A CROSS-VALIDATED STUDY OF MODELLING STRATEGIES FOR AUTOMATIC CHORD RECOGNITION IN AUDIO

John Ashley Burgoyne Laurent Pugin Corey Kereliuk Ichiro Fujinaga
Centre for Interdisciplinary Research in Music and Media Technology
Schulich School of Music of McGill University
Montréal, Québec, Canada H3A 1E3
{ashley,laurent,corey,ich}@music.mcgill.ca

ABSTRACT

Although automatic chord recognition has generated a number of recent papers in MIR, nobody to date has done a proper cross validation of their recognition results. Cross validation is the most common way to establish baseline standards and make comparisons, e.g., for MIREX competitions, but a lack of labelled aligned training data has rendered it impractical. In this paper, we present a comparison of several modelling strategies for chord recognition, hidden Markov models (HMMs) and conditional random fields (CRFs), on a new set of aligned ground truth for the Beatles data set of Sheh and Ellis (2003). Consistent with previous work, our models use pitch class profile (PCP) vectors for audio modelling. Our results show improvement over previous literature, provide precise estimates of the performance of both old and new approaches to the problem, and suggest several avenues for future work.

1 INTRODUCTION

The task of automatic, continuous chord recognition is an area of active study in the MIR community. When working with audio, chord recognition is an especially difficult task because a chord represents such a wide range of possible musical events. Recent studies have shown the benefit of applying stochastic modelling to this task [1, 7, 8, 11, 13]. The most commonly used model is the hidden Markov model (HMM) [10], but more recent work has also explored discriminative models [9] such as the conditional random field (CRF) [14].

In this paper, we use the work of Sheh and Ellis as our departure point [13]. These authors used HMMs to perform chord recognition on a set of 20 Beatles songs. Although their recognition rates were poor, they laid a foundation for future study. We use the same data set, but in addition to their tests, we perform a 10-fold cross validation to verify the validity of our results, training on 18 songs for each run and testing on the remaining 2. Cross validation is essential for obtaining unbiased estimates of model performance when data is limited [5], but because it is so time-consuming to label audio files, no previous studies of audio chord recognition have tried it. Cross-validated recognition rates will be lower than the best possible test rate on a single song, e.g., the metric used in [7], but they give a more realistic depiction of the state of the art and are the only fair way to compare different models.

Another problem that is often encountered when building HMMs is a lack of aligned, labelled training data. When the training data is not aligned, the model must be initialised using a so-called *flat start*. In a flat start, training audio is uniformly segmented based on an unaligned transcription. One hopes that enough of the uniformly segmented labels in the flat start will match the correct alignment so that the model parameters will improve during successive training iterations, but this is unlikely in musical applications because chord lengths vary so widely. Using a flat-start with this training data has indeed been shown to result in poor recognition performance [13], and so for our training, we avoided it.

Our results demonstrate the usefulness of stochastic modelling and highlight the benefits of CRFs, which until now have received very little attention in the MIR community.

2 PITCH CLASS PROFILE VECTORS

Pitch class profile (PCP) vectors, which were introduced by Fujishima in 1999 [3], have been widely used in chord recognition systems [1, 4, 6, 13]. In essence, PCP vectors represent a logarithmically warped and wrapped version of the short time frequency spectrum. The following equations show the computation of a PCP vector:

$$\mathrm{PCP}[i] \quad \triangleq \sum_{\{k\,:\,p[k]=i\}} \|X[k]\|^2 \tag{1}$$

$$p[k] \triangleq \left\{ \mathrm{round}\left[D \log_2 \left(\frac{k}{N} \cdot \frac{f_s}{f_{\mathrm{ref}}} \right) \right] \right\}_{\mathrm{mod}\ D} \tag{2}$$

In equation 2, k is an FFT bin, f_{ref} is the reference pitch, N is the FFT window size, f_s is the sampling rate, and D is the dimensionality of the PCP vector. The $\log_2$ operation *warps* the frequency spectrum to a logarithmic scale, while the modulo operation *wraps* the frequency spectrum at integer multiples (octaves) of the reference frequency. The frequency components in each of the warped

and wrapped frequency bands are then summed (equation 1). In this study, we used $D = 12$ and $f_{\text{ref}} = 261.6$ Hz (C4). Thus, each PCP vector represents 12 semi-tones of a chromatic scale under the same modulus as most Western music theorists as well as the MIDI note number specification.

