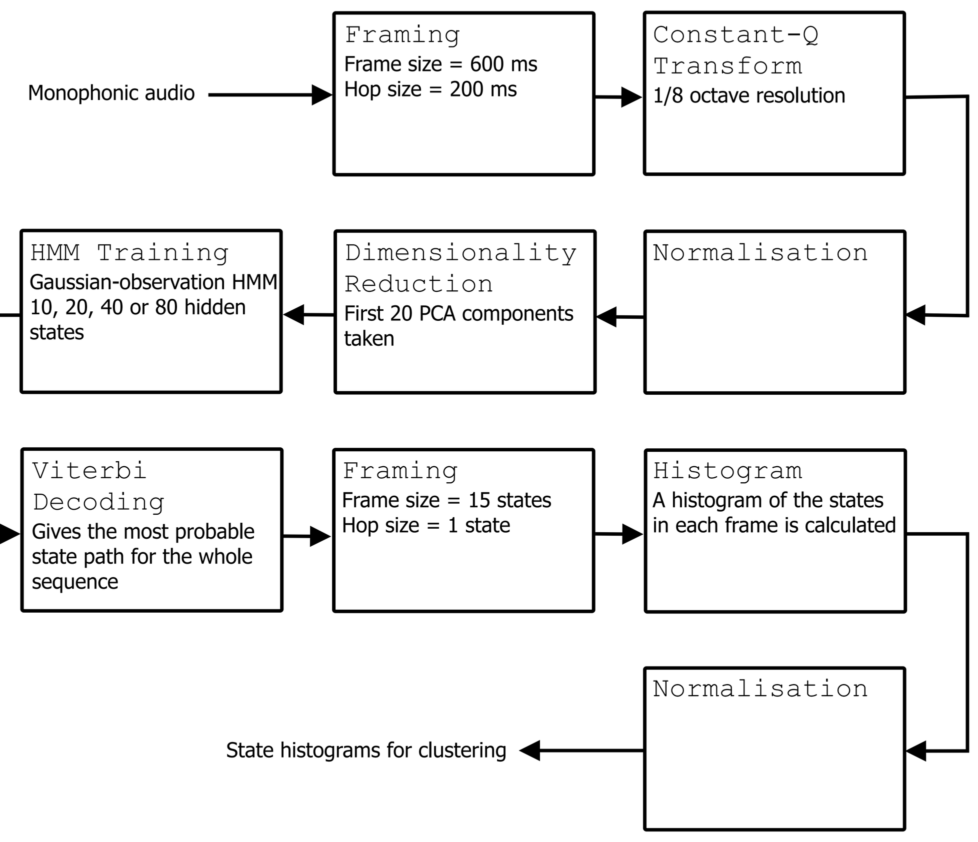
Overview of a Bayesian Music Segmenter

Samer Abdallah, Katy Noland, Mark Sandler, Michael Casey, Christophe Rhodes

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.

Feature Extraction



Feature extraction process. One of two possible clustering methods is applied to the extracted HMM state histograms.

(1) Pairwise Clustering

i) Measure the distance between all possible pairs of histograms using either cosine distance or a symmetrized Kullback-Leibler divergence:

$$d_{\mathrm{kl}}(\mathbf{x},\mathbf{x}') = \sum_{i=1}^{M} \left[x_i \log(x_i/q_i) + x_i' \log(x_i'/q_i) \right]$$

where $q_i = \frac{1}{2}(x_i + x_i')$ and M is the number of bins in the histograms.

ii) Using these pairwise distances D_{ij} derive the cost function

$$\mathcal{H}(m) = \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \frac{D_{ij}}{L} \left(\sum_{\nu=1}^{K} \frac{m_{i\nu} m_{j\nu}}{p_\nu} - 1 \right)$$

where $m_{k\nu}$ is an assignment of histogram i to one of K clusters ν, and L is the number of histogram frames.

iii) Optimise this cost function using a form of mean-field annealing.

(2) Central Clustering

i) Model the histograms as the result of drawing samples from probability distributions determined by K underlying classes. A_{jk} is the probability of observing the jth HMM state in the kth class, and $\mathbf{C}$ is the sequence of class assignments for a given sequence of L histograms $\mathbf{X}$.

