

EXTRACTION OF HIGH-LEVEL MUSICAL STRUCTURE FROM AUDIO DATA AND ITS APPLICATION TO THUMBNAIL GENERATION

Mark Levy, Mark Sandler

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

Michael Casey

Centre for Cognition, Computation and Culture
Goldsmiths College, University of London
New Cross Gate, London SE14 6NW, UK

ABSTRACT

A method for segmenting musical audio with a hierarchical timbre model is introduced. New evidence is presented to show that music segmentation can be recast as clustering of timbre features, and a new clustering algorithm is described. A prototype thumbnail-generating application is described and evaluated. Experimental results are given, including comparison of machine and human segmentations.

1. INTRODUCTION

This paper describes an approach to extracting the overall structure of a piece of music directly from an audio recording. Knowledge of this structure has various useful practical applications, for example: music summarisation, such as the thumbnail-generating application presented here; automatic section-by-section alignment of audio tracks to aid retrieval algorithms; the development of features for use in audio editing or synchronization to video, such as ‘jump to start of next section’, ‘double-click to select current phrase’, etc.

Previous research [1, 2] has explored partial extraction of high-level musical structure by a self-similarity search for repeated sections. With such methods only some sections are identified and labelled, the pairwise similarity between all analysis frames within a track has to be evaluated, which is computationally expensive, and the choice of distance metrics and similarity thresholds is somewhat ad hoc. More recent work [3, 4] extends this approach to extract the entire structure of musical tracks by adding a set of heuristics based on an estimation of the beat, bar and phrase-length of the piece in question, and also by making some extremely limiting assumptions about the nature of the structure being sought, which hold only for conventional pop music. The approach in [5] is less restrictive, using self-similarity to find an initial set of candidate segment boundaries and segment templates which are refined by unsupervised clustering.

Our work [6] recasts the problem of extracting structure as one purely of clustering suitable features. These correspond perceptually to the overall timbre of the music over two different time-scales, and the result of our clustering procedure is a hierarchical segmentation by timbre. This naturally yields a structural analysis of the entire audio track, and also allows the novel use of a range of unsupervised clustering algorithms. The absence of restrictive assumptions, and the perceptually meaningful nature of the features, means that the method can usefully be applied to a wide range of musical styles. Timbre features have been widely used for supervised genre classification (for a recent summary see [7]). Our work aims to extend

this to the unsupervised identification of sections of similar overall timbre within a single audio track.

The organisation of the rest of this paper is as follows: section 2 describes the hierarchical model and the features used; section 3 justifies the use of clustering for segmentation and introduces a new clustering algorithm; section 4 illustrates some output segmentations and gives an evaluation of the performance of our approach in relation to some manual segmentations; section 5 describes an application that generates music thumbnails from machine segmentations; section 6 outlines further work.

2. A HIERARCHICAL MODEL OF TIMBRE

Traditional musicological approaches to structure extraction place little weight on timbre as a structural dimension, depending heavily on repetition, both exact and approximate, as a structural principle. On the other hand, while studies from the audio analysis perspective [8, 9] have successfully clustered short frames into a given number of timbre-types, no high-level structure emerges. Although short sequences of neighbouring frames may be assigned to the same timbre-type, the overall timbre changes frequently during the course of any section of significant length. Our approach builds on previous work by drawing on the following insights. Although timbre changes from frame to frame, the distribution of timbre-types remains fairly consistent over the course of structural segments, and can be used to characterise segment-types (typically ‘chorus’, ‘verse’, etc.). Discernible timbre-types (loosely corresponding to specific combinations of instruments) are nonetheless shared between segment-types. This suggests the following model:

