BAYESIAN IDENTIFICATION OF CLOSELY-SPACED CHORDS FROM SINGLE-FRAME STFT PEAKS

Randal J. Leistikow, Harvey D. Thornburg, Julius O. Smith III, and Jonathan Berger

Center for Computer Research in Music and Acoustics
Stanford University
{randal|harv23|jos|brg}@ccrma.stanford.edu

ABSTRACT

Identifying chords and related musical attributes from digital audio has proven a long-standing problem spanning many decades of research. A robust identification may facilitate automatic transcription, semantic indexing, polyphonic source separation and other emerging applications. To this end, we develop a Bayesian inference engine operating on single-frame STFT peaks. Peak likelihoods conditional on pitch component information are evaluated by an MCMC approach accounting for overlapping harmonics as well as undetected/spurious peaks, thus facilitating operation in noisy environments at very low computational cost. Our inference engine evaluates posterior probabilities of musical attributes such as root, chroma (including inversion), octave and tuning, given STFT peak frequency and amplitude observations. The resultant posteriors become highly concentrated around the correct attributes, as demonstrated using 227 ms piano recordings with -10 dB additive white Gaussian noise.

1. INTRODUCTION

Chord identification has proven to be a long-standing problem in computer music research, despite the innate ability of musically trained humans to readily accomplish the task. A variety of historically successful approaches offer mostly rule-based or "blackboard" schema; of note are [1], [2], [3], among others.

When addressing the problem of automatic chord identification, it is important to consider the extent to which an approach attempts to model the human auditory system. The auditory modeling may be implicit, as in Klapuri's use of the spectral smoothness principle [3], or more explicit, as in Martin's use of a modified Meddis and Hewitt pitch perception model [1]. Rather than modeling the auditory mechanism, our goal is simply to identify chords and related musical attributes as well as possible.

Moreover, the problem becomes as much *cognition* as *perception*. Gang and Berger [4], for instance, emphasize the role of *musical expectations* in learning chord sequences in functional tonal music. In a Bayesian probabilistic framework, musical expectations may be easily encoded as prior and/or conditional distributions involving hidden musical *attributes*, e.g., note, chroma, root, octave, tuning, key, and mode. By so doing, we represent, in a purely statistical framework, pseudocognitive capacities such as temporal integration and the incorporation of knowledge from musical structure.

Several emerging approaches, for instance [5], [6], pursue probabilistic as opposed to rule-based schema. In [6], Sheh and Ellis facilitate temporal fusion via hidden Markov model (HMM) inference, using Fujishima's pitch class profiles [7] as feature observations. Pitch class profiles discard octave information at the front end; however, for applications such as polyphonic transcription, it may be desired to retain the absolute pitches. We adopt short-time Fourier transform (STFT) peaks as features, enabling attributes such as octave to be retained or discarded on the back end, whichever the user may decide.

In [8], Goldstein introduces a probabilistic maximum likelihood pitch inference using STFT peak frequencies as feature observations. A related approach, developed by Thornburg and Leistikow [9], handles also spurious peaks from interference events, additionally incorporating timbral knowledge by statistically modeling joint frequency and amplitude peak observations. Furthermore, the Thornburg-Leistikow method may explicitly evaluate the likelihood of *any* candidate pitch component, rather than merely identifying the most likely component.

In this paper, we extend Thornburg and Leistikow's single pitch likelihood evaluation to the multipitch case. Subsequently, the latter is embedded in a Bayesian chord identification schema inferring musical root, inversion, octave, and tuning as well as the individual pitch components comprising a chord.

2. PROBABILISTIC MODEL

Our probabilistic model is shown in Figure 1:

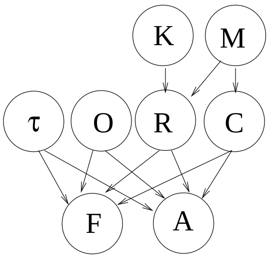

Figure 1: *Probabilistic model.*

Here, F consists of a list of observed STFT peak frequencies, and A the parallel list of amplitudes. Observed quantities are directly influenced by the following hidden *attributes*: tuning τ, octave O, root R, and chroma C. The latter are further influenced by higher-level contextual attributes: key K; mode M. Though key and mode inference proves difficult for single-frame data, the additional structure may facilitate multiframe extensions, as key and mode are likely to be constant across long segments of consecutive frames.

