Melody Extraction and Musical Onset Detection from Framewise STFT Peak Data

Harvey D. Thornburg, Student Member, IEEE, Randal J. Leistikow, Student Member, IEEE, and Jonathan Berger

Abstract—We propose a probabilistic method for the joint segmentation and melody extraction for musical audio signals which arise from a monophonic score. The method operates on framewise short-time Fourier transform (STFT) peaks, enabling a computationally efficient inference of note onset, offset, and pitch attributes while retaining sufficient information for pitch determination and spectral change detection. The system explicitly models note events in terms of transient and steady-state regions as well as possible gaps between note events. In this way, the system readily distinguishes abrupt spectral changes associated with musical onsets from other abrupt change events. Additionally, the method may incorporate melodic context by modeling note-to-note dependences. The method is successfully applied to a variety of piano and violin recordings containing reverberation, effective polyphony due to legato playing style, expressive pitch variations, and background voices. While the method does not provide a sample-accurate segmentation, it facilitates the latter in subsequent processing by isolating musical onsets to frame neighborhoods and identifying possible pitch content before and after the true onset sample location.

I. INTRODUCTION

MUSICAL signals are typically organized as sequences of note events. These events often arise from a performer’s action, and the pitch content is expected to be similar for the duration of each event. The beginning of each note event is termed a note onset, and the end a note endpoint. For the melody extraction task, we aim to estimate the onset time, endpoint time, and pitch of each event. This extraction is so far restricted to nominally monophonic signals, which are signals that may arise from a monophonic score. Of course, the actual recording may contain significant overlaps in the pitch content over time, especially at note transition points, where reverberation or performance techniques sustain notes until after subsequent ones have begun.

Due to computational considerations, the extraction operates on framewise sequences of short-time Fourier transform (STFT) peaks. In most cases these peaks suffice for pitch determination [10], [27], [30]. When tracking pitches over time, however, problems arise where the frame boundaries do not exactly coincide with the pitched portions of note events. Frames may, for instance, contain samples from the transient portion of a note event in which pitch information is not yet salient, cross over the boundary between two notes, contain only background noise, or include information due to other interfering sound sources.

Fortunately, musical signals are highly structured, both at the signal level, in terms of the expected timbral evolution of each note event, and at higher levels, in terms of melodic and rhythmic tendencies. This structure creates context which can be used to predict pitch values and note event boundaries. For instance, prior knowledge of musical key generates expectations favoring certain pitch sequences. Although the key itself is seldom known a priori, we can still benefit from the knowledge that the key is likely to be consistent for long sequences of frames, while being robust to situations in which the key happens to fluctuate more rapidly than expected. To best exploit contextual knowledge in ways that are robust to inherent uncertainties in this knowledge, we propose a probabilistic model which satisfies the following objectives:

- It segments the signal into discrete note events, possibly punctuated by regions of silence or spurious phenomena. Each event is further segmented into transient and steady-state regions. Transient regions account as well for corrupted peak information caused by frame boundary misalignment at the true onset location.
- It distinguishes abrupt spectral changes associated with musical onsets from other, spurious abrupt-change events, such as the performer knocking the microphone, clicks and pops, etc.
- It improves pitch detection across all frames containing significant pitch content in a note event, because the information in all frames is used to infer the content in a given frame.

The desire to fuse information from raw signal observations with prior and contextual knowledge leads naturally to a Bayesian approach, especially when the goal is to minimize the probability of making incorrect decisions about a particular attribute (e.g., the pitch of a note event).

The decision-theoretic optimality of Bayesian methods is briefly reviewed in Section II-B.

1 Perhaps the closest comparable work (in terms of modeling aspects) is that by Cemgil et al., which proposes a generative probabilistic model for note identification in both monophonically and polyphonic cases. Cemgil’s model explicitly encodes note segmentation via a binary (note-off/note-on) indicator variable, the latter influencing the evolution of signal characteristics (frequencies, decay rates, and amplitudes of sinusoidal components). By contrast, the proposed method yields not only a segmentation into individual note events, but also a sub-segmentation of each event into transient and steady-state regions. The sub-segmentation proves useful in applications.
ranging from analysis-synthesis transformations to window switching for audio data compression [18], [7], [29], [8]. In time-scaling, for instance, it is often desirable to dilate/contract the steady-state portions of notes without altering the transient portions [7], [18].

As in the proposed model, the approaches of Cemgil [4], [5] and Hainsworth [11], among others, allow the segmentation to influence the signal characteristics’ evolution, the latter varying smoothly during note events and abruptly across event boundaries. However, there exists an important distinction between these methods and the present approach in terms of the encoding of these characteristics which has profound implications on the computability of exact Bayesian inference. The proposed model, by discretizing relevant signal characteristics (i.e., note pitch as measured by integer value and fractional tuning endpoint, as well as amplitudes of pitched and transient information), obtains an exact inference solution which is computable in linear time, while the exact solution for the models of [4], [5] are at best best quadratic time. Although the authors [11], [5] develop linear-time approximate inference strategies which seem to work well in practice, it may not be possible to check the validity of these approximations in cases where the cost of computing exact inference is prohibitive. The computational cost of the proposed method is analyzed in Section VI. The loss of resolution imposed by the discretization of signal characteristics in the proposed approach fails to be problematic in this sense: first, the primary quantities of interest for segmentation and melody extraction (note values, onset incidences, region labels) are themselves discrete; second, the effects of the discretization may be checked by increasing the grid resolutions and examining whether there is any appreciable change in the result.

The proposed method obtains additional computational savings by operating on a highly reduced, yet quite general feature set, that of framewise STFT peaks. While this approach does not provide a sample-accurate segmentation like some existing methods, [4], [5], [11], sample accuracy may be recovered in a subsequent pass which obtains the precise onset sample locations by preselecting just those frames in which onsets are detected (plus surrounding frames) and performing sample-accurate processing on just those frames. Since onset frame neighborhoods are of limited duration, it may be possible to apply even quadratic-time algorithms to perform the sample-accurate processing. The latter may also be facilitated by partial knowledge, obtained from the framewise pitch content analysis, of signal models before and after change.

Another key innovation of the proposed method is the use of two distinct criteria to perform transcription: global error rate across all frames is minimized for the segmentation variable, then, given that optimal sequence, symbol error rates are minimized in each frame for the signal characteristic variables. Applying a global maximum a posteriori criterion to the segmentation variable yields segmentations in which sequential integrity is preserved, avoiding sequences which are undesirable in transcription, such as those with onsets in successive frames. Section II-B provides a detailed discussion motivating these criteria as the correct ones for the melody extraction and onset detection task.

II. PROBABILISTIC MODEL AND EXTRACTION GOALS

A. Probabilistic Model

The proposed model accounts for the evolution of monophonic signal characteristics governed by a succession of note events punctuated by regions of silence or spurious phenomena. Each note event is further divided into two principal regions, a transient region in which pitch content is ambiguous, followed by a steady-state region in which pitch content is salient. The result is a cyclic succession of regimes (transient→pitch→null), as displayed in Figure 1. The duration of transient or null regimes may be zero or more frames; however, a pitched regime is expected to occupy at least one frame.

We define a musical onset as a function of two successive frames indicating that the transient boundary has been crossed. To represent the cyclic succession of regimes, as well as the onset incidence, we define the following modes which serve to label each frame:

- ‘OT’ – the beginning frame of a transient region, of which there can be at most one per note event.
- ‘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.
- ‘CT’ – the continuation of a transient region in the event the region occupies more than one frame; must follow a previous ‘CT’ or ‘OT’.
- ‘CP’ – the continuation of a pitched region; must follow ‘OP’, ‘CP’, ‘OT’, or ‘CT’.
- ‘N’ – a silent or spurious frame which occurs anytime after the last frame of a note event. A ‘N’ is followed by either another ‘N’ or an onset (‘OT’ or ‘OP’).

The onset incidence is represented by the event that either $M_t = ‘OT’$ or $M_t = ‘OP’$. Such an explicit representation facilitates the modeling of abrupt-change behaviors which characterize the beginnings of note events. For instance, the amplitude of percussive sounds generally undergoes an abrupt increase at onset locations; this behavior may not occur at other regime transitions, such as the transient/pitched boundary. In general, grouping modes into sets with common properties proves useful in modeling the evolution of signal characteristics; Table I summarizes these sets.

Additionally, we define $\mathcal{M}$ as the set of all modes,

$$\mathcal{M} \triangleq \mathcal{P} \cup \mathcal{Q} = \mathcal{O} \cup \mathcal{C} \cup \{‘N’\}$$

![fig1](image-url)
We associate with the frame at time $t$ a variable $M_t \in \mathcal{M}$, which describes the mode associated with that frame. Therefore, we can represent an onset by $M_t \in \mathcal{O}$, and by the constraints of the mode succession, segment the passage into contiguous note event regions.

Recall that in the melody extraction task, we wish at minimum to estimate the pitch of each note region, as well as that region’s onset time and duration. Additionally, we can also extract information potentially useful for a more refined musical transcription by tracking amplitudes and tuning endpoints. Even though note pitches are of primary interest, tunings and amplitudes are likely to exhibit structure across frames in a way that helps both pitch determination and onset detection, hence it becomes necessary to model them regardless of the specific transcription objectives. Under normal circumstances, tuning is likely to be constant or exhibit only local variations due to drifts associated with the onset transients of percussive sounds or expressive modulations such as vibrato. Of course, we must also be robust to the situation where tuning experiences a global shift, for instance, when the playback of a recording from analog tape undergoes a sudden speed change.

