

NEW METHODS IN STRUCTURAL SEGMENTATION OF MUSICAL AUDIO

Mark Levy, Mark Sandler

Centre for Digital Music
Queen Mary, University of London
Mile End Road, London E1 4NS, UK
mark.levy@elec.qmul.ac.uk, mark.sandler@elec.qmul.ac.uk

ABSTRACT

We describe a simple model of musical structure and two related methods of extracting a high-level segmentation of a music track from the audio data, including a novel use of hidden semi-Markov models. We introduce a semi-supervised segmentation process which finds musical structure with improved accuracy given some very limited manual input. We give experimental results compared to existing methods and human segmentations.

1. INTRODUCTION

This paper introduces a simple model of musical structure, and shows how it can be either formalised directly as a segment model, or used to inform a clustering approach to audio segmentation. Although the model is very general, and could be applied to symbolic data (notes, chords, etc.) drawn from score representations of music, our focus here is on its application directly to audio data, in order to extract high-level musical structure from audio tracks with at most a small amount of human intervention. Knowledge of this structure has immediate practical applications in the context of audio or video editing, enabling the development of features such as ‘jump to start of next phrase’, ‘align with current musical phrase’, etc. It leads easily to the automatic extraction of musical summaries or ‘thumbnail’ segments, for use in browsing and searching the large audio collections which are rapidly becoming commonplace through the popularity of download sites, MP3 players and associated technologies. Comparison of segment models also offers possible new methods for content-based audio retrieval, similarity search and music recommendation.

Early research into automatic segmentation of musical audio [1, 2] had some success in the partial extraction of high-level musical structure by a self-similarity search over spectral features to find repeated sections. This approach has some drawbacks: only some sections are identified and labelled, the choice of distance metrics and thresholds for similarity is somewhat ad hoc, and the search is computationally expensive, requiring the calculation of pairwise distances between all analysis frames within a track. Self-similarity search has been extended in [3, 4] to extract the structure of complete pop tracks by incorporating information from beat-tracking and applying a set of heuristics to identify segments as being one of a very small number of specific types (instrumental, verse or chorus). These methods make some highly limiting assumptions about the structure being sought that apply only to conventional pop music. Hidden Markov

models (HMMs), where the hidden states correspond to segment types, are used for segmentation in [5], and a two-pass method, in which candidate segment boundaries are detected and the intervening segments then used to initialise an HMM, is outlined in [6]. These approaches are much less restrictive, but unfortunately the chosen model implicitly defines a geometric probability distribution for segment durations, which is not what we observe in real music, where segment lengths are typically multiples of some basic phrase length, for example 8, 16 or 32 bars, only very rarely taking other values.

In our own work [7] we have partially addressed the issue of modelling expected segment lengths realistically within a clustering framework, by including a term expressing the relative unlikelihood of segments being shorter than an experimentally-determined minimum duration. In this paper we introduce a method for estimating the underlying base phrase length of a piece of music from audio data, and show how this can either be passed to our clustering method, or used to initialise a different model for high-level segmentation, which allows us to specify a full probability distribution for segment lengths.

Although unsupervised approaches to segmentation have been shown to achieve results similar to human judgement in some cases, research into parallel problems in image segmentation [8] suggests that a small amount of supervision may offer large gains in segmentation accuracy. In the case of musical audio segmentation, a major cause of fragility in unsupervised methods is the difficulty of providing good initial parameter values to models such as HMMs, whose training algorithms are well known to be susceptible to poor initialisation [9]. We address this problem here with a simple semi-supervised approach to segmentation.

The organisation of the rest of this paper is as follows: section 2 describes the musical model and the audio features used; section 3 describes how a segment-length distribution is estimated; section 4 introduces the new segmentation methods; section 5 describes the semi-supervised approach and illustrates output segmentations; section 6 evaluates the performance of the methods in relation to previous work and some manual segmentations; section 7 outlines further work.

2. A MODEL OF MUSICAL STRUCTURE

2.1 The nature of musical structure

We can describe the structure of most music as follows. Activity at the level of the beat, i.e. notes to be sung or played by particular instruments, is organised into regular phrases, typically in Western music of 4 or 8 bars, where each bar contains 3 or 4 beats. These phrases are concatenated, whether by the composer on manuscript paper, or the producer or mu-

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

sician in the recording studio, to form structural sections, or segments, in accordance with the stylistic norms of the particular musical genre in question. A piece is then constructed from a sequence of segments of various types, in an order again largely determined by its genre.