For a given track, we divide the space of possible timbres into N timbre-types $\{q_1, \dots, q_N\}$, each of which generates timbre features \vec{x} according to a Gaussian distribution $b_j(x) = P(x|q_j) = \mathcal{N}[x, \mu_j, U]$, where the covariance matrix U is shared by all timbre-types. We model the sequence of timbre features through the track with an N -state Hidden Markov Model [10], where the hidden states correspond to the N timbre-types, i.e. we assume fixed transition probabilities between timbre-types $\{a_{ij}\}$, where $a_{ij} = P(q_j(t+1)|q_i(t))$. Given the trained HMM parameters, we can then decode the most likely sequence of timbre-types $\hat{Q} = \{q(1), \dots, q(T)\}$ to have generated the features. We assume there are a fixed number M of segment-types $\{s_1, \dots, s_M\}$, where each segment-type generates timbre-types according to a fixed distribution $h_m(j) = P(q_j|s_m)$. Note that the characteristic distribution h_m corresponds to the mixture weights for a Gaussian mixture model for each segment-type:

$$P(x|s_m) = \sum_{j=1}^N h_m P(x|q_j) = \sum_{j=1}^N h_m \mathcal{N}[x, \mu_j, U] \quad (1)$$

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

We find both these characteristic distributions, and the most likely corresponding segmentation, by clustering histograms of timbre-types from \hat{Q} , as described below.

We extract timbre features by computing constant- Q spectra at $\frac{1}{8}$ th-octave resolution over large overlapping analysis windows, using a window size three times the hop size. We use a hop equal to the beat length of the music (typically 300-400ms), as estimated by a beat-tracking algorithm [11]. A resolution of one beat is appropriate because this is the natural granularity for musical segmentation. The spectra are normalised and subjected to Principal Component Analysis. Finally we combine the first 20 PCA components and the normalised envelope to yield 21-dimensional feature vectors. We train a 40-state HMM on the entire sequence of feature vectors for a given track, with a single Gaussian output distribution for each state, and a single covariance matrix tied across all states. We then Viterbi-decode the features using the trained model. The resulting state sequence gives the most likely sequence of assignments for each beat of the music to one of 40 possible timbre-types.

To capture timbre variation over a longer time-scale, we histogram the decoded sequence of timbre-types over a sliding window of 7 beats in length, and then normalise the resulting histograms. We then estimate the characteristic mixture of timbre-types for each segment by clustering the histograms into M clusters. The reference histograms for each cluster give distributions of timbre-types $\{h_m\}$ and the cluster assignments give the corresponding segmentation $S = \{s(1), \dots, s(T)\}$, where $s(t)$ gives the segment-type assignment for frame t . In [6] we found that our segmentations varied significantly with the number of clusters M chosen. Using the algorithm presented here, however, redundant clusters are left unoccupied and so we can set M to some arbitrary large number (we frequently use $M = 10$).

3. SEGMENTATION AS A CLUSTERING PROBLEM

Because clustering throws away all information about temporal contiguity, it will not necessarily produce a meaningful segmentation. If different segment-types emit mutually exclusive sets of timbre-types, then clustering should work easily as a segmentation method, but if segment-types have many timbre-types in common, it is likely to fail. More generally, clustering histograms will only produce a usable segmentation if the histograms belonging to frames within segments of a given type do genuinely fall into clusters.

For audio tracks where a manual ‘‘groundtruth’’ segmentation $S_0 = \{s_0(1), \dots, s_0(T)\}$ is available, we can test this with the following simple experiment. We create a sequence of normalised histograms $\{y_1, \dots, y_T\}$ of timbre-types from the audio, as described above. We calculate reference histograms $\{\hat{h}_m\}$ for each segment-type by taking the mean over histograms corresponding to all frames assigned to that segment-type in the manual segmentation: $\hat{h}_m = \frac{\sum_{\{t:s_0(t)=m\}} y_t}{|\{t:s_0(t)=m\}|}$. We then make segment-type assignments $\hat{S} = \{\hat{s}(1), \dots, \hat{s}(T)\}$ for each frame according to the closest reference histogram: $\hat{s}(t) = \operatorname{argmin}_m d(y_t, \hat{h}_m)$, where $d(y_t, \hat{h}_m)$ is a simple Euclidean distance. The degree of agreement between \hat{S} and S_0 gives a clear indication of whether or not the histograms are genuinely clustered by segment-type, while \hat{S} itself gives a reasonable ‘upper bound’ on the quality of segmentation we can get by clustering.

