A Markov-Chain Monte Carlo Approach to Musical Audio Segmentation

Christophe Rhodes, Michael Casey, Samer Abdallah, Mark Sandler

Introduction

A new unsupervised Bayesian clustering model extracts classified structural segments, *intro*, *verse*, *chorus*, *break* etc., from recorded music. This extends previous work by identifying all the segments in a song, not just the chorus or longest section, and by incorporating prior information on the temporal extent of a segment. We present experimental results demonstrating that this method can produce accurate labelled segmentations for popular music.

Signal Preprocessing

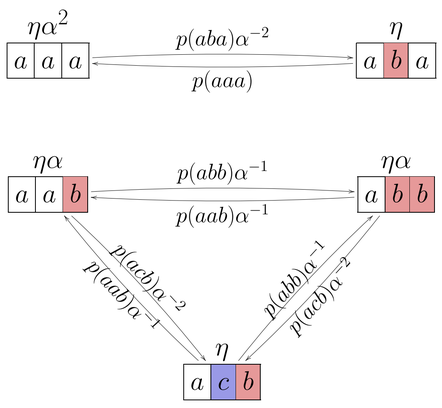
We generate a sequence of low-level symbolic states from the audio by first generating a cepstrum, then modelling each song with a 60-state HMM (using Baum-Welch training and generating the maximum-likelihood decoding). The cepstrum is generated by taking a constant-Q power spectrum with $\frac{1}{8}$th-octave resolution, taking the logarithm of the spectral power and performing PCA on the cepstral shape, retaining the 20 principal components.

Wolff algorithm

Wolff's algorithm was designed to simulate Ising, Potts and $x - y$ systems near critical temperatures without the critical slowing down: it does so by doing block updates, rather than single-site changes. Additionally, it is tuned so that proposed steps are always accepted.

Our Wolff-Gibbs algorithm involves a Wolff domain growing phase, followed by a block-Gibbs sampling step updating the entire domain: first choose a 'seed' site; then grow from the seed with exponential probability in both directions (with a cut-off at current boundary positions); finally, choose a segment label from the lexicon (or, equivalently, one segmentation from a clique) with probabilities proportional to the prior multiplied by a boundary counting factor.

Cliques

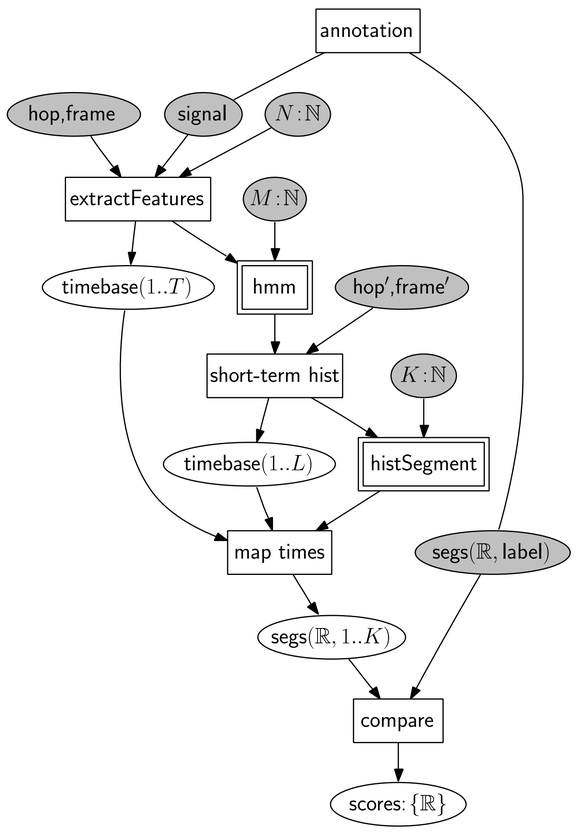


Histogram Clustering

By considering each segment in the signal to represent a distinct process for generating the low-level features, we are drawn towards the notion of observing distinct frequency distributions of our preprocessed features (the HMM states). Taking windowed histograms of the HMM state sequence, we can then compute the likelihood of the ith histogram with class c_i and class-conditional state distribution A_{jc_i}, leading to the model energy for assignments $\mathbf{c}$ of

$$\varepsilon(\mathbf{c}, \theta) = \sum_i^L \sum_j^M \sum_k^K \delta_{kc_i} X_{ji} \log \frac{X_{ji}}{A_{jk}} - \log p(\mathbf{c}). \qquad (1)$$

System



Block Segmentation

We consider an energy function $\varepsilon_{\mathcal{H}}$ for segment durations x, parametrized by ν and γ (generalizing Gamma-like distributions)

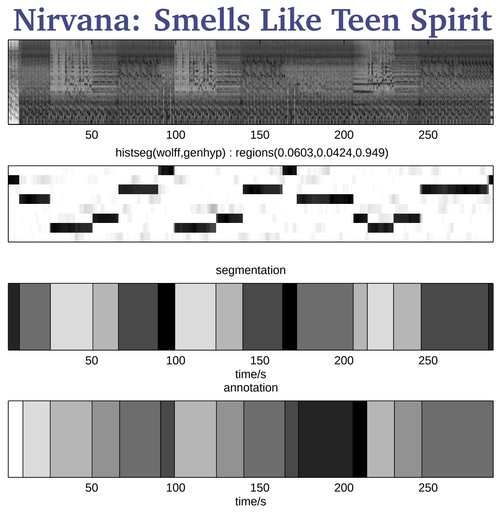
$$\varepsilon_{\mathcal{H}}(x, \nu, \gamma) = \frac{1}{|\nu|} x^{-\nu} + (\gamma + 1) \log x, \qquad (2)$$

and take the prior probability for a segmentation $\mathbf{c}$ into i segments with length $\mathsf{len}_i(\mathbf{c})$ as