2.1. Attribute definitions

- **Tuning**: τ, in units of semitones, is discretized to a set of N_τ equally spaced values $\tau \in \{-0.5 + (l-1)/N_\tau\}_{l=1}^{N_\tau}$.
- **Key**: K takes the value of one of the twelve semitones in an octave $K \in \{0, \ldots, 11\}$.
- **Mode**: The pair $\{K, M\}$ designates what is usually called a "key" in music theory. For example: $M \in \{\mathrm{Major},\ \mathrm{minor}\}$.
- **Root**: R occupies one of the twelve semitones in an octave. The root is relative to a given key: $R \in \{0, \ldots, 11\}$.
- **Octave**: O belongs to a consecutive integer range: $O \in \{O_{min}, \ldots, O_{max}\}$.
- **Chroma**: C is represented by a set of intervals from the root. For instance, a minor triad is expressed: $C = \{0, 3, 7\}$. Similarly a Major-minor seventh chord admits the representation: $C = \{0, 4, 7, 10\}$. C may belong to a standard codebook consisting of major, minor, augmented, and diminished triads, as well as the latter with major and minor sevenths added, with all standard inversions represented (three inversions in case of triads, four inversions in case of seventh chords). In total 44 chromas are represented of which 42 are unique thanks to the inherent ambiguity of augmented triad inversions.

2.2. Bayesian attribute inference

The factorization of the joint $P(\tau, O, R, C, K, M, F, A)$, represented by the directed acyclic graph in Figure 1, is given as follows.

$$\begin{aligned} P(\tau, O, R, C, K, M, F, A) &= P(M)P(K)P(C|M) \\ &\times\ P(R|K, M)P(\tau)P(O) \\ &\times\ P(F, A|\tau, O, R, C) \end{aligned} \quad (1)$$

Priors, $P(M)$,$P(K)$,$P(C|M)$,$P(R|K,M)$,$P(O)$ and $P(\tau)$, encode domain-specific knowledge for a single STFT frame. The *peak likelihood* is given by $P(F, A|\tau, O, R, C)$. Such constitutes the raw information needed to perform any *attribute inference* query.

Suppose we wish to identify some attribute, (say, a chroma C from the 44 possibilities), given only the peak observations F, A. Our criterion is to construct a classifier $\mathcal{T}(F, A) = \hat{C}$ such that the probability of error ($\hat{C} \neq C$) is minimized. Formally, we desire:

$$\mathcal{T}^*(F, A) \quad = \quad \underset{\mathcal{T}(F,A)}{\operatorname{argmin}}\ P(\mathcal{T}(F, A) \neq C) \quad (2)$$

It is readily shown [10] that $\mathcal{T}^*(F, A)$ optimizing (2) maximizes the posterior $P(C|F, A)$:

$$\mathcal{T}^*(F, A) \quad = \quad \underset{C}{\operatorname{argmax}}\ P(C|F, A) \quad (3)$$

The classifier (3) is called *maximum a posteriori* (MAP).

As optimal decisions involving *any* collection of musical attributes derive from analogous MAP rules, the key inference step involves the associated posterior. The latter may be derived by marginalizing irrelevant attributes from $P(\tau, O, R, C, K, M|F, A)$. For instance, the associated posterior for chord recognition concerns chroma and root:

$$P(C, R|F, A) = \sum_{\tau, O, K, M} P(\tau, O, R, C, K, M|F, A) \quad (4)$$

2.3. Specification of priors

Priors $P(M)$, $P(K)$, $P(C|M)$, $P(R|K, M)$, $P(O)$, $P(\tau)$, encoding domain-specific knowledge, become particularly concentrated when conditioning across frames, factoring across two levels: literal frame-to-frame continuities, and structural dependences across note transitions. Little can be said, however, when considering a single frame. Where we lack apriori knowledge altogether (M, K, O, τ), maximum entropy arguments indicate the use of a uniform prior.

However, $P(C|M)$ and $P(R|K, M)$ may encode information specific to a given corpus. In functional tonal music, certain chroma are more common than others in a given mode. For instance, a major triad is far more likely than an augmented triad in Major mode. The latter influences $P(C|M)$; for example: $P(C = \{0, 4, 7\}|M = \mathrm{Major}) > P(C = \{0, 4, 8\}|M = \mathrm{Major})$. Similarly, certain roots prove more common than others in a given key and mode. For instance, a chord rooted on the tonic of the key is more common than a chord rooted an augmented fourth higher. The latter influences $P(R|K, M)$; e.g.: $P(R = 0|K = 0, M = \mathrm{Major}) > P(R = 6|K = 0, M = \mathrm{Major})$.

Specification of $P(C|M)$ and $P(R|K, M)$ by such commonsense reasoning may be suitable; however, we prefer to formally train these distributions from a corpus of musical data. We have not done so as of this writing; instead, we install uniform priors as placeholders, effectively removing K and M from the network.

3. MULTIPITCH LIKELIHOOD EVALUATION

The peak likelihood, $P(F, A|\tau, O, R, C)$, is difficult to evaluate without knowing which observed peaks correspond to pitch components (and their associated harmonic numbers) and which peaks are altogether spurious. As such, we condition first upon a hidden layer, displayed in Figure 1. This layer consists of a set of *pitch components*, a *template* and a *descriptor*.