Although audio tracks with corresponding human segmentations are not available in sufficient numbers to allow large-scale tests, results of this experiment on a small test set of 14 pop songs, containing some 196 hand-labelled segments, strongly support the idea

that segmentation can be reduced to clustering of timbre-type histograms. Typically the overall structure of the segmentation is well preserved, as illustrated in Figure 1(a), with only short runs of misclassified frames, which can easily be identified and corrected, and small differences in the position of segment boundaries. The mean number of correctly classified frames per segment was 87%, over all segments in the test set. Significantly, we have also observed the same clustering of timbre features by segment-type directly in the feature vectors themselves, when averaged over a moving window of the same length as our histogram window. The effectiveness of segmentation by clustering timbre features is therefore not dependent on our particular approach via histograms of decoded timbre-types, but rather depends simply on modelling the dynamic evolution of timbre over a suitable time-scale.

While segmentation by clustering timbre is not a perfect analogue of human segmentation, it clearly has the potential to produce results in many cases that are meaningful in musical terms. Simple clustering techniques, however, are not able to find the reference histograms. We explored a variety of clustering methods within a Bayesian framework in [6, 12]. We have also had good results with the following adaptation of soft k-means clustering:

Initialization. Set M reference histograms $\{h_m\}$ to random values. Set $\beta = \beta_0$.

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

Assignment step. Calculate the responsibilities of each reference histogram for each data histogram:

$$r_m(t) = \frac{\exp(-\beta d_{KL}(h_m, y(t)))}{\sum_{m'} \exp(-\beta d_{KL}(h_{m'}, y(t)))} \quad (2)$$

where $d_{KL}(h_m, y(t))$ is a symmetrised Kullback-Leibler divergence.

Assign each data histogram to a reference histogram:

$$s(t) = \operatorname{argmax}_m r_m(t) \quad (3)$$

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

$$r'_m(t) = r_m(t) \exp(-\lambda n_m(t)) \quad (4)$$

where $n_m(t)$ measures the number of non-matching cluster assignments in the neighbourhood of $y(t)$. More precisely, $n_m(t) = B - A_m(t)$ where B defines the size of the neighbourhood, $A_m(t) = |\{t' \in N(t) : s(t') = m\}|$ is the number of matching assignments, and $N(t) = \{t' : t - (B - 1)/2 \leq t' \leq t + (B - 1)/2\}$.

Reassign each data histogram:

$$s(t) = \operatorname{argmax}_m r'_m(t) \quad (5)$$

Update step. Adjust the reference histograms:

$$h_m = \frac{\sum_t r'_m(t) y(t)}{\sum_t r'_m(t)} \quad (6)$$

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

Set $\beta = \alpha\beta$.

This incorporates a simple duration prior, reflecting our preference for neighbouring frames to be clustered together, defined by a neighbourhood size B and a weighting parameter λ , and a decreasing inverse-temperature parameter β . We use experimentally determined values of $B = 41$, $\lambda = 0.02$, $\beta_0 = 100$, $\alpha = 0.7$ and $\beta_{final} = 0.1$. Some segmentations produced by this algorithm are given in Figure 1(b), showing how the output is largely independent of the value chosen for M .

