MUSIC ANALYSIS WITH A BAYESIAN DYNAMIC MODEL

*Lu Ren**, *David B. Dunson*†, *Scott Lindroth*†† *and Lawrence Carin**

*Department of Electrical and Computer Engineering
†Department of of Statistical Science
††Department of Music
Duke University, Durham, NC 27708-0291

ABSTRACT

A Bayesian dynamic model is developed to model complex sequential data, with a focus on audio signals from music. The music is represented in terms of a sequence of discrete observations, and the sequence is modeled using a hidden Markov model (HMM) with time-evolving parameters. The model imposes the belief that observations that are temporally proximate are more likely to be drawn from HMMs with similar parameters, while also allowing for "innovation" associated with abrupt changes in the music texture. Segmentation of a given musical piece is constituted via the model inference and the results are compared with other models and also to a conventional music-theoretic analysis.

Index Terms— Music, Sequence, Bayesian, Dynamic model, Nonparametric

1. INTRODUCTION

The analysis of music is of interest to music theorists, for aiding in music teaching, for analysis of human perception of sounds [1], and for design of music search and organization tools [2]. A typical goal of music analysis is to segment a given piece, with the objective of inferring interrelationships among motive and themes within the music. We wish to achieve this task without *a priori* setting the number of segments or their length, motivating a non-parametric framework. In this paper we are interested in processing the acoustic waveform directly, and the proposed techniques are also applicable for analysis of general acoustic signals.

Analyzing sequential data has been a longstanding problem in statistical modeling. With music as an example, Paiement [3] proposed a generative model for rhythms based on the distributions of distances between subsequences; to annotate the changes in mixed music, Plotz [4] used stochastic models based on the Snip-Snap approach, by evaluating the Snip model for the Snap window at every position within the music. However, these methods are either based on one specific factor (rhythm) of music [3] or need prior knowledge of the music's segmentation [4]. Recently, a hidden Markov model (HMM) [5] was used to model monophonic music by assuming all the subsequences are drawn i.i.d. from one HMM [6]; alternatively, an HMM mixture [7] was applied to model the variable time-evolving properties of music, within a semi-parametric Bayesian setting. In both of these models the music was divided into subsequences, with an HMM employed to represent each subsequence; such an approach does not account for the expected statistical relationships between temporally proximate subsequences.

A key aspect of our proposed model is an explicit imposition of the belief that the likelihood that two subsequences of music are similar (contained within the same or related segments) increases as they become more proximate temporally. Additionally, with respect to the application of interest here, music has the property that characteristics of a given piece may repeat over time. Based on these considerations, we propose a nonparametric dynamic mixture model with varying mixture weights, while sharing the same set of components (atoms) at different time points. The model is used in this paper to analyze music structure by inferring the time-evolving relationships within the music sequence.

2. NONPARAMETRIC DYNAMIC MIXTURE

2.1. DP-Based Hidden Markov Mixture Model

The standard tool for analysis of sequential data is the hidden Markov model (HMM) [5]. For the discrete sequence of interest, given an observation sequence $\mathbf{x} = \{x_t\}_{t=1}^T$ with $x_t \in \{1, \ldots, M\}$, the corresponding hidden state sequence is $\mathbf{S} = \{s_t\}_{t=1}^T$, from which $s_t \in \{1, \ldots, I\}$. A discrete HMM is represented by parameters $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}\}$, defined as:

- $\mathbf{A} = \{a_{\rho\xi}\}$, $a_{\rho\xi} = Pr(s_{t+1} = \xi | s_t = \rho)$: state transition probability;
- $\mathbf{B} = \{b_{\rho m}\}$, $b_{\rho m} = Pr(x_t = m | s_t = \rho)$: emission probability;
- $\boldsymbol{\pi} = \{\pi_\rho\}$, $\pi_\rho = Pr(s_1 = \rho)$: initial state distribution.

