BAYESIAN SPECTRAL MATCHING:
TURNING YOUNG MC INTO MC HAMMER VIA MCMC SAMPLING
Matthew D. Hoffman†, Perry R. Cook†‡, David M. Blei†
Princeton University
† Department of Computer Science, ‡Department of Music, Princeton, NJ, USA

ABSTRACT
In this paper, we introduce an audio mosaicing technique based on performing posterior inference on a probabilistic generative model. Whereas previous approaches to concatenative synthesis and audio mosaicing have mostly tried to match higher-level descriptors of audio or individual STFT frames, we try to directly match the magnitude spectrogram of a target sound by combining and overlapping a set of short samples at different times and amplitudes. Our use of the graphical modeling formalism allows us to use a standard Markov Chain Monte Carlo (MCMC) posterior inference algorithm to find a set of time shifts and amplitudes for each sample that results in a layered composite sound whose spectrogram approximately matches the target spectrogram.

1. INTRODUCTION
Concatenative synthesis and audio mosaicing techniques take databases of recorded sounds and attempt to combine them to produce a sound matching a target specification [5, 7, 3].

In this paper, we propose an audio mosaicing technique that attempts to solve the following problem: given a set of (short) recorded source sounds, how can we match a (longer) target sound as closely as possible by repeating and combining our source sounds at different times and amplitudes? More formally, we have a set of \( K \) source sounds \( x_k \), and we want to find a set of \( K \) functions \( g(t, k) \) with which to convolve each sound \( x_k \) such that the sum of these convolutions \( z \) is perceptually similar to our target sound:

\[
z(t) = \sum_{k=1}^{K} \sum_{u=0}^{\infty} g(t-u, k)x_k(u)
\]

(1)

Since we allow our source sounds to overlap in time, the dimensionality of the space of possible output sounds grows exponentially with the number of source sounds, and finding a globally optimal solution becomes difficult. We take a probabilistic modeling approach that allows us to apply standard techniques from Bayesian statistics.

We define a probabilistic generative model, the Shift-Invariant Mixture of Multinomials (SIMM), corresponding to a process by which we will generate our output sound from our source sounds, and assume that this model generated our target sound. SIMM has a matrix of hidden variables \( \omega \) that correspond to the functions \( g(t, k) \) that we want to find. We can find a good set of functions \( g(t, k) \) by finding a value for \( \omega \) with high posterior likelihood given the target sound—that is, a value for \( \omega \) that could plausibly have led to our model generating our target sound. Our probabilistic framework allows us to use a Gibbs sampling algorithm to perform approximate posterior inference [4].

In the sequel, we describe our generative model, define a Gibbs sampler to infer the model’s hidden variables, show how those hidden variables tell us how to produce our output sound, and present the results of applying our approach to various combinations of input sources and target sounds.

2. THE SIMM MODEL
Our SIMM model is adapted from the Shift-Invariant Hierarchical Dirichlet Process (SIHDP) [2]. It can be interpreted as a fully Bayesian variant on Shift-Invariant Probabilistic Latent Component Analysis [6].

2.1. Data Representation
We begin by computing the magnitude spectrogram of our target audio using \( W \) non-overlapping windows of \( S \) samples each (multiplied by a Hanning window), yielding \( B = \frac{W}{2} + 1 \) frequency bins per window\(^1\). We will refer to the magnitude in bin \( b \) of window \( w \) as \( y_{wb} \). We normalize the magnitude spectrogram \( \gamma \) so that \( \sum_{b=1}^{B} \sum_{w=1}^{W} y_{wb} = 1 \).

We compute a scaled and quantized version of \( \gamma \), \( \gamma' \), which we will treat as a histogram giving the counts of amplitude quanta at each time \( w \) and frequency bin \( b \):

\[
y'_{wb} = \text{round}(WBVy_{wb})
\]

(2)

\[
N = \sum_{b=1}^{B} \sum_{w=1}^{W} y_{wb}
\]

(3)

\( \nu \) is a constant controlling how finely we quantize the spectrogram. Choosing \( \nu = 1 \) gives us an average of about one

---

\(^1\)A shorter hop size can be used, but using non-overlapping windows is simpler and reduces computational overhead. A lack of time resolution has not been a problem in our experiments.
quantum per window/bin; higher values of \( v \) yield a closer approximation to the continuous spectrogram and more expense. The order of these quanta is arbitrary, so we can model them as being drawn independently from our model.

### 2.2. Generative Process

We assume we are given a set of \( K \) normalized magnitude spectrogram matrices \( \phi_k \) of size \( C \times B \), such that \( \phi_{kcb} \) is the magnitude in frequency bin \( b \) at window \( c \) in sound source \( k \), and \( \sum_{c=1}^{C} \sum_{b=1}^{B} \phi_{kcb} = 1 \) for each \( k \in \{1, \ldots, K\} \). These spectrograms come from the sound sources we will use to reconstruct the target sound. The normalized spectrograms can also be interpreted as joint multinomial distributions over base times \( c \) and bins \( b \). That is, \( \phi_{kcb} \) gives the probability of drawing a quantum \( i \) with base time \( c \) and frequency \( b \) given that the quantum is coming from the \( k \)th source sound.