We represent the state quantities of note value, tuning, and amplitude as follows:

- $N_t \in \mathcal{S}_N = \{N_{\text{min}}, N_{\text{min}} + 1, \ldots, N_{\text{max}}\}$, where each element of $\mathcal{S}_N$ is an integer representing the MIDI note value (e.g., the note C4, middle C, 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 $M_t \in \mathcal{P}$.
- $A_t^Q \in \mathcal{S}_{AQ}$, where $\mathcal{S}_{AQ}$ is an exponentially spaced set of reference amplitudes active when $M_t \in \mathcal{Q}$.

We define $S_t$, the state at time $t$, to be the collection of valid possibilities for all state quantities:

$$S_t = S_t \cap S_{Tt} \cap (S_A \cup S_{AQ}).$$

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 $M_t \in \mathcal{Q}$. State information is expected to be continuous during the steady-state region of each note event, but may vary abruptly across the incidence of an onset. In this way, we may integrate information across sequences of frames which associate with the same pitch event, while responding as quickly as possible to changes in this information.

Finally, we represent the observations $Y_t$ as the set of all STFT peak frequencies and amplitudes in frame $t$. Peaks are chosen from overlapping, Hamming-windowed, zero-padded frames following the quadratic interpolation method described in [25].

The joint distribution of all variables of interest, $P(M_{0:K}, S_{0:K}, Y_{1:K})$, where $K$ is the number of frames, factors over the directed acyclic graph shown in Figure 2, i.e.:

$$P(M_{0:K}, S_{0:K}, Y_{1:K}) =$$

$$P(M_0, S_0) \prod_{t=1}^K P(M_t|M_{t-1}, M_t, S_{t-1})P(Y_t|S_t),$$

![Directed acyclic graph for probabilistic model](image)

**Fig. 2. Directed acyclic graph for probabilistic model**

**B. Inference and Estimation Goals**

The primary goals are to make optimal decisions concerning successions of onsets, endpoints, note pitches, tunings, and amplitudes over frames. However, the exact optimality criteria differ according to each attribute’s interpretation, as well as its role in the melody extraction task.

Two common criteria are as follows. First, we simply minimize the probability of any error in the attribute sequence. That is, letting $Z_{1:K}$ denote the true attribute sequence, and $\hat{Z}_{1:K}(Y_{1:K})$ any estimate of the true sequence based on the observations $Y_{1:K}$, the optimal estimate, $\hat{Z}_{1:K}^*$, satisfies the following,

$$\hat{Z}_{1:K}^* = \arg\min_{\hat{Z}_{1:K}(Y_{1:K})} P(\hat{Z}_{1:K}(Y_{1:K}) \neq Z_{1:K})$$

It is easily shown [3] that the $\hat{Z}_{1:K}^*$ satisfying (4) maximizes the posterior $P(Z_{1:K}|Y_{1:K})$:

$$\hat{Z}_{1:K}^* = \arg\max_{\hat{Z}_{1:K}} P(Z_{1:K}|Y_{1:K})$$

This choice is commonly called the MAP (maximum a posteriori) estimate.

For long enough sequences and under sufficiently noisy conditions, the probability of $\hat{Z}_{1:K}(Y_{1:K}) \neq Z_{1:K}$ may approach unity. That is, it may be impossible to decode the attribute sequence in an errorless fashion. Now, since an estimate $\hat{Z}_{1:K}(Y_{1:K})$ having nothing to do with the recording being analyzed is valued the same as an estimate that differs only in one symbol from the true sequence $Z_{1:K}$, there seems little incentive for the MAP scheme to favor sequences somehow
"close to" the true sequence. To remedy this, we propose to minimize the symbol error rate, which is the same as minimizing the expected number of symbol errors. This second criterion is given as follows.

$$\hat{Z}^*_1:K = \arg\min_{\hat{Z}_1:K} \sum_{t=1}^{K} P(\hat{Z}_t \neq Z_t)$$  \hspace{1cm} (6)

One may show that $\hat{Z}^*_1:K$ in (6) is equivalently obtained by maximizing the framewise posterior, $P(\hat{Z}_t|Y_1:K)$, for each $t \in 1:K$. This posterior is traditionally called the smoothed posterior.

$$\hat{Z}^*_1:K = \{\hat{Z}^*_t\}_{t=1}^{K}$$

$$\hat{Z}^*_t = \arg\max_{\hat{Z}_t} P(\hat{Z}_t|Y_1:K)$$  \hspace{1cm} (7)

A key disadvantage of the maximized smoothed posterior is that it may fail to preserve the integrity of the entire attribute sequence. This becomes most relevant when considering segmentation (onset/endpoint detection; characterization of contiguous note regions), which is tantamount to estimating the mode sequence $M^*_{1:K}$. For instance, suppose the true sample location of an onset 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 occurred, or to the subsequent frame, for which the change in the spectral content regarding the previous frame is most salient. Now, it is clearly less costly for the estimator $M^*_{1:K}$ to hedge in favor of declaring onsets in both frames, which incurs at most one symbol error, as opposed to declaring a single onset in the wrong frame, which incurs at least two errors. However, the introduction of an extraneous onset is disastrous for the melody extraction task, whereas the shift of the estimated onset location by one frame usually has a negligible effect. In this case the MAP criterion (4) is preferable, since by construction, it is unlikely that the joint posterior $P(M^*_{1:K}|Y_1:K)$ concentrates in sequences containing onsets in adjacent frames. In fact, these sequences are guaranteed to have zero probability under our choice of transition prior $P(M_{t+1}|M_t)$.

As a result, the optimal mode sequence is estimated via

$$M^*_1:K = \arg\max_{M_{1:K}} P(M_{1:K}|Y_1:K)$$  \hspace{1cm} (8)

although as Section IV discusses, adherence to (8) is only approximate due to computational cost considerations.

As for the remaining state sequences (i.e., $N_{1:K}$, $T_{1:K}$, $A_{1:K}$, $A^Q_{1:K}$) there seem to be no special reasons to apply the MAP criterion as opposed to estimating the sequences which minimize symbol error rates, as long as these estimates synchronize in some fashion with $M^*_1:K$. A natural objective is to minimize the expected number of symbol errors given $M^*_1:K$. The latter minimization is achieved by maximizing $P(Z_t|M^*_{1:K},Y_1:K)$ where $Z_t$ is one of $N_t$, $T_t$, or $A_t \cup A^Q_t$. The overall melody extraction and segmentation process (including the postprocessing steps which are necessary to convert the identified mode and state sequences to MIDI data) are summarized in the block diagram of Figure 3. Preprocessing is discussed in Section II-A, estimation of distributional parameters in Sections III and IV-B, primary inference in Section IV-A, and postprocessing in Section V.

### III. DISTRIBUTIONAL SPECIFICATIONS

For the model in Figure 2, it remains to specify the prior $P(S_0,M_0)$, the transition distribution across frames: $P(S_{t+1},M_{t+1}|S_t,M_t)$, and the observation likelihood $P(Y_t|S_t)$.

#### A. PRIOR

The prior encodes information about the frame immediately preceding the recording. When this information is absent, we desire a solution which applies in the widest variety of musical contexts. The prior admits the factorization:

$$P(S_0,M_0) = P(M_0)P(S_0|M_0)$$  \hspace{1cm} (9)

In the most general case we specify $P(M_0)$ as uniform, and $P(S_0|M_0)$ as factorizing independently among the components of $S_0$:

$$P(S_0|M_0 \in \mathcal{P}) = P(T_0)P(N_0)P(A_0)$$

$$P(S_0|M_0 \in \mathcal{Q}) = P(T_0)P(N_0)P(A^Q_0)$$  \hspace{1cm} (10)

where $P(T_0)$, $P(N_0)$, $P(A_0)$, and $P(A^Q_0)$ are uniform.\(^2\)

However, there do exist situations in which specific knowledge about the preceding frame is available. It is common,

\(^2\)In many recordings of Western music, we expect the tuning to be roughly A440, meaning that the note A4 \approx 440 Hz; however, it is dangerous to assume this for all applications, because recordings are commonly transposed in web-based music applications, for example, to maintain consistent tempo within a collection of pieces.
for instance, that the beginning of the recording coincides with the beginning of the first note event. In this case, the frame immediately preceding the recording should not contain information associated with a note event; i.e., \( P(M_0 = \text{‘N’}) = 1 \). All other conditional distributions remain unchanged.

### B. Transition distribution

The transition distribution factors accordingly:

\[
P(S_{t+1}, M_{t+1} | S_t, M_t) = P(M_{t+1} | M_t) P(S_{t+1} | M_t, M_{t+1}, S_t)
\]  

(11)

The mode transition dependence, \( P(M_{t+1} | M_t) \), develops according to the transition diagram displayed in Figure 4. In the figure, solid lines indicate transitions with positive probability under the standard note evolution hypothesis modeling the cyclic succession in Figure 1, while dotted lines indicate additional transitions due to other, spurious incidents, for instance an attack transient followed immediately by silence.

