
Finding Latent Sources in Recorded Music With a Shift-Invariant HDP

Matthew D. Hoffman, David M. Blei, and Perry R. Cook

MDHOFFMA@CS.PRINCETON.EDU

Computer Science Dept., Princeton University 35 Olden St., Princeton, NJ 08540 USA

1. Introduction

Much interesting work has been done in recent years on analyzing and finding good representations of music audio data for tasks such as content-based recommendation, audio fingerprinting, and automatic metadata generation. However, popular “bag-of-feature-vector” approaches fail to take into account the way that individual sounds evolve over time, and can only model the qualities of the mixed audio signal, not of individual sounds that occur simultaneously. In this paper, we will present the Shift-Invariant Hierarchical Dirichlet Process (SIHDP), a generative model that allows us to represent songs in terms of the instruments and other sounds that generated them.

The same instruments tend to appear in multiple recordings in different combinations, and without hand-generated metadata there is no way of knowing a priori how many or which sources will appear in a given recording. This suggests that a model based on the Hierarchical Dirichlet Process (HDP) would be ideally suited to modeling groups of songs, since it represents groups of observations (such as songs) as being generated by an initially unspecified number of shared latent components (Teh et al., 2007).

However, the HDP requires that our observations be directly comparable, which is not the case for audio data. Human listeners need to hear how a sound evolves over time to recognize and interpret that sound, but computers cannot directly observe when events in audio tracks begin and end. We therefore modified the HDP to make it invariant to shifts in time by explicitly modeling when in each song time-varying latent sources appear.

This allows us to discover a shared vocabulary of latent sources that describe different events in our set of songs, and to produce a rough transcription of each song in terms of that shared vocabulary. This transcription provides a rich representation of our songs with which we can compare and analyze our songs.

We perform posterior inference on the SIHDP using Gibbs sampling. To make it feasible to do inference on large data sets in a reasonable amount of time, we also develop an exact parallel Gibbs sampler for the SIHDP that can also be applied to the original HDP.

We evaluate the SIHDP’s ability to model audio using a

dataset of real popular music, and measure its ability to accurately find patterns in music using a set of synthesized drum loops. Ultimately, our model produces a rich representation of a set of songs consisting of a set of short sound sources and when they appear in each song.

2. A Shift-Invariant Nonparametric Bayesian Model

We define a probabilistic generative model for recorded songs. We assume that a song is generated by repeatedly selecting a sonic component from a set of available components and then choosing when and how strongly it should appear. Such a component might be, for example, a snare drum or the note middle C on a piano. Components may overlap in time. (This resembles the process by which a sample-based sequencer produces audio.)

2.1. Data Representation

We represent each song j using a quantized spectrogram representation. First, we compute the magnitude S -point magnitude spectrogram of our audio using the DFT and a Hanning window. This gives us a $B \times W$ matrix \hat{y}_j of non-negative real numbers, where W is the number of windows and $B = \frac{S}{2} + 1$ is the number of frequency bins. \hat{y}_{jbw} denotes the amplitude of DFT frequency bin b at time step w . Since the overall amplitude scale of digital audio is arbitrary, we can normalize our spectrograms \hat{y}_j so that $\sum_{b=0}^B \sum_{w=1}^W \hat{y}_{jbw} = 1$, and \hat{y}_j defines a multinomial probability distribution over times w and frequencies b .

Finally, we transform each normalized \hat{y}_j into quantized counts data whose empirical distribution approximates \hat{y}_j . We multiply the normalized \hat{y}_j by a constant $\nu \times W \times B$ and round the result to get the number of observed magnitude “quanta” \bar{y}_{jbw} in bin b at time w of song j :

$$\bar{y}_{jbw} = \text{round}(\hat{y}_{jbw}) \quad (1)$$

$$N_j = \sum_{b=1}^B \sum_{w=1}^W \bar{y}_{jbw} \quad (2)$$

ν is roughly the average number of quanta per time/bin pair, and N_j is the total number of observed quanta in song j . We use the notation $y_{ji} = \{w_{ji}, b_{ji}\}$ to refer to the i th

(exchangeable) quantum of energy in song j as occurring at time w_{ji} and bin b_{ji} , for $i \in \{1, \dots, N_j\}$.

2.2. Generative Process

We present the Shift-Invariant Hierarchical Dirichlet Process (SIHDP), a generative model for our quantized spectrogram data that is an extension of the Hierarchical Dirichlet Process (HDP) (Teh et al., 2007).

Our SIHDP model extends the HDP by modeling each observed time w_{ji} as a sum of two terms: a base time $c_{ji} \in \{1, \dots, C\}$ and a time offset $l_{ji} \in \{-C + 1, \dots, W - 1\}$. We define $L = W + C - 1$ to be the number of possible time offsets. We set C to be the length of the latent components that we wish to model (which will be short relative to the song). The time offsets l can take on any range of values such that there is some c for which $l + c \in \{1, \dots, W\}$.

