Bayesian Model Selection for Harmonic Labelling

Christophe Rhodes*, David Lewis, Daniel Müllensiefen
Department of Computing
Goldsmiths, University of London
SE14 6NW, United Kingdom

April 29, 2008

Abstract

We present a simple model based on Dirichlet distributions for pitch-class proportions within chords, motivated by the task of generating 'lead sheets' (sequences of chord labels) from symbolic musical data. Using this chord model, we demonstrate the use of Bayesian Model Selection to choose an appropriate span of musical time for labelling at all points in time throughout a song. We show how to infer parameters for our models from labelled ground-truth data, use these parameters to elicit details of the ground truth labelling procedure itself, and examine the performance of our system on a test corpus (giving 75% correct windowing decisions from optimal parameters). The performance characteristics of our system suggest that pitch class proportions alone do not capture all the information used in generating the ground-truth labels. We demonstrate that additional features can be seamlessly incorporated into our framework, and suggest particular features which would be likely to improve performance of our system for this task.

1 Introduction

This paper introduces a straightforward model for labelling chords based on pitch-class proportions within windows, and using this model not only to generate chord labels given a symbolic representation of a musical work, but also to infer the relevant level of temporal granularity for which a single label is justified.

The generation of these chord labels was initially motivated by the desire to perform automated musical analysis over a large database of high-quality MIDI transcriptions of musical performances, as part of a larger study investigating musical memory. While the MIDI transcriptions are of high-fidelity with respect to the performances they represent, they do not include any analytic annotations, such as song segmentation, principal melody indications, or significant rhythmic or harmonic motifs; all of these must be generated if desired, but it is not practical to do so manually over the collection of some 14,000 pop song transcriptions.

A time sequence of chord labels, as a compact representation of the harmony of the musical work, can not only be used as the basis for the detection of larger-scale harmonic features (such as cadences, clichés and formulae), but can also inform a structural segmentation of the music, since harmony

*c.rhodes@gold.ac.uk

is an indicator of structure in many popular music styles. Such segmentation is a necessary first step for other feature extraction tools – it is, for example, a prerequisite for the melody similarity algorithms presented in Müllensiefen and Frieler (2004).

A second use for these chord labels is the automatic generation of lead sheets. A lead sheet is a document "displaying the basic information necessary for performance and interpretation of a piece of popular music" (Tagg 2003b). The lead sheet usually gives the melody, lyrics and a sequence of short chord labels, usually aligned with the melody, allowing musicians to accompany the singer or main melody instrument without having a part written out for them.

An advantage of the model we present in this paper is that the overall framework is independent of the type of harmony scheme that it is used with: for example, it can be adapted to generate labels based on tertial or quartal harmonic classification (Tagg 2003a). Furthermore, a similar model selection stage can be used to choose which harmonic classification is most appropriate for a given work, a decision which can be guided by information not present in the observed musical data (such as a genre label) by incorporating that information into a prior probability model.

The rest of this paper is organized as follows: after a discussion of previous related work in section 2, we present our model for the dependency of pitch-class content on the prevailing chord, forming the core of our simple model, and discuss its use in window size selection in section 3. We discuss implementation of parameter inference and empirical results in section 4, and draw conclusions and suggest further work in section 5.

2 Previous Work

Most previous work on chord label assignment from symbolic data is implemented without an explicit model for chords: instead, preference rules, template matching and neural network approaches have been considered (Temperley 2001, Chapter 6 and references therein); an alternative approach involving knowledge representation and forward-chaining inference has also been applied to certain styles of music (Pachet 1991; Scholz et al. 2005). One attempt to use probabilistic reasoning to assign chord labels uses a Hidden Markov Model approach with unsupervised learning (Raphael and Stoddard 2004) of chord models; however, the authors note that they do not provide for a way of making decisions about the appropriate granularity for labelling: *i.e.* how to choose the time-windows for which to compute a chord label.

