1}, S_{t+1}|Y_{1:t}) \end{aligned} \tag{B.15}$$

where

$$\begin{aligned} P\left(M_t, M_{t+1}, S_{t+1}|Y_{1:t}\right) &= \sum_{S_t} P(M_t, S_t, M_{t+1}, S_{t+1}|Y_{1:t}) \\ &= \sum_{S_t} P(M_t, S_t|Y_{1:t})\Big[P(M_{t+1}|M_t, S_t, Y_{1:t}) \\ &\qquad \times P(S_{t+1}|M_t, S_t, M_{t+1}, Y_{1:t})\Big] \\ &= P(M_{t+1}|M_t)\sum_{S_t} P(M_t, S_t|Y_{1:t})P(S_{t+1}|S_t, M_t, M_{t+1}) \\ &= P(M_{t+1}|M_t)\sum_{S_t} \mu(M_t, S_t|Y_{1:t})P(S_{t+1}|S_t, M_t, M_{t+1}) \end{aligned} \tag{B.16}$$

The third step in (B.16) follows from the conditional independence relations indicated by the factorization (B.4).

Then, by definition (Table B.1)

$$\Psi(M_t, S_{t+1}, M_{t+1}) \triangleq \sum_{S_t} \mu(M_t, S_t|Y_{1:t})P(S_{t+1}|S_t, M_t, M_{t+1}) \tag{B.17}$$

the time update (B.15) may be written in terms of $\Psi(M_t, S_{t+1}, M_{t+1})$ and $P(M_{t+1}|M_t)$:

$$\tau(M_{t+1}, S_{t+1}) = \sum_{M_{t+1}} P(M_{t+1}|M_t)\Psi(M_t, S_{t+1}, M_{t+1}) \tag{B.18}$$

The measurement update follows Bayes' rule:

$$\begin{aligned} \mu(M_{t+1}, S_{t+1}) &= P(M_{t+1}, S_{t+1}|Y_{1:t}) \\ &= \frac{P\left(M_{t+1}, S_{t+1}, Y_{t+1}|Y_{1:t}\right)}{\sum_{M_{t+1}, S_{t+1}} P\left(M_{t+1}, S_{t+1}, Y_{t+1}|Y_{1:t}\right)} \end{aligned} \tag{B.19}$$

where

$$\begin{aligned} P\left(M_{t+1}, S_{t+1}, Y_{t+1}|Y_{1:t}\right) &= P\left(M_{t+1}, S_{t+1}|Y_{1:t}\right) P\left(Y_{t+1}|M_{t+1}, S_{t+1}, Y_{1:t}\right) \\ &= \tau(M_{t+1}, S_{t+1})P(Y_{t+1}|M_{t+1}, S_{t+1}) \end{aligned} \tag{B.20}$$

As a result, the measurement update becomes:

$$\mu\left(M_{t+1}, S_{t+1}\right) = \frac{\tau\left(M_{t+1}, S_{t+1}\right) P\left(Y_{t+1} | M_{t+1}, S_{t+1}\right)}{\sum_{M_{t+1}, S_{t+1}} \tau\left(M_{t+1}, S_{t+1}\right) P\left(Y_{t+1} | M_{t+1}, S_{t+1}\right)} \quad \text{(B.21)}$$

This completes the recursion for the filtering pass.

For the smoothing pass, we initialize the posterior $\sigma(M_N, S_N)$, defined in Table B.1, with the final-stage filtered posterior:

$$\begin{aligned} \sigma(M_N, S_N) &= P(M_N, S_N | Y_{1:N}) \\ &= \mu(M_N, S_N) \end{aligned} \quad \text{(B.22)}$$