As in the HDP, we begin by drawing a set of latent components ϕ from a symmetric Dirichlet prior with parameter ϵ , but ϕ is now a two-dimensional joint distribution over base times c and frequency bins b . Each ϕ can be interpreted as a normalized spectrogram of a short audio source. The global component proportion vector β is drawn from a stick-breaking process with concentration parameter γ , and the song-level component proportion vector π_j for each song j is drawn from a DP with concentration parameter α and base distribution $\text{Mult}(\beta)$.

Each component k in song j in the SIHDP has a set of multinomial distributions ω_{jk} over time offsets drawn from a symmetric Dirichlet prior with parameter η .

Each observed quantum of energy y_{ji} consists of a time w_{ji} and a frequency bin b_{ji} at which the quantum appears. To generate y_{ji} , we first select a component k_{ji} to generate the quantum. We draw k_{ji} from $\text{Mult}(\pi_j)$, the song-level distribution over components. We then draw a base time c_{ji} and frequency b_{ji} jointly from $\phi_{k_{ji}}$, and draw a time offset l_{ji} from the distribution over time offsets $\omega_{jk_{ji}}$ for component k_{ji} in song j .

The observed quantum appears at time $w_{ji} = c_{ji} + l_{ji}$ and frequency b_{ji} .

The full generative process for the SIHDP is:

$$\begin{aligned}
 \phi_k &\sim \text{Dir}(\epsilon, \dots, \epsilon) & \omega_{jk} &\sim \text{Dir}(\eta, \dots, \eta) \\
 \beta &\sim \text{GEM}(\gamma) & \pi_j &\sim \text{DP}(\alpha, \text{Mult}(\beta)) \\
 k_{ji} &\sim \text{Mult}(\pi_j) & l_{ji} &\sim \text{Mult}(\omega_{jk_{ji}}) \\
 c_{ji}, b_{ji} &\sim \text{Mult}(\phi_{k_{ji}}) & w_{ji} &= c_{ji} + l_{ji} \\
 y_{ji} &= \{w_{ji}, b_{ji}\} & &
 \end{aligned} \tag{3}$$

The SIHDP is a hierarchical nonparametric Bayesian version of Shift-Invariant Probabilistic Latent Component Analysis (SI-PLCA) (Smaragdis et al., 2008). It signifi-

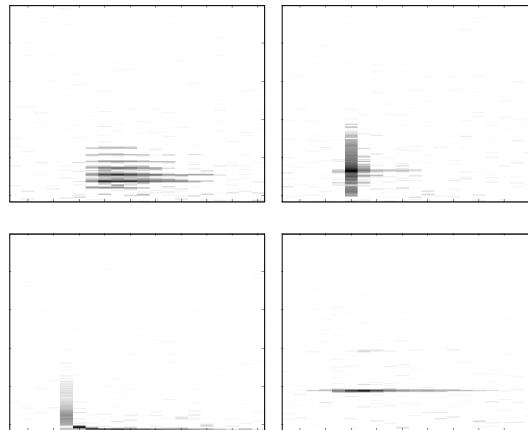


Figure 1. Four latent components discovered from 48 songs taken from the CAL500 corpus of popular music. Qualitatively, these sound like (clockwise from bottom left) a bass drum, a male voice singing “aah,” a snare drum, and a high-pitched whistle.

cantly improves on SI-PLCA by allowing components to be shared across multiple songs, and by automatically determining the number of latent components that are needed to explain the data.

3. Evaluation

We conducted several experiments to test the SIHDP on music audio data using synthetic drum loops and songs taken from the CAL500 dataset (Turnbull et al., 2007). Our model was able to extract good transcriptions of randomly generated drum loops and spectrograms of the isolated drum sounds from mixed spectrogram data. It discovered identifiable instrument sounds from the CAL500 data and basic transcriptions of the songs in terms of these latent sounds. It accomplished this despite having no prior knowledge about music (aside from the maximum allowable length of an instrument sound) or even how many components would be present.

References

- Smaragdis, P., Raj, B., & Shashanka, M. (2008). Sparse and shift-invariant feature extraction from non-negative data. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* (pp. 2069–2072).
- Teh, Y., Jordan, M., Beal, M., & Blei, D. (2007). Hierarchical Dirichlet processes. *Journal of the American Statistical Association, 101*, 1566–1581.
- Turnbull, D., Barrington, L., Torres, D., & Lanckriet, G. (2007). Towards musical query-by-semantic description using the CAL500 data set. *ACM Special Interest Group on Information Retrieval Conference (SIGIR '07)*.