A Robust Mid-level Representation for Harmonic Content in Music Signals

Juan P. Bello and Jeremy Pickens
Centre for Digital Music
Queen Mary, University of London
London E1 4NS, UK
juan.bello-correa@elec.qmul.ac.uk

ABSTRACT

When considering the problem of audio-to-audio matching, determining musical similarity using low-level features such as Fourier transforms and MFCCs is an extremely difficult task, as there is little semantic information available. Full semantic transcription of audio is an unreliable and imperfect task in the best case, an unsolved problem in the worst. To this end we propose a robust mid-level representation that incorporates both harmonic and rhythmic information, without attempting full transcription. We describe a process for creating this representation automatically, directly from multi-timbral and polyphonic music signals, with an emphasis on popular music. We also offer various evaluations of our techniques. Moreso than most approaches working from raw audio, we incorporate musical knowledge into our assumptions, our models, and our processes. Our hope is that by utilizing this notion of a musically-motivated mid-level representation we may help bridge the gap between symbolic and audio research.

Keywords: Harmonic description, segmentation, music similarity

1 Introduction

Mid-level representations of music are measures that can be computed directly from audio signals using a combination of signal processing, machine learning and musical knowledge. They seek to emphasize the musical attributes of audio signals (e.g. chords, rhythm, instrumentation), attaining higher levels of semantic complexity than low-level features (e.g. spectral coefficients, MFCC, etc), but without being bounded by the constraints imposed by the rules of music notation. Their appeal resides in their ability to provide a musically-meaningful description of audio signals that can be used for music similarity applications, such as retrieval, segmentation, classification and browsing in musical collections.

Previous attempts to model music from complex audio signals concentrate mostly on the attributes of timbre and rhythm (Aucouturier and Pachet, 2002; Yang, 2002). These methods are usually limited by the simplicity of their selected feature set, which can be often regarded as low-level. Dixon et al. (2004) demonstrated that it is possible to successfully characterize music according to rhythm by adding higher-level descriptors to a low-level feature set. These descriptors are more readily available for rhythm than for harmony as the state-of-the-art in beat, meter tracking and tempo estimation has had more success than similar efforts on chord and melody estimation.

Pickens et al. (2002) showed success at identifying harmonic similarities between a polyphonic audio query and symbolic polyphonic scores. The approach relied on automatic transcription, a process which is partially effective within a highly constrained subset of musical recordings (e.g. mono-timbral, no drums or vocals, small polyphonies). To effectively retrieve despite transcription errors, all symbolic data was converted to harmonic distributions and similarity was measured by computing the distance between two distributions over the same event space. This is an inefficient process that goes to the unnecessary step of transcription before the construction of an abstract representation of the harmony of the piece.

In this paper we propose a method for semantically describing harmonic content directly from music signals. Our goal is not to do a formal harmonic analysis but to produce a robust and consistent harmonic description useful for similarity-based applications. We do this without attempting to estimate the pitch of notes in the mixture. By avoiding the transcription step, we also avoid its constraints, allowing us to operate on a wide variety of music. The approach combines a chroma-based representation and a hidden Markov model (HMM) initialized with musical knowledge and partially trained on the signal data. The output, which is a function of beats (tactus) instead of time, represents the sequence of major and minor triads that describe the harmonic character of the input signal.

The remainder of this paper is organized as follows: Section 2 reviews previous work on this area; Section 3 gives details about the construction of the feature vector; Section 4 explains the used model and justifies our ini-

tialization and training choices; Section 5 evaluates the representation against a database of annotated pop music recordings; Section 6 discusses the application of our representation to long-term segmentation; and finally, Section 7 presents our conclusions and directions for future work.

2 Background

We are by no means the first to use either chroma-based representations or HMMs for automatically estimating chords, harmony or structure from audio recordings. Previous systems (Gomez and Herrera, 2004; Pauws, 2004) correlate chromagrams[1], to be explained in 3.1, with cognition-inspired models of key profiles (Krumhansl, 1990) to estimate the overall key of music signals. Similarly Harte and Sandler (2005) correlate tuned chromagrams with simple chord templates for the frame-by-frame estimation of chords in complex signals. While differing in their goals, these studies identified the lack of contextual information about chord/key progressions as a weakness of their approaches, as at the level of analysis frames there are a number of factors (e.g. transients, arpeggios, ornamentations) that can negatively affect the local estimation.

In their research on audio thumbnailing, Bartsch and Wakefield (2001) found that the structure of a piece, as seen by calculating a similarity matrix, is more salient when using beat-synchronous analysis of chromas. Longer analysis frames help to overcome the noise introduced by transients and short ornamentations. However, this solution still does not make use of the fact that in a harmonic progression certain transitions are more likely to occur than others.

