THE 2010 LABROSA CHORD RECOGNITION SYSTEM

Daniel P. W. Ellis
LabROSA, Columbia University
New York, USA
dpwe@ee.columbia.edu

Adrian Weller
LabROSA, Columbia University
New York, USA
adrian@cs.columbia.edu

ABSTRACT

For the MIREX 2010 Audio Chord Extraction task, we submitted a total of four systems. Our base system is a trainable chord recognizer based on two-band chroma representations and using a Structured SVM classifier to replace the more familiar hidden Markov model. We submit two versions of this system, one which transposes all training data through all 12 possible chords to maximize the training data available for each chord (and hence improve generalization to rarely-seen chords and keys), and one which simply trains on the chords in their original transposition, leading to a smaller model and possible learning of key-specific features. We also submit two pre-trained models, based on these two frameworks, trained in-house on the 180 Beatles and 20 Queen tracks for which ground-truth chord labels have been made available.

1. INTRODUCTION

Audio chord recognition takes a full musical signal, such as a commercial pop music recording, and returns a sequence of labels indicating the chords in the piece and the times during which they are active. Although the vocabulary of possible chords can become quite complex, for many purposes – such as searching for common musical patterns – it is sufficient to define a smaller subset of basic chords. In this work, we use a vocabulary of 25 chords – one major and one minor chord for each of the 12 root chroma (C, C#, D . . . B), plus one "no chord" symbol.

Figure 1 shows the basic structure of our system. We use our beat-synchronous, instantaneous-frequency chroma features, originally developed for cover song detection [2]. These features are fed into the SVM^{hmm} package of Joachims [4] which generates a model then used at recognition time. We submitted an earlier version of this system to MIREX 2009 [6]. Within this structure, a number of variants were investigated, as described below.

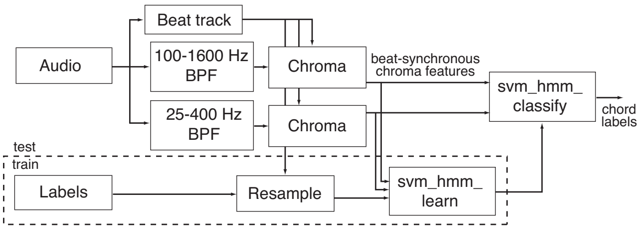

Figure 1. Block diagram of structured SVM audio chord recognition system.

2. DATA & EVALUATION

We evaluated our models on the 180 tracks of the 12-album Beatles corpus, using the hand-marked labels from Chris Harte [3]. Our audio apparently came from different digital masterings of the albums, because not only were there variable time offsets compared to the labels, but our audio actually had slightly different speeds than those that can be inferred from the labels. Moreover, the speed difference is not even constant throughout the track, as illustrated in figure 2 which shows residual systematic time differences between the time of each ground-truth chord label and the nearest beat time from our beat tracker, even after linear speed compensation is applied. We surmise that when a new digital master is made from the original studio tapes, variations in the tape speed lead to slightly different values. Although most speed variations were below 0.5% – too small to be perceived as a pitch shift – this is enough to result in grossly misaligned chord labels by the end of a 3 minute track. We manually corrected time and speed offsets for all 180 tracks to obtain our training labels. Because the chord labels are effectively quantized onto the grid defined by our beat tracker, the remaining nonlinear deviations have little or no effect on system training, and only minor impact on scoring.

We divided the data into four cuts, each consisting of 3 albums, with the release dates interspersed in order to get a range of styles in each cut. Our cuts were {Please Please Me, Help!, Magical Mystery Tour}, {With The Beatles, Rubber Soul, The White Album}, {A Hard Day's Night, Revolver, Abbey Road}, and {Beatles For Sale, Sgt. Pepper's Lonely Hearts Club Band, Let It Be}. To evaluate each system, we trained on three cuts and tested on the fourth, repeating this four times until every track had been used as a test target; the results are the scoring over all 180 tracks. In order to cast the detailed ground truth la-

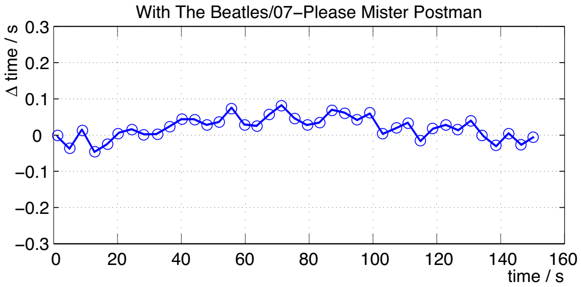

Figure 2. Differences in timing between manual chord labels and automatic beat times. Even after correcting for a 0.3 s timing skew and a 0.5% speed difference, the plot shows there is a residual nonlinear timing drift, with labels, relative to the beats extracted from the audio, early at the start and end of the track, and late in the middle.