ii) The overall log-likelihood of the model reduces to

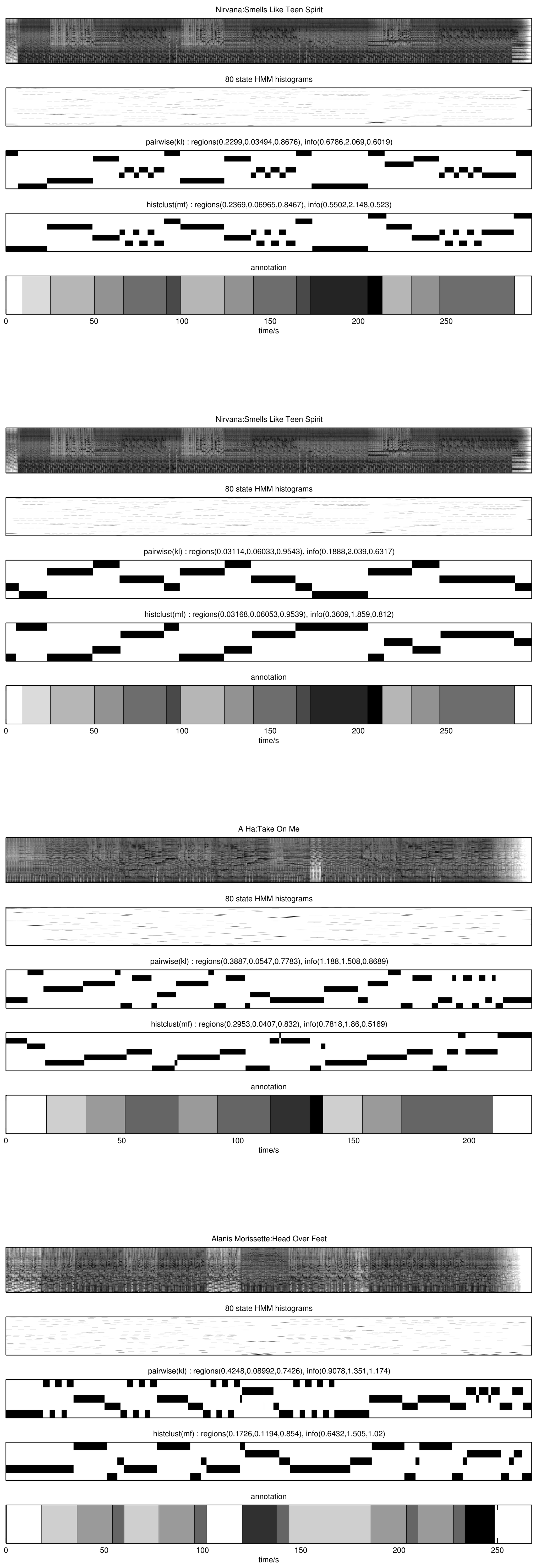
$$\mathcal{H}_h = \sum_{i=1}^{L} \sum_{j=1}^{M} \sum_{k=1}^{K} \delta(k, C_i) X_{ji} \log \frac{X_{ji}}{A_{jk}}$$

iii) Optimise this cost function using a form of deterministic annealing, equivalent to expectation maximisation with a 'temperature' parameter which gradually falls to zero.

Segmentations

- Segmentations were performed on 14 popular music songs, downsampled to 11.025 kHz mono.
- Both the pairwise and central clustering algorithms were tested with between 2 and 10 final segment types.
- Annotations given by an expert listener were used as ground truth.
- Internal structure is frequently visible where repeated sections have been split between two clusters.