An alternative way of embedding the idea of harmonic progression into the estimation is by using HMMs. The work by Raphael and Stoddard (2003) is a good example of successfully using HMMs for harmonic analysis; although their analysis is done from MIDI data, they do adopt beat-synchronous observation vectors.

Perhaps the approach which is most similar to ours is that proposed by Sheh and Ellis (2003) for chord estimation. In this approach an HMM is used on Pitch Class Profile features (PCP) estimated from audio. Both the models for chords (147 of them) and for chord transitions, are learned from random initializations using the expectation maximization (EM) algorithm. Importantly, this approach differs from ours on that no musical knowledge is explicitly encoded into the model, something that, as will be demonstrated in future sections, has a notable impact on the robustness of the estimation. Also, our choice of feature set and use of a beat-synchronous analysis frame minimizes the effect of local variations. Finally, our proposal differs in scope, we are not trying to achieve chord transcription but to generate a robust harmonic blueprint from audio, and to this end we limit our chord lexicon to the 24 major and minor triads, a symbolic alphabet that we consider to be sufficient for similarity-based applications.

[1] also referred to as Harmonic Pitch Class Profiles: HPCP

3 Features

The first stage of our analysis is the calculation of a sequence of suitable feature vectors. The process can be divided into four main steps: 36-bin chromagram calculation, chromagram tuning, beat-synchronous (tactus) segmentation and 12-bin chromagram reduction.

3.1 Chromagram calculation

A standard approach to modeling pitch perception is as a function of two attributes: *height* and *chroma*. Height relates to the perceived pitch increase that occurs as the frequency of a sound increases. Chroma, on the other hand, relates to the perceived circularity of pitched sounds from one octave to the other. The musical intuitiveness of the chroma makes it an ideal feature representation for note events in music signals. A temporal sequence of chromas results in a time-frequency representation of the signal known as chromagram.

In this paper we use a common method for chromagram generation known as the constant Q transform (Brown, 1991). It is a spectral analysis where frequency-domain channels are not linearly spaced, as in DFT-based analysis, but logarithmically spaced, thus closely resembling the frequency resolution of the human ear. The constant Q transform X_{cq} of a temporal signal $x(m)$ can be calculated as:

$$X_{cq}(k) = \sum_{n=0}^{N(k)-1} w(n,k)x(n)e^{-j2\pi f_k n} \quad (1)$$

where both, the analysis window $w(k)$ and its length $N(k)$, are functions of the bin position k. The center frequency f_k of the k^{th} bin is defined according to the frequencies of the equal-tempered scale such that:

$$f_k = 2^{k/\beta} f_{min} \quad (2)$$

where β is the number of bins per octave, thus defining the resolution of the analysis, and f_{min} defines the starting point of the analysis in frequency. From the constant Q spectrum X_{cq}, the chroma for a given frame can then be calculated as:

$$Chroma(b) = \sum_{m=0}^{M} |X_{cq}(b + m\beta)| \quad (3)$$

where $b \in [1, \beta]$ is the chroma bin number, and M is the total number of octaves in the constant Q spectrum. In this paper, the signal is downsampled to 11025Hz, $\beta = 36$ and analysis is performed between $f_{min} = 98$Hz and $f_{max} = 5250$Hz. The resulting window length and hop size are 8192 and 1024 samples respectively.

3.2 Chromagram tuning

Real-world recordings are often not perfectly tuned, and slight differences between the tuning of a piece and the expected position of energy peaks in the chroma representation can have an important influence on the estimation of chords.

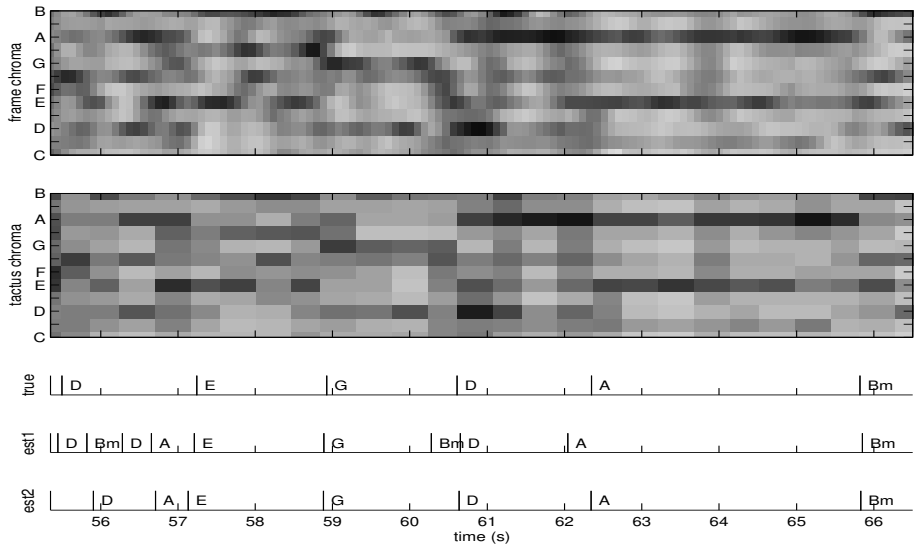


Figure 1: Frame and tactus-based feature vectors for *Eight days a week* by The Beatles. At the bottom the estimated chord labels can be observed: "true" corresponds to the ground-truth chord annotation, "est1" corresponds to the chord labels estimated using frame-based features, and "est2" corresponds to the chords estimated using tactus-based features.