There has been substantial work in the symbolic domain on the related task of keyfinding. For instance, Krumhansl (1990, Chapter 4) presents a decision procedure based on Pearson correlation values of observed pitch-class profiles with profiles generated from probe-tone experiments. Another class of algorithms used for keyfinding is a geometric representation of keys and tones, attempting to capture the perceived distances between keys by embedding them in a suitable space (Chuan and Chew 2005). The profile-based model has been refined (Temperley 2001, Chapter 7) by making several modifications: altering details of the chord prototype profiles; dividing the piece into shorter segments; adjusting the pitch-class observation vector to indicate merely presence or absence of that pitch class within a segment, rather than the proportion of the segment's sounding tones, and thus avoiding any attempt at weighting pitches based on their salience; and imposing a change penalty for changing key label between successive segments.

There are existing explicit models for keys and pitch-class profiles: one such (Temperley 2004) is defined such that for each key, the presence or absence of an individual pitch class is a Bernoulli distribution (so that the pitch-class profile is the product of twelve independent Bernoulli distribu-

tions); in this model, there are also transition probabilities between successive chords. This model was further refined in Temperley (2007) by considering not just pitch classes but the interval between successive notes. These models are based on the notion of a fixed-size 'segment', which has two effects: first, the key models are not easily generalized to windows of different sizes, as the occurrence of a particular scale degree (*i.e.* pitch relative to a reference key) is not likely to be independent in successive segments; second, unless the segment length is close to the right level of granularity, a postprocessing stage will be necessary to smooth over fragmented labels.

There has been more work towards chord recognition in the audio domain, where the usual paradigm is to model the chord labels as the hidden states in a Hidden Markov Model generating the audio as observation vectors (Bello and Pickens 2005; Sheh and Ellis 2003). One problem in training these models is in the lack of ground truth, of music for which valid chord labels are known (by 'valid' here, we mean sufficient for the purposes for which automated chord labelling is intended, though of course these may vary between users); approaches have been made to generate ground truth automatically (Lee and Slaney 2006), but such automatic ground truth generation depends on a reliable method of generating labels from the symbolic data or from something that can be mapped trivially onto it; without such a reliable method, hand-annotated ground truth must be generated, as for example in Harte et al. (2005).

One feature of the method presented in this paper in contrast to most existing harmony or key identification techniques is that it has an explicit musically-motivated yet flexible model for observable content (*i.e.* pitch-class distributions) at its core, rather than performing some *ad-hoc* matching to empirical prototypes. This flexibility confers two modelling advantages: first, the parameters of the model can be interpreted as a reflection of musical knowledge (and adjusted, if necessary, in a principled way); second, if evidence for additional factors influencing chord labels surfaces, in general or perhaps for a specific style of music under consideration, these additional factors can be incorporated into the model framework without disruption.

3 Model

The repertoire of chords that we represent is triad-based (though an extension to include other bases is possible with some care over the dimensionality of the relevant spaces); motivated by their prevalence in western popular music, we aim to distinguish between major, minor, augmented, diminished and suspended (sus4 and sus9) triads with any of the twelve pitch classes as the root, and we will infer probability distributions over these chord labels given the musical content of a window. Of the six, it should be noted that augmented and diminished chords are much rarer in popular music, and that suspended chords, despite their names, are frequently treated in popular music as stable and not as needing to resolve, and so require categories of their own – *e.g.* in soul or country music where they form independent sonorities; see Tagg (2003a). We introduce the Dirichlet distribution on which our chord model is based, give our explicit model for the dependence of pitch-class proportions on the chord, and then explain how we can use this to perform selection of window size in a Bayesian manner.

3.1 Dirichlet distributions

The Dirichlet distribution is a model for proportions of entities within a whole. Its density function is

$$p(\mathbf{x}|\alpha) = \frac{1}{B(\alpha)} \prod_i x_i^{\alpha_i - 1} \tag{1}$$

with support on the simplex $\sum_i x_i = 1$. The normalizing constant $B(\alpha)$ is defined as

$$B(\alpha) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma\left(\sum_i \alpha_i\right)} \tag{2}$$

where Γ is the gamma function $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t$.