Assuming that $\sigma(M_{t+1}, S_{t+1})$ has been computed, the update for $\sigma(M_t, S_t)$ is

$$\begin{aligned} \sigma\left(M_t, S_t\right) &= P\left(M_t, S_t | Y_{1:N}\right) \\ &= \sum_{M_{t+1}, S_{t+1}} P\left(M_t, S_t, S_{t+1}, M_{t+1} | Y_{1:T}\right) \\ &= \sum_{M_{t+1}, S_{t+1}} P(M_{t+1}, S_{t+1} | Y_{1:T}) P(M_t, S_t | M_{t+1}, S_{t+1}, Y_{1:t}) \end{aligned} \quad \text{(B.23)}$$

where

$$\begin{aligned}
P(M_t, S_t|S_{t+1}, M_{t+1}, Y_{1:t}) &= \sum_{M_{t+1},S_{t+1}} P\left(M_{t+1}, S_{t+1}|Y_{1:N}\right) P\left(M_t, S_t|M_{t+1}, S_{t+1}, Y_{1:t}\right) \\
&= \sum_{M_{t+1},S_{t+1}} \Big[P\left(S_{t+1}, M_{t+1}|Y_{1:N}\right) \\
&\qquad \times \frac{P\left(M_{t+1}, S_{t+1}|M_t, S_t, Y_{1:t}\right) P\left(M_t, S_t|Y_{1:t}\right)}{P\left(M_{t+1}, S_{t+1}|Y_{1:t}\right)}\Big] \\
&= P(M_t, S_t|Y_{1:t}) \sum_{M_{t+1}} \Big[P(M_{t+1}|M_t) \\
&\qquad \times \sum_{S_{t+1}} P(S_{t+1}|S_t, M_t, M_{t+1}) \frac{P\left(M_{t+1}, S_{t+1}|Y_{1:N}\right)}{P\left(M_{t+1}, S_{t+1}|Y_{1:t}\right)}\Big] \\
&= \mu(M_t, S_t) \sum_{M_{t+1}} \Big[P(M_{t+1}|M_t) \sum_{S_{t+1}} \phi(M_{t+1}, S_{t+1}) \\
&\qquad \times P(S_{t+1}|S_t, M_t, M_{t+1})\Big] \qquad \text{(B.24)}
\end{aligned}$$

where $\phi(M_{t+1}, S_{t+1})$ is as defined in Table B.1.

Similarly, we obtain the pairwise mode posterior:

$$\begin{aligned}
\sigma^{(2)}(M_t, M_{t+1}) &= P(M_t, M_{t+1}|Y_{1:N}) \\
&= \sum_{S_t,S_{t+1}} P(M_{t+1}, S_{t+1}|Y_{1:N}) \frac{P(M_{t+1}, S_{t+1}|M_t, S_t, Y_{1:t}) P(M_t, S_t|Y_{1:t})}{P(M_{t+1}, S_{t+1}|Y_{1:t})} \\
&= P(M_{t+1}|M_t) \sum_{S_{t+1}} \phi(M_{t+1}, S_{t+1}) \Psi(M_t, S_{t+1}, M_{t+1}) \qquad \text{(B.25)}
\end{aligned}$$

To summarize, the filtering and smoothing passes consist of the following:

- **Filtering pass**

  Initialize:

$$\mu(M_1, S_1) = \frac{P(S_1|M_1) P(Y_1|S_1)}{\sum_{S_1} P(S_1|M_1) P(Y_1|S_1)} \qquad \text{(B.26)}$$

For $t \in 1:N-1$, compute:

$$\begin{aligned}
\Psi(M_t, S_t, M_{t+1}) &= \sum_{S_t} \mu(M_t, S_t) P(S_{t+1}|S_t, M_t, M_{t+1}) \\
\tau(M_{t+1}, S_{t+1}) &= \sum_{M_t} P(M_{t+1}|M_t) \Psi(M_t, S_t, M_{t+1}) \\
\mu(M_{t+1}, S_{t+1}) &= \frac{P\left(M_{t+1}, S_{t+1}, Y_{t+1}|Y_{1:t}\right)}{\sum_{M_{t+1}, S_{t+1}} P\left(M_{t+1}, S_{t+1}, Y_{t+1}|Y_{1:t}\right)}
\end{aligned} \tag{B.27}$$

For $t \in 1:N$, store $\mu(M_t, S_t)$; for $t \in 1:N-1$, store $\mu(M_t, S_{t+1}, M_{t+1})$; for $t \in 2:N$; store $\tau(M_t, S_t)$.