The 36-bin per octave resolution is intended to clearly map spectral components to a particular semitone regardless of the tuning of the recording. Each note in the octave is mapped by 3 bins in the chroma, such that bias towards a particular bin (i.e. sharpening or flattening of notes in the recording) can be spotted and corrected. To do this we use a simpler version of the tuning algorithm proposed by Harte and Sandler (2005). The algorithm starts by picking all peaks in the chromagram. Resulting peak positions are quadratically interpolated and mapped to the $[1.5, 3.5]$ range. A histogram is generated with this data, such that skewness in the distribution is indicative of a particular tuning. A corrective factor is calculated from the distribution and applied to the chromagram by means of a circular shift. Finally, the tuned chromagram is low-pass filtered to eliminate sharp edges.

3.3 Beat-synchronous segmentation

As mentioned before, beat-synchronous analysis of the signal helps to overcome the problems caused by transient components in the sound, e.g. drums and guitar strumming, and short ornamentations, often introduced by vocals. Both these cases are quite common in pop music recordings, hence the relevance of this processing step. Furthermore, harmonic changes often occur at a longer time span than that defined by the constant Q analysis, thus the default temporal resolution results unnecessary and often detrimental.

In our approach we use the beat tracking algorithm proposed by Davies and Plumbley (2005). This method has proven successful for a wide variety of signals. Using beat-synchronous segments has the added advantage that the resulting representation is a function of beat, or "tactus", rather than time. These facilitates comparison with songs in different tempos.

3.4 Observation Vectors

Finally, the chromagram is averaged within beat segments and further reduced from 36 to 12 bins by simply summing within semitones. A piece of music is thus represented as a sequence of these 12 dimensional vectors.

4 Chord Labeling

Let us turn our attention to the chord labeling of the chroma sequence. Recall, however, that our goal is not true harmonic analysis, but a mid-level representation which we believe will be useful for music similarity and music retrieval tasks. For this we apply the HMM framework (Rabiner, 1989). As mentioned in section 2, we are not the first to use this framework, but we utilize it in a relatively new way, based largely on music theoretic considerations.

4.1 Chord lexicon

The first step in labeling the observations in a data stream is to establish the lexicon of labels that will be used. We define a *lexical chord* as a pitch template. Of the 12 octave-equivalent (mod 12) pitches in the Western canon, we select some n-sized subset of those, call the subset a *chord*, give that chord a name, and add it to the lexicon. Not all possible chords belong in a lexicon and we must therefore restrict ourselves to a musically-sensible subset. The chord lexicon used in this work is the set of 24 major and minor triads, one each for all 12 members of the chromatic scale: C Major, c minor, C♯ Major, c♯ minor ... B♭ Major, b♭ minor, B Major, b minor. Assuming octave-invariance, the three members of a major triad have the relative semitone values n, $n + 4$ and $n + 7$; those of a minor triad n, $n + 3$ and $n + 7$. No distinction is made between enharmonic equivalents (C♯/D♭, A♯/B♭, etc.).

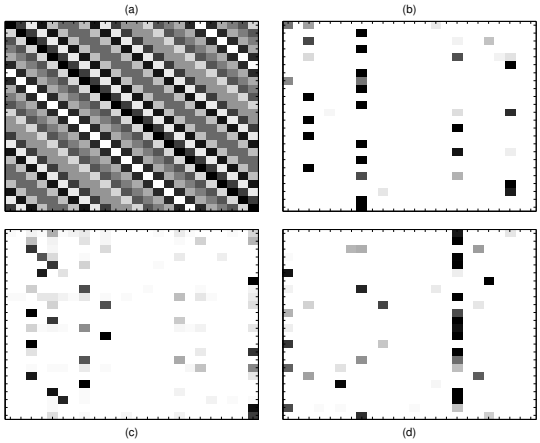


Figure 2: State-transition distribution A: (a) initialization of A using the circle of fifths, (b) trained on *Another Crossroads* (M. Chapman), (c) trained on *Eight days a week* (The Beatles), and (d) trained on *Love me do* (The Beatles). All axes represent the 24 lexical chords (C→B then c→b)

We have chosen a rather narrow space of chords. We did not include dyads nor other more complex chords such as augmented, diminished, 7^{th} or 9^{th} chords. Our intuition is that by including too many chords, both complex and simple, we run the risk of "overfitting" our models to a particular piece of music. As a quick thought experiment, imagine if the set of chords were simply the entire $\sum_{n=1..12} \binom{12}{n} = 2^{12} - 1$ possible combinations of 12 notes. Then the set of chord labels would be equivalent to the set of 12-bin chroma and one would not gain any insight into the harmonic "substance" of a piece, as each observation would likely be labeled with itself. This is an extreme example but it illustrates the intuition that the richer the lexical chord set becomes, the more our feature selection algorithms might overfit one piece of music and not be useful for the task of determining music similarity.