Example segmentations



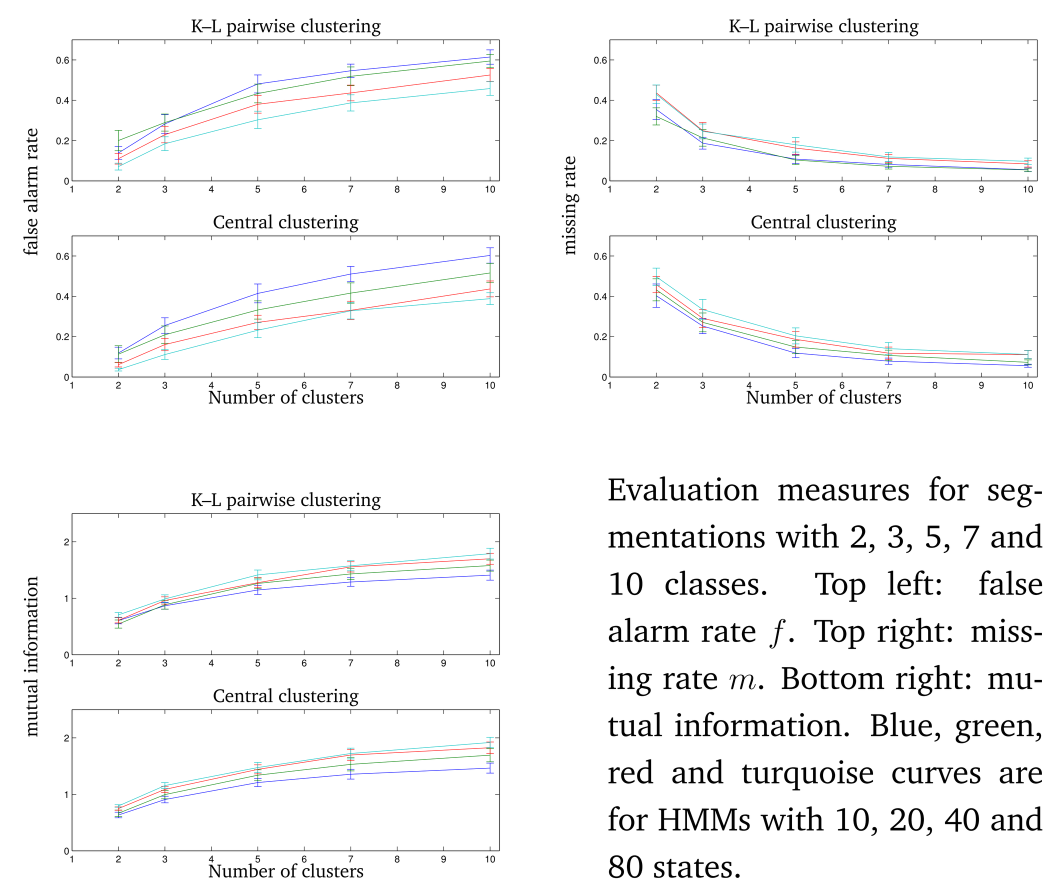
Four sets of example machine segmentations, with the constant-Q spectrogram (top), HMM state histograms (second) and ground truth segmentations (bottom) for comparison. The ground truth segments are shown using different shades of grey for the different segment labels.

Evaluation

i) Compare each output segment with the most closely corresponding ground truth segment using a directional Hamming distance. This measures the number of missed and falsely identified segment boundaries.

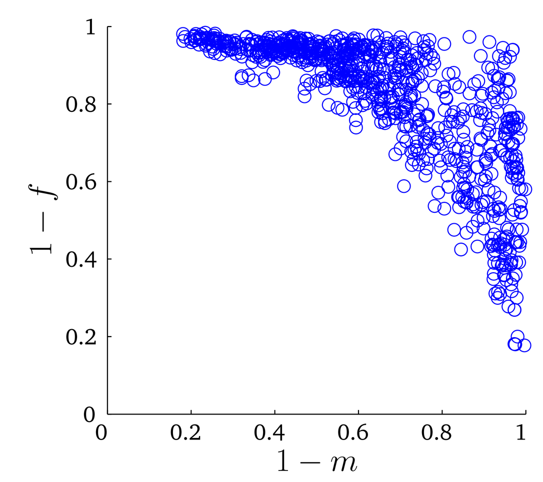
ii) Calculate the mutual information between the ouput and expert segment labels for each frame. This measures the quality of the sequence of labels.

Fragmentation tradeoff



Evaluation measures for segmentations with 2, 3, 5, 7 and 10 classes. Top left: false alarm rate f. Top right: missing rate m. Bottom right: mutual information. Blue, green, red and turquoise curves are for HMMs with 10, 20, 40 and 80 states.

Overall performance



Values of $1-f$, corresponding loosely to precision, plotted against values of $1-m$, analogous to recall, over all songs and segmentation methods presented. The optimal average tradeoff point is approximately (0.8,0.8).

Conclusions

i) Both algorithms can produce segmentations similar to the ones provided by a human expert.

ii) The numbers of missed and false boundaries increase with number of segment types requested, but so does the mutual information, showing that extra classes are put to good use.

iii) Over-segmentation often reveals the internal structure of segments in a consistent way, revealing a sort of 'abstract score'.

iv) Subsequent work solves the fragmentation problem by incorporating an explicit prior on segment durations.

References

[1] Q. Huang and B. Dom, "Quantitative methods of evaluating image segmentation," in *Proc. IEEE Intl. Conf. on Image Processing (ICIP'95)*, 1995.

[2] T. Hofmann and J. M. Buhmann, "Pairwise data clustering by deterministic annealing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 1, 1997.

[3] J. Puzicha, T. Hofmann, and J. M. Buhmann, "Histogram clustering for unsupervised image segmentation," *Proceedings of CVPR '99*, 1999.

Queen Mary University of London

samer.abdallah@elec.qmul.ac.uk
katy.noland@elec.qmul.ac.uk
mark.sandler@elec.qmul.ac.uk
Centre for Digital Music,
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.

Goldsmiths UNIVERSITY OF LONDON