A COMPARISON OF TIMBRAL AND HARMONIC MUSIC SEGMENTATION ALGORITHMS

Mark Levy, Katy Noland, Mark Sandler

Centre for Digital Music,
Queen Mary, University of London,
Mile End Road,
London, E1 4NS,
United Kingdom.

ABSTRACT

Four music segmentation algorithms are presented, one based on purely timbral features, one on purely harmonic features, and two on different combinations of features. They are compared against each other and against human annotations of two albums by *The Beatles*. Example segmentations are given together with a quantitative measure of boundary accuracy. No algorithm is found to be clearly superior, although examples suggest that the combined algorithms can offer improved boundary detection.

Index Terms— music, segmentation, boundary evaluation, information retrieval

1. INTRODUCTION

A high level musical structure is created by either a repetition of a sequence of features, or a change in a constant feature. This paper investigates ways of dividing a musical extract into high level sections such as *verse* or *chorus*, using timbre and harmony. An ideal segmentation would correctly extract the start and end times of each segment, and assign equivalent segments to the same class.

Such a segmentation has applications in musical summary generation, in music retrieval systems and in audio editing software for functions such as "skip to next section".

Previous techniques for partial structure extraction include [1] and [2], which use a self-similarity search to find repeated sections, assuming that a frequently repeated section is likely to be a chorus. However, the self-similarity search is computationally expensive and an ad hoc choice of distance metrics and thresholds has to be made. More recently extensions to give a complete structural analysis have been developed using heuristics based on rhythmic information [3, 4], but they include limiting assumptions that are only valid for conventional pop music. The multipass approach described in [5] is much less restrictive, where segment templates are found using self-similarity then refined by unsupervised clustering.

We explore the theory that segmentation using a single musical feature can be improved by using information from additional features. An existing segmentation algorithm based on harmonic features and one based on timbral features are investigated, and combined at an early and at a late stage in the extraction process to give a total of four segmentation techniques, which are compared against each other.

Section 2 describes the segmentation algorithms, the experiments are explained in section 3, example segmentations and boundary evaluation results are presented in section 4 and section 5 concludes the paper.

2. SEGMENTATION ALGORITHMS

2.1. Algorithm 1: Timbre

The timbral segmentation method is described in [6]. The timbral features correspond to the MPEG-7 AudioSpectrumProjection descriptor. We extract constant-Q log-power spectra over large (roughly 1.5s) overlapping analysis windows and normalise. We extract the first 20 principal components and add the feature norm as an extra dimension, then model the resulting timbral features with a 40-state HMM. Finally we Viterbi-decode the sequence of features against the trained model, to give a state path where the state at each time represents a particular local 'timbre-type'.

To capture variation over a longer timescale, we histogram the sequence of timbre types over a sliding window of 7 states. We then use the adapted soft K-means algorithm given in box 1 to cluster the histograms into $M = 10$ clusters, where each cluster represents a segment-type. The algorithm adapts to the data by leaving redundant clusters un-occupied. The resulting sequence of cluster assignments gives us our timbral segmentation.

2.2. Algorithm 2: Harmony

The harmonic segmentation technique is that described in [7]. A 24-state HMM is defined such that each state represents a key, which emits a chord transition at each time step, as shown in figure 1. The model is initialised using the results of perceptual tests that relate the chord transitions and keys, then

trained using an expectation-maximisation algorithm. The Viterbi algorithm is used to decode the most likely key sequence for each song, and contiguous frames in the same key are treated as a segment.

This algorithm is currently reliant on hand annotations of chords for the input observations, but can be extended to audio with the addition of a chord recognition step, such as [8].

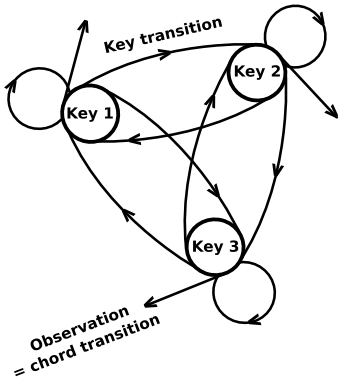


Fig. 1. Harmonic model showing 3 of the 24 states.

BOX 1: CLUSTERING ALGORITHM

Initialization. Set M reference histograms $\{h_m\}$ to random values. Assign all histograms $y(t)$ arbitrarily to the first reference histogram i.e. set $s(t) = 1, \forall t$. Set $\beta = \beta_0$.