While it is clear that the harmony of only the crudest music can be reduced to a mere succession of major and minor triads, as this choice of lexicon might be thought to assume, we believe that this is a sound basis for a probabilistic approach to labeling. In other words, the lexicon is a robust mid-level representation of the salient harmonic characteristics of many types of music, notably popular music.

4.2 HMM initialization

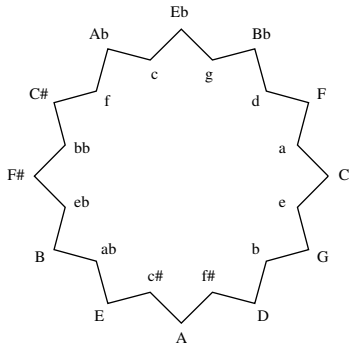
In this paper we are not going to cover the basics of hidden Markov modeling. This is far better covered in works such as (Rabiner, 1989) and even by previous music HMM papers cited above. Instead, we begin by describing the initialization procedure for the model. As labeled training data is difficult to come by, we forgo supervised learning and instead use the unsupervised mechanics of HMMs for parameter estimation. However, with unsupervised training it is crucial that one start the model off in a reasonable state, so that the patterns it learns correspond with the states over which one is trying to do inference.

4.2.1 Initial state distribution [π]

Our estimate of π is $\frac{1}{24}$ for each of the 24 states in the model. We have no reason to prefer, a priori, any state above any other.

4.2.2 State transition matrix [A]

Prior to observing an actual piece of music we also do not know what states are more likely to follow other states. However, this is where a bit of musical knowledge is useful. In a song, we might not yet know whether a C major triad is more often followed by a B$^\flat$ major or a D major. But it is reasonable to assume that both hypotheses are more likely than an F$^\sharp$ major. Most music tends not to make large, quick harmonic shifts. One might gradually wander from the C to the F$^\sharp$, but not immediately. We use this notion to initialize our state transition matrix.



The figure above is a doubly-nested circle of fifths, with the minor triads (lower case) staggered throughout the major triads (upper case). Triads closer to each other on the circle are more consonant, and thus receive higher initial transition probability mass than triads further away. Specifically, the transition C→C is given a probability $\frac{12+\epsilon}{144+24\epsilon}$, where ϵ is a small smoothing constant, C→e = $\frac{11+\epsilon}{144+24\epsilon}$ and then clockwise in a decreasing manner, until C→F$^\sharp$ = $\frac{0+\epsilon}{144+24\epsilon}$. At that point, the probabilities begin increasing again, with C→b$^\flat$ = $\frac{1+\epsilon}{144+24\epsilon}$ and C→a = $\frac{11+\epsilon}{144+24\epsilon}$.

The entire 24×24 transition matrix, as seen in Figure 2(a), is constructed in a similar manner for every state, with a state's transition to itself receiving the highest initial probability estimate, and the remaining transitions receiving probability mass relative to their distance around the 24-element circle above.

4.2.3 Observation (output) distribution [B]

Each state in the model generates, with some probability, an observation vector. We assume a continuous observation distribution function modeled using a single multivariate Gaussian for each state, each with mean vector μ and covariance matrix Σ.

Sheh and Ellis (2003) use random initialization of μ and a Σ covariance matrix with all off diagonal elements set to 0, reflecting their assumption of completely uncorrelated features. We wish to avoid this assumption. One of the main purposes of this paper is to argue that musical

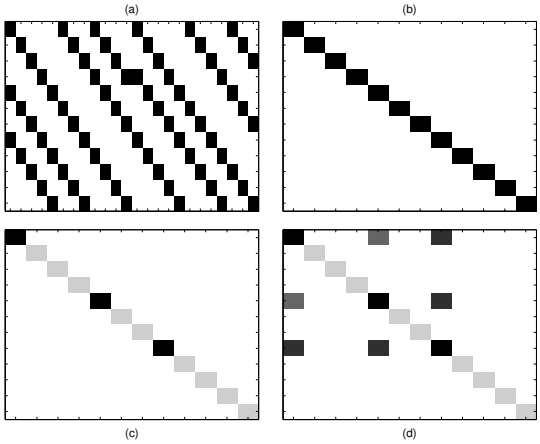


Figure 3: Initializations for μ and Σ. Top-left is μ for all states (a). Then for a C major chord: diag-only (b), weighted-diag (c), and off-diag(d) initializations of Σ. The x axis in (a) corresponds to the 24 lexical chords. All other axes refer to the 12 notes in the chroma circle.

knowledge needs to play an important role in music information retrieval tasks. Thus if we are using triads as our hidden state labels, μ and Σ should reflect this fact.