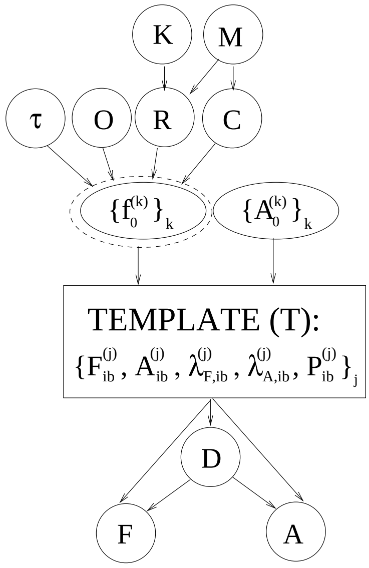
K
M
τ
O
R
C
$\{f_0^{(k)}\}_k$
$\{A_0^{(k)}\}_k$
TEMPLATE (T):
$\{F_{ib}^{(j)}, A_{ib}^{(j)}, \lambda_{F,ib}^{(j)}, \lambda_{A,ib}^{(j)}, P_{ib}^{(j)}\}_j$
D
F
A

Figure 2: *Probabilistic model with exposed hidden layer.*

3.1. Pitch components

For each set of attributes, there exist Q pitch components, where Q is the number of intervals in the chroma. Each (k^{th}) pitch component, $k \in \{1, \ldots, Q\}$, is assigned a *fundamental frequency* $f_0^{(k)}$ and a *reference amplitude* $A_0^{(k)}$. Since the latter is unknown, $A_0^{(k)}$ constitutes a nuisance parameter to be marginalized. Each fundamental frequency, $f_0^{(k)}$, is computed accordingly:

$$f_0^{(k)} \quad = \quad \frac{2\pi \ \cdot \ 2^{\left[C^{(k)}+12\ O+R+\tau\right]/12+c_{440}}}{SR} \tag{5}$$

where $c_{440} = 4.0314$ ensures $A4 \leftrightarrow 440$ Hz, and SR is the sampling rate in Hz.

3.2. Template

First consider the situation where every observed STFT peak results from, and hence may be linked to, exactly one harmonic of a single pitch component.

Define $k^{(j)}$ to be the index of the pitch component to which the j^{th} STFT peak corresponds, and let $h^{(j)}$ denote the harmonic number. In the absence of inharmonicity and noise, the ideal observed frequency would be $h^{(j)} f_0^{\left(k^{(j)}\right)}$.

Let $F^{(j)}$ denote the peak frequency observation corresponding to the j^{th} observed STFT peak. We take the latter's distribution to be Gaussian with mean $F_{ib}^{(j)}$ and variance $\lambda_{F,ib}^{(j)}$:

$$F^{(j)} \quad \sim \quad \mathcal{N}\left(F_{ib}^{(j)}, \lambda_{F,ib}^{(j)}\right) \tag{6}$$

where

$$\begin{aligned} F_{ib}^{(j)} &= h^{(j)} f_0^{\left(k^{(j)}\right)} \\ \lambda_{F,ib}^{(j)} &= \Lambda_{F,ib} F_{ib}^{(j)} F_{ib}^{(1)} \end{aligned} \tag{7}$$

where $\Lambda_{F,ib}$ is a user-specified noise variance scaling.

Let $A^{(j)}$ denote the peak amplitude observation corresponding to the j^{th} observed peak. In the absence of noise, $A^{(j)} = A_0^{\left(k^{(j)}\right)} c_A^{h^{(j)}}$, where c_A is a user-specified spectral decay parameter. To allow for noise, $\left[A^{(j)}\right]^2$ is most appropriately modeled as a scaled noncentral χ_2^2; following [9]:

$$P\left(2\left[A_{ib}^{(j)}\right]^2 / \lambda_{A,ib}^{(j)}\right) \quad \sim \quad \chi^2_{2,\left(2\left[A_{ib}^{(j)}\right]^2/\lambda_{A,ib}^{(j)}\right)} \tag{8}$$

where

$$\begin{aligned} A_{ib}^{(j)} &= A_0^{(j)} c_A^{h^{(j)}} \\ \lambda_{A,ib}^{(j)} &= \Lambda_{A,ib} A_{ib}^{(j)} A_{ib}^{(1)} \end{aligned} \tag{9}$$

Of course, not all template peaks may appear in the STFT. Each template peak has a prior probability of being detected, denoted as $P_b^{(j)}$. The latter is modeled as decaying geometrically with the harmonic number:

$$P_b^{(j)} \quad = \quad \phi_b^{h^{(j)}} \tag{10}$$

The aforementioned distributional parameters are organized into a *template* T, contaning all information necessary to evaluate
$P\left(F, A | \tau, O, R, C\right)$.