Note that for each individual component of the whole, represented by an individual random variable x_i, the corresponding α_i controls the dependence of the density (1) for small values of this component: if $\alpha_i > 1$, the probability density tends towards zero in the limit $x_i \to 0$; if $\alpha_i < 1$, the density increases without limit as $x_i \to 0$.

3.2 The Chord Model

Our introductory chord model is triad-based, in that for each chord we consider the tones making up the triad separately from the other, non-triadic tones. The proportion of a region made up of triad tones is modelled as a Beta distribution (a Dirichlet distribution with only two variables), and the triad tone proportion is then further divided into a Dirichlet distribution over the three tones in the triad.

Denoting the proportion of non-triadic tones as $\bar{t}$, and that of triadic tones as t, where the latter is made up of root r, middle m and upper u, we can write our chord model as for tone proportions given a chord label c as

$$p(rmut\bar{t}|c) = p(t\bar{t}|c)p(rmu|t\bar{t}c) \tag{3}$$

with support on the simplexes $t + \bar{t} = 1$, $r + m + u = 1$; each of the terms on the right-hand side is a Dirichlet distribution. We simplify the second term on the right-hand side by asserting that the division of the harmonic tones is independent of the amount of harmonic tones in a chord, so that $p(rmu|t\bar{t}c) = p(rmu|c)$. In principle, each chord model has two sets of independent Dirichlet parameters α; in practice we will consider many chords to be fundamentally similar, effectively tying those parameters.

This simple chord model does not allow for certain common harmonic labels, such as seventh chords or open fifths (as these are not triadic); we leave this extension for further work. Additionally, there is a possible confusion even in the absence of noise between the suspended chords, as the tones present in a sus4 chord are the same as those in a sus9 chord four scale degrees higher.

3.3 Bayesian Model Selection

We start with a set of possible models for explaining some data, where each individual model is in general parameterized by multiple parameters. Given this set of distinct models, and some observed data, we can make Bayesian decisions between models in an analogous fashion to selecting a particular set of parameters for a specific model; in general, we can generate probability distributions over models (given data) in a similar way to the straightforward Bayesian way of generating probability

distributions over the parameter values of a given model. For a full exposition of Bayesian Model Selection, see *e.g.* MacKay (2003, Chapter 28).

In the context of our problem, of chord labelling and window size selection, we choose a metrical region of a structural size: in our investigation for popular music, we choose this region to be one bar, the basic metrical unit in that style. The different models for explaining the musical content of that bar, from which we will aim to select the best, are different divisions of that bar into independently-labelled sections. For example, one possible division of the bar is that there is no segmentation at all; it is all one piece, with one chord label for the whole bar. Another possible division is that the bar is made up of two halves, with a chord label for each half bar. These divisions of the bar play the rôle of distinct models, each of which has Dirichlet parameters for each independently-labelled section of the bar. In our experiment described in section 4, the corpus under consideration only contains works in common time, with four quarter beats in each bar, and we consider all eight possible divisions of the bar that do not subdivide the quarter beat (*i.e.* , 1+1+1+1, 1+1+2, 1+2+1, 2+1+1, 2+2, 1+3, 3+1, 4).

The Bayesian Model Selection framework naturally incorporates the Occam factors in a quantitative manner: if there is evidence for two different chord labels, then the whole-bar model will not be a good fit to the data; if there is no evidence for two distinct chord labels, then there are many more different poor fits for a more fine-grained model than for the whole-bar model.

To be more precise, we can write the inference over models M given observed data D as

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)} \tag{4}$$

where

$$p(D|M) = \sum_c p(D|cM)p(c|M) \tag{5}$$

is the normalizing constant for the inference over chord labels c for a given model M. Note that there is an effective marginalization over chord labels for each model – when considering the evidence for a particular model, we add together contributions from all of its possible parameters, not simply the most likely. We can use the resulting probability distribution (4) to select the most appropriate window size for labelling.