To model the whole music piece with one HMM [6], one may divide the sequence into J successive subsequences $\{\mathbf{x}_j\}_{j=1}^J$, each of length T with $\mathbf{x}_j = \{x_{jt}\}_{t=1}^T$ and $x_{jt} \in$

$\{1, ..., M\}$. The joint distribution of the observation subsequences given the model parameters $\boldsymbol{\theta}$ yields

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^{J} \left\{ \sum_{\mathbf{s}_j} \pi_{s_{j,1}} \prod_{t=1}^{T-1} a_{s_{j,t},s_{j,t+1}} \prod_{t=1}^{T} b_{s_{j,t},x_{j,t}} \right\} \quad (1)$$

However, rather than employing a single HMM for a given piece, which is clearly overly-simplistic, one may allow the model parameters to vary with time by letting

$$\mathbf{x}_j \sim HMM(\boldsymbol{\theta}_j), \quad \boldsymbol{\theta}_j \sim \sum_{k=1}^{K} p_k \boldsymbol{\theta}_k^*, \quad j = 1, \ldots, J, \quad (2)$$

which denotes that the subsequence $\mathbf{x}_j$ is drawn from an HMM with parameters $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_j$ is drawn from a unique set of mixture components $\{\boldsymbol{\theta}_k^*\}_{k=1}^K$ with respective mixture weights $\{p_k\}_{k=1}^K$.

Instead of setting the number of the mixture components a *priori*, we can apply a Dirichlet process by assuming $\boldsymbol{\theta}_j \sim G$, with $G \sim DP(\alpha_0 G_0)$, where G_0 is a base probability measure and α_0 is a non-negative real number [8]. Sethuraman [9] showed that

$$G = \sum_{k=1}^{\infty} p_k \delta_{\boldsymbol{\theta}_k^*}, \quad p_k = \tilde{p}_k \prod_{i=1}^{k-1} (1 - \tilde{p}_i) \quad (3)$$

where $\{\boldsymbol{\theta}_k^*\}_{k=1}^{\infty}$ represent a set of atoms drawn i.i.d. from G_0 and $\{p_k\}_{k=1}^{\infty}$ represent a set of weights, with the constraint $\sum_{k=1}^{\infty} p_k = 1$; each $\tilde{p}_k$ is drawn i.i.d. from the beta distribution $Be(1, \alpha_0)$. Since in practice the $\{p_k\}_{k=1}^{\infty}$ diminish quickly with increasing k (for reasonable choices of α_0), a truncated stick-breaking process [10] is often employed, with a large truncation level K, to approximate the infinite stick breaking process (in this approximation $\tilde{p}_K = 1$). We note that a draw G from a $DP(\alpha_0 G_0)$ is discrete with probability one.

2.2. Nonparametric Dynamic Structure

Placing a DP on the distribution of the subsequence-specific HMM parameters, $\boldsymbol{\theta}_j$, allows for borrowing of information across the subsequences, but does not incorporate information that subsequences from proximal times should be more similar. Hence, motivated by [11], we propose a more flexible dynamic mixture model in which

$$\boldsymbol{\theta}_j \sim G_j, \quad G_j = \sum_{k=1}^{\infty} p_{jk} \delta_{\boldsymbol{\theta}_k^*}, \quad \boldsymbol{\theta}_k^* \sim H, \quad (4)$$

where the subsequence-specific mixture distribution G_j has weights that vary with j, represented as $\mathbf{p}_j$. Including the same atoms for all j allows for repetition in the music structure across subsequences, with the varying weights allowing substantial flexibility.

Based on these considerations, we propose a dynamic hierarchical Dirichlet process (dHDP) with the following structure:

$$G_j = (1 - \tilde{w}_{j-1}) G_{j-1} + \tilde{w}_{j-1} H_{j-1} \quad (5)$$

where $G_1 \sim DP(\alpha_{01} G_0)$, H_{j-1} is called an innovation measure drawn from $DP(\alpha_{0j} G_0)$, and $\tilde{w}_{j-1} \sim Be(a_w, b_w)$. To impose sharing of the same components across all time, $G_0 \sim DP(\gamma H)$, as in a hierarchical Dirichlet process (HDP) [12]. The measure G_j is modified from G_{j-1} by introducing a new innovation measure H_{j-1}, and the random variable $\tilde{w}_{j-1}$ controls the probability of innovation (*i.e.*, it defines the mixture weights).

A draw $G_0 \sim DP(\gamma H)$ may be expressed as

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\boldsymbol{\theta}_k^*} \quad (6)$$

and the weights are drawn $\boldsymbol{\beta} \sim Stick(\gamma)$, where $Stick(\gamma)$ corresponds to letting $\beta_k = \tilde{\beta}_k \prod_{i=1}^{k-1} (1 - \tilde{\beta}_i)$ with $\tilde{\beta}_k \overset{iid}{\sim} Be(1, \gamma)$. Since the same atoms $\boldsymbol{\theta}_k^* \overset{iid}{\sim} H$ are used for *all* G_j, it is also possible to share data between subsequences widely separated in time; this latter property may be of interest when the music has temporal repetition.

