A MARKOV-CHAIN MONTE-CARLO APPROACH TO MUSICAL AUDIO SEGMENTATION

Christophe Rhodes, Michael Casey

Goldsmiths College, University of London
New Cross, London SE14 6NW

Samer Abdallah, Mark Sandler

Queen Mary, University of London
Mile End Road, London E1 4NS

ABSTRACT

This paper describes a method for automatically segmenting and labelling sections in recordings of musical audio. We incorporate the user's expectations for segment duration as an explicit prior probability distribution in a Bayesian framework, and demonstrate experimentally that this method can produce accurate labelled segmentations for popular music.

1. INTRODUCTION

This paper describes a method for incorporating our prior expectations about the size of musical structures or segments into a system for producing labelled segmentations of musical audio. Automatically-generated segmentations have application in audio fingerprinting, thumbnailing (see e.g. [1]), content-based retrieval systems, summary generation and user interface provision for navigation in audio editors.

Previous studies in segmentation have used various spectral features such as timbre [2, 3, 4] or chroma [5, 6, 7, 8] to generate time series of feature vectors. We do not address the choice of audio features here, but instead examine the typical subsequent use of the series of vectors.

Several studies [2, 5, 6, 7, 8] compute pairwise similarity matrices between feature vectors with some distance measure for individual frames, then apply a filter of some form. Others (e.g. [4]) perform k-means clustering between frames, and then post-process this clustering by using an HMM with fewer states, or generate labels by using an HMM directly on the feature vectors and then average over a window [3].

These filtering or post-processing stages are introduced to reduce noise in the classification, and to reintroduce the notion of temporal closeness which was lost in the clustering; ideally, the classification would be informed of our expectations and so not produce noise in the first place, and would have temporal coherence built in.

In our previous work [9] we have avoided the need for a post-processing stage by performing clustering with large (of the order of 3s) analysis windows, but this is an equally *ad hoc* way to address the expected scale of segment size, and does not address the issue of temporal closeness beyond the coherence of an individual feature frame. We therefore introduce a prior probability distribution on the sizes of segments, and adjust the classification algorithms to incorporate this prior; this probability distribution explicitly encodes our assumptions regarding the segmentation.

This research was supported by EPSRC grant GR/S84750/01 (Hierarchical Segmentation and Semantic Markup of Musical Signals).

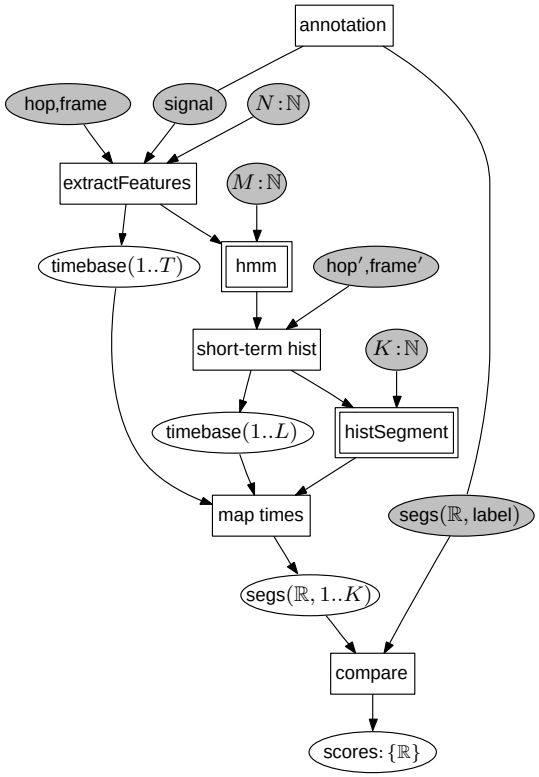

Fig. 1. Overview of our segmentation method.

This paper continues with a description of our method for generating labelled segmentations of audio in section 2; we present some empirical results in section 3, and draw conclusions in section 4.

2. METHOD

Figure 1 shows an overview of our segmentation method; nodes corresponding to input variables are shaded, and the

double-bordered boxes represent probabilistic models. We discuss the individual stages in more detail below.

2.1. Extracting features

The processing chain begins with a uniformly sampled monophonic audio signal (from a single channel, as far as that is possible) and breaks it into a sequence of short overlapping fragments; our sample data was 16-bit mono at a 11.025kHz sample rate, and we used a window size of 400ms with a hop size of 200ms over these audio samples to generate a constant-Q power spectrum with $\frac{1}{12}$th-octave resolution. This power-spectrum is log-normalized and 20 principal components are extracted, which along with the envelope magnitude form 21-dimensional feature vectors, corresponding to the result of the extractFeatures node in fig. 1.