$$T \quad \stackrel{\Delta}{=} \quad \{F_{ib}^{(j)}, \lambda_{F,ib}^{(j)}, A_{ib}^{(j)}, \lambda_{A,ib}^{(j)}, P_b^{(j)}\}_{j=1}^{N_{ib}} \tag{11}$$

3.3. Merging overlapped harmonics

A significant complication arises in the multipitch case where harmonics from different pitch components fail to be resolved by the STFT. The minimum frequency distance Δf between harmonics sufficient to resolve peaks depends on the analysis window's length and shape, via the mainlobe width of the latter's discrete-time Fourier transform (DTFT). For the length-M Hamming window used in our STFT front end, $\Delta f = 8\pi/M$.

As each template peak is meant to describe the distribution of at most one observed STFT peak, we *merge* template peaks into clusters for which each peak frequency exists in a Δf-neighborhood of some other frequency within the cluster. As this clustering forms an equivalence relation, each peak in the original template is assigned to exactly one cluster.

Upon merging, each cluster, indexed by $\{(j,l)\}_{l=1}^{N_{ib}^{(j)}}$, is replaced by a single template peak. In the above, we let l^* refer to the index for which the amplitude noncentrality parameter $A_{ib}^{(j,l)}$ is largest; we call the corresponding peak the *primary* peak.

Merged peak template parameters are obtained as follows.

- The merged frequency mean adopts the mean-amplitude-weighted average over the frequency means for each peak in the cluster:

$$F_{ib}^{(j)} \quad = \quad \frac{\sum_{l=1}^{N_{ib}^{(j)}} A_{ib}^{(j,l)} F_{ib}^{(j,l)}}{\sum_{l=1}^{N_{ib}^{(j)}} A_{ib}^{(j,l)}} \tag{12}$$

- The natural frequency variance is given via (7) as a continuous function of frequency. Hence, the frequency variances among all peaks in a given cluster should be roughly the same. We obtain the natural variance from the primary peak, and add to this the square of the spread of the frequency means within the cluster:

$$\lambda_{F,ib}^{(j)} \quad = \quad \lambda_{F,ib}^{(j,l^*)} + \left[\max_l \ F_{ib}^{(j,l)} - \min_l \ F_{ib}^{(j,l)}\right]^2 \tag{13}$$

- Peaks may overlap in any phase relationship. As such, the merged peak's noncentrality parameter is taken to equal that of the primary peak, while the scale parameter adds to that of the primary peak, the squared amplitudes of the other peaks within the cluster. This scale parameter specification becomes exaggerated, accounting for only the worst cases: where all peaks interfere exactly in phase, *or* where all peaks but the primary peak interfere with the latter exactly 180° out of phase:

$$A_{ib}^{(j)} \quad = \quad A_{ib}^{(j,l^*)} \tag{14}$$

$$\lambda_{A,ib}^{(j)} \quad = \quad \lambda_{A,ib}^{(j,l^*)} + \sum_{l=1, l\neq l^*}^{N_{ib}^{(j)}} \left[A_{ib}\right]^2 \tag{15}$$

- The merged peak's survival probability is taken to equal that of the primary peak:

$$P_b^{(j)} \quad = \quad P_b^{(j,l^*)} \tag{16}$$

3.4. Descriptor and linkmap representation

Of course, the template only accounts for *potential* STFT peaks, as excessive noise and other types of interference may prevent the detection of some peaks associated with the template. Furthermore, interference events may generate *spurious* peaks in the observed peaklist which have no relation to those in the template.

To evaluate the likelihood of the observed peaklist, then, we condition upon a *descriptor* D encoding linkage between template and observed peaks. The desired unconditional likelihood $P(F,A|T)$ evaluates by summing over conditional likelihoods $P(F,A|D,T)$ weighted by an appropriate prior $P(D,T)$:

$$P(F,A|T) \;=\; \sum_{D} P(F,A|D,T)P(D|T) \qquad (17)$$

In [9], the authors propose a symbolic encoding of D enabling distributions analogous to $P(F,A|D,T)$ for the single-pitch case to be written in closed algebraic form. In this paper, we adopt only the graphical *linkmap* representation, shown in [9] to be equivalent. Figure 3 illustrates an example linkmap. The scenario represents a perfect fifth interval where the third harmonic from the root and the second harmonic from the fifth are merged. Three pairs of linked peaks, two undetected peaks, and two spurious peaks are evidenced by the Figure. Evaluating $P(F,A|D,T)$ and $P(D|T)$ requires a probabilistic model for the generation of spurious peaks. The spurious peak frequency distribution is modeled according to a Poisson process, while the squared-amplitude distribution is modeled according to a scaled *central* χ^2_2 with scaling parameter $\sigma^2_{A,o}$ (see [9], Section 2).