Loop while $\beta \geq \beta_{final}$

Assignment step. Calculate the responsibilities of each reference histogram $r_m(t)$ for each data histogram $y(t)$:

$$r_m(t) = \frac{exp(-\beta d_{KL}(h_m, y(t))}{\sum_{m'} exp(-\beta d_{KL}(h_{m'}, y(t)))}$$

Adjust responsibilities by a term expressing local quality of current segmentation:

$$r'_m(t) = r_m(t) exp(-\lambda n_m(t))$$

Update step. Adjust the reference histograms:

$$h_m = \frac{\sum_t r'_m(t) y(t)}{\sum_t r'_m(t)}$$

Assign each data histogram to a reference histogram:

$$s(t) = \mathrm{argmax_m} r'_m(t)$$

Repeat assignment and update steps a fixed number of times or until the assignments do not change.

Set $\beta = \alpha\beta$.

$d_{KL}(.,.)$ is a symmetrised Kullback-Leibler divergence. $n_m(t)$ measures the number of non-matching cluster assignments within a fixed neighbourhood of $s(t)$.

2.3. Algorithm 3: HMM trained with combined features

The first combined method is similar to algorithm 1, and creates large timbre-and-harmony feature vectors early in the process. The timbral features are calculated as in algorithm 1 by applying PCA to a constant-Q spectrum. A chromagram is calculated for the same frames and appended to the timbral features after they are scaled to the same range of [0, 1]. The timbral segmentation is then implemented as already described, but using these larger features to train the HMM.

2.4. Algorithm 4: Clustering with Combined Features

The second combined algorithm joins the two techniques just before the clustering stage. The chord recognition algorithm described in [8] is employed to begin with, and used to create the observations for the harmonic model of algorithm 2. The HMM is trained as for the harmony-only segmentation, then the posterior state probabilities are calculated to give the likelihood of each key at each time frame.

The HMM stage of the timbral extraction is carried out as in algorithm 1, but using a 24-state HMM to give state histograms with the same number of dimensions as the harmonic state probabilities. The harmonic posterior state probability is appended to the histogram of timbral features after normalisation to give both features equal weighting, then the clustering algorithm is applied to the large feature vectors.

3. EXPERIMENTS

3.1. Test Data

Segmentations of two albums by the Beatles, *With The Beatles* and *Sgt. Pepper's Lonely Hearts Club Band* were completed by hand, with all segments beginning on the first beat of the bar and labels largely following those given in [9], but it should be noted that opinions vary regarding the ideal segmentation. All four algorithms were tested against the hand annotations. Visualisations such as those in figure 2 can be used for a subjective evaluation, and a quantative measure of boundary accuracy was also calculated.

3.2. Boundary Evaluation

Precision and recall values were used to measure the boundary accuracy.

$$precision = \frac{number\ of\ correct\ machine\ boundaries}{total\ number\ of\ machine\ boundaries}$$

$$recall = \frac{number\ of\ ground\ truth\ boundaries\ detected}{total\ number\ of\ ground\ truth\ boundaries}$$

If a detected boundary occurred within 3 seconds of a ground truth boundary it was considered correct. Similarly, if a ground truth boundary was within 3 seconds of a detected boundary it was considered to have been correctly recalled.

4. RESULTS

Figure 2 shows examples for 3 songs from the Beatles collection. These examples show that the different algorithms cannot be simply ranked, and no algorithm stands out as always better than the others. In example (a) algorithm 4 shows improved boundary accuracy over the other techniques, since it is able to make use of the improved precision of the harmonic features, but the clustering algorithm does not allow the short segments created by algorithm 2 so these are merged. Although algorithm 2 shows the worst precision value, it does show repetition in the structure that is not present in any of the other machine segmentations, and can be used to identify that there are two verses at the start of the song.

In example (b) algorithm 3 performs best, while algorithm 1 appears to be over-segmenting. However, listening to the song reveals that it has found a lower level structure: each of the verse/refrain segments consists of a 4-bar verse-like phrase followed by an 8-bar refrain-like passage. The harmony changes very little in this song, which in the combined segmentations has helped to smooth out the over-segmentation.