These vectors are used to train a Gaussian-observation HMM with 60 states for each song, which then generate the most probable state sequence for that song (the output of the hmm probabilistic model node). These states are collected into short-time histograms (short-term hist) over windows of 4 states with a hop size of 2 states.

2.2. Generating segments

Our segmentation algorithm models a segment as a sequence of samples of HMM state histograms drawn from a class-specific probability distributions, with the boundaries of the segment being where the probability distribution changes; following [10], we can perform segmentation given a flat prior probability by performing deterministic annealing on an Expectation-Minimization optimization over frame label assignments and the class probability distributions.

The energy function for this optimization is

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

where M and K are parameters as in fig. 1, and L is the length of the signal in histogram frames; $\mathbf{A}$ contains the prototype class probability distributions, $\mathbf{X}$ the sequence of observed histograms; and $p(\mathbf{c})$ is the prior probability of segmentation $\mathbf{c}$.

However, except for the case of overly large analysis windows (of the order of 3 seconds in [9]), this method with a uniform prior $p(\mathbf{c})$ fails to find long segments corresponding to sections such as the verse or the chorus of songs; instead the segmentations generated correspond to changes in low-level audio features themselves.

In order to encode our interest in higher-level segments into the segmentation algorithm, we incorporate an explicit non-flat prior probability distribution $p(\mathbf{c})$ on segmentations into the segmentation procedure. We propose a probability distribution for segment lengths $p_{\mathcal{H}}(x)$ with a wide spread

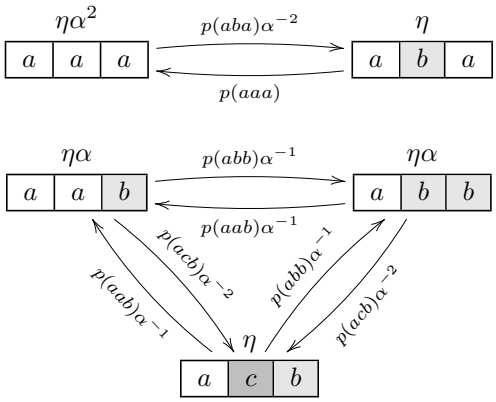

Fig. 2. Two 'cliques' including all local configurations and all central interval reassignments. Each a, b, or c represents a *sequence* of sites classified a a, b, or c, so each arrow represents a *block* assignment of the central section. The expressions by the boxes show the relative probabilities of proposing the central section. The arrows are labelled by the relative probability of the step according to the sampling distribution.

about a single peak, in terms of

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

as

$$p_{\mathcal{H}}(x) = \frac{e^{-\beta \varepsilon_{\mathcal{H}}(x,\nu,\gamma)}}{\int e^{-\beta \varepsilon_{\mathcal{H}}(x,\nu,\gamma)} \mathrm{d}x}, \qquad (3)$$

where in this investigation we set ν to be 2, penalizing short segments strongly, and γ to 0. With $\gamma = 0$, the mode of the distribution is always 1; we can arrange this to correspond to any given time by measuring in those units, in this case 20s. This then gives us the prior probability of a given segmentation as

$$p(\mathbf{c}) = \prod_{i} p_{\mathcal{H}}(\mathsf{len}_i(\mathbf{c})) \qquad (4)$$

where $\mathsf{len}_i(\mathbf{c})$ denotes the length of the ith segment in $\mathbf{c}$.

The presence of this prior probability distribution over segment lengths has the effect of strongly coupling label assignments over the frames, so we can no longer use the usual EM algorithm; there is no closed form for improving the cost function by changing the segmentation $\mathbf{c}$ given class probability distributions $\mathbf{A}$. Instead, we explore the state space using a block-Gibbs sampler allocating class labels to a domain selected by a Wolff [11] algorithm, modified to ensure that detailed balance holds given a non-uniform prior (see fig. 2); we use these samples as probabilistic enhancements to the cluster assignments, and perform deterministic annealing of the assignments over the energy (eq. 1) as in [10] (this process is denoted by to histSegment in fig. 1). A full description of this step is in preparation; similar work for image analysis can be found in [12].