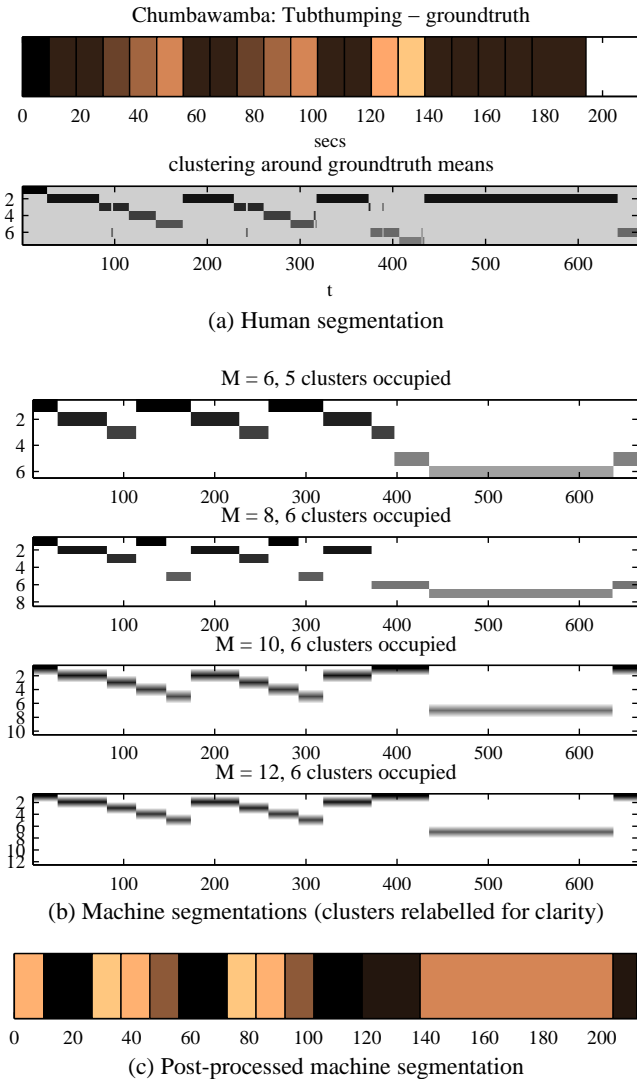


Fig. 1. Segmentations of a pop song

4. EVALUATION

If our intention is to imitate human judgement then the significant measures are perceptual: do we hear something change at segment boundaries? do segments of the same type sound similar? etc. In the case of popular music, recent research¹ suggests that different

¹Personal communication from M. Bruderer, Technical University of Eindhoven.

listeners share a consistent view of where the major segment boundaries fall, and we can therefore reasonably construct quantitative tests against a human “groundtruth”. Figure 2(a) gives an idea of how well groundtruth segments are reproduced in machine segmentations of our small test set, using measures of boundaries missed m and segments fragmented f developed in [6], and showing a modest improvement on our earlier results. The distances of found boundaries from the nearest groundtruth boundary are shown in Figure 2(b).

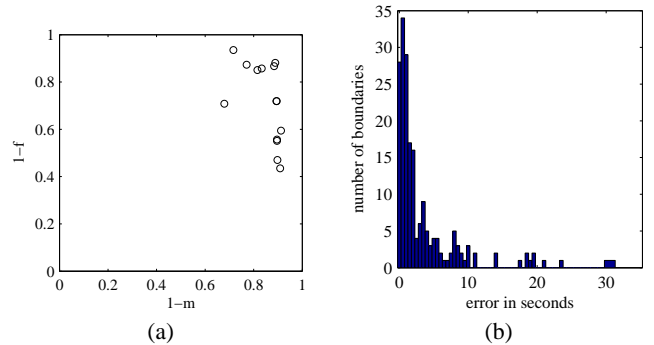


Fig. 2. Evaluation measures over test set: (a) $1 - f$ against $1 - m$ (loosely ‘precision’ vs ‘recall’) (b) Accuracy of machine segment boundaries

We can improve segmentations in a post-processing phase by noting that conventional popular music follows an extremely simple structure, dictated by the verse-chorus form of the lyrics and very predictable phrase-lengths, so that segments are simple multiples of a basic eight-bar phrase. These lengths are quite precisely observed in performance, and hence in the audio signal, due to the ubiquitous use of machine-generated click-tracks or drum loops in the recording studio. This implies an extremely strong constraint on segment lengths, leading to a simple optimisation problem which we implement as a brute-force search for the best-fitting basic phrase-length. Given machine-produced segment boundaries at times $\{t_0, \dots, t_L\}$ we define a distortion measure $z = \sum_{i=1}^{L-1} (\hat{t}_i(u, d) - t_i)^2 + d^2$, where $\hat{t}_i(u, d) = d + u \cdot \text{round}(\frac{t_i - d}{u})$ is the corrected position of t_i to match a fixed phrase-length u and an offset d , representing any silence or unmeasured music at the very start of the track. We find optimal values for $\{\hat{t}_i\}$ by minimising z over small non-negative values of d and a range of values of u close to half of the median segment length. A post-processed segmentation is shown in Figure 1(c).

Although the segmentations produced by timbre clustering will not always agree with human judgement, it should be noted that there are applications which can perfectly well use machine segmentations that do not reflect a typical human notion of musical structure. These include both some of the practical applications mentioned in section 1, and musicological applications, such as the structural analysis of less familiar musical repertoires, in which traditional approaches break down.

5. A THUMBNAIL-GENERATING APPLICATION

In contrast to the partial segmentations extracted by self-similarity searching, the clustering approach automatically yields a complete segmentation of the supplied audio track. The better musical overview