Let us take for example the C major triad state. Instead of initializing μ randomly, we initialize it to 1.0 in the C, E, and G dimensions, and 0.0 elsewhere. This reflects the fact that the triad is grounded in those dimensions. Initializations of μ for all states can be seen in Fig. 3(a).

The covariance matrix should also reflect our musical knowledge. Covariance is a measure of the extent to which two variables move up or down together. Thus, for a C major triad, it is reasonable that pitches which comprise the triad are more correlated than pitches which do not belong to the triad. Naturally, the pitches C, E, and G are strongly correlated with themselves. Furthermore, these pitches are also strongly correlated with each other. We symmetrically use the knowledge, gained both from music theory as well as empirical evidence (Krumhansl, 1990), that the dominant is more important than the mediant in characterizing the root of the triad. We set the covariance of the tonic with the dominant to 0.8, the mediant with the dominant to 0.8, and the tonic with the mediant to 0.6. The actual values are heuristic, but the principle we use to set them is not.

The remainder of covariances in the matrix are set to zero, reflecting the fact that from the perspective of a C major triad there is little useful correlation between, say, an F$^\sharp$ and an A$^\sharp$. The non-triad member diagonals are set to 0.2 both to indicate that non-triad pitches need not be as strongly self-correlated, as well as to insure that the matrix is positive, semi-definite. Figure 3(d) shows the covariance matrix used for the C major triad state.

The covariance for C minor is constructed almost exactly the same way, but with the mediant on D$^\sharp$/E$^\flat$ rather than on E, as would be expected. The remainder of the matrices for all the states are constructed by circularly shifting the major/minor matrix by the appropriate number of semitones.

4.3 HMM Training

A key difference between our approach and previous systems is our use of musical knowledge for model initialization. There are two important pieces of information that we are providing the system: a template for every chord in the lexicon, as given by μ and Σ, and cognitive-based knowledge about likely chord progressions, as given by the state transition probability matrix A.

It is relatively safe to say that the template for a chord is almost universal, e.g. a C major triad is always supposed to have the notes C, E and G. If we were to change our chord models from song to song we cannot longer assume that a certain state will always map to the same major or minor triad. Our labels would not have universal value. Furthermore, it is very unlikely that all chords in our lexicon will be present in any given song (or on any reasonably sized training set), and in training, this situation gives rise to the undesirable effect of different instances of existing chords being mapped to different (available) states, usually those that are initialized closely, e.g. relative and parallel minors and majors.

On the other hand, chord progressions are not universal, changing from song to song depending on style, composer, etc. Our initial state transition probability matrix provides a reference, founded in music cognition and theory, on how certain chord transitions are likely to occur in most western tonal music, especially pop music. We believe that this knowledge captures the *a-priori* harmonic intuition of a human listener. However, we want to provide the system with the adaptability to develop models for the particular chord progression of a given piece (see Fig. 2), much as people do when exposed to a piece of music they have never heard before.

We therefore propose selectively training our model using the standard expectation maximization (EM) algorithm for HMM parameter estimation (Rabiner, 1989), such that we disallow adjustment of $B = \{\mu, \Sigma\}$, while π and A are updated as normal. We believe this kind of selective training to provide a good trade-off between the need for a stable reference for chords, and a flexible, yet principled, modeling of chord sequences.

4.4 Chord Labeling (Inference)

Once we have both a trained model and an observation sequence, we can apply standard inference techniques (Rabiner, 1989) to label the observations with chords from our lexicon. The idea is that there are many sequences of hidden states that could have been responsible for generating the chroma vector observation sequence.

The goal is to find that sequence that maximizes the likelihood of the data without having to enumerate the exponentially many (24^n, for a sequence of length n, in our model) number of sequences. To this end a dynamic programming algorithm known as Viterbi is used (Forney, 1973). This algorithm is well covered in the literature and we do not add any details here.