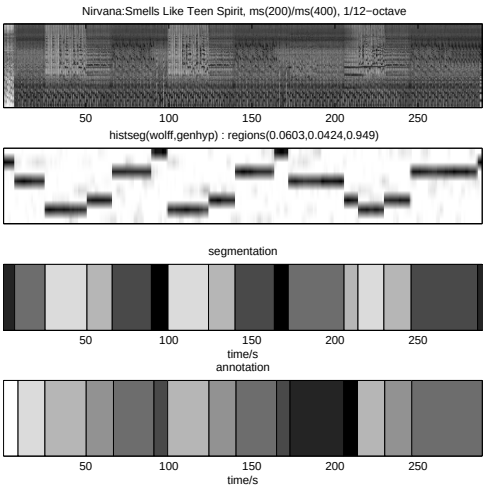

Fig. 3. Results for 'Smells Like Teen Spirit' by Nirvana. From top: spectrogram, cluster assignments, corresponding segmentation, manual annotation. F (eq. 5) for this segmentation is 0.95.

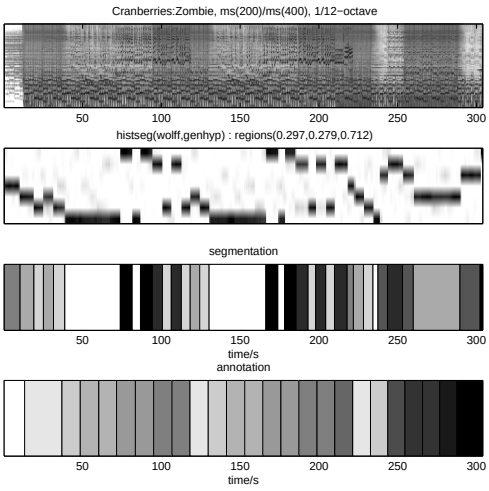

Fig. 4. Results for 'Zombie' by The Cranberries. From top: spectrogram, cluster assignments, corresponding segmentation, manual annotation. F (eq. 5) for this segmentation is 0.71.

We then generate the final, labelled, segmentation by considering contiguous regions of the same class label as a segment, and inverting (maptimes) the effect of the windowing operations to generate a segmentation over time. This segmentation can then be examined (compare) against the annotation provided by a human expert, producing numerical scores according to various evaluation metrics.

3. RESULTS

The segmentations performed for the results we present were performed with the parameter for the number of classes set to the number of classes in each song's ground truth. Two segmentations resulting from this process are displayed in fig. 3 (for 'Smells Like Teen Spirit' by Nirvana) and fig. 4 (for 'Zombie' by The Cranberries). The segmentation displayed in fig. 3 is clearly very close to the ground truth annotation: there are very minor disagreements over boundary positions, but otherwise the regions are in close correspondence and the labelling is consistent.

The segmentation in fig. 4 shows some interesting effects of our algorithm when contrasted with the expert annotation. Firstly, there is a deficiency in our system in that it is incapable of detecting the difference between two adjacent identically-labelled segments and one large segment encompassing the same time. Thus, the large segment around 50s (and repeated at 150s) corresponds to an ABB pattern in the annotation (where the A section appears similar to the B section). Additionally, our algorithm has found some microstructure in the second region of the annotation, dividing it (and its repeats) into two alternating segments.

We compute precision and recall statistics based on an adaptation of a performance measure in image segmentation [13], computing the maximum overlapping segment in the ground truth and the machine segmentation. This is similar to the P and R statistics in [7] – the differences are that we compute these measures over all segment labels, not just the chorus (though we ignore the identity of the label) and that for each section in the comparison, we consider only the largest contiguously labelled section in the corresponding region, not the sum. Figure 5 demonstrates the results for our corpus of 14 songs.

Similarly, to compare our results to previous results as far as that is possible, following [7], we compute one particular F measure [14]

$$F = \frac{2PR}{P+R} \tag{5}$$

from our P and R statistics. The results for seven of the songs lie above the curve for $F = 0.75$, used as a threshold in [7], and the other seven have $F \geq 0.70$; the mean F for the 14 songs is 0.78.