2.1 Gaussian Distributions

The choice of parameters used to model a PCP vector is not trivial. Each dimension in the normalised PCP vector has a continuous output and thus can be modelled as a probability distribution. In [13], Sheh and Ellis modelled PCP vectors using single Gaussians. Using our labelled training data, it is possible to show that single Gaussians do not provide an adequate model of the PCP vectors. Figure 1 shows 3 examples of a single Gaussian superimposed on 1 dimension of an A-minor PCP vector. Clearly a mixture of Gaussians would be more suitable to accurately model the shape of these distributions. In this paper, we used a mixture of Gaussians to model the probability distributions of the PCP vectors, and as can be seen from the results, the models trained using a mixture of Gaussians outperform those using single Gaussians.

2.2 Dirichlet Distributions

Because all values in a normalised PCP vector must be non-negative and sum to one, one may think of them as the parameters to a hypothetical multinomial distribution. Although mixtures of Gaussians can theoretically be used to model any probability distribution, for multinomials, there is another common model known as the Dirichlet distribution. This distribution is the so-called conjugate prior of the multinomial distribution, i.e., given a set of parameters that represent an archetypal multinomial distribution, it represents the probability that any other multinomial distribution might arise instead. Unlike mixtures of Gaussians, Dirichlets enforce the constraint that the output distributions be valid multinomial distributions, which is equivalent to the constraints on valid normalised PCP vectors. Moreover, they require fewer parameters to train:

$$\text{Dir}(\boldsymbol{p}, \boldsymbol{u}) \triangleq \frac{1}{Z(\boldsymbol{u})} \prod_{i=1}^{D} p_i^{u_i - 1} \tag{3}$$

The p_i correspond to the bins of the multinomial distribution, which are the 12 values of the PCP vector in our case. The parameter vector $\boldsymbol{u} = \{u_1, u_2, \ldots, u_D\}$, where D is the number of bins in the multinomial, determines both the mean and the variance of the distribution. The normalisation term $Z(\boldsymbol{u})$ is beyond the scope of this paper, but more information can be found in [2], an earlier application of Dirichlet distributions to chord recognition.

3 CHORD SEQUENCE MODELS

3.1 Hidden Markov Models

HMMs are generative stochastic models that attempt to model a hidden first-order Markov process based on a set of observable outputs. Nonetheless, it should be noted that in reality chord progressions are high-order Markov processes, and so the first-order Markov assumption may result in model deficiencies.

HMMs for chord recognition have been used with two different approaches. Sheh and Ellis [13] have experimented with a model-discriminant (MD) approach, which means that every potential chord is modelled by its own left-right, single-state HMM. Each model is trained either individually or by using an embedded version of the Baum-Welch algorithm that concatenates the HMMs. In order to perform this step, one needs labelled—but not necessarily aligned—training data. Chord recognition is performed using the Viterbi algorithm on the network of component HMMs to yield the most likely sequence of component *models*. Bello and Pickens [1] and Lee and Slaney [6], on the other hand, have experimented with a path-discriminant (PD) approach, in which every chord is modelled by one state in a larger, fully connected HMM, which can be trained with the expectation-maximisation (EM) algorithm and does not require labelled data. Chord recognition with the PD approach uses the standard Viterbi algorithm to obtain the most likely sequence of *states*.

3.2 Conditional Random Fields

Given a sequence of observations $\boldsymbol{X}$, HMMs seek to maximise the joint probability $P(\boldsymbol{X}, \boldsymbol{Y})$ for a hidden state sequence $\boldsymbol{Y}$. This method works well in practise, but from a theoretical perspective, it is not quite the question that one ought to be asking at recognition time. During recognition, the observation sequence is always fixed, and so it may make more sense to model only the conditional probability distribution $P(\boldsymbol{Y}|\boldsymbol{X})$. This frees the recogniser from needing to enforce any particular model $P(\boldsymbol{X})$ of the data, and thus it is no longer necessary for the components of the observation vectors to be conditionally independent, as they must be for HMMs. Such a model may include thousands or even millions of observation features.

The closest analogue to the HMM that uses this modelling technique is known as the linear-chain CRF, one of the most commonly used member of the larger CRF family [14]. Besides the probabilistic characteristics mentioned above, CRFs differ from the HMMs in that each hidden state depends not just on the current observation but on the complete observation sequence. At decoding time, linear-chain CRFs are quite similar to HMMs, using a variant of the Viterbi algorithm. Unlike an HMM, however, in order to train a linear-chain CRF, one must have access to fully labelled and aligned training data. Training is considerably slower for CRFs than it is for an HMMs regardless of whether one uses a path-discriminant or model-discriminant approach, but fortunately, there are