The rationale behind the standard note evolution hypothesis is as follows. A primary governing principle is that onsets, as they indicate the beginnings of note events, may not occur in adjacent frames; i.e., an onset mode must be followed immediately by a continuation or null mode: \( P(M_{t+1} \in C \cup N | M_t \in O) = 1 \). The latter ensures that the segmentation is well defined, especially when the attack transient occupies 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) discussed in Section II-A.

Nonzero elements of \( P(M_{t+1} | M_t) \) corresponding to the the standard hypothesis are estimated via the EM algorithm [6]. The latter encompasses iterations which, if properly initialized, converge to the (constrained) maximum-likelihood estimate \( P(M_{t+1} | M_t) \). A favorable initialization ensures not only a correct convergence, but also that the convergence within specified limits takes as few iterations as possible. Further details appear in Section IV-B.

Our initialization of \( P(M_{t+1} | M_t) \) proceeds via a generative, heterogeneous Poisson process model, as shown in Figure 1. This model represents the cyclic succession of a transient region of expected length \( \tau_T \), followed by a pitched region of expected length \( \tau_P \), followed by a null region of expected length \( \tau_N \). Individual lengths are modeled as independent, exponentially distributed random variables.

Table II gives the Poisson model in algebraic form. Each term \( p_{j,k}^{(m)} \) denotes the probability that the beginning of the next frame lies in a region of type \( k \) 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 \( p_{P,P}^{(1)} \). Finally, \( p_s \) represents the probability of spurious transition corresponding to any of the dotted lines in Figure 1; we set \( p_s \) to some small, fixed nonzero value; i.e., \( p_s = .001 \) is used to generate the results in Section VII.

<table>
<thead>
<tr>
<th>( M_{t+1} = \text{‘OT’} )</th>
<th>( M_{t+1} = \text{‘OP’} )</th>
<th>( M_{t+1} = \text{‘CP’} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_t = \text{‘OT’} )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( M_t = \text{‘OP’} )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( M_t = \text{‘CT’} )</td>
<td>( p_s )</td>
<td>( p_s )</td>
</tr>
<tr>
<td>( M_t = \text{‘CP’} )</td>
<td>( p_{P,P} ) (^{(0)})</td>
<td>( p_{P,P} ) (^{(1)})</td>
</tr>
<tr>
<td>( M_t = \text{‘N’} )</td>
<td>( p_{T,N} ) (^{(0)})</td>
<td>( p_{T,N} ) (^{(1)})</td>
</tr>
<tr>
<td>( M_{t+1} = \text{‘CP’} )</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

TABLE II

Generative Poisson model for the initialization of \( P(M_{t+1} | M_t) \).

### C. State transition behavior

The state transition behavior, \( P(S_{t+1} | M_t, M_{t+1}, S_t) \) describes the prior uncertainty in the next state, \( S_{t+1} \), given the current state, \( S_t \), over different values of \( M_t \) and \( M_{t+1} \). We note that this distribution depends on \( M_t \) at least through \( M_t \in P \) or \( M_t \in Q \), as the relation between two temporally adjacent pitched states is fundamentally different than the relations between a pitched state following a non-pitched state. However, we assume no additional dependence on \( M_t \).

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. For instance, given \( M_t \in P \), \( M_{t+1} = \text{‘CP’} \) 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, \( M_{t+1} = \text{‘OP’} \) signifies that frame \( t+1 \) contains 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. The latter models specific tendencies in melodic transitions. It may be difficult to say much about these tendencies, however, unless musical key is also introduced as a hidden state variable.

For fixed \( M_t \) and \( M_{t+1} \), the transition behavior factors independently over the components of \( S_t \):

\[
P(S_{t+1}|S_t, M_{t+1} \in \mathcal{P}, M_t \in \mathcal{P}) = 
\frac{P(T_{t+1}|T_t, M_{t+1} \in \mathcal{P}, M_t \in \mathcal{P})}{P(T_{t+1}|T_t, M_{t+1} \in \mathcal{P}, M_t \in \mathcal{P})} \times P(N_{t+1}|N_t, M_{t+1} \in \mathcal{P}, M_t \in \mathcal{P}) 
\times P(A_{t+1}|A_t, M_{t+1} \in \mathcal{P}, M_t \in \mathcal{P})
\]

(12)

and similarly for \( M_t \in \mathcal{Q} \) or \( M_{t+1} \in \mathcal{Q} \), although in these cases \( A_{t+1} \) is replaced by \( A_{t+1}^{\mathcal{Q}} \) or \( A_t \) is replaced by \( A_t^{\mathcal{Q}} \).

The factorization (12) assumes no interdependence between state components when \( M_t \) and \( M_{t+1} \) are in evidence. Though situations may arise which do indicate interdependence (e.g., \( \{T_t = 0.49, N_t = 60\} \) and \( \{T_t = -0.5, N_t = 61\} \) refer to the same pitch hypothesis), such occurrences are infrequent enough that in most practical situations, it is unnecessary to model this interdependence.

We discuss now the individual distributions on the r.h.s. of (12), considering note, tuning, and amplitude, in that order. To begin, if \( M_{t+1} = \text{‘CP’}, M_t \in \mathcal{P} \) 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 \( N_t \) in such a way that large changes in note value are less likely than small changes or no change. To express this concentration, we define the double-sided exponential distribution:

\[
E_2(X_1|X_0, \alpha_+, \alpha_-) = \begin{cases} 
  c, & X_1 = X_0 \\
  \frac{\kappa(X_1)-\kappa(X_0)}{\alpha_+}, & X_1 > X_0 \\
  \frac{\kappa(X_0)-\kappa(X_1)}{\alpha_-}, & X_0 > X_1 
\end{cases}
\]

(13)

where \( c \) is chosen such that the distribution sums to unity, and \( \kappa(X) = k \) 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:

\[
P(N_{t+1}|N_t, M_{t+1} = \text{‘CP’}, M_t \in \mathcal{P}) = 
E_2(N_{t+1}|N_t, \alpha_N, \alpha_N)
\]

(14)

Ideally \( \alpha_N = 0 \), but we allow some small deviation for robustness to the case where the tuning endpoint approaches \( \pm 0.5 \), as here some ambiguity may result as to the note value. On the other hand, allowing for more extreme tuning variations (e.g., vocal vibrato greater than \( \pm 0.5 \)) requires that \( \alpha \) should remain very close to zero. 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 concerning note-to-note dependences. Thus:

\[
P(N_{t+1}|N_t, M_{t+1} \in \mathcal{Q}, M_t \in \mathcal{M}) = 
E_2(N_{t+1}|N_t, \alpha_N, \alpha_N)
\]

(15)

Let \( P_{\text{note-trans}}(N_1|N_0) \) be the dependence where \( N_0 \) and \( N_1 \) are the values of adjacent note events; if such information is absent, the dependence is simply 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 \( P_{\text{note-trans}}(N_1|N_0) \):

\[
P(N_{t+1}|N_t, M_{t+1} = \text{‘OP’}, M_t \in \mathcal{M}) = 
P_{\text{note-trans}}(N_1|N_0)
\]

(16)

The remaining cases involve certain \((M_t, M_{t+1})\) combinations which occur with zero probability due to the mode transition mechanism of Table II. These distributions can be set to anything without affecting the inference outcome, so we specify them as to minimize computational effort (see Table III).

The conditional distribution of the tuning, \( T_{t+1} \), reflects the assumption that tuning is expected to be constant, or vary only slightly throughout the recording, independently of onsets, endpoints, and note events. Of course, this is not entirely true; as some instruments exhibit a dynamic pitch envelope, we should allow for some degree of variation in order to be robust.

Hence:

\[
P(T_{t+1}|T_t, M_{t+1} \in \mathcal{M}, M_t \in \mathcal{M}) = 
E_2(T_{t+1}|T_t, \alpha_{T+}, \alpha_{T-})
\]

(17)

where \( \alpha_{T+} = \alpha_{T-} = \Delta \) indicates symmetry of the expected tuning variation.

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

\[
P(A_{t+1}|A_t, M_{t+1} = \text{‘CP’}, M_t \in \mathcal{P}) = 
E_2(A_{t+1}|A_t, \alpha_{A+}, \alpha_{A-})
\]

(18)

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 marimba and piano. On the other hand, setting \( \alpha_{A+} = \alpha_{A-} \) may be more appropriate for violin, voice, and other sustained tones, or tones with lengthy attack regions. In all other cases, \( A_{t+1} \) is independent of \( A_t \) (or \( A_t^{\mathcal{Q}} \)). Where \( M_{t+1} = \text{‘OP’}, A_{t+1} \) 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:

\[
P(A_{t+1}|A_t, M_{t+1} = \text{‘OP’}, M_t \in \mathcal{P}) = 
E_1(A_{t+1}, \beta_A, \text{‘OP’})
\]

(19)

\[
P(A_{t+1}|A_t^{\mathcal{Q}}, M_{t+1} = \text{‘OP’}, M_t \in \mathcal{Q}) = 
E_1(A_{t+1}, \beta_A, \text{‘OP’})
\]

(20)
where, using the notation of (13),
\[ E_1(X, \beta) = e^{\beta V(X)} \quad (21) \]

The constant \( c \) is chosen such that for fixed \( \beta \), \( E_1(X, \beta) \) sums to unity over values of \( X \). Setting \( \beta_A, OP > 1 \) means that the pitched onset amplitude favors higher amplitudes. Care must be taken that \( \beta_A, OP \) is not too large, however, to allow for quiet notes. Where \( M_{t+1} \in \{OT\} \) or \( M_{t+1} \in \{CT\} \) (i.e., \( M_{t+1} \in T \)), the distribution is similar, but concerning \( A_{t+1}^Q \) in place of \( A_{t+1}^N \):
\[ P \left( A_{t+1}^Q | A_t, M_{t+1} \in T, M_t \in P \right) = E_1 \left( A_{t+1}^Q, \beta_A, T \right) \]
\[ P \left( A_{t+1}^Q | A_t^Q, M_{t+1} \in T, M_t \in Q \right) = E_1 \left( A_{t+1}^Q, \beta_A, T \right) \quad (22) \]

where \( \beta_A, T > 1 \).