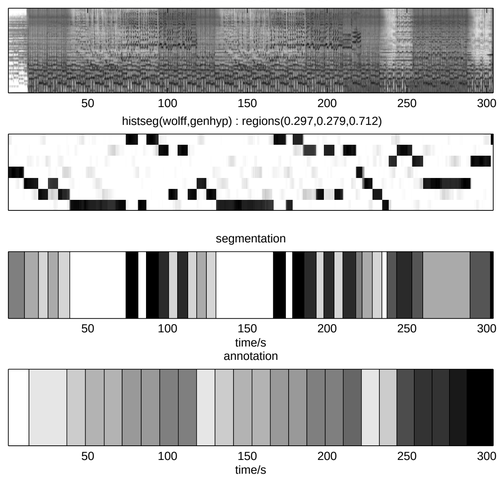
$$p(\mathbf{c}) \propto \prod_i e^{-\beta \varepsilon_{\mathcal{H}}(\mathsf{len}_i(\mathbf{c}), \nu, \gamma)}. \qquad (3)$$

This prior represents the segment duration model, and augments the energy function from the probability distribution model (1) to produce the final target distribution.

Nirvana: Smells Like Teen Spirit



Cranberries: Zombie



Performance measures

We can use a directional Hamming distance d_{GM} by finding for each segment in the machine segmentation S_M^i the segment in the ground truth S_G^j with the maximum overlap, and then summing the difference,

$$d_{GM} = \sum_{S_M^i} \sum_{S_G^k \neq S_G^j} |S_M^i \cap S_G^k|, \qquad (4)$$

and taking $1 - \frac{d_{GM}}{L}$ as a measure of recall. Taking the other direction of the Hamming distance d_{MG} analogously gives a precision-like measure in terms of frame classification, although note that we have considered only the correspondence between the extents of each segment, not their labelling.

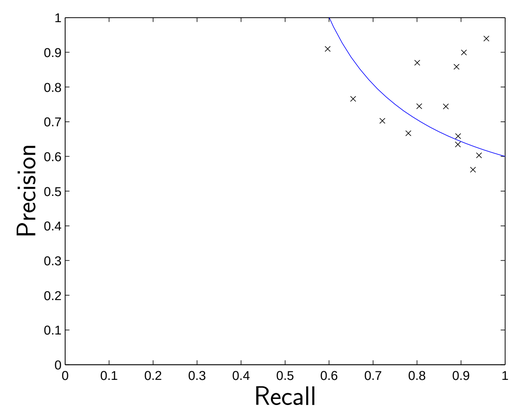
Summary statistics

Given precision P and recall R, it is common to use a measure which combines the two. One way, balancing precision and recall and treating them as equally important, is to use the F statistic given by

$$F = \frac{2PR}{P + R} \qquad (5)$$

Results

Segment precision vs. recall for our test corpus. The line corresponds to $F = 0.75$.



Future Work

We can refine the treatment of the inter-class dynamics by allowing each class to have its own characteristic timescale – chorus and verse sections typically last longer than bridges – or by having non-uniform transition probabilities between classes. The temporal dynamics within a segment are presently unmodelled; only the overall distribution of HMM states is accounted for. This prevents us from detecting repeated segments within the model, and leads us to consider explicit modelling of the multiple levels of hierarchy.

We could improve performance and reliability by introducing information from different modes of preprocessing the audio: using information from an onset or beat-detector to bias domain growing towards onsets or even eight-bar boundaries.

References

[1] U. Wolff, "Collective Monte Carlo Updating for Spin Systems," *Physical Review Letters*, vol. 62, no. 4, pp. 361–364, 1989.

[2] J. Puzicha, T. Hofmann, and J. M. Buhmann, "Histogram clustering for unsupervised image segmentation," *Proc. CVPR*, 1999.

[3] S. Abdallah, K. Noland, M. Sandler, M. Casey, and C. Rhodes, "Theory and evaluation of a Bayesian music structure extractor," in *Proc. ISMIR*, 2005.

[4] M. Goto, "A chorus-section detecting method for musical audio signals," in *Proc. ICASSP*, 2003.

[5] A. Barbu and S.-C. Zhu, "Cluster Sampling and Its Applications in Image Processing," Tech. Rep. 409, Department of Statistics, UCLA, 2004.

Queen Mary University of London

m.casey@gold.ac.uk
c.rhodes@gold.ac.uk
Centre for Cognition, Computation and Culture,
Goldsmiths College, University of London.

samer.abdallah@elec.qmul.ac.uk
mark.sandler@elec.qmul.ac.uk
Centre for Digital Music,
Queen Mary, University of London.

Goldsmiths University of London