The flexibility of this approach is evident in equation (5): the chord models $p(D|cM)$ can differ in parameter values or even in their identity between window sizes, and that the prior probabilities for their generation $p(c|M)$ can also be different for different models of the bar M.

4 Experiment

4.1 Parameter estimation

In order to test our chord model, (see equation 3), we must choose values for the α parameters of the Dirichlet distributions. We summarize the maximum-likelihood approach (from a labelled 'training set') below, noting also the form of the prior for the parameters in the conjugate family for the Dirichlet distribution); in addition, we performed a large search over the parameter space for the training set, attempting to maximize performance of our model at the labelling task with a binary loss function.

We can rewrite the Dirichlet density function (1) as $e^{-\sum_i [(1-\alpha_i)\log x_i] - \log B(\alpha)}$, demonstrating that it is in the exponential family, and that $\sum_i \log x_i$ is a sufficient statistic for this distribution; additionally, there is a conjugate prior for the parameters of the form

$$\pi(\alpha|\mathbf{A}^0, B^0) \propto e^{-\sum_i \left[(1-\alpha_i)A_i^0\right] - B^0 \log B(\alpha)} \tag{6}$$

with support $\alpha_i \in \mathbb{R}_0^+$.

Given N observations $\mathbf{x}^{(k)}$, the posterior density is given by $p(\alpha|\mathbf{x}^{(k)}) \propto p(\mathbf{x}^{(k)}|\alpha)\pi(\alpha)$, which is

$$e^{-\sum_i \left[(1-\alpha_i)\left(A_i^0 + \sum_k^N \log x_i^{(k)}\right)\right] - (B^0+N)\log B(\alpha)}; \tag{7}$$

that is, of the same form as the prior in equation (6), but with the hyperparameters $\mathbf{A}^0$ and B^0 replaced by $\mathbf{A} = \mathbf{A^0} + \sum_{\mathbf{k}} \log \mathbf{x^{(k)}}$ (with the logarithm operating componentwise) and $B = B^0 + N$. The likelihood is of the form of equation (7), with $\mathbf{A}^0$ and B^0 set to 0.

The maximum likelihood estimate for parameters is then obtained by equating the first derivatives of the log likelihood to zero; from equation (2), we see that

$$\frac{\partial \log B(\alpha)}{\partial \alpha_i} = \frac{\partial}{\partial \alpha_i}\left[\sum_k \log\Gamma(\alpha_k) - \log\Gamma\left(\sum_k \alpha_k\right)\right] = \Psi(\alpha_i) - \Psi\left(\sum_k \alpha_k\right), \tag{8}$$

where Ψ is the digamma function; therefore,

$$\frac{\partial \log \mathcal{L}}{\partial \alpha_i} = A_i - B\frac{\partial \log B(\alpha)}{\partial \alpha_i} = A_i - B\left[\Psi(\alpha_i) - \Psi\left(\sum_k \alpha_k\right)\right], \tag{9}$$

giving $\Psi\left(\sum_k \alpha_k\right) = \Psi(\alpha_i) - \frac{A_i}{B}$ for the maximum point, which we solve numerically for α_i using the bounds discussed in Minka (2003). In addition, performing a quadratic (Gaussian) approximation around the maximum, we can obtain estimates for the error bars on the maximum likelihood estimate from $\left.\frac{\partial^2 \log \mathcal{L}}{\partial \alpha_i^2}\right|_{\max} = -\sigma_{\alpha_i}^{-2}$, giving

$$\sigma_{\alpha_i} = \left(B\left[\Psi'(\alpha_i) - \Psi'\left(\sum_k \alpha_k\right)\right]\right)^{-\frac{1}{2}}; \tag{10}$$

for the purpose of the confidence interval estimates in this paper, we disregard covariance terms arising from $\frac{\partial^2 \log \mathcal{L}}{\partial \alpha_i \partial \alpha_j}$.