- **Smoothing pass**
  Initialize:

$$\sigma(M_N, S_N) = \mu(M_N, S_N) \tag{B.28}$$

Then for $t = N-1$ down to 1, compute:

$$\begin{aligned}
\phi(M_{t+1}, S_{t+1}) &= \frac{\sigma(M_{t+1}, S_{t+1})}{\tau(M_{t+1}, S_{t+1})} \\
\sigma(M_t, S_t) &= \mu(M_t, S_t) \sum_{M_{t+1}} P(M_{t+1}|M_t) \sum_{S_{t+1}} \Big[\phi(M_{t+1}, S_{t+1}) \\
&\qquad \times P(S_{t+1}|S_t, M_t, M_{t+1})\Big] \\
\sigma^{(2)}(M_t, M_{t+1}) &= P(M_{t+1}|M_t) \sum_{S_{t+1}} \phi(M_{t+1}, S_{t+1}) \Psi(M_t, S_{t+1}, M_{t+1})
\end{aligned} \tag{B.29}$$

The pairwise mode posterior, $\sigma^{(2)}(M_t, M_{t+1})$, may be substituted into (3.84) and (3.85), to complete the EM iteration, as desired.

Bibliography

[1] M. Abe and J.O. Smith III. Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT. In *Proceedings of the 117th AES Convention*, San Francisco, 2004.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[3] P. Allen and R. Dannenberg. Tracking musical beats in real time. In *Proceedings of the International Computer Music Conference*, pages 140–143, Glasgow, 1990.

[4] J. Amuedo. Personal communication, 2005.

[5] R. Andre-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(1), 1988.

[6] B.S. Atal and S.L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50:637–655, 1971.

[7] M. Basseville. Edge detection using sequential methods for change in level - Part II - Sequential detection of change in mean. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(1):32–50, 1981.

[8] M. Basseville and A. Benveniste. Sequential detection of abrupt changes in spectral characteristics of digital signals. *IEEE Transactions on Information Theory*, 29(5):709–723, 1983.

[9] M. Basseville, B. Espiau, and J. Gasnier. Edge detection using sequential methods for change in level - Part I - A sequential edge detection algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(1):24–31, 1981.

[10] J.P. Bello. Phase-based note onset detection for music signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, 2003.

[11] J. Bensa, S. Bilbao, R. Kronland-Martinet, and J.O. Smith III. Piano string modeling: from partial differential equations to digital wave-guide model. *Journal of the Acoustical Society of America*, 112(5):2239–2259, 2002.

[12] J. Berger. Personal communication, 2004.

[13] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, University of Oxford, UK, 1995.

[14] C. Bishop. *Introduction to Digital Audio Coding and Standards*. Kluwer, Dordrecht, the Netherlands, 2003.

[15] K. Brandenburg. MP3 and AAC explained. In *Proceedings of the 17th AES International Conference on High Quality Audio Coding*, Florence, Italy, 1999.

[16] K. Brandenburg and M. Bosi. Overview of MPEG audio: current and future standards for low bit rate audio coding. *Journal of the Audio Engineering Society*, 45(1/2):4–21, 1997.

[17] M.A. Casey. *Auditory Group Theory with Applications to Statistical Basis Methods for Structured Audio*. PhD thesis, Massachusetts Institute of Technology, Media Laboratory, Cambridge, MA, 1998.

[18] A.T. Cemgil. *Bayesian Music Transcription*. PhD thesis, Radboud University, Nijmegen, the Netherlands, 2004.

[19] A.T. Cemgil. Polyphonic pitch identification and Bayesian inference. In *Proceedings of the International Computer Music Conference*, Miami, FL, 2004.

[20] A.T. Cemgil, H.J. Kappen, and D. Barber. Generative model based polyphonic music transcription. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003.

[21] A.T. Cemgil, H.J. Kappen, P. Desain, and H. Honing. On tempo tracking: tempogram representation and Kalman filtering. In *Proceedings of the International Computer Music Conference*, pages 352–355, Berlin, 2000.