Parameters						TP %		
Feature scope	π	A	B μ	B Σ	Training	CD1	CD2	TOTAL
tactus	$\frac{1}{24}$	random	template	diag-only	π, A, B	22.88	29.83	26.36
tactus	$\frac{1}{24}$	random	template	weighted-diag	π, A, B	34.14	36.24	35.19
tactus	$\frac{1}{24}$	random	template	off-diag	π, A, B	33.13	44.36	38.74
tactus	$\frac{1}{24}$	circle of 5^{ths}	template	off-diag	π, A, B	38.09	47.75	42.93
frame	$\frac{1}{24}$	circle of 5^{ths}	template	off-diag	π, A	58.96	74.78	66.87
tactus	$\frac{1}{24}$	circle of 5^{ths}	template	off-diag	π, A	**68.55**	**81.54**	**75.04**

Figure 4: Results for various model parameters

5 Evaluation and Analysis

In summary, our system, for a single piece of music, is:

1. Compute the 36-bin chromagram for the music piece.
2. Tune the chromagrams (globally) to remove slight sharpness or flatness and avoid energy leaking from one pitch class into another
3. Segment the signal frames into tactus-sized windows, average the chroma within each window, and finally reduce each chroma from 36 to 12 bins by summing all three bins for each pitch class
4. Selectively train the HMM to get a sense of the harmonic movement of the piece
5. Decode the HMM (do inference) to give a good mid-level harmonic characterization of the piece

Despite our stated goal of harmonic description rather than analysis, we found that it is still useful to attempt quantitative evaluation of the goodness of our representation by comparing the generated labels to an annotated collection of music. We use the test set proposed and annotated by Harte and Sandler (2005). It contains 28 recordings (mono, $f_s = 44.1kHz$) from the Beatles albums: *Please Please Me* (CD1) and *Beatles for Sale* (CD2). Note that all recordings are polyphonic and multi-instrumental containing drums and (multi-part) vocals.

The majority of chords (89.51%) in the manually labeled test set belong to our proposed lexicon of major and minor triads. However, the set also contains more complex chords such as major and minor 6^{ths}, 7^{ths} and 9^{ths}. For simplicity, we map any complex chord to its root triad, so for example C#m7sus4 becomes simply C#m. If anything, this mapping has the effect of overly penalizing our results, as chords of 4 or more notes could contain triads other than its root triad, e.g. Fm7 (F, G#, C, D#) has 100% overlap with G# (G#, C, D#) and Fm (F, G#, C). Comparisons are made on a frame-by-frame basis, such that a true positive is defined as a one-to-one match between estimation and annotation.

To quantitatively demonstrate some of the hypotheses put forward on this paper, we evaluate a series of incremental improvements to our approach. Figure 4 shows the model parameters for each experiment and its corresponding results for the test set (in percentage of true positives). Results are presented per CD and in total. The considered model parameters are:

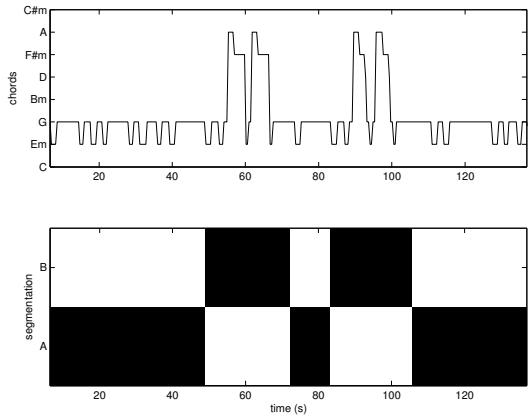
- Feature scope: Whether it is a frame-by-frame (time-based) or a beat-synchronous (tactus-based) chroma feature set.
- Initialization of A: Whether it is randomly initialized or initialized according to the circle of fifths.
- Initialization of B: Whether Σ is initialized as a diagonal matrix with elements equal to 1.0 (diag-only, Fig. 3(b)), whether it is the diagonal with weighted triad elements, as in Fig. 3(c), and off-diagonal elements set to 0.0 (weighted-diag), or whether it includes the mediant and dominant off-diagonal elements, i.e. the Fig. 3(d) matrix (off-diag).
- Training: Whether π, A and B are updated in the expectation-maximization step of HMM training or whether B is left fixed and only π and A are adjusted.

Results in Figure 4 clearly support the choices made in this paper. The first three rows show how initializing Σ with a weighted diagonal and off-diagonal elements outperforms diagonal-only initializations. This supports the view that the feature set is highly correlated along the dimensions of the elements of a chord. The weighted diagonal in itself introduces a noticeable amount of improvement over the unitary diagonal, a further indication of the strong correlation between the tonic, mediant and dominant of a chord.

The initialization of A using the circle of fifths brings about more than 10% relative improvement when compared to the random initialization. This shows how the use of musical knowledge is crucial. .

From the analysis of the last two rows in Figure 4 two more observations can be made. The first is that selective training introduces considerable benefits into our approach. The huge accuracy increase (from 42.93% to 75.04%) supports the view that the knowledge about chords encoded in B is universal, and as such it should not be modified during training. This accuracy increase occurs for every song, showing the generality of this assertion.