While this is clearly a weak account of the real act of composition, it does suggest a straightforward formal model of musical structure that we can reasonably expect to fit a great many pieces of music, for example most popular, folk, world and much classical music. We assume further that we have some way of labelling each beat or frame of the music in such a way that beats or frames which are musically similar are assigned the same label. A piece can then be represented as a sequence of labels $\{y_t\}, t = 1, 2, \dots, T$, and the task of segmentation consists of assigning each of the y_t to one of a set of segment-types $Q = \{q_1, q_2, \dots, q_M\}$, subject to suitable expectations, which we can express as conditional probability distributions, on the resulting segment durations and the particular sequences of labels observed in each segment-type.

2.2 The observation sequence

Given an audio track, we aim to label each beat as belonging to one of N possible timbre-types dividing the overall space of timbre used in the track. We first estimate the beat using a beat-tracking algorithm [10] and then extract constant- Q spectra at $\frac{1}{8}$ -octave resolution, using a hop equal to the beat-length (typically 300-400ms) and a window of three times the hop size. 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 an N -state HMM on the sequence of feature vectors, 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 to give the most likely sequence of timbre-types. In music with a relatively small overall timbre space, such as simple verse-chorus songs, we observe that the labelled timbre-types correspond clearly to particular combinations of notes performed with a similar instrumentation. Figure 1 shows a typical sequence of timbre-type labels. The number of timbre-types N should be large enough to show clear variation in the labels observed in segments of different type, but small enough for the computational demands of (and the number of parameters to be learned by) a formal model to remain manageable. Following experiments over a small test set of tracks we use a value of $N = 40$. Note that although we use approximately beat-length analysis frames in this paper, this is not a requirement of our approach, but rather is intended i) to aid clarity of discussion and ii) to limit computational requirements by keeping the maximum segment length to a modest value. In music where full beat-tracking is possible, we are able to use strictly beat-length frames, but we reserve discussion of this for a future paper.

3. ESTIMATING EXPECTED SEGMENT LENGTHS

Periodicity and rhythmic structure at the beat and bar level have previously been estimated from the autocorrelation or ‘beat spectrum’ of suitable extracted features [11, 12]. We extend this to search for periodicity at the phrase level by a direct analysis of the sequence of timbre-type labels. We first create normalised histograms $\{x_t\}$ of timbre-types over a

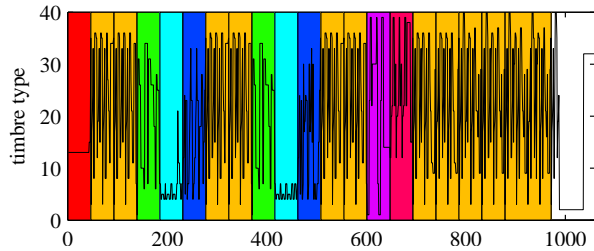


Figure 1: Sequence of timbre-types against manual segmentation. Note how related segments (shown in same background shade) contain similar sequences of states.

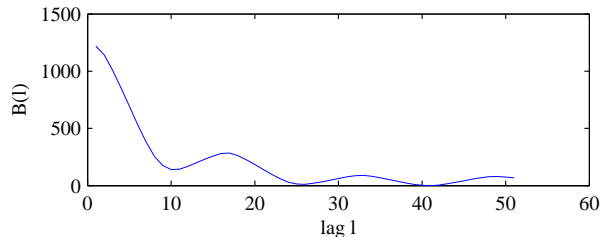


Figure 2: Approximate autocorrelation of timbre-type sequence, showing first strong peak at 4-bar phrase length.

moving window of length w . We then sum pairwise distances between histograms at lags $1 \leq l \leq D$ to give an approximate autocorrelation $B(l) = \sum_{t=1}^{T-l} -d_{KL}(x_t, x_{t+l})$ where $d_{KL}(x, x')$ is a symmetrised Kullback-Liebler divergence, reflecting the relative likelihood of the two histograms being drawn from two separate or one combined distribution of timbre-types, and is given by

$$d_{KL}(x, x') = \sum_{i=1}^N x(i) \log \frac{x(i)}{q(i)} + x'(i) \log \frac{x'(i)}{q(i)}$$

where $q(i) = \frac{x(i) + x'(i)}{2}$.