[22] E. Chew and Y. Chen. Mapping midi to the spiral array: disambiguating pitch spellings. In *Proceedings of the Eighth INFORMS Computer Society Conference*, pages 259–275, Chandler, AZ, 2003.

[23] E. Chew and Y. Chen. Real-time pitch spelling using the spiral array. *Computer Music Journal*, 29(2), 2005.

[24] M. Cooper and J. Foote. Audio retrieval by rhythmic similarity. In *Proceedings of the Third International Symposium on Musical Information Retrieval*, pages 81–85, Paris, 2002.

[25] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Interscience, Yorktown Heights, NY, 1991.

[26] H. Cramer. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ, 1946.

[27] D. Crisan and A. Doucet. Convergence of sequential Monte Carlo methods. Technical Report CUED/F-INFENG/TR381, Cambridge University Department of Engineering, 2000.

[28] A.P. Dempster, J. Laird, and J. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, B(39):1–38, 1977.

[29] W. D'Haes, X. Rodet, and D. V. Dyck. Control parameter estimation for a physical model of a trumpet using pattern recognition. In *Proceedings of the 1st IEEE Benelux Workshop on Model-based Processing and Coding of Audio (MPCA-2002)*, Leuven, Belgium, 2002.

[30] S. Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1):39–58, 2001.

[31] M. Dolson. The phase vocoder: a tutorial. *Computer Music Journal*, 10(4):14–27, 1986.

[32] A. Doucet. On sequential simulation-based methods for Bayesian filtering. Technical Report CUED/F-INFENG/TR310, Cambridge University Department of Engineering, 1998.

[33] C. Duxbury, J.P. Bello, M. Davies, and M. Sandler. A combined phase and amplitude based approach to onset detection for audio segmentation. In *Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services*, Munich, 2003.

[34] C. Duxbury, J.P. Bello, M. Davies, and M. Sandler. Complex domain onset detection for musical audio signals. In *Proceedings of the 6th International Conference on Digital Audio Effects*, London, 2003.

[35] C. Duxbury, M. Davies, and M. Sandler. Improved time-scaling of musical audio using phase locking at transients. In *Proceedings of the 112th AES Convention*, Munich, 2002.

[36] B. Edler. Codierung von audiosignalen mit uberlappender transformation und adaptiven fensterfunktionen. *Frequenz*, 43(9):252–256, 1989.

[37] P.A. Esquef, M. Karjalainen, and V. Valimaki. Frequency-zooming ARMA modeling for analysis of noisy string instrument tones. *EURASIP Journal on Applied Signal Processing: Special Issue on Digital Audio for Multimedia Communications*, 10:953–967, 2003.

[38] P. Fearnhead. *Sequential Monte Carlo Methods in Filter Theory*. PhD thesis, University of Oxford, 1998.

[39] K. Fitz, L. Haken, and P. Christiansen. Transient preservation under transformation in an additive sound model. In *Proceedings of the International Computer Music Conference*, Berlin, 2000.

[40] W.J. Fitzgerald. Markov chain Monte Carlo methods with applications to signal processing. *Elsevier Signal Processing*, 81(1):3–18, 2001.

[41] J. Flanagan and R. Golden. Phase vocoder. *Bell System Technical Journal*, 45:1493–1509, 1966.

[42] H. Fletcher, E.D. Blackham, and R. Stratton. Quality of piano tones. *Journal of the Acoustical Society of America*, 34(6):749–761, 1961.

[43] J. Foote, M. Cooper, and U. Nam. Audio retrieval by rhythmic similarity. In *Proceedings of the Third International Symposium on Musical Information Retrieval*, pages 265–272, Paris, 2002.

[44] D. Gabor. Theory of communication. *Journal of the Institute of Electronic Engineers*, 93(26):429–457, 1946.

[45] D. Gang and J. Berger. Modeling the degree of realized expectation in functional tonal music: a study of perceptual and cognitive modeling using neural networks. In *Proceedings of the International Computer Music Conference*, pages 454–457, Hong Kong, 1996.

[46] S. Godsill and M. Davy. Bayesian harmonic models for musical pitch estimation and analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, 2002.