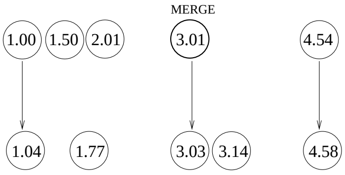

Figure 3: *Linkmap example: perfect fifth interval. All frequencies are expressed in ratios to the root's fundamental frequency.*

Robustness in the presence of spurious peaks is primarily due to the difference in conditional likelihoods between spurious and linked peaks. If one of the peak amplitudes/frequencies is highly inconsistent with respect to all template peak distributions, the contribution $P(F,A|D,T)P(D|T)$ will be quite small for those D for which the peak is linked, relative to those D for which this peak is spurious. Hence, most of the unconditional likelihood will concentrate in D for which this peak is spurious.

We now consider the benefits of merging. Figure 4 shows a hypothetical situation prior to merging template peaks. Since the amplitude of the overlap-interference peak is highly inconsistent with the template peak distributions (9), the present solution ends up discarding information from overlap-interference peaks altogether, effectively labeling them as spurious[1]. While overlap-interference peak amplitudes may be unreliable, the associated frequencies are especially likely to carry useful information as they correspond to template peaks from multiple pitch components. We therefore desire a merge operation which induces a distributional model concentrating on linked overlap-interference peaks. In this manner, merging enables the effective discounting of amplitude observations while retaining useful frequency information.

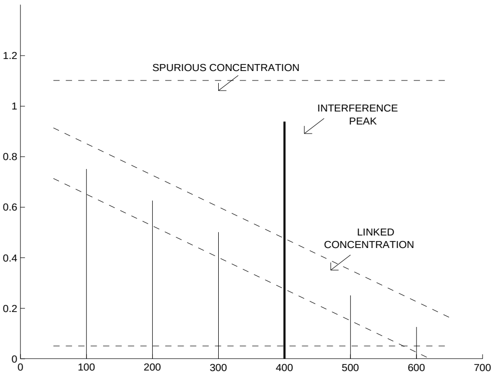

Figure 4: *Illustration of STFT interference peak due to overlapping harmonics in the context of linked and spurious amplitude distributions.*

3.5. MCMC unconditional likelihood evaluation

The number of possible linkmaps and hence descriptors, with N_{ib} template peaks and N observed peaks, is as follows:

$$\#\{D\} \;=\; \sum_{n=0}^{\min(N_{ib},N)} \binom{N_{ib}}{n}\binom{N}{n} \qquad (18)$$

If $N_{ib} = N$, (18) and Stirling's approximation [11] yield:

$$\begin{aligned}\#\{D\} \;&=\; \binom{2N}{N} \\ &\approx\; \frac{4^N}{\sqrt{\pi N}}\left[1-\mathcal{O}\left(\frac{1}{N}\right)\right] \qquad (19)\end{aligned}$$

A further complication arises in that $P(F,A|D,T)$, via (9,11), requires knowledge of the reference amplitudes $A_{0,Q} \triangleq \{A_0^{(k)}\}_{k=1}^{Q}$, where Q denotes the number of pitch components. Our solution is to marginalize the $A_{0,Q}$. Here, each $A_0^{(k)} \in A_{0,Q}$ is discretized to a grid $\mathcal{A}$, the latter consisting of N_A amplitudes uniformly positioned in dB space on a [-9 dB, +9 dB] interval relative to the highest amplitude peak. Thus $A_{0,Q}$ belongs to the product space $\mathcal{A}^Q$ consisting of N_A^Q discrete possibilities. A uniform prior $P(A_{0,Q})$ is placed on these possibilities. Marginalization of $A_{0,Q}$ and (17) yield:

$$P(F,A|T) = \sum_{D\in\mathcal{D},A_{0,Q}\in\mathcal{A}^Q} P(F,A|D,T(A_{0,Q}))P(A_{0,Q})P(D|T(A_{0,Q})) \qquad (20)$$

The number of terms in the summation (20) grows exponentially with the common number of template and observed peaks, as well as the number of pitch components.

[1] Technically speaking, the conditional likelihood concentrates in descriptors D for which interference peaks are unlinked.

To facilitate the computation, we introduce a Markov chain Monte Carlo (MCMC) approximate enumeration, analogous to Section 4 of [9]. In practice, virtually all of the unconditional likelihood concentrates in a few $\{D, A_{0,Q}\}$-possibilities. Our goal then becomes to construct a Markov chain traversing the $\{D, A_{0,Q}\}$ with the highest contributions to the sum (20). To this end, we specify the stationary distribution $\pi(D, A_{0,Q})$, as proportional to $[P(F, A|D, T(A_{0,Q}))P(A_{0,Q})P(D|T(A_{0,Q}))]^\gamma$, where $\gamma > 1$.