The measures $G_1, H_1, \ldots, H_{J-1}$ have their own mixture weights, yielding

$$G_1 = \sum_{k=1}^{\infty} \zeta_{1k} \delta_{\boldsymbol{\theta}_k^*}, H_1 = \sum_{k=1}^{\infty} \zeta_{2k} \delta_{\boldsymbol{\theta}_k^*}, \ldots, H_{J-1} = \sum_{k=1}^{\infty} \zeta_{Jk} \delta_{\boldsymbol{\theta}_k^*}$$
$$\boldsymbol{\zeta}_j \overset{ind}{\sim} DP(\alpha_{0j} \boldsymbol{\beta}), \quad j = 1, \ldots, J \quad (7)$$

where, analogous to the discussion at the end of Section 2.1, the different weights $\boldsymbol{\zeta}_j = \{\zeta_{jk}\}_{k=1}^{\infty}$ are independent given $\boldsymbol{\beta}$ since $G_1, H_1, \ldots, H_{J-1}$ are independent given G_0.

To further develop the dynamic relationship from G_1 to G_J, we extend the mixture structure in (5) from group to group:

$$\begin{aligned} G_j &= (1 - \tilde{w}_{j-1}) G_{j-1} + \tilde{w}_{j-1} H_{j-1} \\ &= \prod_{l=1}^{j-1} (1 - \tilde{w}_l) G_1 + \sum_{l=1}^{j-1} \{ \prod_{m=l+1}^{j-1} (1 - \tilde{w}_m) \} \tilde{w}_l H_l \\ &= w_{j1} G_1 + w_{j2} H_1 + \ldots + w_{jj} H_{j-1} \end{aligned} \quad (8)$$

where $w_{11} = 1$, $\tilde{w}_0 = 1$, and for $j > 1$ we have $w_{jl} = \tilde{w}_{l-1} \prod_{m=l}^{j-1} (1 - \tilde{w}_m)$, for $l = 1, 2, \ldots, j$. It can be easily verified that $\sum_{l=1}^{j} w_{jl} = 1$ for each j, with w_{jl} the prior probability that parameters for subsequence j are drawn from the lth component distribution, where $l = 1, \ldots, j$ indexes $G_1, H_1, \ldots, H_{j-1}$, respectively.

Based on the dependent relation induced here, we have an explicit form for each $\{\mathbf{p}_j\}_{j=1}^{J}$ in (4):

$$\mathbf{p}_j = \sum_{l=1}^{j} w_{jl} \boldsymbol{\zeta}_l. \quad (9)$$

If all $\tilde{w}_j = 0$, all of the groups share the same mixture distribution related to G_1 and the model reduces to the Dirichlet mixture model described in Section 2.1. If all $\tilde{w}_j = 1$ the model instead reduces to the HDP [12], in which there is no temporal dependence between the adjacent subsequences. Therefore, the dynamic HDP is more general than both DP and HDP, with each a special case. In the posterior computation, we treat the $\tilde{\mathbf{w}}$ as random variables and add beta priors on them for more flexibility. To encourage the adjacent groups to be shared, the prior $Be(\tilde{w}_j|a_w, b_w)$ for all $j = 1, \ldots, J-1$ should be specified to have $E(\tilde{w}_j) < 0.5$.

2.3. Posterior Computation

A modification of the block Gibbs sampler [10] is proposed for dHDP HMM mixture inference. As discussed in Section 2.1, a truncated stick-breaking process [10] is employed to reduce the computational complexity. For easier inference. we introduce two indicator vectors $\mathbf{r}_j$ and $\mathbf{z}_j$ for each subsequence $\mathbf{x}_j$. The $\mathbf{r}_j$ has only one component equal to 1 with others equal to zero: if $r_{jl} = 1$, then $\boldsymbol{\theta}_j$ is drawn from the lth component distribution in (8), where l might be equal to one of the values from 1 to j. The $\mathbf{z}_j$ is another indicator vector of length equal to K (K represents the truncation level), with $z_{jk} = 1$ if the subsequence $\mathbf{x}_j$ is allocated to the k^{th} atom ($\boldsymbol{\theta}_j = \boldsymbol{\theta}^*_k$) and $z_{jk} = 0$ otherwise.