System	Dims	Time	Acc%
Baseline	12	0:08	67.7
+ tuning fixes	12	0:08	68.6
+ whitening	12	0:08	71.6
+ quad terms	90	0:23	72.5
+ prev. frame	102	0:21	73.2
+ LF chroma	114	0:23	76.0
+ 12×, $C = 400$	114	2:38	77.6
Best HMM	24	0:06	74.9

Table 1. Accuracy, feature dimensionality, and training times of different model variants.

bels into our 25 major/minor/no chord set, any chord label including "min" was considered a minor chord, and everything else was considered a major chord. Although this is perhaps not the best-motivated choice, the prevalence of non-canonical-triad chords is believed small enough to have little influence.

Although the system is based on beat intervals, the evaluation is weighted by the non-quantized temporal overlap. Ground truth labels are cast into 25 categories using the same rules as used to obtain the training labels. Accuracy (the proportion of time for which labels were correct) reflects all correctly-labeled intervals except recognizing "no chord" as "no chord", since this makes the results too dependent on details of lead-in and lead-out. Recognizing "no chord" as some other chord, or labeling valid chords as "no chord", does contribute to error (i.e. the denominator), however.

3. RESULTS

Table 1 shows the overall accuracy on a range of system variants. The table is structured as a series of enhancements moving towards more complex, and better performing system, although these represent a simplified picture of the actual system evolution. Each line is explained in more detail below:

- **Baseline**: Beat-averaged chroma features (12 dimensions per beat) are passed to the structured SVM training routine, along with the corresponding class label (from the 25-entry vocabulary). The parameter C, which controls the tradeoff between training error and margin, Is set to 100, which gave the best results for this training setup.

- **tuning fixes**: On inspecting the results, we noticed that several tracks were getting almost no chords correct. For several of these, it was because the automatic tuning reference was off by one semitone. Instead of modifying the labels, we added a special case to force the automatic tuning to stretch beyond ± 0.5 semitones for these files. This line represents the significant performance gain obtained by changing the tuning reference for just 4 Beatles tracks, as shown in table 2.

- **whitening**: [5] reports a dramatic improvement in trained chord recognition by normalizing the means and variances of their constant-Q spectral features over a local window. We implemented something like this by dividing out the energy of the spectrum smoothed by a window whose bandwidth was proportional to center frequency. We found that an extremely narrow window (a Gaussian with half-width 0.02 octaves) gave the best result, which amounts in most cases to setting all the raw spectral magnitudes to 1. Because the actual spectral peaks are chosen in our chroma calculation on the basis of phase consistency [1], the spectral magnitudes influence only the weight contributed by each harmonic identified. However, it is clear that this "spectral whitening" had a significant beneficial effect on normalizing away timbral variations in the training data.

- **quad terms**: This corresponds to adding quadratic terms to the basic 12 chroma bins i.e. an additional $(12 \times 13)/2 = 78$ dimensions formed as the product of every possible pairing of features. Since the SVM is using only a linear kernel, this increases the effective dimensionality of the space in which the decision boundary is sought, akin to using a quadratic kernel.

- **prev. frame**: We added the features for the previous frame as part of the current frame's feature vector. Note that the quadratic features were not calculated for these additional dimensions. We tried many combinations of multiple past and future frames, and different quadratic term configurations, but this simple setup performed nearly the best.

- **LF chroma**: We added another 12 chroma bins calculated from a spectrum centered around 100 Hz, instead of the 400 Hz center of the main chroma. These features often capture the bass line.