Where \( M_{t+1} = \{'N', \} \), the distribution of \( A_{t+1}^Q \) follows either line of (22), depending on \( M_t \in P \) or \( Q \), but with \( \beta_A, P > 1 \) in place of \( \beta_A, T \), since the null mode favors lower amplitudes.

Table III summarizes the aforementioned state transition behavior, filling in “don’t-care” possibilities.

<table>
<thead>
<tr>
<th>( M_{t+1} )</th>
<th>( M_t \in P )</th>
<th>( M_t \in Q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>{'OT'}</td>
<td>( P(N_{t+1}</td>
<td>N_t) = E_2(N_{t+1}, \alpha_N, \alpha_N) )</td>
</tr>
<tr>
<td></td>
<td>( P(A_{t+1}^Q</td>
<td>A_t^Q) = E_1(A_{t+1}^Q, \beta_A, T) )</td>
</tr>
<tr>
<td>{'OP'}</td>
<td>( P(N_{t+1}</td>
<td>N_t) = P_{\text{note, trans}}(N_{t+1}, N_t) )</td>
</tr>
<tr>
<td></td>
<td>( P(A_{t+1}^Q</td>
<td>A_t^Q) = E_1(A_{t+1}^Q, \beta_A, T) )</td>
</tr>
<tr>
<td>{'CT'}</td>
<td>( P(N_{t+1}</td>
<td>N_t) = E_2(N_{t+1}, \alpha_N, \alpha_N) )</td>
</tr>
<tr>
<td></td>
<td>( P(A_{t+1}^Q</td>
<td>A_t^Q) = E_1(A_{t+1}^Q, \beta_A, T) )</td>
</tr>
<tr>
<td>{'CP'}</td>
<td>( P(N_{t+1}</td>
<td>N_t) = P_{\text{note, trans}}(N_{t+1}, N_t) )</td>
</tr>
<tr>
<td></td>
<td>( P(A_{t+1}^Q</td>
<td>A_t^Q) = E_1(A_{t+1}^Q, \beta_A, T) )</td>
</tr>
<tr>
<td>{'N'}</td>
<td>( P(N_{t+1}</td>
<td>N_t) = E_2(N_{t+1}, \alpha_N, \alpha_N) )</td>
</tr>
<tr>
<td></td>
<td>( P(A_{t+1}^Q</td>
<td>A_t^Q) = E_1(A_{t+1}^Q, \beta_A, T) )</td>
</tr>
</tbody>
</table>

**TABLE III**

State transition table for component distribution of

\[ P(S_{t+1} | S_t, M_{t+1}, M_t) \]

**D. Frame likelihood**

The likelihood for frames with pitch content (where the amplitude is represented by \( A_t \)) is computed by a modification of the method of [30], to which we refer as the canonical evaluation. The latter evaluates \( P_{\text{can}}(Y_t | f_{0,t}, A_{0,t}) \) where \( f_{0,t} \) is the radian fundamental frequency, and \( A_{0,t} \) is an unknown reference amplitude. In [30], [28] this method is shown to be robust to real-world phenomena such as inharmonicity, undetected peaks, and spurious peaks due to noise and other interference phenomena (see also [16] for polyphonic extensions).

Care must be taken in the association of the hypotheses \((f_{0,t}, A_{0,t})\) with those of the state \((N_t, T_t, A_t)\). 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 [30], the reference amplitude is estimated as the maximum amplitude over all peaks in the frame, denoted as \( A_{\text{max},t} \). The latter seems to yield favorable psychoacoustic properties (as borne out by informal listening tests) 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, 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 relevant explanation is that in the absence of other contextual cues, we focus on the loudest of multiple pitch components: \( A_{0,t} = A_{\text{max},t} \).

Unfortunately the latter choice ignores the state variable \( A_t \), allowing little possibility for the conditional distribution of \( A_t \) to be influenced by the signal, except indirectly via the mode layer. This choice disables 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 \( A_{\text{max},t} \) as an independent noisy observation\(^4\) of \( A_t \), as shown in Figure 5. By so doing, we blend the strategy which derives \( A_{\text{max},t} \) from the state \((A_{0,t} = A_t)\) with the strategy incorporating psychoacoustic salience \((A_{0,t} = A_{\text{max},t})\). The conditional distribution for the observation layer becomes as follows:

\[ P(Y_t, A_{\text{max},t} | N_t, T_t, A_t) = P(Y_t | N_t, T_t, A_t) P(A_{\text{max},t} | A_t) \quad (23) \]

\(^4\)It may seem counterintuitive to model \( A_{\text{max},t} \) and \( Y_t \) as conditionally independent of \( A_t \) since, unconditionally speaking, \( A_{\text{max},t} \) is a deterministic function of \( Y_t \). However, we wish not to introduce bias by assuming specific dependencies between the noise on \( A_{\text{max},t} \) and the amplitude/frequency noises on other peaks of \( Y_t \).
Here the dependence of \( Y_t \) is modeled via \( P_{\text{can}}(Y_t|f_0, A_{0,t}) \) using \( A_{0,t} = A_t \), and \( A_{\text{max},t} \) is modeled as \( A_t \) plus Gaussian noise:

\[
P(Y_t|N_t, T_t, A_t) = P_{\text{can}}(Y_t|f_0(N_t, T_t), A_{0,t} = A_t) = \mathcal{N}(A_t, \sigma^2_A)
\]

(24)

We may interpolate between the rigid cases \((A_{0,t} = A_{\text{max},t} vs. A_{0,t} = A_t)\) simply by varying \( \sigma^2_A \) between zero and infinity. A detailed discussion of this amplitude observation weighting appears in Appendix I.

Another consideration is the computational complexity of the frame likelihood computation. As shown in in Section IV, the inference requires that \( P_{\text{can}}(Y_t|f_0(N_t, T_t), A_t) \) be evaluated for each note, tuning, and amplitude value for each frame. The resultant computational cost may be excessive, especially for real-time operation. Hence, we introduce a a posteriori inference requires that \( \mathcal{A} \) be excessive, especially for real-time operation. Hence, we introduce a a posteriori inference requires that \( A_{\text{max},t} \) and \( A_{\text{max},t} \) simply by varying \( \sigma^2_A \) between zero and infinity. A detailed discussion of this amplitude observation weighting appears in Appendix I.

Another consideration is the computational complexity of the frame likelihood computation. As shown in in Section IV, the inference requires that \( P_{\text{can}}(Y_t|f_0(N_t, T_t), A_t) \) be evaluated for each note, tuning, and amplitude value for each frame. The resultant computational cost may be excessive, especially for real-time operation. Hence, we introduce a a posteriori inference requires that \( A_{\text{max},t} \) and \( A_{\text{max},t} \) simply by varying \( \sigma^2_A \) between zero and infinity. A detailed discussion of this amplitude observation weighting appears in Appendix I.

Another consideration is the computational complexity of the frame likelihood computation. As shown in in Section IV, the inference requires that \( P_{\text{can}}(Y_t|f_0(N_t, T_t), A_t) \) be evaluated for each note, tuning, and amplitude value for each frame. The resultant computational cost may be excessive, especially for real-time operation. Hence, we introduce a a posteriori inference requires that \( A_{\text{max},t} \) and \( A_{\text{max},t} \) simply by varying \( \sigma^2_A \) between zero and infinity. A detailed discussion of this amplitude observation weighting appears in Appendix I.

Another consideration is the computational complexity of the frame likelihood computation. As shown in in Section IV, the inference requires that \( P_{\text{can}}(Y_t|f_0(N_t, T_t), A_t) \) be evaluated for each note, tuning, and amplitude value for each frame. The resultant computational cost may be excessive, especially for real-time operation. Hence, we introduce a a posteriori inference requires that \( A_{\text{max},t} \) and \( A_{\text{max},t} \) simply by varying \( \sigma^2_A \) between zero and infinity. A detailed discussion of this amplitude observation weighting appears in Appendix I.

Another consideration is the computational complexity of the frame likelihood computation. As shown in in Section IV, the inference requires that \( P_{\text{can}}(Y_t|f_0(N_t, T_t), A_t) \) be evaluated for each note, tuning, and amplitude value for each frame. The resultant computational cost may be excessive, especially for real-time operation. Hence, we introduce a a posteriori inference requires that \( A_{\text{max},t} \) and \( A_{\text{max},t} \) simply by varying \( \sigma^2_A \) between zero and infinity. A detailed discussion of this amplitude observation weighting appears in Appendix I.
To prepare for the backward pass, quantities $\mu^*(M_t, S_t)$ and $M^*_t(M_{t+1})$ are stored for $t \geq 1$ as well as $\tau^*(M_t, S_t)$ for $t \geq 2$. The final-stage objective, $J(M_K)$, is also stored. The backward pass is initialized as follows.

$$M^*_K = \text{argmax}_{M_K} J(M_K)$$

(31)

The remainder of the backward pass consists of the following recursions, as $t$ decreases from $K-1$ down to 1:

$$M^*_t = M^*_t(M^*_{t+1})$$

$$\tau^*(M_t, S_t) = \tau^*(M^*_t, S_t, M^*_{t+1})$$

(30)

A complete derivation of the approximate Viterbi recursions, $(28 - 32)$ is similar to that given in Appendix A of [28].