We defer detailed discussion of a suitable form of the prior on these chord parameters to future work. We have derived an approximate noninformative prior (Jaynes 2003, Chapter 12) within the conjugate family, but its use is inappropriate in this setting, where we can bring considerable musical experience to bear (and indeed the maximum *a posteriori* estimates generated using this noninformative prior give inferior performance than the maximum likelihood estimates in our experiment).

4.2 Results

Our corpus of MIDI transcriptions is made up of files each with thousands of MIDI events, with typically over five instruments playing at any given time; each bar typically contains several dozen

Chord, win	$\alpha_{t\bar{t}}$	α_{rmu}
Maj/Min, bar	$\{6.28, 1.45\} \pm \{0.49, 0.099\}$	$\{3.91, 1.62, 2.50\} \pm \{0.23, 0.11, 0.15\}$
Maj/Min, sub-bar	$\{3.26, 0.72\} \pm \{0.32, 0.054\}$	$\{4.04, 2.66, 2.29\} \pm \{0.21, 0.15, 0.13\}$
other	$\{5.83, 1.04\} \pm \{0.82, 0.12\}$	$\{4.08, 2.35, 1.49\} \pm \{0.38, 0.23, 0.16\}$

Table 1: Maximum likelihood estimates and 1σ error bars for Dirichlet distributions, based on labelled ground truth.

notes. We selected 16 songs in broadly contrasting styles, and ground-truth chord labels for those transcriptions of performances were generated by a human expert, informed by chord labels as assigned by song books to original audio recordings. We then divided our corpus of 640 labelled bars into "training" and "testing" sets of 233 and 407 bars respectively. Based on an initial inspection of the training set, we performed maximum likelihood parameter estimation for the chord models for three different sets of labels: major or minor chord labels for an entire bar; major or minor labels for windows shorter than a bar; and all other labels.

From the inferred parameters for major and minor chords at different window sizes in table 1, there was clear evidence for qualitatively different label generation at sub-bar window sizes from the behaviour of labelling whole bars: the sub-bar window sizes have high probability density for small non-triadic tones, while whole-bar windows have a vanishing probability density near a zero proportion of non-triadic tones (from the different qualitative behaviour of distributions with Dirichlet parameters below and above 1.0: 0.72 and 1.45 in our case). We interpret this as showing that the ground-truth labels were generated such that a sub-bar window is only labelled with a distinct chord if there is strong evidence for such a chord – *i.e.* only small quantities of non-triadic tones. If no sub-bar window is clearly indicated, then a closest-match chord label is applied to the whole bar, explaining the only slight preference for chord notes in the whole-bar distribution. There was insufficient ground-truth data to investigate this issue over the other families of chords (indeed, there was only one example of an augmented chord in the training data set).

Using the maximum likelihood estimates of table 1, we performed inference over window sizes and chord labels over the testing set, obtaining 53% of correct windows and 75% of correct labels given the window. Additionally, we performed a large (but by no means exhaustive) search over the parameter space on the training data, and obtained parameter values which performed better than these maximum likelihood estimates on the testing set, giving 75% windows and 76% chords correctly. It should be noted that the training and testing sets are quite similar in character, being individual bars drawn from the same pieces; it would be difficult to justify claims of independence between the sets. Validation on an independent testset (*i.e.* music excerpts drawn from different pieces) is currently being undertaken.

We interpret these trends as suggesting that the model for chords based simply on tone proportions is insufficiently detailed to capture successfully enough of the process by which ground-truth labels are assigned. The fact that maximum likelihood estimates perform noticeably worse than a set of parameters from training indicates that there is structure in the data not captured by the model; we conjecture that inclusion of a model for the chord label conditioned on the functional bass note in a window would significantly improve the performance of the model.

Another musically-motivated refinement to the model would be to include an awareness of context, for instance by including transition probabilities between successive chord labels (in addition to the

implicit ones from the musical surface). This corresponds to removing the assumption that the labels are conditionally independent given the musical observations: an assumption that is reasonable as a first approximation, but in actuality there will be short-term dependence between labels as, for instance, common chord transitions (such as IV-V-I) might be favoured over alternatives in cases where the observations are ambiguous; similarly, enharmonic decisions will be consistent over a region rather than having an independent choice made at the generation of each label.