The generative process for SIMM is:

1. Draw a \( K \times L \) matrix \( \omega \) defining a joint multinomial distribution over sound sources \( k \) and time offsets \( l \) from a symmetric Dirichlet distribution with parameter \( \eta \):

   \[
   \omega \sim \text{Dir}(\eta, \ldots, \eta) \quad (4)
   \]

   \( \omega_{kl} \) is the joint probability of drawing a quantum from source \( k \) with time offset \( l \).

2. For each quantum \( i \in \{1, \ldots, N\} \):
   - Draw a source ID \( k_i \) and a time offset \( l_i \) jointly from \( \text{Mult}(\omega) \):
     \[
     \{k_i, l_i\} \sim \text{Mult}(\omega) \quad (5)
     \]
   - Draw a base time \( c_i \) and a frequency bin \( b_i \) jointly from the spectrogram/joint distribution \( \phi_k \) :
     \[
     \{c_i, b_i\} \sim \text{Mult}(\phi_k) \quad (6)
     \]
   - Set the observed time \( w_i \) for quantum \( i \) based on the base time \( c_i \) and the time offset \( l_i \):
     \[
     w_i = c_i + l_i \quad (7)
     \]

3. For each time \( w \) and frequency \( B \), count the quanta appearing at \( w \) and \( b \) to yield \( \hat{y}_{wb} \), the magnitude in the quantized spectrogram at \( w \) and \( b \).

Each observed quantum \( i \) appears at time \( w_i \) and frequency \( b_i \). We can avoid sampling \( \omega \) since we have placed a conjugate Dirichlet prior on \( \omega \) and can therefore compute the posterior predictive likelihood of \( \{k_i, l_i\} \) given the other \( k \)'s and \( l \)'s (denoted \( k_{-i} \) and \( l_{-i} \)) and the hyperparameter \( \eta \). We therefore resample only the values for the source indicators \( k \) and the time offsets \( l \). This leads to faster convergence, since it lets us work in a lower-dimensional space. Once we have estimates for \( k \) and \( l \), we can compute the Maximum A Posteriori (MAP) value for \( \omega|k, l, \eta \).

To resample each pair \( k_i, l_i \) we need to compute the joint posterior likelihood that the quantum \( i \) appearing at time \( w_i \);
and bin $b_i$ was drawn from a source $k$ at a time offset $l$:

$$P(k_i = k, l_i = l|w_i, b_i, k_{-i}, l_{-i}, \phi, \eta) \propto P(c_i = w_i - l, b_i|k_i = k, l_i = l, \phi) \times P(k_i = k, l_i = l|k_{-i}, l_{-i}, \eta)$$  \(8\)

The joint likelihood of the base time $c_i = w_i - l$ and the frequency bin $b_i$ is given by the component distribution $\phi_k$:

$$P(c_i = w_i - l, b_i|k_i = k, l_i = l, \phi) = \phi_{kc_ib_i}$$  \(9\)

The likelihood of the pair $k, l$ conditioned on $\eta$ and the other source indicators $k_{-i}$ and time offsets $l_{-i}$ is

$$P(k_i = k, l_i = l|k_{-i}, l_{-i}, \eta) = \int_\omega P(\omega|\eta) \omega_{kl} d\omega$$  \(10\)

$$= \frac{n_{kl} + \eta}{N - 1 + KL\eta}$$  \(11\)

Where $n_{kl}$ is the number of other quanta coming from source $k$ with time offset $l$. We can compute the integral in equation 10 analytically because the Dirichlet distribution is conjugate to the multinomial distribution.

Using equations 9 and 10, equation 8 becomes:

$$P(k_i = k, l_i = l|w_i, b_i, k_{-i}, l_{-i}, \phi, \eta) \propto \phi_{kc_ib_i} \frac{n_{kl} + \eta}{N - 1 + KL\eta}$$  \(12\)

We repeatedly resample the source indicator $k_i$ and time offset $l_i$ for each observed quantum $i$ conditioned on the other indicators $k_{-i}$ and $l_{-i}$ until 20 iterations have gone by without the posterior likelihood $P(k, l|w, b, \eta, \phi)$ yielding a new maximum. At this point we assume that the Gibbs sampler has converged and that we have found a set of values for $k$ and $l$ that is likely conditioned on the data.

Once we have drawn values from the posterior for $k$ and $l$, we compute the MAP estimate $\hat{\omega}$ of the joint distribution over sources and times $\omega$ conditioned on $k, l,$ and the hyperparameter $\eta$. Since the prior on $\omega$ is a Dirichlet distribution, the MAP estimate $\hat{\omega}$ of $\omega|k, l, \eta$ is given by:

$$\hat{\omega}_{kl} \propto \max(0, n_{kl} + \eta - 1)$$  \(13\)

Here $n_{kl}$ is the total number of observed quanta that came from source $k$ at time $l$.