### B. Estimation of $P(M_{t+1}|M_t)$

The distribution $P(M_{t+1}|M_t)$ depends on the collection of free parameters in Table II. These parameters are identified via the EM algorithm. To begin, define

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

(33)

Let for each $j \in \mathcal{M}$, the set $S_j \in \mathcal{M}$ denote the set of possibilities for $k$ for which $p_{k|j}$ is unknown. That is, we characterize the collection of free parameters, $\theta$, as follows.

$$\theta = \bigcup_{j \in \mathcal{M}} \bigcup_{k \in S_j} \{p_{k|j}\}$$

(34)

The EM algorithm begins with an initial guess for $\theta$, i.e., $\theta^{(0)}$, and proceeds over iterations $i$, updating $\theta = \theta^{(i)}$. Each iteration comprises two steps: the first, or expectation step, computes the expected log likelihood $\log P(M_{0,K}, S_{0,K}, Y_{1:K} | \theta)$ where $M_{0,K}$ and $S_{0,K}$ are generated according to the distribution $P(M_{0,K}, S_{0,K}, Y_{1:K} | \theta)$. That is, we form

$$Q(\theta^{(i)}) = E_{P(M_{0,K}, S_{0,K}, Y_{1:K} | \theta^{(i)})} \left[ \log P(M_{0,K}, S_{0,K}, Y_{1:K} | \theta) \right]$$

(35)

The subsequent maximization step chooses $\theta^{(i+1)}$ as a value of $\theta$ maximizing $Q(\theta^{(i)})$.

The EM steps are as follows ([28], Appendix B):

- **Expectation step**: Compute

$$\sigma^{(2)}(M_t, M_{t+1}) = P(M_t, M_{t+1}|Y_{1:K}, \theta^{(i)})$$

(36)

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

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

$$p^{(i+1)}_{k|j} = \frac{\sum_{t=1}^{K-1} \sigma^{(2)}(M_t = j, M_{t+1} = k)}{\sum_{k \in \mathcal{M}} \sum_{t=1}^{K-1} \sigma^{(2)}(M_t = j, M_{t+1} = k)}$$

(37)

Computation of the posterior quantities $\sigma^{(2)}(M_t, M_{t+1})$ for $t \in 1 : K - 1$ is the result of a standard Bayesian smoothing inference. The latter consists of a forward, filtering pass, followed by a backward, smoothing pass analogous to the Rauch-Tung-Striebel method for Kalman smoothing [19], [12]. The associated recursions are given in Appendix III.

### V. Postprocessing

The goal of postprocessing is to take the maximum a posteriori mode sequence, $M^*_{1:K}$, and the smoothed note posterior $P(N_t|M^*_{1:K}, Y_{1:K})$, and produce a string of distinct note events, which can be stored in a standard MIDI file\footnote{Note that $P(N_t|M^*_{1:K}, Y_{1:K})$ may be derived from the primary inference output, $P(S_t|M^*_{1:K}, Y_{1:K})$, by marginalization.} The MIDI file output contains a sequence of note events, each consisting of an onset time, note value, and duration. Additionally, we provide a sub-segmentation into transient and pitched regions by specifying the duration of each note event’s transient region.
Table V specifies the quantities propagated in the postprocessing recursions for the \( m^{th} \) note event.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( o^{(m)} )</td>
<td>Onset frame for note event</td>
</tr>
<tr>
<td>( p^{(m)} )</td>
<td>First pitched frame in note event</td>
</tr>
<tr>
<td>( q^{(m)} )</td>
<td>Note event duration</td>
</tr>
<tr>
<td>( e_+^{(m)} )</td>
<td>One frame beyond end of note event</td>
</tr>
<tr>
<td>( N^{+}(m) )</td>
<td>MIDI note value</td>
</tr>
</tbody>
</table>

**TABLE V**

Transcription output quantities

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

\[
\begin{align*}
^+ \min(S) & \triangleq \begin{cases} 
\infty, & S = \emptyset \\
\min(S), & \text{otherwise}
\end{cases} \\
\end{align*}
\]

The postprocessing algorithm iterates over note events \( k \), stopping only when either the onset frame, pitch boundary, or the advanced end point \( (o^{(m)}, p^{(m)}, e_+^{(m)}) \), 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.

\[
o^{(1)} = ^+ \min \{ t \geq 1 : M_t^r \in \mathcal{O} \}
\]

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.

\[
o^{(m)} = ^+ \min \{ t \geq e_+^{(m-1)} : M_t^r \in \mathcal{O} \}
\]

\[
p^{(m)} = ^+ \min \{ t \geq o^{(m)} : M_t^r \in \mathcal{P} \}
\]

\[
e_+^{(m)} = ^+ \min \{ t \geq p^{(m)} : M_t^r \notin \mathcal{C} \}
\]

If \( k = 1 \), the initialization (39) is used in place of (40) in case of \( o^{(m)} \). As indicated, \( e_+^{(m)} \), lies one frame beyond the last frame of the note event. The duration of note \( k \) is just the simple difference between \( e_+^{(m)} \) and \( o^{(m)} \), unless \( e_+^{(m)} \) has been truncated by the end of the signal. In the latter case, the duration is \( N - o^{(m)} + 1 \).

\[
d^{(m)} = \min(e_+^{(m)}, N + 1) - o^{(m)}
\]

To obtain the MIDI note value, we extract

\[
N^{+}(m) = \arg\max_n P(N_{\mu^{(1)} + c} = n | Y_{1:K})
\]

where

\[
c = \min(c_0, e_+^{(m)} - p^{(1)} - 1)
\]

Here \( c \) is a margin variable ensuring that the maximum \textit{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 \( c_0 \); \( c_0 = 3 \) suffices in the extraction of MIDI information from the examples of Section VII.

VI. COMPUTATIONAL COST

As discussed in Section II-B, the overall melody extraction and segmentation algorithm operates by estimating the transition matrix \( P(M_{t+1} | M_t) \) by the EM approach of Section IV-B, followed by the primary inference discussed in Section IV-A, followed by postprocessing to convert the MAP mode sequence \( M_{t,N}^* \) and posteriors \( P(S_{1:N} | Y_{1:N}, M_{t,N}^*) \) into MIDI data as discussed in the previous section. We assess computational cost with respect to the following variables: \( K \), the number of signal frames, \( N_M \), the number of different modes, \( N_T \), the number of tunings, \( N_A \), the number of pitched amplitudes, and \( N_A^0 \), the number of null amplitudes.

The cost of estimating \( P(M_{t+1} | M_t) \) is the combined cost of the expectation and maximization steps (36, 37) times the number of EM iterations (typically fixed at two for the Poisson initialization; see Figure 8 for typical results). The expectation step consists of the standard Bayesian smoothing inference detailed in Appendix III. From (66,67) the inference appears linear in \( K \) and quadratic in \( N_M \). Additionally, the summations

\[
\sum_{S_t} [\mu(M_t, S_t|Y_{1:t}) P(S_{t+1} | S_t, M_{t+1}, M_t)]
\]

\[
\times \sum_{S_{t+1}} \phi(M_{t+1}, S_{t+1}) P(S_{t+1} | S_t, M_{t+1}, M_t)
\]

(44)

appear quadratic in \( N_S \), the number of distinct state possibilities, which equals \( N_N N_T (N_A + N_A^0) \). However, a more careful analysis of the factorization of \( P(S_{t+1}|S_t, M_{t+1}, M_t) \), as presented in Section III enables the interchange of orders of summation among the components of \( S_t \) or \( S_{t+1} \) to reduce the overall cost. Let us consider the first summation and the case \( M_t \in \mathcal{P}, M_{t+1} \in \mathcal{P} \); all other summations and cases are similar. From (12), it follows:

\[
\sum_{S_t} [\mu(M_t \in \mathcal{P}, S_t | Y_{1:t}) P(S_{t+1} \in \mathcal{P}, M_{t+1} \in \mathcal{P}, M_t \in \mathcal{P})]
\]

\[
\times \sum_{T_t} [P(T_{t+1} | T_t, M_{t+1} \in \mathcal{P}, M_t \in \mathcal{P})]
\]

\[
\times \sum_{N_t} [P(N_{t+1} | N_t, M_{t+1} \in \mathcal{P}, M_t \in \mathcal{P})]
\]

\[
\times \sum_{A_t} [P(A_{t+1} | A_t, M_{t+1} \in \mathcal{P}, M_t \in \mathcal{P}, \mu(M_t, S_t|Y_{1:t}))]
\]

(45)

The cost of (45) becomes \( O(N_N N_T (N_A + N_A^0) (N_N + N_T + N_A) \). The cost for all required summations of this form hence becomes \( O(N_N N_T (N_A + N_A^0) (N_N + N_T + N_A + N_A^0)) \), meaning that the overall cost for estimation of \( P(M_{t+1} | M_t) \) is \( O(K N_M^2 N_N N_T (N_A + N_A^0) (N_N + N_T + N_A + N_A^0)) \).