The parametric form for each of the components $\boldsymbol{\theta}^*_k$ for $k = 1, \ldots, K$ is a hidden Markov model (HMM) with $\boldsymbol{\theta}^*_k = (\mathbf{A}^*_k, \mathbf{B}^*_k, \boldsymbol{\pi}^*_k)$. As in [7], the component parameters $\mathbf{A}^*_k$, $\mathbf{B}^*_k$ and $\boldsymbol{\pi}^*_k$ are assumed to be independent, with the base measure having a product form with Dirichlet components for each of the probability vectors.

The posterior distribution of the model parameters is expressed as $Pr(\boldsymbol{\theta}^*, \tilde{\mathbf{w}}, \boldsymbol{\zeta}, \boldsymbol{\beta}, \boldsymbol{\alpha}_0, \gamma, \mathbf{z}, \mathbf{r}|\mathbf{X})$. In each iteration of the Gibbs sampler, each variable is drawn from the full conditional posterior distribution given all other samples. We choose initial values for these variables and run through the sequence of the Markov chain until it converges to a stationary distribution. The samples are collected to approximately represent the parameters' full posterior distribution underlying the data.

Since the indicator vector $\mathbf{z}_j$, for $j = 1, \ldots, J$, represents the membership of sharing across all the subsequences, we use this information to segment the music, by assuming that the subsequences possessing the same membership should be grouped together. In order to overcome the issue of label switching that exists in Gibbs sampling, instead of using the membership $\mathbf{z}$ to represent the result, we use the similarity measure $E(\mathbf{z}'\mathbf{z})$, in which $\mathbf{z}'$ represents the transpose of $\mathbf{z}$. Here $E(\mathbf{z}'\mathbf{z})$ is approximated by averaging the quantity $\mathbf{z}'\mathbf{z}$ from multiple iterations, and in each iteration $\mathbf{z}'_j\mathbf{z}_{j'}$ measures the sharing degree of $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_{j'}$ by integrating out the index of atoms. Related clustering representations of non-parametric models have been considered in [13].

3. MUSIC EXPERIMENTS

Before the music is modeled with the dHDP HMM mixture, the acoustic signal is sampled at 22.05 KHz and divided into 100 ms contiguous frames; 40-dimensional Mel frequency cepstral coefficients (MFCCs) were extracted from each frame, these being effective for extracting perceptually important parts of the spectral envelope of audio signals. Each frame was quantized into discrete symbols using vector quantization (VQ) (the codebook size $M = 16$).

We consider a relatively well known musical piece: "A Day in the Life" from the Beatle's album *Sgt. Peppers Lonely Hearts Club Band*. This Beatle's song has many distinct sections (vocals, along with clearly distinct instrumental parts). After the piece is processed, the sequence is represented as a series of discrete observations, shown in Figure 1. We divided the piece into 88 subsequences and each subsequence includes 75 observations. The music segmentation results with dHDP HMM mixture is shown in Figure 2. The results in Figure 2 (a) quantify how inter-related any one subsequence of the music is to all others. We observe that the music is decomposed into clear contiguous segments of various lengths, and segment repetitions are evident. To evaluate the segmentation results, we provide a music-theoretic analysis (via the third author) in Table 1

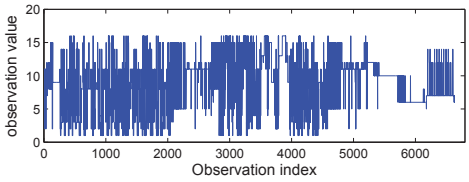
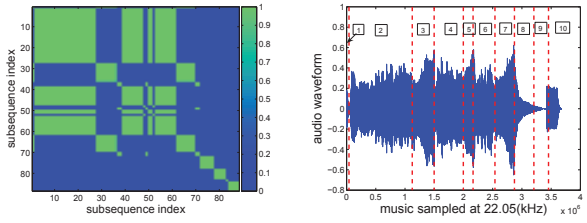


Fig. 1. Sequence of code indices for the Beatle's music using a codebook of dimension $M = 16$.



(a) (b)

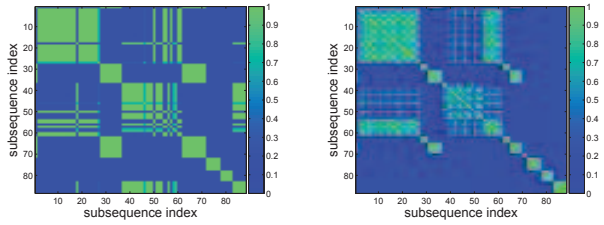
Fig. 2. Segmentation results of the dHDP HMM modeling for the Beatles music. (a) The similarity matrix $E(\mathbf{z}'\mathbf{z})$. (b) The segmentation result on the Beatles audio waveform (blue curves represent the audio waveform, red dashed lines represent segment positions and the number in a box labels the partition index).