The performance of our approach, without any of the above refinements, is at least comparable to techniques which do relax the assumption of conditional independence between labels; for example, the algorithm of Temperley (2001), which infers chord labels over the entire sequence (using dynamic programming to perform this inference efficiently), achieves a comparable level of accuracy (around 77%) on those pieces from our dataset for which it correctly computes the metrical structure.

5 Conclusions

We have presented a simple description of the dependence of chord labels and pitch-class profile, with an explicit statistical model at its core; this statistical model can be used not only to infer chord labels given musical data, but also to infer the appropriate granularity for those labels. Our empirical results demonstrate that adequate performance can be achieved, while suggesting that refinements to the statistical description could yield significant improvements. The model presented ignores all context apart from the bar-long window in question, and operates only on pitch-class profile data; incorporation of such extra information can simply be achieved by extending the statistical model. Similarly, we can incorporate available metadata into our model, for instance by defining a genre-specific chord label prior; and we can change the repertoire of chords under consideration without alteration of the framework, simply by replacing one component of the observation model.

Acknowledgments

C.R. is supported by EPSRC grant GR/S84750/01; D.L. and D.M. by EPSRC grant EP/D038855/1.

References

Bello, Juan P. and Jeremy Pickens. 2005. A Robust Mid-Level Representation for Harmonic Content in Musical Signals. In *Proc. ISMIR*, 304–311.

Chuan, Ching-Hua and Elaine Chew. 2005. Polyphonic Audio Key Finding Using the Spiral Array CEG Algorithm. In *Proc. ICME*, 21–24.

Harte, Christopher, Mark Sandler, Samer Abdallah, and Emilia Gómez. 2005. Symbolic Representation of Musical Chords: A Proposed Syntax for Text Annotations. In *Proc. ISMIR*, 66–71.

Jaynes, Edwin T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

Krumhansl, Carol L. 1990. *Cognitive Foundations of Musical Pitch*. Oxford University Press.

Lee, Kyogu and Malcolm Slaney. 2006. Automatic Chord Recognition from Audio Using an HMM with Supervised Learning. In *Proc. ISMIR*.

MacKay, David J. C. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Minka, Thomas. 2003. Estimating a Dirichlet Distribution. `http://research.microsoft.com/~minka/papers/dirichlet/`.

Müllensiefen, Daniel and Klaus Frieler. 2004. Cognitive Adequacy in the Measurement of Melodic Similarity: Algorithmic vs. Human Judgments. *Computing in Musicology 13*, 147–176.

Pachet, François. 1991. A meta-level architecture applied to the analysis of Jazz chord sequences. In *Proc. ICMC*.

Raphael, Christopher and Joshua Stoddard. 2004. Functional Harmonic Analysis Using Probabilistic Models. *Computer Music Journal 28*(3), 45–52.

Scholz, Ricardo, Vìtor Dantas, and Geber Ramalho. 2005. Funchal: a System for Automatic Functional Harmonic Analysis. In *Proc. SBCM*.

Sheh, Alexander and Daniel P. W. Ellis. 2003. Chord Segmentation and Recognition using EM-trained Hidden Markov Models. In *Proc. ISMIR*, 185–191.

Tagg, Philip. 2003a. *Harmony* entry. In J. Shepherd, D. Horn, and D. Laing (Eds.), *Continuum Encyclopedia of Popular Music of the World*. Continuum, New York.

Tagg, Philip. 2003b. *Lead sheet* entry. In J. Shepherd, D. Horn, and D. Laing (Eds.), *Continuum Encyclopedia of Popular Music of the World*. Continuum, New York.

Temperley, David. 2001. *The Cognition of Basic Musical Structures*. MIT Press.

Temperley, David. 2004. Bayesian Models of Musical Structure and Cognition. *Musicae Scientiae 8*, 175–205.

Temperley, David. 2007. *Music and Probability*. MIT Press.