[47] J. Goldstein. An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, 54:1496–1516, 1973.

[48] M.M. Goodwin and J. Laroche. Audio segmentation by feature-space clustering using linear discriminant analysis and dynamic programming. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003.

[49] M. Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, 30(2):159–171, 2001.

[50] F. Gouyon. Detection and modeling of transient regions in musical signals. Master's thesis, 1999. Report CCRMA, Stanford University/ DEA SIC thesis, ENSEEIHT Toulouse, France.

[51] S.W. Hainsworth. *Techniques for the Automated Analysis of Musical Audio*. PhD thesis, University of Cambridge, 2003.

[52] T. Helie, C. Vergez, J. Levine, and X. Rodet. Inversion of a physical model of a trumpet. In *Proceedings of the 1999 IEEE Conference on Decision and Control*, Phoenix, AZ, 1999.

[53] H.L. Helmholtz. *Die Lehre von dem Tonempfindungen als physiologische Grundlage fr die Theorie der Musik*. Braunschweig: F. Vieweg, 1870.

[54] C. Hory, N. Martin, and A. Chehikian. Spectrogram segmentation by means of statistical features for non-stationary signal interpretation. *IEEE Transactions on Signal Processing*, 50(12):2915–2925, 2002.

[55] E.T. Jaynes. On the rationale of maximum entropy methods. *Proceedings of the IEEE*, pages 939–952, 1992.

[56] T. Jehan. Musical signal parameter estimation. Master's thesis, 1997. Report CNMAT, University of California at Berkeley/ MS Thesis in Electrical Engineering and Computer Sciences from IFSIC, University of Rennes 1, Rennes, France.

[57] T. Kailath, A. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall, Englewood Cliffs, NJ, 2000.

[58] M. Karjalainen, P.A. Esquef, P. Antsalo, A. Makivirta, and V. Valimaki. Frequency-zooming ARMA modeling of resonant and reverberant systems. *Journal of the Audio Engineering Society*, 50(12):1012–1039, 2002.

[59] K. Kashino and S. Godsill. Bayesian estimation of simultaneous musical notes based on frequency domain modeling. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, 2004.

[60] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka. Application of Bayesian probability network to music scene analysis. In *Working Notes of ICJAI Workshop of Computational Auditory Scene Analysis*, Montreal, 1995.

[61] S. Kay. *Fundamentals of Statistical Signal Processing II: Detection Theory*. Prentice Hall, Englewood Cliffs, N.J., 1998.

[62] D. Knuth, I. Vardi, and R. Richberg. 6581 (the asymptotic expansion of the middle binomial coefficient). *American Mathematical Monthly*, 97(7):626–630, 1990.

[63] L.H. Koopmans. *The spectral analysis of time series*. Academic Press, New York, NY, 1974.

[64] C.L. Krumhansl. Music psychology and music theory: problems and prospects. *Music Theory Spectrum*, 17(1):53–80, 1995.

[65] D. Lang and N. de Freitas. Beat tracking the graphical model way. In *Proceedings of Neural Information and Processing Systems (NIPS-17)*.

[66] J. Laroche. A new analysis/synthesis system of musical signals using Prony's method: application to heavily damped percussive sounds. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 2053–2056, Glasgow, 1989.

[67] J. Laroche and M. Dolson. Phase-vocoder: about this phasiness business. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 1997.

[68] J. Laroche and M. Dolson. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, 7(3):323–332, 1999.

[69] S. Larson and S. McAdams. Musical forces and melodic expectations: comparing computer models and experimental results. *Music Perception*, 21(4):457–498, 2004.

[70] S.L. Lauritzen. Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–1108, 1992.

[71] R. Leistikow. *Bayesian Modeling of Musical Expectations using Maximum Entropy Stochastic Grammars*. PhD thesis, Stanford University, Department of Music, Stanford, CA, 2006. To be published.

[72] R. Leistikow, H. Thornburg, J.O. Smith III, and J. Berger. Bayesian identification of closely-spaced chords from single-frame stft peaks. In *Proceedings of the 7th International Conference on Digital Audio Effects*, Naples, Italy, 2004.