The Metropolis-Hastings rule [12] is a general method for constructing a Markov chain admitting a desired $\pi(D, A_{0,Q})$, with a user-specified *sampling distribution* $q(D', A'_{0,Q}|D, A_{0,Q})$. As long as q is irreducible, the algorithm will converge, though arbitrarily slowly unless this sampling distribution is well chosen. Our choice mixes alterations in $A_{0,Q}$ (via $q_A(A'_{0,Q}|A_{0,Q})$) independently with alterations in D (via $q_D(D'|D)$), each chosen with probability 0.5.

In case of q_A, we select one $A_0^{(k)}$ uniformly among the Q possibilities and move it up or down one gridpoint (ΔdB level), except for the boundary cases where only one adjacency exists. In case of q_D, we exploit a similar notion of adjacency on the space of linkmaps. The following types of moves exist, as shown in Figure 5: 1.) *Remove a link*; 2.) *Add a nonintersecting link*; 3.) *Switch a link to the adjacent position on input*; 4.) *Switch a link to the adjacent position on output*.

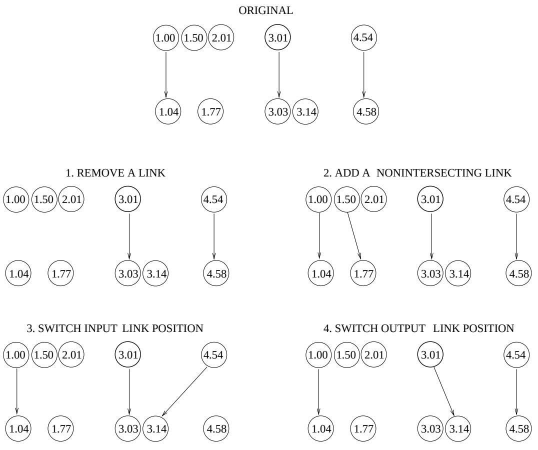
ORIGINAL
1.00
1.50
2.01
3.01
4.54
1.04
1.77
3.03
3.14
4.58
1. REMOVE A LINK
2. ADD A NONINTERSECTING LINK
3. SWITCH INPUT LINK POSITION
4. SWITCH OUTPUT LINK POSITION

Figure 5: *Example linkmap moves.*

A move type is selected with uniform probability over the types with at least one move possibility, then a move is selected uniformly among those possibilities.

The general Metropolis-Hastings rule proceeds over iterations i as follows. As a shorthand, define the state $S^{(i)} = \{D^{(i)}, A_{0,Q}^{(i)}\}$, and define $S'^{(i)}, S^{(i+1)}$ analogously.

1. Sample $S'^{(i)} \sim q(S'^{(i)}|S^{(i)})$
2. Select

$$S^{(i+1)} = \begin{cases} S'^{(i)} & \text{w. prob} \quad \alpha(S^{(i)}, S'^{(i)}) \\ S^{(i)} & \text{w. prob} \quad 1 - \alpha(S^{(i)}, S'^{(i)}) \end{cases}$$

where

$$\alpha(S^{(i)}, S'^{(i)}) = \min\left(1, \frac{\pi(S'^{(i)})q(S^{(i)}|S'^{(i)})}{\pi(S^{(i)})q(S'^{(i)}|S^{(i)})}\right)$$

The irreducibility of q follows from the irreducibilities of q_D and q_A and the fact either distribution may be selected with probability 0.5, which is strictly positive. The irreducibility of q_A obtains since one may traverse any configuration of grid points in $\mathcal{A}$ for each $A_{0,Q}$ by accumulating adjacent steps, each with probability at least $1/(2Q)$. Similarly, the irreducibility of q_D follows since one can reach any linkmap from any other linkmap by removing and adding links one-by-one. There are only finitely many such possibilities each of which occurs with strictly positive probability.

The initialization of $A_{0,Q}$ proceeds by taking all elements $A_0^{(k)}$ equal to the maximum STFT peak amplitude. The initialization of D follows by McAulay-Quatieri peak matching [9], [13]. In practice, we obtain rapid convergence using only ~ 600 MCMC iterations. Subsequent optimizations are as follows: first, we hash intermediate computations for previously visited D and/or $A_{0,Q}$ values; second, we vary γ according to the annealing schedule $\gamma^{(i)} = 0.1, \gamma^{(i+1)} = \min\left(1.005\gamma^{(i)}, 5\right)$, allowing a wider range of $\{D, A_{0,Q}\}$ possibilities to be visited at the outset.