The second observation is that the use of a tactus-based feature set clearly outperforms the frame-by-frame estimation. This point is further illustrated by the chord estimation example in Fig. 1, where the frame-by-frame estimation is subject to small variations due to phrasing or ornamentation (as shown by the spurious estimations of B minor chords between 56 and 60.5 seconds), while the



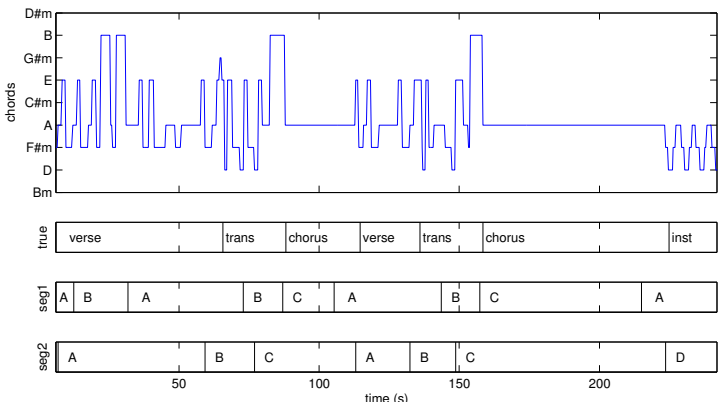


Figure 5: (left) *Love me do* by The Beatles: estimated chord sequence (top) and estimated segments, showing the long-term structure "ABABA"

Figure 6: (right) Estimated chord sequence (top) and long-term segment boundaries from *Wonderwall* by Oasis: "true" refers to ground-truth annotation, "seg1" to segments obtained using our raw chord label sequence and "seg2" to segments obtained by collapsing our chord label sequence into a simple chord sequence by removing contiguous duplicates

tactus-based estimation shows more stability and, therefore, accuracy when compared to the ground-truth annotation. Furthermore, chord changes are more likely to occur on the beat, thus chords detected using the tactus-based feature set tend also to be better localized.

Our results compare favorably to those reported by Sheh and Ellis (2003) and Harte and Sandler (2005). The maximum true positives rate in the collection is 90.86% for *Eight days a week*. Conversely, the worst estimation is for *Love me do*, with only 49.27% of chords correctly identified. For the latter case almost all errors are due to relative minor confusions: C being confused with E minor consistently through the song. As we will see in the next section the consistency of the representation, even when wrong, can be useful for certain applications.

6 Application to Segmentation

To show the applicability of our chord labels to long-term segmentation of songs we use a histogram clustering algorithm developed by Abdallah et al. (2005). The algorithm calculates a sequence of unlabeled states (e.g. A and B) that represent the long-term sections of a song (e.g. chorus, verse, bridge, etc) from a sequence of histograms computed from our labeled sequence. It consists of a phase of simulated annealing to learn the state transition probability matrix (Puzicha et al., 1999) and a second phase of combined annealing and Gibbs sampling to compute the posterior probabilities of segments belonging to given states, and thus the sequence of states. See (Robert and Casella, 1999) for an introduction.

The top plot of Fig. 5 shows the resulting chord labeling for *Love me do*, the song on which our labeling performed the worst. The bottom plot shows, for each time step, the marginal posterior probabilities obtained from the segmentation algorithm, such that white indicates zero probability and black indicates a probability of 1. From both these plots we can clearly see the simple structure of the song, of the form "ABABA". This demonstrates how, even when imperfect, our representation is consistent, allowing for successful clustering of its symbols. To our knowledge, this success is the first example of long-term segmentation using a mid-level harmonic feature set.

Figure 6 shows segmentation results for a more complicated structure, that of "Wonderwall" by Oasis. The top plot shows our calculated sequence of chord labels ("chords"). The next line ("true") shows the manually annotated segments of the song. The middle line depicts the automatically segmented sections using our chord labels ("seg1"). Finally, the bottom line ("seg2") shows the automatically segmented sections obtained after first collapsing our tactus-based chord labels (e.g. CCG-GFFFEAAAA) into a simple sequence of chords (e.g. CGFEA) by removing contiguous duplicates.

As can be seen in "seg1", there are some problems with the segmentation: the verse is segmented as to include parts of the transition, the chorus section and a final instrumental *Coda*, creating some confusion between them, and thus resulting in errors. On the other hand, segmentation on the collapsed chord sequence is more accurate, both in terms of temporal localization and segregation between states. We suggest that this is because the resulting chord groupings can be thought of as equivalent to musical phrases. Indeed, some informal testing seems to support the idea that when the number of segmentation states is increased and the length of our histograms is reduced, we start to pick up segments that are related to sections at a shorter temporal scale (e.g. phrases). While a proper study on segmentation is beyond the scope of this paper, we suggest that this increased granularity is potentially a major asset of harmonic-based segmentation, in opposition to timbre-based segmentation, where short-term structures are not necessarily indicative of musical gestures.