While \( N_M = 5 \) for the present algorithm, one may consider adding modes which model more sophisticated envelope behavior (attack-decay-sustain-release) or which encode certain types of pitch fluctuations such as vibrato. Hence, it is useful to consider \( N_M \) as a variable for the cost analysis.
From (30, 32), we see that terms of the same type as in (45) dominate the cost of the primary inference; hence the computational cost of the latter is also $O(K N_{M} NT (N_A + N_{A^O}))$. Finally, the postprocessing steps require the marginalization of $P(N_t | M_{t+1}, Y_{1:K})$ from $P(N_t | M_{t+1}, Y_{1:K})$, a process which requires $O(N_S) = O(NN_T (N_A + N_{A^O}))$ computations per frames, plus the iterations (40) which are at most linear in the number of frames and constant with respect to other quantities. The postprocessing computations thus appear negligible with respect to the other stages. As such, the overall computational cost of the proposed method remains as $O(K N_{M} NT (N_A + N_{A^O}))$.

VII. RESULTS

A. Primary inference

Our joint melody extraction and onset detection algorithm has been applied to a small collection of violin and piano recordings. In each case, the score is monophonic; however, the presence of reverberation, note overlaps due to legato playing style, and even the performer’s tendency to sing accompanying phrases are evidenced. A representative piano recording is the introductory motive of Bach’s Invention 2 in C Minor (BWV 773), performed by Glenn Gould. The recording is fairly clean, with minimal reverberation; however, legato playing and the performer’s tendency to vocalize approximately two octaves below the melody line, pose particular difficulties for the pitch estimation. Onset locations, however, are clearly retrievable by eye or by a heuristic following of amplitude jumps as in Schloss’s “surfboard” algorithm [23]. A representative violin recording, several notes from the third movement of Bach’s solo violin Sonata No. 1 in G minor (BWV 1001), performed by Nathan Milstein, is awash in reverberation, making the note transitions extremely difficult to locate by time-domain methods which monitor only amplitude. The reverberation also induces effective polyphony during significant portions of the note duration. Moreover, the presence of portamento renders ambiguous the exact moments of several note transitions. Nevertheless, the melody is clearly audible.

Primary inference results for the piano example are shown in Figure 6. Directly beneath the time-domain waveform are shown the onset indication and the MAP mode determination, $M_{t+1}$. The MAP mode determination indicates the segmentation according to onset locations and note durations, as well as the sub-segmentation into transient and pitched regions within each note event (see Section V for details about how this information is extracted via postprocessing). Lower sections of the figure display posteriors for note, tuning, amplitude, and null amplitude values for $t = 1 : K$ given $M_{t+1}$ and STFT peaks $Y_{1:K}$.

We observe that the concentration of the note posterior on the correct value continues after the endpoint of the note event and into the initial transient portion of the next event. This continuation is by design: to model note-to-note dependences via $P_{\text{note,trans}}(N_t+1 | N_t)$, we memorize the note value until the beginning of the first pitched region of the next event (see Section III). For melody extraction, we select the value maximizing the note posterior for the initial frame within the first pitched region. This value is assigned to the entire note event. In the example, there appears virtually no pitch ambiguity, despite the occasional presence of the performer’s voice. Amplitude posteriors trace out appropriately decaying envelopes, and tuning remains relatively constant throughout.

Results for the violin example are shown in Figure 7. In this example, we validate the segmentation by extracting portions corresponding to identified note events and repeatedly listening to them, since it is extremely difficult to detect the majority of onset locations by eye. The reverberation and expressive pitch variations in this example tend toward the worst case among the range of similar recordings tested; however, such effects are common enough that it becomes necessary to assess the performance under such conditions. We note that the initial grace note is incorrectly detected an octave below the true pitch; however, the surrounding notes are correctly segmented. The tuning posterior is less consistent than in the piano example, thanks to the range of pitch variations (portamento, vibrato, etc.) available to the performer. A clear illustration of tuning variability appears in the form of a portamento transition between the fourth and fifth notes of the violin example (Figure 7). As the performer slides smoothly down one semitone, the tuning posterior tracks the expressive gesture by first concentrating on pitches corresponding to the lower tuning values relative to the fourth note, then concentrating on higher-pitched tunings at the beginning of the fifth note. Indeed, despite the highly reverberant context and expressive pitch variations, we observe that onset locations are identified in the places that make the best sense for the determination of note boundaries.

B. Estimation of $P(M_{t+1} | M_t)$

The convergence of the constrained EM approach for the estimation of $P(M_{t+1} | M_t)$ is shown in Figure 8 for the uniform and Poisson initializations. While the uniform case nearly converges after five iterations, the Poisson converges much faster, after only two iterations. We note empirically that the melody extraction and segmentation results remain relatively insensitive to deviations in the specification of $P(M_{t+1} | M_t)$; however, the exact distribution may itself be of interest in characterizing recording qualities and performance tendencies.

VIII. CONCLUSIONS AND FUTURE RESEARCH

As the included piano and violin examples demonstrate, our method provides a robust segmentation and melody extraction in the presence of significant reverberation, note overlaps across onset boundaries, expressive pitch variations, and background voices. The violin example demonstrates the capacity of the algorithm to correctly segment audio even when the locations of many onsets in the time-domain waveform are ambiguous.

\footnote{In cases where pitch fluctuation within one note may be greater than a half-step (e.g., vocal vibrato), the space of tunings may be extended to encompass a wider range of values.}
Fig. 6. Piano example: Introductory motive of Bach’s Invention 2 in C-minor (BWV 773), performed by Glenn Gould. Rectangle sizes vary logarithmically according to probability, with the smallest visible rectangle corresponding to a probability of 0.03, and the largest, 1.0.

Fig. 7. Violin example: several notes from the third movement of Bach’s solo violin Sonata No. 1 in G minor (BWV 1001), performed by Nathan Milstein. Rectangle sizes vary logarithmically according to probability, with the smallest visible rectangle corresponding to a probability of 0.03, and the largest, 1.0.
Forthcoming extensions include the integration of additional layers of musical context and the simultaneous sample-accurate segmentation and demixing of individual note events overlapping at onset boundaries. To incorporate more general models of melodic expectation, we augment the state $S_t$ to encode higher-level contextual attributes such as key and additional pitch history via N-grams. Such encoding allows us to capture melodic tendencies such as inertia, gravity, magnetism, and others addressed in functional tonal analysis [15], [17]. To perform sample-accurate segmentation, we recognize that the dependence on $A_t$, the pitch inference becomes impulsively concentrated about $A_t$; i.e.:

$$P(\text{A}_{\text{max},t}|A_t) \sim \delta(\text{A}_{\text{max},t}, A_t) \quad (47)$$

It suffices to show that, given (47), $P(N_t, T_t|Y_t, \text{A}_{\text{max},t})$ becomes identical to the inference $P'(N_t, T_t|Y_t)$ where

$$P'(Y_t|N_t, T_t, A_t) = P_{\text{can}}(Y_t|f_0(N_t, T_t), A_{0,t} = \text{A}_{\text{max},t}) \quad (48)$$

Expanding $P(N_t, T_t|Y_t, \text{A}_{\text{max},t})$ according to Bayes’ rule yields the following:

$$P(N_t, T_t|Y_t, \text{A}_{\text{max},t}) \propto \sum_{\nu \tau} \int_{A_t} \rho(A_t, \nu, \tau, Y_t, \text{A}_{\text{max},t}) \, dA_t \quad (49)$$

where

$$\rho(A_t, N_t, T_t, Y_t, \text{A}_{\text{max},t}) \overset{\Delta}{=} P(A_t, N_t, T_t, Y_t, \text{A}_{\text{max},t}) \quad (50)$$

and

$$P(A_t, N_t, T_t, Y_t, \text{A}_{\text{max},t}) = P(A_t) P(N_t, T_t|A_t) \times P_{\text{can}}(Y_t|f_0(N_t, T_t), A_{0,t} = A_t) \delta(\text{A}_{\text{max},t}, A_t) \quad (51)$$

Substituting (51) into (49) obtains integral expressions with impulsive terms. These expressions, and hence (49), simplify as follows.

$$\int_{A_t} P(A_t, N_t, T_t, Y_t, \text{A}_{\text{max},t}) \, dA_t = P(N_t, T_t|A_t = \text{A}_{\text{max},t}) \times P_{\text{can}}(Y_t|f_0(N_t, T_t), A_{0,t} = \text{A}_{\text{max},t}) \quad (52)$$

APPENDIX I

INDEPENDENT AMPLITUDE OBSERVATION WEIGHTING

Recall from Section III-D that we may interpolate between the rigid cases ($A_{0,t} = \text{A}_{\text{max},t}$ vs. $A_{0,t} = A_t$) simply by varying $\sigma^2_A$ between zero and infinity.

