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 $\boldsymbol{x}_k$, and we want to find a set of K functions $g(t,k)$ with which to convolve each sound $\boldsymbol{x}_k$ such that the sum of these convolutions $\boldsymbol{z}$ is perceptually similar to our target sound:

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

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 $\boldsymbol{\omega}$ 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 $\boldsymbol{\omega}$ with high posterior likelihood given the target sound—that is, a value for $\boldsymbol{\omega}$ 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 $B = \frac{S}{2} + 1$ frequency bins per window [1]. We will refer to the magnitude in bin b of window w as $\hat{y}_{wb}$. We normalize the magnitude spectrogram $\hat{\boldsymbol{y}}$ so that $\sum_{b=1}^{B}\sum_{w=1}^{W}\hat{y}_{wb} = 1$.

We compute a scaled and quantized version of $\hat{\boldsymbol{y}}$, $\bar{\boldsymbol{y}}$, which we will treat as a histogram giving the counts of amplitude quanta at each time w and frequency bin b:

$$\bar{y}_{wb} = \text{round}(WB\nu\hat{y}_{wb}) \tag{2}$$

$$N = \sum_{b=1}^{B}\sum_{w=1}^{W}\bar{y}_{wb} \tag{3}$$

ν 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.

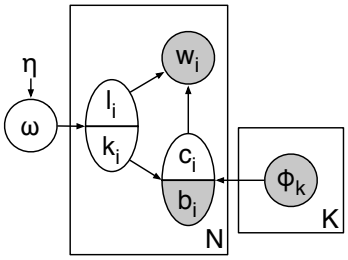

Figure 1. The graphical model for SIMM. Shaded nodes represent observed data, unshaded nodes represent hidden variables. Directed edges denote that the variable pointed to depends on the variable from which the edge originates. Nodes with two variable names denote tuples drawn jointly—for example, c_i and b_i are drawn jointly from a multinomial distribution with parameter ϕ_{k_i}, and depend on both k_i and ϕ_{k_i}. Only b_i is directly observed, so only that half of the node is shaded. Plates denote replication of each variable within the plate by the number at the lower right.

quantum per window/bin; higher values of ν 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 ϕ_k of size $C \times B$, such that ϕ_{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, ϕ_{kcb} gives the probability of drawing a quantum i with base time c and frequency b given that the quantum is coming from the kth source sound.

The generative process for SIMM is:

1. Draw a $K \times L$ matrix $\boldsymbol{\omega}$ defining a joint multinomial distribution over sources k and time offsets l from a symmetric Dirichlet distribution with parameter η:

$$\boldsymbol{\omega} \sim \text{Dir}(\eta,\ldots,\eta) \tag{4}$$

   ω_{kl} is the joint probability of drawing a quantum from source k with time offset l.

2. For each quantum $i \in \{1,\ldots,N\}$:

   (a) Draw a source ID k_i and a time offset l_i jointly from Mult($\boldsymbol{\omega}$):

$$\{k_i, l_i\} \sim \text{Mult}(\boldsymbol{\omega}) \tag{5}$$

   (b) Draw a base time c_i and a frequency bin b_i jointly from the spectrogram/joint distribution ϕ_{k_i}:

$$\{c_i, b_i\} \sim \text{Mult}(\phi_{k_i}) \tag{6}$$

   (c) Set the observed time w_i for quantum i based on the base time c_i and the time offset l_i:

$$w_i = c_i + l_i \tag{7}$$

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

Each observed quantum i appears at time w_i and frequency bin b_i, which are selected according to the process above. We assume that quanta always add constructively. This assumption ignores the possibility of phase cancellation between sources, but it makes our simple mixture modeling approach possible. We leave building a more complicated phase-aware model as future work.

Figure 1 shows SIMM as a graphical model, which summarizes the dependencies between the variables. Given this generative process and an observed spectrogram $\hat{\boldsymbol{y}}$, we will infer values for the process's hidden parameters $\boldsymbol{k}, \boldsymbol{l}, \boldsymbol{\omega}$.

3. INFERENCE AND SYNTHESIS

Our primary objective is to find a good value for the matrix $\boldsymbol{\omega}$, which defines the joint distribution over time offsets l and sources k. Once we have inferred $\boldsymbol{\omega}$ from the data, it will tell us by how much to time-shift and scale each short component to recreate the target sound.

3.1. Gibbs Sampler

We use Gibbs sampling, a Markov Chain Monte Carlo (MCMC) technique that allows us to approximate a sample from the posterior distribution $P(\boldsymbol{k}, \boldsymbol{l} | \boldsymbol{w}, \boldsymbol{b}, \phi, \eta)$, since this distribution is difficult to compute analytically. In Gibbs sampling, we repeatedly sample new values for each variable conditioned on the values of all other variables. After an initial "burn-in" period, the distribution of the sampled $\boldsymbol{k}$ and $\boldsymbol{l}$ converges to their true posterior distribution [4].