[73] F. Lerdahl. *Tonal Pitch Space*. Oxford University Press, Oxford, 2001.

[74] S. Levine. *Audio representations for data compression and compressed domain processing*. PhD thesis, Stanford University, Department of Electrical Engineering, Stanford, CA, 1998.

[75] S. Levine and J.O. Smith III. A sines+transients+noise audio representation for data compression and time-pitch-scale modifications. In *Proceedings of the 105th Audio Engineering Society Convention*, San Francisco, CA, 1998.

[76] S. Levine and J.O. Smith III. A switched parametric and transform audio coder. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, 1999.

[77] S. Levine, T. Verma, and J.O. Smith III. Multiresolution sinusoidal modeling for wideband audio with modifications. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, 1998.

[78] Y. Liu and J.O. Smith III. Watermarking sinusoidal audio representations by quantization index modulation in multiple frequencies. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Montreal, 2004.

[79] H.L. Lu and J.O. Smith III. Joint estimation of vocal tract filter and glottal source waveform via convex optimization. In *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 1999.

[80] S. Malone. Much ado about humming: the Gould descant. *Glenn Gould Magazine*, 6(1):35–38, 2000.

[81] R.J. McAulay and T.F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. ASSP*, 34(4):744–754, 1986.

[82] B.C. Moore. *An Introduction to the Psychology of Hearing.* Academic Press, London, 1997.

[83] J. Moorer. The use of the phase vocoder in computer music applications. *Journal of the Audio Engineering Society*, 26(1/2):42–45, 1978.

[84] K. Murphy. Filtering, smoothing, and the junction tree algorithm. `http://citeseer.nj.nec.com/361819.html`, 1998.

[85] E. Narmour. *The Analysis and Cognition of Basic Melodic Structures: the Implication-Realization Model*. University of Chicago Press, Chicago, 1990.

[86] J. Pampin. ATS – a Lisp environment for spectral modeling. In *Proceedings of the International Computer Music Conference*, Beijing, 2000.

[87] V. Pavlovic, J.M. Rehg, and T. Cham. A dynamic Bayesian network approach to tracking using learned switching dynamic models. In *Proceedings of the International Workshop on Hybrid Systems*, Pittsburgh, PA, 2000.

[88] B. Porat. *Digital Processing of Random Signals*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[89] M.R. Portnoff. Time-scale modification of speech based on short-time Fourier analysis.

[90] M.R. Portnoff. Implementation of the digital phase vocoder using the fast Fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(2):243–248, 1976.

[91] H. Purnhagen, B. Edler, and C. Ferekidis. Object-based analysis/synthesis audio coder for very low bit rates. In *Proceedings of the 104th Audio Engineering Society Convention*, Amsterdam, 1998.

[92] Y. Qi, T. Minka, and R. Picard. Bayesian spectrum estimation of unevenly sampled nonstationary data. Technical Report Vismod-TR-556, MIT Media Lab, 2002.

[93] T. Quatieri, R. Dunn, and T. Hanna. A subband approach to time-scale expansion of complex acoustic signals. *IEEE Transactions on Speech and Audio Processing*, 3(6):515–519, 1995.

[94] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[95] C. Raphael. Automatic segmentation of acoustic musical signals using hidden markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:360–370, 1999.

[96] C. Raphael. Automatic transcription of piano music. In *Proceedings of the Third International Symposium on Musical Information Retrieval*, pages 81–85, Paris, 2002.

[97] D.C. Rife and R.R. Boorstyn. Single-tone parameter estimation from discrete-time observations. *IEEE Transactions on Information Theory*, 20(5):591–598, 1974.

[98] J. Risset and M. Mathews. Analysis of musical instrument tones. *Physics Today*, 22(2), 1969.

[99] X. Rodet. Stability/instability of periodic solutions and chaos in physical models of musical instruments. In *Proceedings of the International Computer Music Conference*, pages 352–355, Copenhagen, 1994.