7 Conclusion

The main contribution of this work is the creation of an effective mid-level representation for music audio signals. We have shown that by considering the inherent musicality of audio signals one achieves results far greater than raw signal processing and machine learning techniques alone (Figure 4). Our hope is that these ideas and their results will encourage those in the field working on raw audio to build more musicality into their techniques. At the same time, we hope it also encourages those working on the symbolic side of music retrieval to aide in the creation of additional musically sensible mid-level representations without undue concern over whether such representations strictly adhere to formal music theory guidelines.

In support of this goal, we have integrated into a single framework a number of state-of-the-art music processing algorithms. Specifically, we build our algorithms upon a musical foundation in the following ways: (1) The audio signal is segmented into tactus windows rather than time-based frames. (2) Pitch chroma are tuned. (3) A lexicon of 24 triads is used, which is neither too specific or too general, in an attempt to *describe* harmonic movement in a piece rather than doing a formal harmonic analysis. (4) Initialization of the machine learning (HMM) algorithm is done in a manner that respects the dependency between tonic, mediant, and dominant pitches in a triad, as well as the consonance between neighboring triads in a sequence. Finally, (5) the machine learning algorithm itself is modified with an eye toward musicality; updates to model parameters are done so as to maintain the relationship between pitches in a chord, but be amenable to changing chord transitions in a sequence.

In the future we are planning a series of audio-to-audio music retrieval experiments to further show the validity of our approach. We will also continue to develop and integrate techniques that emphasize the musical nature of the underlying source. We believe that this mindset is vital to continuing development in the field.

8 Acknowledgments

The authors wish to thank Chris Harte, Matthew Davies, Katy Noland and Samer Abdallah for making their code available. We also wish to thank Geraint Wiggins and Christopher Raphael for their insights regarding the training and music-based initializations of HMMs. This work was partially funded by the European Commission through the SIMAC project IST-FP6-507142.

References

S. Abdallah, K. Noland, M. Sandler, M. Casey, and C. Roads. Theory and evaluation of a Bayesian music structure extractor. In *Proceedings of the 6th ISMIR Conference, London, UK*, 2005.

J.-J. Aucouturier and F. Pachet. Music similarity measures: What's the use? In *Proceedings of the 3rd ISMIR, Paris, France.*, pages 157–163, 2002.

M. A. Bartsch and G. H. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, NY,*, pages 15–18, 2001.

J. Brown. Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America,*, 89(1): 425–434, 1991.

M. E. P. Davies and M. D. Plumbley. Beat tracking with a two state model. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, Penn., USA*, pages 241–244, 2005.

S. Dixon, F. Gouyon, and Gerhard Widmer. Towards characterisation of music vian rhythmic patterns. In *Proceedings of the 5th ISMIR, Barcelona, Spain.*, pages 509–516, 2004.

G. D. Forney. The viterbi algorithm. *Proc. IEEE*, 61:268–278, 1973.

E. Gomez and P. Herrera. Estimating the tonality of polyphonic audio files: Cognitive versus machine learning modelling strategies. In *Proceedings of the 5th ISMIR, Barcelona, Spain.*, pages 92–95, 2004.

C. A. Harte and M. B. Sandler. Automatic chord identification using a quantised chromagram. In *Proceedings of the 118th Convention of the Audio Engineering Society*, Barcelona, Spain, May 28-31 2005.

C. L. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, New York, 1990.

S. Pauws. Musical key extraction from audio. In *Proceedings of the 5th ISMIR, Barcelona, Spain.*, pages 96–99, 2004.

J. Pickens, J. P. Bello, G. Monti, T. Crawford, M. Dovey, M. Sandler, and D. Byrd. Polyphonic score retrieval using polyphonic audio queries: A harmonic modeling approach. In *Proceedings of the 3rd ISMIR*, pages 140–149, Paris, France, October 2002.

J. Puzicha, J. M. Buhmann, and T. Hofmann. Histogram clustering for unsupervised image segmentation. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Ft. Collins, CO, USA*, pages 2602–2608, 1999.

L. R. Rabiner. A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

C. Raphael and J. Stoddard. Harmonic analysis with probabilistic graphical models. In *Proceedings of the 4th ISMIR*, pages 177–181, Baltimore, Maryland, October 2003.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 1999.

A. Sheh and D. P. W. Ellis. Chord segmentation and recognition using em-trained hidden markov models. In *Proceedings of the 4th ISMIR*, pages 183–189, Baltimore, Maryland, October 2003.

C. Yang. MACSIS: A scalable acoustic index for content-based music retrieval. In *Proceedings of the 3rd ISMIR, Paris, France.*, pages 53–62, 2002.