Figure 2 shows an example of the resulting function, in which candidate phrase lengths appear as peaks. The degree of smoothing of $B(l)$, and therefore the smallest phrase length to be considered, is controlled by the histogram window size w . To estimate a base phrase length d_{pl} , we first calculate a moving baseline $B_0(l)$ by smoothing $B(l)$ with a median filter of length 5, and then pick the larger of the first two peaks in $B(l) - B_0(l)$. In experiments over a test set of popular music, d_{pl} was reliably found to be a four-bar phrase length when using a histogram window size of $w = 7$ (with beat-length frames). Given this base phrase length, it is simple to estimate the overall distribution of segment lengths for a track, because in most music the length of the great majority of segments is some multiple of d_{pl} , as illustrated in Figure 3, which shows the distribution of segment lengths over the test set according to expert human segmentations.

4. SEGMENTATION METHODS

4.1 The segmental or hidden semi-Markov model

In an HMM [9], at each time t the system is in one of M states $\{q_1, q_2, \dots, q_M\}$ and generates an observation y_t according to a distribution $P(y_t | Q_t = q_i) = b_i(y_t)$. The sys-

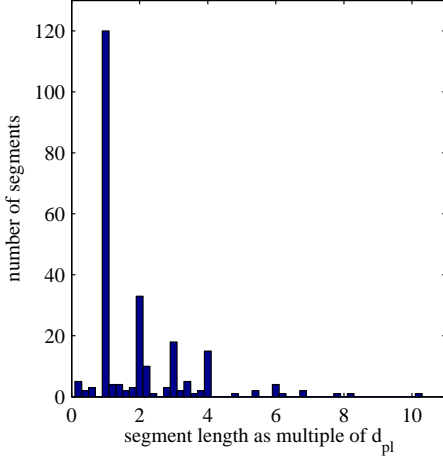


Figure 3: Distribution of segment lengths over test set.

tem then makes a transition to another state with probability $P(s_{t+1} = q_j | Q_t = q_i) = a_{ij}$. Consequently i) the duration d for which the system stays in any given state i has an implicit geometric distribution $P(d) = a_{ii}^{d-1}(1 - a_{ii})$; and ii) the observations generated while in this state are independent of one another and identically distributed. Although HMMs have proved effective in modelling processes where i) and ii) do not necessarily hold, including sequences of musical audio features considered at the frame timescale (see [13] for a recent example), a different model is clearly required for high-level musical structure.

The segmental or hidden semi-Markov model (HSMM) [14, 15] extends the HMM to remove these constraints. In an HSMM, the state duration d takes values $1, 2, \dots, D$ according to an explicit distribution, and the d observations generated by a given state i ending at time t are modelled by a joint distribution $P(y_{t-d+1}, \dots, y_t | i)$. The system enters an initial state i according to a distribution $P(i) = \pi_i$ and selects a duration d according to $P(d|i)$. It then generates d observations according to $P(y_1, \dots, y_{d-1} | i)$, before making a transition to a different state $j, j \neq i$ according to $P(j|i) = a_{ij}$. A duration is chosen, and the corresponding number of observations generated, from the distributions for the new state j , before making a transition to another state, and so on. The process continues until T observations have been generated. Inference in the HSMM can be performed with the following forwards and backwards recursions, where we write $F_t = 1$ if there is a change of state at time $t + 1$ (see [15] for a full derivation):

$$\begin{aligned} \alpha_t(j) &\doteq P(Q_t = j, F_t = 1, y_1, \dots, y_t) \\ &= \sum_d P(y_{t-d+1}, \dots, y_t | j, d) P(d|j) \alpha_{t-d}^*(j) \\ \alpha_t^*(j) &= \sum_i \alpha_t(i) a_{ij} \end{aligned}$$

initialised by $\alpha_0^*(j) = \pi_j$ and

$$\begin{aligned} \beta_t(i) &\doteq P(y_{t+1}, \dots, y_T | Q_t = i, F_t = 1) \\ &= \sum_j \beta_t^*(j) a_{ij} \\ \beta_t^*(i) &= \sum_{d=1}^D \beta_{t+d}(i) P(d|i) P(y_{t+1}, \dots, y_{t+d} | i) \end{aligned}$$

initialised by $\beta_T(i) = 1$.