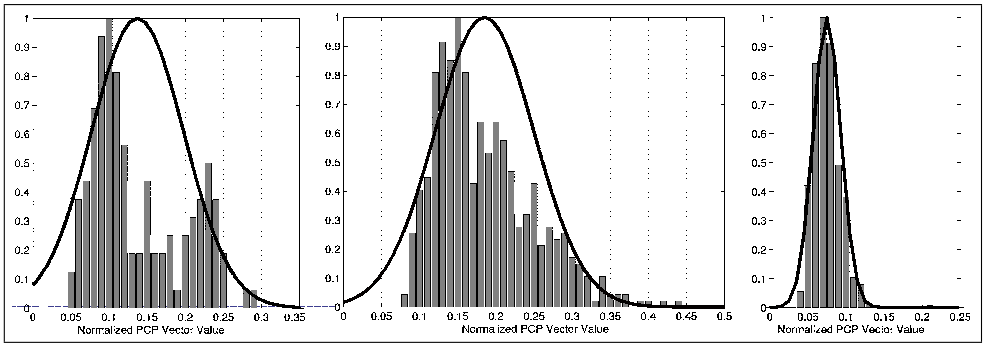

Figure 1. Single Gaussian (smooth curve) plotted over three different dimensions of a PCP vector (histogram).

Roots:	A, A♭, B, B♭, C, D, D♭, E, E♭, F, F♯, G
Families:	maj, min, aug, dim
Examples:	A, Bm, E+, Fdim

Table 1. List of chords used in recognition experiments.

optimisation techniques to improve the training speed. We chose the limited-memory Broyden–Fletcher–Goldfarb–Shanno method, a variant of Newton's method [12].

4 EXPERIMENTS AND RESULTS

The Beatles recordings in our data set were filtered and resampled at 11 025 Hz to remove the high-frequency content. An STFT was calculated on each song using an FFT size of 2048 samples and a hop size of 1024 samples (92 ms). From the STFT, 12-dimensional PCP vectors were calculated (for simplicity, we did not attempt the tuning adjustments used in [1] or [4]), and given knowledge about the original key of each song, a second set of PCP vectors was was generated by transposing (rotating) the first set to C major so as to reduce the chances of learning key-dependent harmonic relationships. (In a large-scale application, an automatic key-finding algorithm could be used for this purpose.) Each song was hand-transcribed with chord labels, and these labels were then simplified to triads only: major, minor, augmented, and diminished (see table 1). These labels are a departure from Sheh and Ellis, who attempted to recognise 7th chords as well, but we felt that the data set was insufficient to estimate so many models properly. When used with the rotated PCP vectors, the labels were also transposed to C major. After simplification, both the rotated and unrotated sets of chord labels contained 24 distinct chord symbols out of the 48 that would have been theoretically possible.

The HMM-MD model was implemented using the Hidden Markov Model Toolkit (HTK) [1] with single-state component models and 1, 6, or 12 Gaussians to model each PCP bin. For the HMM-PD, an ergodic (fully-connected) HMM with one state for each chord of the list was trained

[1] http://htk.eng.cam.ac.uk/

Model	Gau.	Recognition rate (%)	
		Rotated	Unrotated
HMM-PD	1	24.2	28.8
HMM-PD	6	31.9	34.1
HMM-PD	12	37.9	36.1
HMM-PD	20	45.1	40.5
HMM-MD	1	37.1	39.7
HMM-MD	6	48.8	45.5
HMM-MD	12	48.8	47.1
HMM-PD	24	48.7	31.6
CRF-D	–	39.5	45.3
CRF-G	1	34.7	29.9
CRF-DG	1	39.9	42.4

Table 2. Frame-by-frame recognition results for all models with varying numbers of Gaussians.

using the Torch machine learning library. [2] During initialisation, every portion of the labelled data was assigned to its corresponding state. We experimented using 1, 6, 12 and 20 Gaussians per state. Training took a couple of seconds on a 2.7 GHz PowerPC G5 processor.

The linear-chain CRFs [3] were trained with transition features between all chords in the training sets and special features denoting starting and ending symbols. We tried three versions of the emission features, the first equivalent to a single Gaussian (CRF-G), the second equivalent to a Dirichlet distribution (CRF-D), and the third a combination of both sets of emission features (CRF-GD), taking advantage of the fact that the features in CRFs need not be independent for the model to run properly. The L-BFGS optimisation routine was allowed to run for 250 iterations with a constraint on the parameters that their standard deviation be no more than 10. The purpose of these limits was to avoid over-training, which is a particular risk with CRFs. It took four to six hours to train each run of each model on a 2.7 GHz PowerPC G5 processor.

Our results are summarised in table 2. Our evaluation was done by carrying out a frame-by-frame comparison

[2] http://www.torch.ch/

[3] http://crf.sourceforge.net/

of the recognised labels with the hand marked labels. The number of correct frames overall was divided by the total number of frames overall in order to give a percentage score to the recognition. Unlike some other papers, e.g., [7], we did not allow for any fuzziness in recognition at the boundaries, and so the figures represented here will be lower but more precise. Results are presented for several mixture sizes of Gaussians.