Prior and Posterior Chroma Probabilities

True Chroma	*Cand. Chroma*	*Prior*	*Posterior*
min_inv0 {0,3,7}	**min_inv0**	**0.1666667**	**1**
min_inv0	min_inv1	0.1666667	0
min_inv0	min_inv2	0.1666667	2.261584e-56
min_inv0	Maj_inv0	0.1666667	0
min_inv0	Maj_inv1	0.1666667	0
min_inv0	Maj_inv2	0.1666667	0
min_inv1{3,7,12}	min_inv0	0.1666667	9.525349e-70
min_inv1	**min_inv1**	**0.1666667**	**1**
min_inv1	min_inv2	0.1666667	4.509738e-98
min_inv1	Maj_inv0	0.1666667	0
min_inv1	Maj_inv1	0.1666667	0
min_inv1	Maj_inv2	0.1666667	0
min_inv2 {7,12,15}	min_inv0	0.1666667	2.676833e-145
min_inv2	min_inv1	0.1666667	4.012468e-06
min_inv2	**min_inv2**	**0.1666667**	**0.999996**
min_inv2	Maj_inv0	0.1666667	0
min_inv2	Maj_inv1	0.1666667	0
min_inv2	Maj_inv2	0.1666667	0
Maj_inv0 {0,4,7}	min_inv0	0.1666667	0
Maj_inv0	min_inv1	0.1666667	0
Maj_inv0	min_inv2	0.1666667	0
Maj_inv0	**Maj_inv0**	**0.1666667**	**1**
Maj_inv0	Maj_inv1	0.1666667	6.865183e-83
Maj_inv0	Maj_inv2	0.1666667	1.348039e-57
Maj_inv1 {4,7,12}	min_inv0	0.1666667	9.678233e-263
Maj_inv1	min_inv1	0.1666667	4.088142e-222
Maj_inv1	min_inv2	0.1666667	5.335909e-321
Maj_inv1	Maj_inv0	0.1666667	7.255914e-19
Maj_inv1	**Maj_inv1**	**0.1666667**	**1**
Maj_inv1	Maj_inv2	0.1666667	6.160874e-39
Maj_inv2 {7,12,16}	min_inv0	0.1666667	0
Maj_inv2	min_inv1	0.1666667	0
Maj_inv2	min_inv2	0.1666667	0
Maj_inv2	Maj_inv0	0.1666667	2.125255e-128
Maj_inv2	Maj_inv1	0.1666667	6.687174e-14
Maj_inv2	**Maj_inv2**	**0.1666667**	**1**

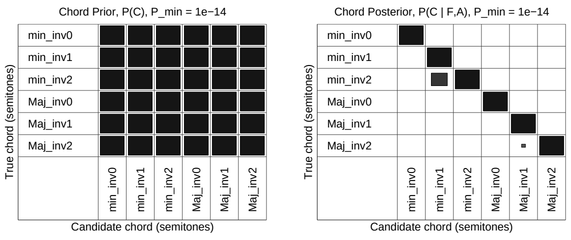

Table 1: *Prior and posterior chroma probabilities.*

4. RESULTS

Tables 1–3 display confusion maps showing associated posterior probabilities. Here, all attributes not displayed in each table are

Prior and Posterior Octave Probabilities			
True Octave	*Cand. Octave*	*Prior*	*Posterior*
2	**2**	**0.2**	**1**
2	3	0.2	0
2	4	0.2	0
2	5	0.2	0
3	2	0.2	1.553992e-186
3	**3**	**0.2**	**1**
3	4	0.2	0
3	5	0.2	0
4	2	0.2	0
4	3	0.2	6.920196e-44
4	**4**	**0.2**	**1**
4	5	0.2	0
5	2	0.2	3.219936e-177
5	3	0.2	5.635765e-85
5	4	0.2	4.813075e-11
5	**5**	**0.2**	**1**

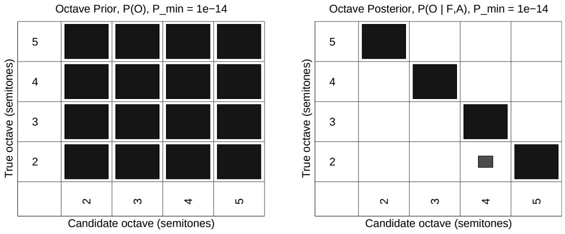

Table 2: *Prior and posterior octave probabilities.*

Prior and Posterior Tuning Probabilities			
True Tuning	*Cand. Tuning*	*Prior*	*Posterior*
-0.25	**-0.25**	**0.3333333**	**1**
-0.25	0	0.3333333	2.7864e-216
-0.25	0.25	0.3333333	3.868828e-283
0	-0.25	0.3333333	3.94324e-258
0	**0**	**0.3333333**	**1**
0	0.25	0.3333333	8.221319e-218
0.25	-0.25	0.3333333	3.664227e-308
0.25	0	0.3333333	6.476317e-252
0.25	**0.25**	**0.3333333**	**1**