4. CONCLUSIONS

We have described a method explicitly incorporating a prior distribution of segment lengths for generating labelled segmentations of musical audio, and provided evidence (fig. 5) that it can produce segmentations similar to annotations produced by a human expert (fig. 3) and that where the segmentation differs significantly (as in fig. 4) from the expert annota-

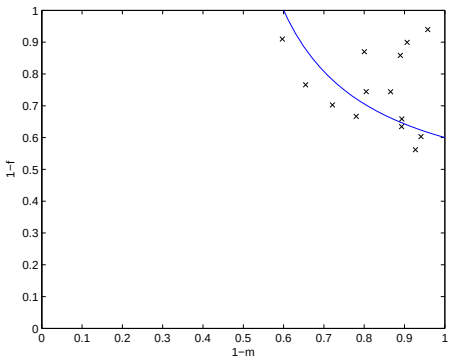

Fig. 5. Precision *vs* recall for our segmentation over a corpus of 14 songs, measured against a ground truth annotation provided by a human expert. The solid line corresponds to $F = 0.75$.

tion provided there is nevertheless a correspondence between the structure of the segmentation and that of the music.

We believe that there is a need for a more sophisticated method of evaluating segmentations of music. The evaluation statistic we have used discards label information (though experimentally our segmentations have a high degree of label correspondence with the annotations). We have suggested [9] evaluations based on Mutual Information between segmentation and annotation, which does capture label information; however, while this assists in comparing different segmentation methods or parameter sets over the same corpus, it is not clear how to aggregate this measure over disparate tracks into a meaningful figure; nor is it clear how to map a particular requirement into a threshold for application-specific evaluation.

This method for generating a segmentation, with its explicit prior probability for segment durations, can naturally be extended to have label-specific prior distributions; while a general prior should be broad, it is possible to incorporate more domain knowledge – such as the features of a known genre, or the output of other signal-processing tools such as a beat extractor – into the prior distribution; for signals known to come from Western popular music, having multiple priors with modes at times corresponding to four bars, eight bars and sixteen bars might be of interest.

We have not chosen tailored audio features for this investigation, preferring to concentrate on the treatment of those features. It is worth investigating the sensitivity of this method to the corpus it is applied to and the features used; in particular, it would be interesting to use the same features and corpus as in [7]. Performance might also improve with the use of a mixture of explicitly timbral and harmonic features, in contrast with the $\frac{1}{12}$th-octave features which we have used, so as to be able to differentiate segments both on instrumentation and on harmony.

5. REFERENCES

[1] Mark Levy, Mark Sandler, and Michael Casey, "Extraction of High-Level Musical Structure From Audio Data and its Application to Thumnail Generation," Accepted for publication in ICASSP 2006.

[2] Jonathan Foote, "Visualizing music and audio using self-similarity," in *ACM Multimedia (1)*, 1999, pp. 77–80.

[3] B. Logan and S. Chu, "Music summarization using key phrases," in *Proc. ICASSP*, Istanbul, 2000.

[4] Geoffroy Peeters, Amaury La Burthe, and Xavier Rodet, "Toward automatic music audio summary generation from signal analysis," in *Proc. ISMIR*, Paris, 2002.

[5] Mark A. Bartsch and Gregory H. Wakefield, "To Catch a Chorus: Using Chroma-Based Representations for Audio Thumbnailing," in *Proc. WASPAA*, Mohonk, New York, 2001.

[6] R. Dannenberg and N. Hu, "Discovering musical structure in audio recordings," in *Proc. ICMAI*, Edinburgh, 2002.

[7] Masataka Goto, "A chorus-section detecting method for musical audio signals," in *Proc. ICASSP*, Hong Kong, 2003, vol. V, pp. 437–440.

[8] L. Lu, M. Wang, and H. Zhang, "Repeating pattern discovery and structure analysis from acoustic music data," in *6th ACM SIGMM MIR Workshop*, New York, 2004.

[9] Samer Abdallah, Katy Noland, Mark Sandler, Michael Casey, and Christophe Rhodes, "Theory and evaluation of a Bayesian music structure extractor," in *Proc. Sixth ISMIR*, London, 2005.

[10] Jan Puzicha, Thomas Hofmann, and Joachim M. Buhmann, "Histogram clustering for unsupervised image segmentation," *Proceedings of CVPR '99*, Fort Collins, Colorado, 1999.

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

[12] Adrian Barbu and Song-Chun Zhu, "Cluster sampling and its applications in image processing," Tech. Rep. 409, University of California, Los Angeles, 2004.

[13] Qian Huang and Byron Dom, "Quantitative methods of evaluating image segmentation," in *Proc. IEEE Intl. Conf. on Image Processing (ICIP'95)*, Washington DC, 1995.

[14] C. J. van Rijsbergen, *Information Retrieval*, Butterworth, 1979.