State initial and transition probabilities can then be re-estimated as

$$\begin{aligned} \hat{\pi}_i &= \frac{\pi_i \beta_0^*(i)}{\sum_{i'} \pi_{i'} \beta_0^*(i')} \\ \hat{a}_{ij} &= \frac{\sum_{t=1}^T \alpha_t(i) a_{ij} \beta_t^*(j)}{\sum_{j'} \sum_{t=1}^T \alpha_t(i) a_{ij'} \beta_t^*(j')} \end{aligned}$$

and state duration probabilities as

$$\hat{P}(d|i) = \frac{\sum_t \alpha_{t-d}^*(i) P(d|i) P(y_{t-d+1}, \dots, y_t | i) \beta_t(i)}{\sum_{d'} \sum_t \alpha_{t-d'}^*(i) P(d'|i) P(y_{t-d'+1}, \dots, y_t | i) \beta_t(i)}$$

In the case of discrete observations, if we continue to treat observations as conditionally independent, i.e.

$$P(y_{t-d+1}, \dots, y_t | i) = \prod_{t'=t-d+1}^t P(y_{t'} | i) \quad (1)$$

the observation probabilities can be re-estimated as

$$\hat{P}(k|i) = \frac{\sum_{t:y_t=k} \sum_{\tau < t} [\gamma_\tau^*(i) - \gamma_\tau(i)]}{\sum_t \sum_{\tau < t} [\gamma_\tau^*(i) - \gamma_\tau(i)]}$$

where $\gamma_t(i) \doteq P(Q_t = i, F_t = 1 | y_1, \dots, y_T) \propto \alpha_t(i) \beta_t(i)$ and $\gamma_t^*(i) \doteq P(Q_{t+1} = i, F_t = 1 | y_1, \dots, y_T) \propto \alpha_t^*(i) \beta_t^*(i)$.

The HSMM clearly meets the needs of the model outlined in 2.1, treating our sequence of beat or frame labels as the observations and the underlying structural segments as the succession of hidden states. For the time being we continue to treat observations as conditionally independent (1), i.e. we assume that segment-types can be distinguished by a characteristic distribution of labels (we presented some experimental evidence for this view in [7]). Given the trained HSMM, we can find the most likely sequence of segment-types to have generated the observed labels by decoding with a suitable extension of the Viterbi algorithm, based on the following recursion for $\delta_t(i)$, the posterior probability of the best state sequence ending in state i at time t :

$$\begin{aligned} \delta_t(i) &= \max_d \delta_{t-d}^*(i) P(d|i) P(y_{t-d+1}, \dots, y_t | i) \\ \delta_t^*(j) &= \max_i \delta_t(i) a_{ij} \end{aligned}$$

4.2 Segmentation by clustering

In [7] we introduced a method of clustering histograms of timbre-types (similar to those we use for estimating the base phrase length in section 3), subject to a constraint on the expected minimum segment length. This constraint is expressed as a term based on the number of matching assignments within a given neighbourhood. We previously used an experimentally-determined constant for the neighbourhood size, but now set the neighbourhood to be $B = 2d_{pl} + 1$, i.e. we consider labels within d_{pl} beats to be relevant.

5. SEGMENTATION EXPERIMENTS

We carried out segmentation experiments using both these methods on a small test set of 14 varied pop tracks from the 1980s and 90s for which we have expert manual segmentations prepared for the MPEG-7 working group, which

show a wide range of formal structure. We segment sequences of timbre-types extracted as described in section 2.2 by constrained histogram clustering and using our own HSMM toolbox implemented in Matlab. As the HSMM has to learn a large number of parameters from relatively few observations, and because its training algorithm (Expectation-Maximisation) guarantees a solution that is only a local minimum in the problem space, a good parameter initialisation is essential. As outlined in 3, we have the strong expectation that musical segments will be integer multiples of an underlying phrase length. We therefore initialise the state duration distribution for all states of the HSMM to have strong probabilities for multiples of d_{pl} up to the maximum state duration D , with very small non-zero probabilities for all other durations. We use the following method to initialise the state observation distributions of the HSMM:

5.1 Semi-supervised approach

We imagine that the user has opened a track in an audio editor and now wishes to segment it. The user clicks and drags twice to select two regions within the track, roughly corresponding to segments of different types. We then initialise two HSMM states with the distribution of timbre-types found in each of these regions, and the remainder randomly, to make up a total of M states. We use a fixed value of $M = 6$, suggested by the human groundtruths for our test set, although in a real application the desired value of M could easily be set interactively by the user. In our experiments we simulate the user’s region selection by choosing the first occurrence of each of the two most frequent segment types according to the groundtruth, but with each of their boundary positions subject to a random error of up to 2 seconds.