- **12×, *C*=400**: To increase the training data available for individual chords, we transposed all our training examples through all 12 possible rotations, and trained on all of these (effectively making all chord

Track	Tuning/cents
beatles/Lovely Rita	-66
beatles/Strawberry Fields Forever	-52
beatles/Wild Honey Pie	-55
beatles/Ticket To Ride	-54
queen/Another One Bites The Dust	54

Table 2. Manual exceptions for the automatic tuning.

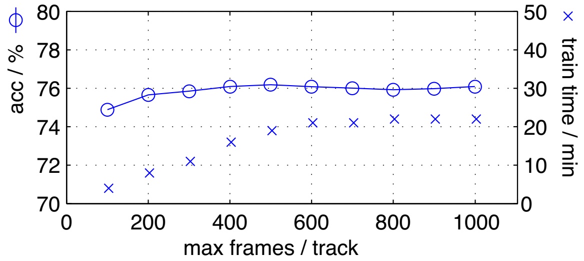

Figure 3. Variation of accuracy and training time with the maximum number of beats allowed for each track.

models identical modulo a rotation). This larger data set required longer training times. Training time was further increased by switching to a higher value of the training parameter C, which improved performance slightly.

- **Best HMM**: We evaluated our MIREX 08/09 HMM-Gaussian system within the same framework, and including all applicable enhancements (tuning, whitening, LF chroma, 12× data). While these helped, the system is still inferior to the structured SVM.

4. DATA TRUNCATION

SVM training memory requirements grow with the size of the training set, the size of the feature vector, and the training parameter C. For our larger models, we found that training on all available data (the 180 Beatles tracks plus 20 Queen tracks, or 94,974 total beats) exhausted the 16GB of RAM on our machine. We therefore experimented with truncating individual tracks after some number of beats, reasoning that we will obtain better diversity of training data by using fewer beats from more tracks rather than vice-versa, while remembering that in order to let the model learn good sequence features, we should retain as far as possible the original contiguity of the data.

Figure 3 shows the result of this experiment (trained on the “LF chroma” model of table 1). Although the average length of a track is a little under 500 beats, with the longest (“I want you/She’s so heavy”) composed of over 1800 beats, we see that accuracy is only affected when we truncate at 300 frames or shorter – at which point the total training time has already been halved. In fact, truncating at 500 beats gives the best performance, possibly because it avoids over-emphasizing some longer tracks (including the infamous “Revolution 9”).

5. MIREX

For the MIREX 2010 Audio Chord Estimation evaluation, we submitted four systems:

- **EW1** – a train/test system using all the enhancements listed in section 3 including 12× data transposition.
- **EW2** – a “lightweight” train-test system that does not use 12× data transposition.
- **EW3** – a pretrained system, equivalent to EW1, trained on the 180 Beatles tracks [3] and 20 Queen tracks [5]. To be able to train this comfortably on our 16GB machine, we also employed a maximum beats/track limit (as discussed in section 4) of 300.
- **EW4** – a pretrained version of EW2, trained without beats/track limitation.

6. CONCLUSIONS

We have described the audio chord recognition system we submitted to MIREX 2010. The full code to run all of these systems is available at `http://labrosa.ee.columbia.edu/projects/chords/`.

7. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) under Grant No. IIS-0713334. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

8. REFERENCES

[1] F. J. Charpentier. Pitch detection using the short-term phase spectrum. In *Proc. ICASSP-86*, pages 113–116, Tokyo, 1986.

[2] D. P. W. Ellis and G. Poliner. Identifying cover songs with chroma features and dynamic programming beat tracking. In *Proc. ICASSP*, pages IV–1429–1432, Hawai’i, 2007.

[3] Christopher Harte, Mark Sandler, Samer A. Abdallah, and Emilia Gómez. Symbolic representation of musical chords: A proposed syntax for text annotations. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 66–71, London, UK, 2005.

[4] T. Joachims, T. Finley, and C.N.J. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, 2009. Available: `http://www.cs.cornell.edu/People/tj/svm_light/svm_hmm.html`.

[5] Matthias Mauch and Simon Dixon. Approximate note transcription for the improved identification of difficult

chords. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, page (to appear), Utrecht, 2010.

[6] Adrian Weller, Daniel P. W. Ellis, and Tony Jebara. Structured prediction models for chord transcription of music audio. In *Proc. International Conference on Machine Learning and Applications*, pages 590–595, Los Alamitos, CA, USA, 2009. IEEE Computer Society.