Assuming $A_t \in \mathbb{R}^+$, as $\sigma^2_A \rightarrow 0$, the pitch inference $P(N_t, T_t|Y_t, \text{A}_{\text{max},t})$, becomes identical to the inference $P'(N_t, T_t|Y_t)$ where $P'(Y_t|N_t, T_t, A_t)$ is defined as the canonical evaluation with $A_{0,t}$ replaced by $\text{A}_{\text{max},t}$. On the other hand, as $\sigma^2_A \rightarrow \infty$, the pitch inference $P(N_t, T_t|Y_t, \text{A}_{\text{max},t})$ converges to $P(N_t, T_t|Y_t)$, which is the canonical evaluation using $A_{0,t} = A_t$, $\text{A}_{\text{max},t}$ being ignored.

To show, we first consider the case $\sigma^2_A \rightarrow \infty$. Here the dependence on $A_t$ vanishes:

$$P(\text{A}_{\text{max},t}|A_t) \rightarrow P(\text{A}_{\text{max},t}) \quad (46)$$

As a result, $\text{A}_{\text{max},t}$ and the collection $\{Y_t, N_t, T_t, A_t\}$ become mutually independent. Then $P(N_t, T_t|Y_t, \text{A}_{\text{max},t})$ converges to $P(N_t, T_t|Y_t)$, as was to be shown.

Now we consider the case $\sigma^2_A \rightarrow 0$. To begin, we note that as $\sigma^2_A \rightarrow 0$, $P(\text{A}_{\text{max},t}|A_t)$ becomes impulsively concentrated about $A_t$; i.e.:

$$P(\text{A}_{\text{max},t}|A_t) \sim \delta(\text{A}_{\text{max},t}, A_t) \quad (47)$$

It suffices to show that, given (47), $P(N_t, T_t|Y_t, \text{A}_{\text{max},t})$ becomes identical to the inference $P'(N_t, T_t|Y_t)$ where

$$P'(Y_t|N_t, T_t, A_t) = P_{\text{can}}(Y_t|f_0(N_t, T_t), A_{0,t} = \text{A}_{\text{max},t}) \quad (48)$$

Expanding $P(N_t, T_t|Y_t, \text{A}_{\text{max},t})$ according to Bayes’ rule yields the following:

$$P(N_t, T_t|Y_t, \text{A}_{\text{max},t}) \propto \sum_{\nu \tau} \int_{A_t} \rho(A_t, \nu, \tau, Y_t, \text{A}_{\text{max},t}) \, dA_t \quad (49)$$

where

$$\rho(A_t, N_t, T_t, Y_t, \text{A}_{\text{max},t}) \overset{\Delta}{=} P(A_t, N_t, T_t, Y_t, \text{A}_{\text{max},t}) \quad (50)$$

and

$$P(A_t, N_t, T_t, Y_t, \text{A}_{\text{max},t}) = P(A_t) P(N_t, T_t|A_t) \times P_{\text{can}}(Y_t|f_0(N_t, T_t), A_{0,t} = A_t) \delta(\text{A}_{\text{max},t}, A_t) \quad (51)$$

Substituting (51) into (49) obtains integral expressions with impulsive terms. These expressions, and hence (49), simplify as follows.

$$\int_{A_t} P(A_t, N_t, T_t, Y_t, \text{A}_{\text{max},t}) \, dA_t = P(N_t, T_t|A_t = \text{A}_{\text{max},t}) \times P_{\text{can}}(Y_t|f_0(N_t, T_t), A_{0,t} = \text{A}_{\text{max},t}) \quad (52)$$

Fig. 8. Comparison of EM convergence with uniform vs. Poisson initialization. States along the horizontal axis of each figure correspond to $M_{t+1}$, and vertical axis $M_t$. The rectangle size varies logarithmically according to the probability, with the smallest visible square corresponding to a probability of 0.03, and the largest, 1.0.
Now, since \( A_t \) and \( \{N_t, T_t\} \) are \textit{a priori} independent, (52) simplifies further:

\[
\int_{A_t} P(A_t, N_t, T_t, Y_t, A_{\text{max,t}}) \, dA_t = P(N_t, T_t) \\
\times \prod_{i=0}^{o} P(f_0(N_t, T_t), A_{0,i} = A_{\text{max,t}}) \tag{53}
\]

It thereby follows that the substitution of (53) into (49) obtains the same relation as the expansion of (48) via Bayes’ rule (in parallel fashion to (49)). As such:

\[
P(N_t, T_t|Y_t, A_{\text{max,t}}) = P'(N_t, T_t|Y_t, A_{\text{max,t}}) \tag{54}
\]
as was to be shown.

In the preceding, 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 \( \sigma^2 \) 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 \( A_t \in \mathbb{R}^+ \).

\textbf{APPENDIX II}

\textbf{DETERMINISTIC LIKELIHOOD APPROXIMATION}

Let \( Y \) denote the set of STFT peak observations for a single frame. \( Y \) consists of a vector of peak frequencies \( F \), and a parallel vector of peak amplitudes \( A \). We denote the radian frequency of the \( j^{\text{th}} \) output peak by \( F(j) \), and the corresponding amplitude by \( A(j) \). The task is to evaluate the likelihood of these observations given hypotheses \( f_0 \) for the fundamental frequency, and \( A_0 \) for the reference amplitude.

In [30], it is shown that this likelihood may be evaluated with respect to a template of fictitious \textit{input peaks} which represent actual sinusoidal components generated by the pitch hypothesis \((f_0, A_0)\). Each input peak governs the frequency and amplitude distribution of a peak potentially observed in the STFT output. Of course, not all input peaks are observed; moreover, spurious peaks due to interference and sidelobe detection may also appear in the output. In [30], the unknown mapping between input and output peaks is marginalized:

\[
P(F, A|f_0, A_0) = \sum_{L \in \mathcal{L}} P(L|f_0, A_0) P(F, A|L, f_0, A_0) \tag{55}
\]

where \( L \), a linkmap in the set of valid maps \( \mathcal{L} \), describes which output peaks correspond to which input peaks, which output peaks are spurious, and (by implication) which input peaks are missing from the output\(^8\). \( L \) can be represented as a mapping \( L : \mathcal{J}_o \rightarrow \mathcal{J}_j \) where \( \mathcal{J}_o \triangleq 1 : N_o \) and \( \mathcal{J}_j \triangleq 1 : N_j \cup \{\text{‘S’}\} \), where \( N_o \) is the number of output peaks, \( N_j \) is the number of input peaks, and ‘S’ denotes the spurious possibility. That is, if \( j \) is the index of an output peak; \( L(j) \) returns the index of the corresponding input peak, except when \( L(j) = \text{‘S’} \), which means that output peak is spurious. The indices of output peaks are sorted in terms of increasing frequency; the indices of input peaks are sorted in terms of increasing mean frequency parameter. An example linkmap is shown in Figure 9.

![Example linkmap, where numerical values represent the radian frequency associated with each peak. Symbolically, \( L(1) = \text{‘S’}, L(2) = 2, L(3) = \text{‘S’}, \) and \( L(4) = 5 \).

From [30], we obtain that the set of valid linkmaps are those for which either of the following statements is true for any two output indices \( j(0) \) and \( j(1) \) in \( \mathcal{J}_j \):

- if both output peaks are linked, the linkages do not cross, i.e., if \( L(j(0)) \in 1 : N_j \) and \( L(j(1)) \in 1 : N_j \), then \( j(1) > j(0) \rightarrow L(j(1)) > L(j(0)) \)
- one or both of the output peaks is spurious, i.e., \( L(j(0)) = \text{‘S’} \) or \( L(j(1)) = \text{‘S’} \)

In practice, we may approximate the summation in (55) with \( P(L|f_0, A_0) \) as a uniform prior. That is:

\[
P(F, A|f_0, A_0) \approx \frac{1}{\#\mathcal{L}} \sum_{L \in \mathcal{L}} P(F, A|L, f_0, A_0) \tag{56}
\]

where \( \#\mathcal{L} \) is the cardinality of \( \mathcal{L} \). Even though the specification of \( P(L|f_0, A_0) \) in [30] varies significantly with \( L \), we expect less variation for those \( L \) for which \( P(F, A|L, f_0, A_0) \) is large, provided \((f_0, A_0)\) is reasonably close to the true pitch hypothesis. This approximation has been tested empirically for a limited number of acoustic examples consisting of clean to noisy and reverberant recordings of piano and violin, including those used to generate the results in Section VII.

As specified in [30], \( P(F, A|L, f_0, A_0) \) factors as a product distribution over individual output peaks:

\[
P(F, A|L, f_0, A_0) = \prod_{j=1}^{N_o} P(F(j), A(j)|L(j), f_0, A_0) \tag{57}
\]

which means

\[
P(F, A|f_0, A_0) \approx \frac{1}{\#\mathcal{L}} \sum_{L \in \mathcal{L}} \prod_{j=1}^{N_o} P(F(j), A(j)|L(j), f_0, A_0) \tag{58}
\]

The form of the individual \( P(F(j), A(j)|L(j), f_0, A_0) \) distributions, as given in [30], takes into account amplitude and frequency deviations in STFT peaks resulting from signals containing sinusoids embedded in additive white Gaussian noise [30].