In example (c) while the two combined algorithms give similar numerical results, there is a difference in boundary locations. In fact, the harmonic segmentation, while missing some boundaries, is the most accurate localiser of those it does find.

The boundary evaluation results of Table 1 verify that there is no clear best algorithm. The harmonic segmentation performs least well overall, but when combined with the timbral technique it can improve accuracy even when it gives a poor segmentation alone, since it is less sensitive to changes of instrumentation that often occur mid-section. Further investigation of these cases is needed, but these initial results suggest that using a combination of features for segmentation can be effective.

Table 1. Average boundary evaluation results

Segmentation algorithm	With The Beatles		Sgt. Pepper	
	Precision	Recall	Precision	Recall
1: timbre	0.67	0.64	0.61	0.72
2: harmony	0.51	0.55	0.53	0.63
3: combined	0.67	0.59	0.65	0.64
4: combined	0.70	0.59	0.57	0.58

5. CONCLUSION

There is no clearly superior segmentation algorithm of these four. Combining harmonic information either before or just after the HMM stage seems to give an improvement in many cases, though this is yet to be fully evaluated. It must be borne in mind that the Beatles' music, although varied, is not representative of all genres, and different features may be more suitable for general useage. However, the information provided by a harmonic or timbral analysis can also be used additional to a structural segmentation, for instance as an additional parameter in a music search engine.

We believe that graphical and tabular results are both valuable and provide different information to the researcher. In our future work we intend to take this one step further by comparing these two objective measures with user-preference listening tests.

6. REFERENCES

[1] M. Goto, "A chorus-section detecting method for musical audio signals," in *Proceedings of the 2003 IEEE Conference on Acoustics, Speech and Signal Processing*, 2003.

[2] J. Foote, "Visualizing music and audio using self-similarity," in *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, 1999, pp. 77–80.

[3] N. C. Maddage, C. Xu, M. S. Kankanhalli, and X. Shao, "Content-based music structure analysis with applications to music semantics understanding," in *Proceedings of the 12th annual ACM international conference on Multimedia*, 2004.

[4] L. Lu, M. Wang, and H.-J. Zhang, "Repeating pattern discovery and structure analysis from acoustic music data," in *Proceedings of the 6th ACM SIGMM international workshop on multimedia information retrieval, New York*, 2004.

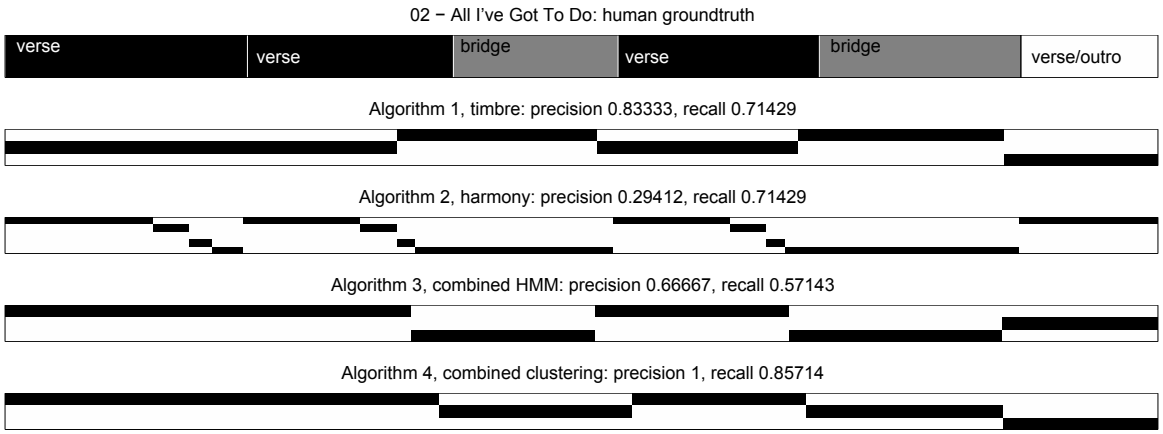
[5] G. Peeters, A. La Burthe, and Xavier Rodet, "Toward automatic music audio summary generation from signal analysis," in *Proceedings of the 3rd International Conference on Music Information Retrieval, Paris*, 2002.

[6] M. Levy, M. Sandler, and M. Casey, "Extraction of high-level musical structure from audio data and its application to thumbnail generation," in *Proceedings of the 2006 IEEE Conference on Acoustics, Speech and Signal Processing*, 2006.