For comparison, we now analyze the same music using both the DP-HMM [7] and HDP-HMM [12], which are two special cases of our model (fixing all $\tilde{w}_j = 0$ and $\tilde{w}_j = 1$ respectively, without changing anything else). The results are

presented in Figure 3 analogous to the dHDP HMM presentation. The performance of the dHDP relative to the other approaches is consistent with the music theoretic analysis and yields "cleaner" segmentation results.

segment index	subsequences included	music theory explanation
1	1^{st}	an instrumental accompaniment with some applause
2	$2^{nd} \sim 27^{th}$	three verses sung by Lennon
3	$28^{th} \sim 36^{th}$	an orchestral crescendo continues
4	$37^{th} \sim 47^{th}$	an interlude ('Woke up,...') sung by McCartney
5	$48^{th} \sim 52^{nd}$	a short transition
6	$53^{rd} \sim 61^{st}$	a verse part sung by Lennon
7	$62^{nd} \sim 69^{th}$	the same orchestral crescendo as the third part
8	$70^{th} \sim 77^{th}$	the "famous" final chords played on three different pianos
9	$78^{th} \sim 82^{nd}$	an almost quiet part
10	$83^{rd} \sim 88^{th}$	the famous "studio chatter" part

Table 1. Segmentation of the Beatles music with musical explanation [14].



(a) (b)

Fig. 3. Results of DP-HMM and HDP-HMM modeling for the Beatles music. (a) The similarity matrix $E(\mathbf{z}'\mathbf{z})$ from DP-HMM. (b) The similarity matrix $E(\mathbf{z}'\mathbf{z})$ from HDP-HMM.

4. CONCLUSIONS

A Bayesian dynamic mixture model has been proposed for music analysis. The model has the following characteristics: (*i*) with inferred probabilities, the underlying parameters associated with data at adjacent times are the same; and (*ii*) since the same underlying atoms are used in the mixtures at all times, it is possible that the same atoms may be used at temporally distant time, allowing the capture of repeated patterns in temporal data. A relatively simple Gibbs sampler is employed for model inference. The performance of the dHDP HMM mixture is demonstrated on real music. Compared with other mixture models, the dHDP HMM mixture yields better segmentation results by considering the time evolving within the music piece (the presented results are representative of tests on many different pieces).

5. REFERENCES

[1] D. Temperley, "A probabilistic model of melody perception," *Cognitive Science*, vol. 32, pp. 418–444, 2008.

[2] K. Ni, J. Paisley, L. Carin, and D. Dunson, "Multi-task learning for sequential data via ihmms and the nested dirichlet process," in *Proc. 24th International Conference on Machine Learning (ICML)*, 2007.

[3] J.-F. Paiement, Y. Grandvalet, S. Bengio, and D. Eck, "A generative model for rhythms," *NIPS'2007 Music, Brain & Cognition Workshop*, 2007.

[4] T. Plotz, G.A. Fink, P. Husemann, S. Kanies, K. Lienemann, T. Marschall, M. Martin, L. Schillingmann, M. Steinrucken, and H. Sudek, "Automatic detection of song changes in music mixes using stochastic models," *18th International Conference on Pattern Recognition (ICPR'06)*, 2006.

[5] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[6] C. Raphael, "Automatic segmentation of acoustic musical signals using hidden markov models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 4, pp. 360–370, 1999.

[7] Y. Qi, J.W. Paisley, and L. Carin, "Music analysis using hidden markov mixture models," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5209–5224, 2007.

[8] T.S. Ferguson, "A bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[9] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 2, pp. 639–650, 1994.

[10] H. Ishwaran and L.F. James, "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 161–173, 2001.

[11] D.B. Dunson, "Bayesian dynamic modeling of latent trait distributions," *Biostatistics*, vol. 7, no. 4, pp. 551–568, 2006.

[12] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, "Hierarchical dirichlet processes," *JASA*, vol. 101, no. 476, pp. 1566–1581, 2006.

[13] M. Medvedovic and S. Sivaganesan, "Bayesian infinite mixture model based clustering of gene expression profiles," *Bioinformatics*, vol. 18, no. 9, pp. 1194–1206, 2002.

[14] M. Lewisohn, *The Beatles Recording Sessions*, New York: Harmony Books, 1988.