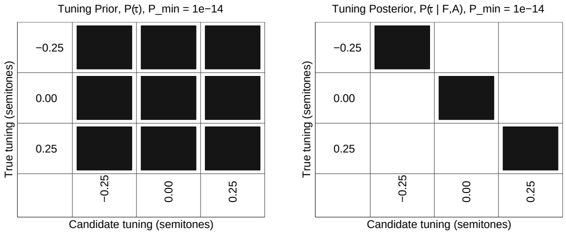

Table 3: *Prior and posterior tuning probabilities.*

marginalized in the corresponding posterior probability computation. To generate these maps, we derive input signals by mixing several single-note piano recordings together with -10 dB additive Gaussian white noise. The STFT analysis operates on a 227 ms frame immediately following the attack region. We specify the following user parameter values: $C_A = 0.8$, $\Lambda_{F,ib} = 0.001$, $\Lambda_{A,ib} = 0.5$, $\phi_b = 0.8$, $\sigma^2_{A,o} = \left(0.008A_0^{(k)}\right)^2$.

Indeed, the resultant posteriors seem highly concentrated about the correct attributes. Concentration results, however, remain somewhat sensitive to specific user parameter settings; once set, however, the parameters seem to generalize well to a variety of signals of similar type. The given settings seem to perform especially well for tuning inference, demonstrating "perfect pitch" within 0.25 semitones, thus exceeding the abilities of most human listeners. For chroma inference, the maximum confusion seems to be between adjacent inversions (on the level of $4.0 \cdot 10^{-6}$); in particular, the confusion is asymmetric: most of the remaining probability concentrates on the lower adjacent inversion. Similarly, the maximum confusion for octave inference, though miniscule $4.8 \cdot 10^{-11}$, concentrates asymmetrically on the suboctave.

Future work necessarily involves learning user-specified parameters, via EM, enlargement of the MCMC space, and/or other techniques.

5. REFERENCES

[1] Keith D. Martin, "Automatic transcription of simple polyphonic music: Robust front end processing," *MIT Media Laboratory Perceptual Computing Technical Report No. 399*, 1996.

[2] Anssi Klapuri, "Number theoretical means of resolving a mixture of several harmonic sounds," in *Proc. EUSIPCO*, Rhodes, Greece, 1998.

[3] Anssi Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," in *Proc. ICASSP*, Salt Lake City, UT, 2001, pp. 3381–3384.

[4] Dan Gang and Jonathan Berger, "Modeling the degree of realized expectation in functional tonal music: A study of perceptual and cognitive modeling using neural networks," in *Proc. ICMC*, Hong Kong, 1996, Int'l Computer Music Association, pp. 454–457.

[5] Kunio Kashino, Kazuhiro Nakadai, Tomoyoshi Kinoshita, and Hidehiko Tanaka, "Application of Bayesian probability network to music scene analysis," in *Working Notes of IJCAI Workshop of Computational Auditory Scene Analysis IJCAI-CASA*, 1995, pp. 52–59.

[6] Alexander Sheh and Daniel P.W. Ellis, "Chord segmentation and recognition using EM-trained hidden Markov models," in *Proc. 4th Int'l Symposium on Music Information Retrieval ISMIR-03*, Baltimore, MD, 2003.

[7] Takuya Fujishima, "Realtime chord recognition of musical sound: A system using Common Lisp Music," in *Proc. ICMC*, Beijing, 1999, Int'l Computer Music Association, pp. 464–467.

[8] Julius Goldstein, "An optimum processor theory for the central formation of the pitch of complex tones," *Journal of the Acoustical Society of America*, vol. 54, pp. 1496–1516, 1973.

[9] Harvey D. Thornburg and Randal J. Leistikow, "A new probabilistic spectral pitch estimator: Exact and MCMC-approximate strategies," Esbjerg, Denmark, 2004, CMMR-2004: Computer Music Modeling and Retrieval.

[10] Christopher M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.

[11] Donald Knuth, Ilan Vardi, and Rolf Richberg, "6581 (the asymptotic expansion of the middle binomial coefficient)," *American Mathematical Monthly*, vol. 97, no. 7, pp. 626–630, 1990.

[12] William J. Fitzgerald, "Markov chain Monte Carlo methods with applications to signal processing," *Elesevier Signal Processing*, vol. 81, no. 1, pp. 3–18, 2001.

[13] Robert J. McAulay and Thomas F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. ASSP*, vol. 34, no. 4, pp. 744–754, 1986.