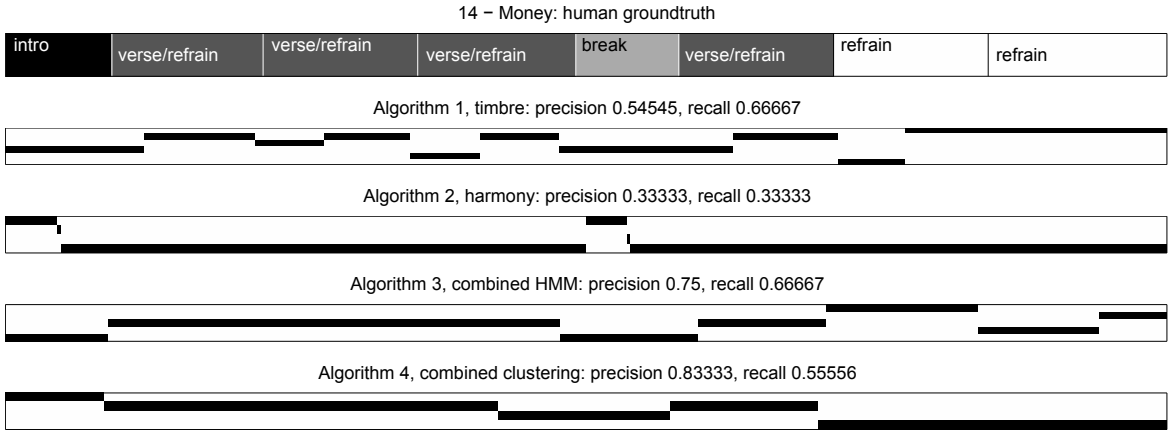
[7] K. Noland and M. Sandler, "Key estimation using a hidden Markov model," in *Proceedings of the 7th International Conference on Music Information Retrieval, Victoria, Canada*, 2006.

[8] C. Harte and M. Sandler, "Automatic chord identification using a quantised chromagram," in *Proceedings of AES 118th Convention, Barcelona*, 2005.

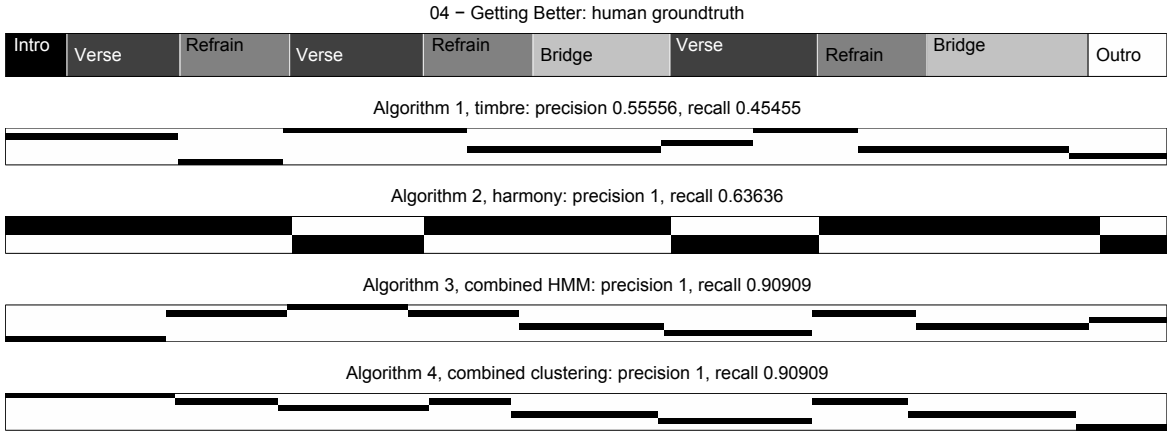
[9] A. W. Pollack, "Notes on ... series," *soundscapes.info*, 2000, [web site], [accessed 2006 Sep 27], Available: *http://www.icce.rug.nl/∼soundscapes/DATABASES/AWP/awp-notes_on.shtml*.



(a) *All I've Got to Do* from *With The Beatles*



(b) *Money* from *With The Beatles*



(c) *Getting Better* from Sgt. Pepper

Fig. 2. Example segmentations of 3 songs, showing the human annotations and 4 machine segmentations.