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

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:

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

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

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

The likelihood of the pair k, l conditioned on η and the other source indicators $\boldsymbol{k}_{-i}$ and time offsets $\boldsymbol{l}_{-i}$ is

$$\begin{aligned} & P(k_i = k, l_i = l | k_{-i}, l_{-i}, \eta) \\ = & \int_{\boldsymbol{\omega}} P(\boldsymbol{\omega} | \eta) \omega_{kl} d\boldsymbol{\omega} \\ = & \frac{n_{kl} + \eta}{N - 1 + KL\eta} \end{aligned} \tag{10}$$

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:

$$P(k_i = k, l_i = l | w_i, b_i, \boldsymbol{k}_{-i}, \boldsymbol{l}_{-i}, \phi, \eta) \propto \phi_{k c_i b_i} \frac{n_{kl} + \eta}{N - 1 + KL\eta} \tag{11}$$

We repeatedly resample the source indicator k_i and time offset l_i for each observed quantum i conditioned on the other indicators $\boldsymbol{k}_{-i}$ and $\boldsymbol{l}_{-i}$ until 20 iterations have gone by without the posterior likelihood $P(\boldsymbol{k}, \boldsymbol{l} | \boldsymbol{w}, \boldsymbol{b}, \eta, \phi)$ 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 $\boldsymbol{k}$ and $\boldsymbol{l}$ that is likely conditioned on the data.

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

$$\hat{\omega}_{kl} \propto \max(0, n_{kl} + \eta - 1) \tag{12}$$

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 $\boldsymbol{\omega}$, we can produce an approximate version of our input audio using only the short sources corresponding to the component distributions ϕ. $\hat{\omega}_{kl}$ 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

$$g(t,k) = \begin{cases} 0 & \text{if } \mathrm{mod}(t,S) \neq 0 \\ \hat{\omega}_{k, \frac{t}{S}+1} & \text{if } \mathrm{mod}(t,S) = 0 \end{cases} \tag{13}$$

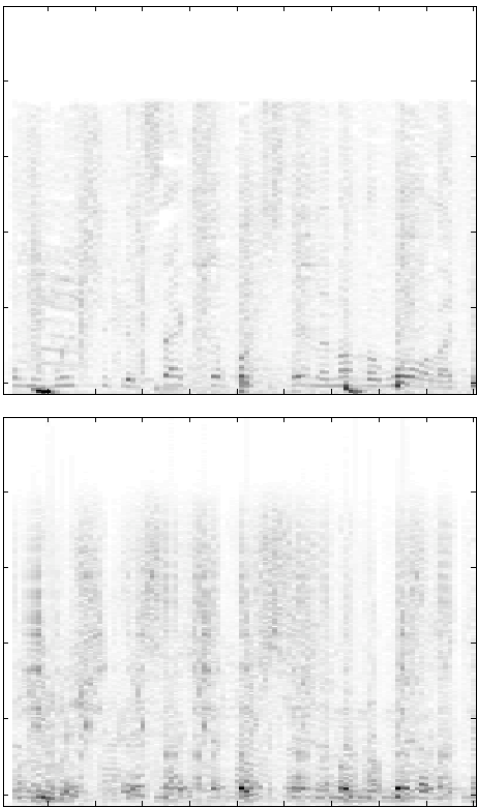

Figure 2. *Top: Spectrogram of 2.3 seconds of Young MC's "Bust a Move." Bottom: Spectrogram of 2.3 seconds of the same song reconstructed from spoken words from the TIMIT corpus using our SIMM model.*

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 η

η controls the sparseness of our joint distribution $\boldsymbol{\omega}$ over times and sources. Rather than specify η a priori, we place a gamma prior on η and adapt the hyperparameter sampling technique in [1] to resample η 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 ϕ to match the target spectrogram $\hat{\boldsymbol{y}}$ 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

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

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.*

differences between the normalized spectrograms of the target sound and resynthesized sound. Let $\hat{z}$ be the normalized spectrogram of the resynthesized sound. Our error metric is

$$err = 0.5 \sum_{b=1}^{B} \sum_{w=1}^{W} |\hat{z}_{wb} - \hat{y}_{wb}| \tag{14}$$

which ranges between 0.0 (perfect agreement between the spectrograms) and 1.0 (no overlap between the spectrograms).

Table 4 presents the errors obtained by our approach when trying to match 23.2 second clips (1000 512-sample windows at 22.05 KHz) from the songs "Dirty Deeds Done Dirt Cheap" by AC/DC and "Bust a Move" by Young MC, using samples selected at random from the songs "Dirty Deeds Done Dirt Cheap," "Blitzkrieg Bop" by the Ramones, and "U Can't Touch This" by MC Hammer. We also used words spoken by various speakers from the TIMIT corpus of recorded speech as source samples. Samples from similar songs tend to produce lower errors, whereas the model had trouble reproducing music using spoken words. The speech samples produce a quantitatively weaker match to the target audio, but the "automatic a cappella" effect of trying to reproduce songs using speech proved aesthetically interesting.

All output sounds are available at the URL given above.

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

[1] M. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, pp. 577–588, 1995.

[2] M. Hoffman, D. Blei, and P. Cook, "Finding latent sources in recorded music with a shift-invariant HDP," in *International Conference on Digital Audio Effects (DAFx) (under review)*, 2009.

[3] A. Lazier and P. Cook, "MOSIEVIUS: Feature driven interactive audio mosaicing," in *International Conference on Digital Audio Effects (DAFx)*, 2003.

[4] R. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Department of Computer Science, University of Toronto, Tech. Rep. CRG-TR-93-1, 1993.

[5] D. Schwarz, "Concatenative sound synthesis: The early years," *Journal of New Music Research*, vol. 35, no. 1, pp. 3–22, 2006.

[6] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 2069–2072.

[7] A. Zils and F. Pachet, "Musical mosaicing," in *International Conference on Digital Audio Effects (DAFx)*, 2001.