We have found that the following procedure improves Viterbi-decoding in the HSMM. After training we set state duration probabilities for all states to zero for short durations d where $2 \leq d < d_{pl}$. This ensures that any unwanted fragmentary segments in the decoded state sequence will be exactly of length 1, and can then easily be merged with a simple smoothing procedure. We observe that in some cases one particular HSMM state becomes a ‘switching’ state, whose occurrences are all of length 1, for example separating two segments of the same type in order to make up a total duration greater than the maximum state duration D allowed by the HSMM.

Examples of machine segmentations using these methods are shown in Figure 3, together with the human “groundtruth”. We observe that even in cases where clustering method successfully finds the structure of the music, the HSMM often gives more accurate segment boundary positions.

6. EVALUATION

Our approach places few hard constraints on output segmentations. Using constant- Q features and the current form of segment observation distributions in the HSMM, we simply expect that segments of the same type will share a characteristic overall pitch/timbre content. Our estimation of a base phrase length encourages the methods to find a solution consistent with the conventions of musical form, but there is no guarantee that our results will agree with expert human

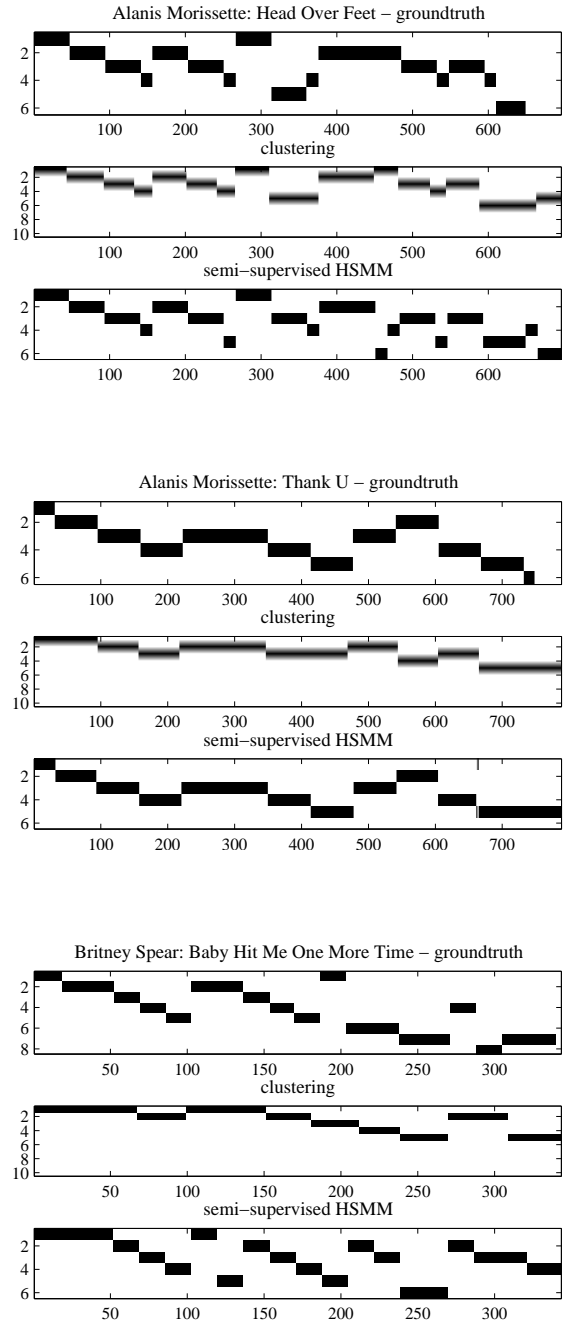


Figure 4: Human (top) and machine segmentations of three pop songs (segment-types relabelled for clarity).

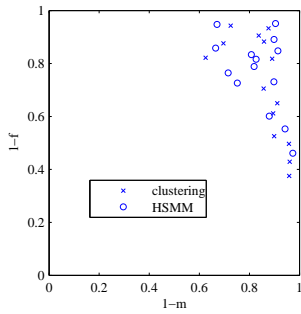


Figure 5: Evaluation of machine segmentations over test set: $1 - f$ against $1 - m$ (loosely ‘precision’ vs ‘recall’).

judgement. In our output we regularly observe details of segmentation that do not accord with the “groundtruth” but which are perfectly reasonable when evaluated subjectively, for example where the last line of a chorus is identified as a separate segment because the vocal line always jumps to a much higher register. In general, the structure revealed by machine segmentation will depend on the relationship of the chosen audio features to the music in question. Whether or not the machine segmentation is satisfactory consequently depends to a large extent on the use to which it will be put, and worthwhile reference groundtruths for testing will reflect this.