The simplest model here, the path-discriminant HMM (HMM-PD), also performs the worst. When PCP bins are modelled as single Gaussians, its best performance is 28.8 percent. The model-discriminant HMM (HMM-MD), in contrast, performs much better, at 39.7 percent even with a single Gaussian and reaching 48.8 percent with 12 Gaussians. At 24 Gaussians, over-training starts to reduce performance, especially for the unrotated vectors. This model is the same as in [13], but our best recognition rates are a more than twofold improvement over theirs using the same training set and evaluation script. We speculate that the large difference is due to the inclusion of a mixture of Gaussians, the exclusion of a flat-start during model training, and a reduction of the classification set to triads. For both HMM-PD and -MD, the unrotated PCP vectors perform better with smaller numbers of Gaussians and the rotated PCP vectors become slightly better as the the number of Gaussians increases. Although the differences between performance on the two PCP sets is never more than five percentage points for the HMMs, this pattern warrants further investigation.

At the single-Gaussian level (CRF-G), CRFs perform better than the PD HMMs but do not quite match the performance of the MD HMMs, perhaps on account of overtraining; unlike the HMMs, the single-Gaussian CRFs perform much better on rotated PCPs than unrotated. The most interesting feature of the CRF results, however, is the large improvement in performance when using Dirichlet distributions (CRF-D and CRF-DG). Although CRF-D is not quite able to match HMM-MD performance at its maximal number of Gaussians, it comes very close to it while using a factor of 40 fewer model parameters. There is no question that these distributions warrant further study for chord recognition.

5 SUMMARY AND FUTURE WORK

We presented a comparison of both traditional and new approaches, HMMs and CRFs, to audio chord recognition using PCP vectors. Overall, our results compare favourably with previous research for this task, but they also suggest that on their own, PCP vectors may not be sufficient for reliable discrimination. Moreover, we were able to perform our experiments with a fully annotated training set of live recordings, which is rare for the field. These annotations allowed us to cross-validate our results for a more accurate representation of the state of the art.

Our results suggest that further research is needed in modelling audio features for chord recognition. One approach would be to incorporate Dirichlet distributions more widely, e.g., in the HMM-based models we used. Another would be to investigate alternatives or supplements to PCP vectors, e.g., [6]. Both should be explored as audio chord recognition enters the MIREX competitions in coming years.

6 ACKNOWLEDGEMENTS

We would like to thank the Canada Foundation for Innovation and the Social Sciences and Humanities Research Council of Canada for financial support, Tristan Matthews for preparing the ground truth, and Aaron Courville for his advice about stochastic models.

7 REFERENCES

[1] J. P. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signals," in *Proc. 6th Int. Conf. Mus. Inf. Ret.*, 2005, pp. 304–11.

[2] J. A. Burgoyne and L. K. Saul, "Learning harmonic relationships in digital audio with Dirichlet-based hidden Markov models," in *Proc. 6th Int. Conf. Mus. Inf. Ret.*, 2005, pp. 438–43.

[3] T. Fujishima, "Realtime chord recognition of musical sound: A system using Common Lisp Music," in *Proc. Int. Comp. Mus. Conf.*, 1999.

[4] C. A. Harte and M. B. Sandler, "Automatic chord identification using a quantised chromagram," in *Proc. 118th Conv. Aud. Eng. Soc.*, 2005.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Berlin: Springer, 2001.

[6] K. Lee, "Automatic chord recognition from audio using enhanced pitch class profile," in *Proc. Int. Comp. Mus. Conf.*, 2006.

[7] K. Lee and M. Slaney, "Automatic chord recognition from audio using an HMM with supervised learning," in *Proc. 7th Int. Conf. Mus. Inf. Ret.*, 2006, pp. 133–7.

[8] J.-F. Paiement, D. Eck, and S. Bengio, "A probabilistic model for chord progressions," in *Proc. 6th Int. Conf. Mus. Inf. Ret.*, 2005, pp. 312–9.

[9] G. E. Poliner and D. P. W. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Advances in Signal Processing*, forthcoming.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–87, 1989.

[11] C. Sailer and K. Rosenbauer, "A bottom-up approach to chord detection," in *Proc. Int. Comp. Mus. Conf.*, 2006, pp. 612–15.

[12] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proc. Human Lang. Tech. Conf.*, 2003, pp. 213–20.

[13] A. Sheh and D. P. W. Ellis, "Chord segmentation and recognition using EM-trained hidden Markov models," in *Proc. 4th Int. Conf. Mus. Inf. Ret.*, 2003, pp. 185–91.

[14] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, Eds. MIT Press, 2006.