[100] M. Saunders and B. Kim. PDCO: primal-dual interior method for convex objectives. `http://www.stanford.edu/group/SOL/software/pdco.html`, 2003.

[101] E.G. Schellenberg. Simplifying the implication-realization model of melodic expectancy. *Music Perception*, 14:295–318, 1997.

[102] W.A. Schloss. *On the Automatic Transcription of Percussive Music: From Acoustic Signal to High-Level Analysis*. PhD thesis, Stanford University, Stanford, CA, 1985.

[103] S. Serafin. *The Sound of Friction: Real-Time Models, Playability and Musical Applications*. PhD thesis, Stanford University, Stanford, CA, 2004.

[104] S. Serafin, J.O. Smith III, H. Thornburg, F. Mazzella, G. Thonier, and A. Tellier. Data-driven identification and computer animation of a bowed string model. In *Proceedings of the International Computer Music Conference*, Havana, Cuba, 2001.

[105] X. Serra. *A System for Sound Analysis-Transformation-Resynthesis Based on a Deterministic Plus Stochastic Decomposition*. PhD thesis, Stanford University, Stanford, CA, 1989.

[106] X. Serra and J.O. Smith III. Spectral modeling synthesis. In *Proceedings of the International Computer Music Conference*, Columbus, OH, 1989.

[107] A. Sheh and D.P. Ellis. Chord segmentation and recognition of EM-trained hidden Markov models. In *Proceedings of the 4th International Symposium on Music Information Retrieval*, Baltimore, MD, 2003.

[108] J.O. Smith III. *Spectral Audio Signal Processing*. W3K Publishing, Stanford, CA, 2006. To be published.

[109] J.O. Smith III and P. Gossett. A flexible sampling-rate conversion method. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 19.4.1–19.4.4, San Diego, CA, 1984. `http://ccrma.stanford.edu/~jos/src`.

[110] J.O. Smith III and X. Serra. PARSHL: an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. San Francisco, CA, 1987. `http://ccrma.stanford.edu/~jos/parshl/parshl.html`.

[111] T. Svendsen and F.K. Soong. On the automatic segmentation of speech signals. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 77–80, Dallas, TX, 1987.

[112] J. Tabrikian, S. Dubnov, and Y. Dickalov. Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model. *IEEE Transactions on Speech and Audio Processing*, 12(1):76–87, 2004.

[113] H. Takeda, T. Nishimoto, and S. Sagayama. Automatic rhythm transcription from multiphonic MIDI signals. In *Proceedings of the International Conference on Music Information Retrieval*, Baltimore, MD, 2003.

[114] H. Terasawa, M. Slaney, and J. Berger. Perceptual distance in timbre space. In *Proceedings of the International Conference on Auditory Display*, Limerick, Ireland, 2005. To appear.

[115] H. Thornburg and F. Gouyon. A flexible analysis-synthesis method for transients. In *Proceedings of the International Computer Music Conference*, pages 400–403, Berlin, 2000.

[116] H. Thornburg and R.J. Leistikow. Analysis and resynthesis of quasi-harmonic sounds: an iterative filterbank approach. In *Proceedings of the 6th International Conference on Digital Audio Effects*, London, 2003.

[117] H. Thornburg and R.J. Leistikow. An iterative filterbank approach for extracting sinusoidal parameters from quasiharmonic sounds. In *Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003.

[118] W. Verhelst and M. Roelands. An overlap-add technique based on waveform similarity (WSOLA) for high-quality time-scale modification of speech. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 554–557, Minneapolis, MN, 1993.

[119] P.J. Walmsley, S.J. Godsill, and P.J. Rayner. Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters. In *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 1999.

[120] Y. Wang and M. Vilermo. The modified discrete cosine transform for audio coding and error concealment. In *Proceedings of the AES International Conference on Virtual, Synthetic, and Entertainment Audio (AES22)*, Espoo, Finland, 2002.

[121] Wikipedia. Occam's razor. `http://en.wikipedia.org/wiki/Occam's_razor`.

[122] E. Wold. *Nonlinear Parameter Estimation of Acoustic Models*. PhD thesis, University of California at Berkeley, Berkeley, CA, 1987.