Previous work in this field divides broadly into research focussed directly on segmenting pop songs into instrumental, verse and chorus sections [3, 4, 2] and more perceptually-motivated approaches which can lead to a more finely-grained segmentation, and are harder to evaluate [5, 6]. Although we also have a broad range of applications in mind for our segmentations (and use relatively detailed reference segmentations), a simple evaluation of segment boundary accuracy against a human reference segmentation is still one useful comparative measure. Over our test set, using clustering some 59% of groundtruth boundaries were found to within 2 seconds. Using the semi-supervised HSMM, 72% were accurate to within 2 seconds, and 82% to within 4 seconds. These figures compare with 55% of boundaries found within 2 seconds against a set of instrumental-verse-chorus groundtruths in [4]. For comparison with our own previous work, Figure 5 gives an idea of how well groundtruth segments are reproduced in machine segmentations, using measures of boundaries missed m and segments fragmented f developed in [16].

7. CONCLUSIONS

The methods described in sections 3 and 4 show how the underlying base phrase length of a piece of music can be estimated directly from audio data and used in a formal model of musical structure to produce automatic segmentations that correspond closely to human judgement in many cases. A semi-supervised method requiring only two ‘click and drag’ operations from the user produces segmentations in which over 80% of the reference boundaries are found within a 4 second threshold, and 72% within 2 seconds.

Future work includes experimenting over a larger test set to produce segmentations that are optimal for various applications, and extending the HSMM to handle continuous data

so that it can be trained directly on audio features.

REFERENCES

- [1] Jonathan Foote, “Visualizing music and audio using self-similarity,” in *ACM Multimedia (1)*, 1999, pp. 77–80.
- [2] Masataka Goto, “A chorus-section detecting method for musical audio signals,” in *Proc. ICASSP*, 2003, vol. V, pp. 437–440.
- [3] N. Maddage, X. Changsheng, M. Kankanhalli, and X. Shao, “Content-based music structure analysis with applications to music semantics understanding,” in *6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, October 2004.
- [4] L. Lu, M. Wang, and H. Zhang, “Repeating pattern discovery and structure analysis from acoustic music data,” in *6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, October 2004.
- [5] Jean-Julien Aucouturier and Mark Sandler, “Segmentation of musical signals using hidden markov models,” in *Proc. AES 110th Convention*, 2001.
- [6] Geoffroy Peeters, Amaury La Burthe, and Xavier Rodet, “Toward automatic music audio summary generation from signal analysis,” in *International Symposium on Music Information Retrieval*, 2002.
- [7] Mark Levy, Mark Sandler, and Michael Casey, “Extraction of high-level musical structure from audio data and its application to thumbnail generation,” in *Proc. ICASSP*, 2006.
- [8] Y. Boykov and M.-P. Jolly, “Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images,” in *International Conference on Computer Vision*, 2001, vol. I, pp. 105–112.
- [9] Lawrence R. Rabiner, “A tutorial on hidden markov models and selection applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [10] M. E. P. Davies and M. D. Plumbley, “Beat tracking with a two state model,” in *Proc. ICASSP*, 2005.
- [11] E. D. Scheirer, “Tempo and beat analysis of acoustic musical signals,” *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 588–601, Jan 1998.
- [12] J. Foote and S. Uchihashi, “The beat spectrum: A new approach to rhythm analysis,” in *Proc. ICME*, 2001.
- [13] Juan P. Bello and Jeremy Pickens, “A robust mid-level representation for harmonic content in music signals,” in *Proc. ISMIR*, 2005, pp. 304–311.
- [14] M. Ostendorf, V. Digalakis, and O. Kimball, “From HMM’s to segment models: A unified view of stochastic modeling for speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996.
- [15] Kevin P. Murphy, “Hidden semi-Markov models (HSMMs),” www.ai.mit.edu/~murphyk, 2002.
- [16] Samer Abdallah, Katy Noland, Mark Sandler, Michael Casey, and Christophe Rhodes, “Theory and evaluation of a bayesian music structure extractor,” in *Proc. ISMIR*, 2005.