### 3.2. Sonifying the MAP Estimate

By sonifying $\hat{\omega}$, the MAP estimate of $\omega$, we can produce an approximate version of our input audio using only the short sources corresponding to the component distributions $\phi$. $\hat{\omega}_{kl}$ gives the amplitude of source $k$ at time offset $l$, which corresponds to sample $S(l - 1)$, where $S$ is the number of samples per window, and samples begin at sample 0. If we convolve each short input source $k$ by a signal $g$ such that

$$g(t, k) = \begin{cases} 0 & \text{if } \mod(t, S) \neq 0 \\ \hat{\omega}_{k, \frac{t}{S} + 1} & \text{if } \mod(t, S) = 0 \end{cases}$$  \(14\)

and add the result for each source, we obtain a signal whose spectrogram approximates the spectrogram of the target. Figure 2 shows an example of the final result of this process.

### 3.3. Resampling $\eta$

$\eta$ controls the sparseness of our joint distribution $\omega$ over times and sources. Rather than specify $\eta$ a priori, we place a gamma prior on $\eta$ and adapt the hyperparameter sampling technique in [1] to resample $\eta$ each iteration.

### 4. EVALUATION

Ultimately the effectiveness of our approach should be evaluated qualitatively. Sound examples generated by the method described in this paper are available at http://www.cs.princeton.edu/~mdhoffma/icmc2009.

We also performed a quantitative evaluation of our approach. We tested SIMM’s ability to find an arrangement of the given components $\phi$ to match the target spectrogram $\hat{y}$ by computing and sonifying a MAP estimate $\hat{\omega}$ of the joint distribution over times and components as described in section 3, then comparing the sum of the magnitudes of the
Table 1. Errors obtained by our approach when trying to match songs by AC/DC and Young MC using various sets of sound sources, and the learned values of the hyperparameter η. In all cases our method outperforms a baseline of white noise. Note that lower errors do not necessarily translate to a more aesthetically interesting result.

<table>
<thead>
<tr>
<th>Sound Source</th>
<th>K</th>
<th>Sample Length</th>
<th>AC/DC Error</th>
<th>η</th>
<th>Young MC Error</th>
<th>η</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise</td>
<td>N/A</td>
<td>N/A</td>
<td>0.6395</td>
<td>N/A</td>
<td>0.6265</td>
<td>N/A</td>
</tr>
<tr>
<td>Ramones</td>
<td>100</td>
<td>116 ms</td>
<td>0.3844</td>
<td>0.004569</td>
<td>0.4039</td>
<td>0.001979</td>
</tr>
<tr>
<td>Ramones</td>
<td>200</td>
<td>116 ms</td>
<td>0.3787</td>
<td>0.002386</td>
<td>0.4100</td>
<td>0.003376</td>
</tr>
<tr>
<td>AC/DC</td>
<td>100</td>
<td>116 ms</td>
<td>0.3455</td>
<td>0.005579</td>
<td>0.3821</td>
<td>0.003689</td>
</tr>
<tr>
<td>AC/DC</td>
<td>200</td>
<td>116 ms</td>
<td>0.3349</td>
<td>0.002382</td>
<td>0.3841</td>
<td>0.001939</td>
</tr>
<tr>
<td>MC Hammer</td>
<td>100</td>
<td>116 ms</td>
<td>0.3838</td>
<td>0.005017</td>
<td>0.3753</td>
<td>0.003875</td>
</tr>
<tr>
<td>MC Hammer</td>
<td>200</td>
<td>116 ms</td>
<td>0.3740</td>
<td>0.002993</td>
<td>0.3732</td>
<td>0.002499</td>
</tr>
<tr>
<td>TIMIT</td>
<td>100</td>
<td>464 ms</td>
<td>0.5898</td>
<td>0.004742</td>
<td>0.6102</td>
<td>0.003262</td>
</tr>
<tr>
<td>TIMIT</td>
<td>200</td>
<td>464 ms</td>
<td>0.5275</td>
<td>0.002097</td>
<td>0.6110</td>
<td>0.001796</td>
</tr>
</tbody>
</table>

5. DISCUSSION

We presented a new audio mosaicing approach that attempts to match the spectrogram of a target sound by combining a vocabulary of shorter sounds at different time offsets and amplitudes. We introduced the SIMM model and showed how to use it to find a set of time offsets and amplitudes that will result in an output sound that matches the target sound. Our probabilistic approach is extensible, and we expect future refinements will yield further interesting results.

6. ACKNOWLEDGMENTS

David M. Blei is supported by ONR 175-6343, NSF CAREER 0745520, and grants from Google and Microsoft.

7. REFERENCES