Now, suppose we consider extending the summation in (58) over the set of \textit{all} possible maps \( \mathcal{J}_o \rightarrow \mathcal{J}_j \), which we denote.
as \(\mathcal{L}^*\). Then \(\mathcal{L}^*\) becomes a Cartesian product space, enabling the interchange of sums and products in (58). Making this substitution yields:

\[
P(F, A|f_0, A_0) \approx \frac{1}{\#\mathcal{L}} \sum_{L \in \mathcal{L}^*} \prod_{j=1}^{N_o} P(F(j), A(j)|L(j), f_0, A_0)
\]

where \(\mathcal{L}^* = I_1^* \otimes I_2^* \otimes \ldots \otimes I_{N_o}^*\), where \(I_j^*\) denotes the set of possible maps from the index \(j\) to \(\mathcal{J}_i\), each of which corresponds to a possibility for \(L(j)\). Thus, the summation (59) may be written:

\[
P(F, A|f_0, A_0) \approx \frac{1}{\#\mathcal{L}} \sum_{L(1) \in \mathcal{J}_1} \ldots \sum_{L(N_o) \in \mathcal{J}_N} \prod_{j=1}^{N_o} P(F(j), A(j)|L(j), f_0, A_0)
\]

Interchanging sums and products yields as follows:

\[
P(F, A|f_0, A_0) \approx \frac{1}{\#\mathcal{L}} \prod_{j=1}^{N_o} \sum_{L(j) \in \mathcal{J}_j} P(F(j), A(j)|L(j), f_0, A_0)
\]

The approximation and product-sum interchange allows the likelihood evaluation to take place using far fewer computations than the exact evaluation of (55). In (61), the number of link evaluations of the form \(P(F(j), A(j)|L(j), f_0, A_0)\) is exactly \(N_o(N_i + 1)\), which for \(N_o = N_i = N\), is \(O(N^2)\). By contrast, the number of terms enumerated in (55) is\(^9\):

\[
\#\mathcal{L} = \min(N_o, N_i) \sum_{j=0}^{\min(N_o, N_i)} \binom{N_o}{j} \binom{N_i}{j}
\]

For \(N_o = N_i = N\), the number of terms (62) is \(O(4^N/\sqrt{N})\) [30]. Hence the deterministic approximation yields a polynomial-time solution to an inherently exponential-time problem.

In practice, approximating the sum (58) by (61) seems to obtain virtually identical results for the acoustic examples considered. The approximation seems to hold best where the output peaks are well separated in frequency and where the pitch content is salient, meaning that the standard deviation of \(F(j)\) given any \(L(j) \in 1: K_i\) is small. We expect output peaks to be well separated in the context of nonreverberant harmonic sounds which lack complex character (beating, chouring, etc.) or when sufficiently long analysis windows are used. To show this, define \(\tilde{\mathcal{L}} = \mathcal{L}^* \setminus \mathcal{L}\); the approximation error comprising the difference between the r.h.s. of (59) and (58), \(\eta(F, A|f_0, A_0)\), may be expressed:

\[
\eta(F, A|f_0, A_0) = \frac{1}{\#\mathcal{L}} \sum_{L \in \mathcal{L}} \prod_{j=1}^{N_o} P(F(j), A(j)|L(j), f_0, A_0)
\]

By definition of \(\mathcal{L}\), we note that each \(L \in \tilde{\mathcal{L}}\) has the property that there exists \(j^{(0)} \in 1: N_i, j^{(1)} \in 1: N_i\), for which \(j^{(1)} > j^{(0)}\), but \(L(j^{(1)}) \leq L(j^{(0)})\). If \(P(F(j^{(0)}), A(j^{(0)}))|L(j^{(0)}), f_0, A_0)\) is negligibly small, this term tends to annihilate the product corresponding to \(L\) in (63). On the other hand, if \(P(F(j^{(0)}), A(j^{(0)}))|L(j^{(0)}), f_0, A_0)\) is non-negligible, by the pitch salience hypothesis this must mean \(F(j^{(0)})\) is close to the mean frequency of the input peak at index \(L(j^{(0)})\). Now by assumption \(j^{(1)} > j^{(0)}\), implying that \(F(j^{(1)}) > F(j^{(0)})\), Furthermore, if we assume that the output peaks are well separated in frequency, the difference \(F(j^{(1)}) - F(j^{(0)})\) is non-negligible. However, because \(L(j^{(1)}) \leq L(j^{(0)})\), by the frequency sorting construction the frequency of the input peak with index \(L(j^{(1)})\) must be less than or equal to the frequency of the input peak with index \(L(j^{(0)})\). Hence \(F(j^{(1)})\) exceeds the frequency of the input peak associated with \(L(j^{(1)})\) by a non-negligible amount. As a result, \(P(F(j^{(1)}), A(j^{(1)}))|L(j^{(1)}), f_0, A_0)\) becomes negligible, annihilating the product corresponding to \(L\) in (63). Thus, for all \(L \in \tilde{\mathcal{L}}\), each product becomes negligible implying that \(\eta(F, A|f_0, A_0)\) is also negligible. Hence

\[
\sum_{L \in \mathcal{L}} \prod_{j=1}^{N_o} P(F(j), A(j)|L(j), f_0, A_0) \approx \sum_{L \in \mathcal{L}^*} \prod_{j=1}^{N_o} P(F(j), A(j)|L(j), f_0, A_0)
\]

as was to be shown.

**APPENDIX III**

**BAYESIAN SMOOTHING INFERENCE RECURSIONS**

Recall from Section IV-B, each iteration of the EM depends on the computation of the pairwise mode posterior (36):

\[
\sigma^{(2)}(M_t, M_{t+1}) = P(M_t, M_{t+1}|Y_{1:t}, \theta_{M}^{(t)})
\]

Quantities propagated in filtered and smoothed passes as well as the necessary inputs are summarized in Table VI.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Quantity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\pi(M_0, S_0))</td>
<td>(P(M_0, S_0))</td>
<td>Prior</td>
</tr>
<tr>
<td>(\pi(M_{t+1}</td>
<td>M_t))</td>
<td>(P(M_{t+1}</td>
</tr>
<tr>
<td>(\pi(M_t, S_t))</td>
<td>(P(M_t, S_t</td>
<td>Y_{1:t-1}))</td>
</tr>
<tr>
<td>(\mu(M_t, S_t))</td>
<td>(P(M_t, S_t</td>
<td>Y_{1:t-1}))</td>
</tr>
<tr>
<td></td>
<td>(P(M_t, S_t</td>
<td>Y_{1:t-1}))</td>
</tr>
<tr>
<td>(\sigma^{(2)}(M_t, M_{t+1}))</td>
<td>(P(M_t, M_{t+1}</td>
<td>Y_{1:t}))</td>
</tr>
<tr>
<td>(\sigma^{(2)}(M_t, M_{t+1}))</td>
<td>(P(M_t, M_{t+1}</td>
<td>Y_{1:t}))</td>
</tr>
<tr>
<td>(\Psi(M_t, S_{t+1}, M_{t+1}))</td>
<td>(\frac{P(M_t, S_{t+1}</td>
<td>M_{t+1})}{P(M_{t+1}</td>
</tr>
<tr>
<td>(\phi(M_t, S_t))</td>
<td>(\frac{P(M_t, S_t</td>
<td>Y_{1:t-1})}{P(M_t, S_t</td>
</tr>
</tbody>
</table>

TABLE VI

Inference inputs and propagated quantities

To initialize the filtering pass, we set \(\mu(M_0, S_0) = \pi(M_0, S_0)\). The remainder of the filtering pass consists of the following recursions for \(t \in 1: K-1\).
The single-frame likelihood evaluation, as well as several suggestions motivating the deterministic approximation for data compression and time/pitch-scale modifications.

\[
\Psi(M_t, S_{t+1}, M_{t+1}) = \sum_{S_t} \left[ \mu(M_t, S_t | Y_{1:t}) \times P(S_{t+1} | S_t, M_{t+1}, M_t) \right] \\
\tau(M_{t+1}, S_{t+1}) = \sum_{M_t} \left[ P(M_{t+1} | M_t) \times \Psi(M_t, S_{t+1}, M_{t+1}) \right] \\
\mu(M_{t+1}, S_{t+1}) = \frac{\tau(M_{t+1}, S_{t+1}) P(Y_{t+1} | M_{t+1}, S_{t+1})}{\sum_{M_{t+1}, S_{t+1}} \tau(M_{t+1}, S_{t+1}) P(Y_{t+1} | M_{t+1}, S_{t+1})} \tag{66}
\]

The smoothing pass begins with the initialization \(\sigma(S_K, M_K) = \mu(S_K, M_K)\). The remainder of the smoothing pass consists of the following recursions, as \(t\) decreases from \(K-1\) down to 1:

\[
\phi(M_{t+1}, S_{t+1}) = \frac{P(M_{t+1}, S_{t+1} | Y_{1:t})}{P(M_{t+1}, S_{t+1} | Y_{1:t})} \\
\sigma(M_t, S_t) = \mu(M_t, S_t) \sum_{M_{t+1}} \left[ P(M_{t+1} | M_t) \times \sum_{S_{t+1}} \phi(M_{t+1}, S_{t+1}) \right] \\
\sigma^{(2)}(M_t, M_{t+1}) = P(M_{t+1} | M_t) \sum_{S_{t+1}} \phi(M_{t+1}, S_{t+1}) \times \Psi(M_t, S_{t+1}, M_{t+1}) \tag{67}
\]

ACKNOWLEDGMENT

The authors would like to thank Brent Townshend for suggestions motivating the deterministic approximation for the single-frame likelihood evaluation, as well as several anonymous reviewers for helpful suggestions and corrections.

REFERENCES