Structured Prediction Models for Chord Transcription of Music Audio

Adrian Weller, Daniel Ellis, Tony Jebara
Columbia University, New York, NY 10027
aw2506@columbia.edu, dpwe@ee.columbia.edu, jebara@cs.columbia.edu

Abstract

Chord sequences are a compact and useful description of music, representing each beat or measure in terms of a likely distribution over individual notes without specifying the notes exactly. Transcribing music audio into chord sequences is essential for harmonic analysis, and would be an important component in content-based retrieval and indexing, but accuracy rates remain low. In this paper, the existing 2008 LabROSA Supervised Chord Recognition System is modified by using different machine learning methods for decoding structural information, thereby achieving significantly superior results. Specifically, the hidden Markov model is replaced by a large margin structured prediction approach (SVMstruct) using an enlarged feature space. Performance is significantly improved by incorporating features from future (but not past) frames. The benefit of SVMstruct increases with the size of the training set, as might be expected when comparing discriminative and generative models. Without yet exploring non-linear kernels, these improvements lead to state-of-the-art performance in chord transcription. The techniques could prove useful in other sequential learning tasks which currently employ HMMs.

1. Background

The Music Information Retrieval Evaluation eXchange [*http://www.music-ir.org/mirex/2008*] organized a contest where entrants were judged on their ability to identify the chords in commercial recordings of popular music. The evaluation was performed over a set of manually-labeled Beatles songs. Chord labels were simplified to 25 possibilities – one for each of the 12 major chords, one for each of the 12 minor chords, and one additional label to represent 'no chord'. The LabROSA Supervised Chord Recognition System [*http://labrosa.ee.columbia.edu/projects/chords/*] obtained the second highest accuracy in the evaluation, scoring about 10% relative worse than the best system. The modifications reported here improved the LabROSA system performance by approximately 8% relative, into the realm of state-of-the-art. By appealing to a well-established large margin discriminative methodology that has been popularized by support vector machines, this performance is achieved without extensive tweaking or domain adaptation. Moreover, this framework may be combined with other approaches, and further increases in performance are certainly possible through the investigation of more elaborate nonlinear kernels without requiring a substantial reformulation of the underlying algorithms.

The main stages of the LabROSA system may be summarized thus: An input song is first converted into beat-synchronous frames (for the Beatles songs used, the average number of frames per song is 459, with a range of 77 to 1806), each with 12 chroma features which are constructed to estimate the intensity of each semitone regardless of octave. Each of these 12 features is in the range [0,1]. It is assumed that the chord is constant within a frame. The remaining task is then a sequence labeling problem, where here we focus on accuracy per frame as the metric.

The baseline LabROSA system uses a Hidden Markov Model (HMM) with Viterbi decoding to compute the most likely sequence of chord labels from a song's chroma features. It is also possible to compute the most likely label for a sequence on a token by token basis, call this 'MaxGamma' decoding. Since the evaluation considers only per frame accuracy, we explored MaxGamma decoding, which generally provided a slight improvement.

We also employed a more recent max-margin discriminative approach, SVMstruct, which has proved very successful in other applications. This has better regularization properties, reducing the risk of over-fitting when adding more features. The LabROSA system's HMM uses Gaussian emissions, which lead to curved (quadratic) decision boundaries between labels. Since here only linear kernels for SVMstruct are considered, to allow comparison and as a first step towards more sophisticated kernels, in

some runs quadratic terms were added, i.e. pairwise products of existing features were added as new features. Features from neighboring frames were also introduced in some models, as suggested in [Y. Altun, I. Tsochantaridis and T. Hofmann, "Hidden Markov Support Vector Machines", *ICML* 2003].

2. Experiments

All experiments were performed on frame-level data, using the 180 labeled Beatles songs, and 25 possible chord labels described above. Ten random permutations of all the songs were selected. For each permutation, every model was trained on the first train% (30%, 60% or 90%) of the 180 permuted songs. The last 10% of the permuted songs was used for testing, and for validation if required, irrespective of the amount of training data used. Since the HMM models do not require a validation set, they were simply tested on the entire final 10%. The SVMstruct models, however, require the estimation of a *C* parameter. This was achieved by splitting the final 10% into two halves – the penultimate 5% of the permuted songs, call this set A, and the last 5%, call this set B. Each model was trained with a broad range of values of *C*. The particular value which gave optimal performance on set A was used for testing the model on set B, and vice versa. The results on A and B were then combined by averaging, weighted by the respective number of frames, to give the accuracy per frame over the entire test set.

T. Joachims' SVMstruct code was used, instantiated as SVM-HMM with the precision constant e (epsilon) set to 0.1, the order of dependencies of transitions in HMM t set to the default 1, and the order of dependencies of emissions in HMM e set to the default 0. With these settings, the interdependency structure of the features and labels in the model is comparable to that of the HMM used in the LabROSA system.

3. Results

Figure 1 displays the accuracy for each model as the amount of training data was varied, averaged over the ten permutations of songs. On the far right, *HMMv* is the baseline HMM approach with Viterbi decoding used in the LabROSA system. To its left, *HMMg* is the same model but using MaxGamma decoding, showing a small improvement.

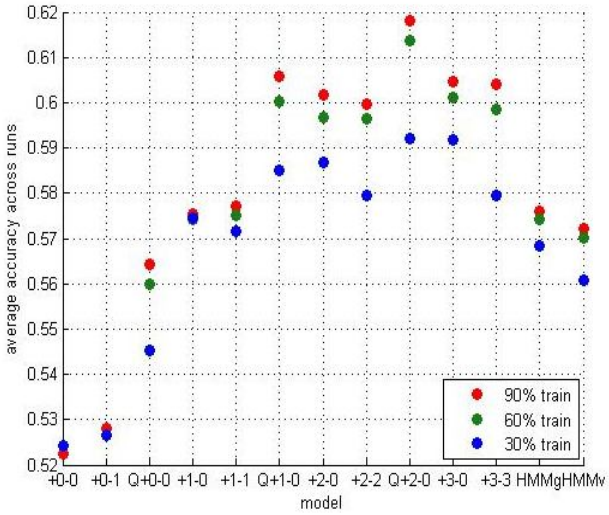

Figure 1. Average accuracies for each model

The models to the left are SVMstruct runs using various feature combinations. *+0-0* on the far left uses the original 12 chroma features for each frame – the same features used by the HMM models. To its right: *+m-n* uses features from the current frame along with those from each of the next *m* frames and each of the previous *n* frames; *Q* at the front means that in addition, all quadratic cross-terms have been added. One exception: *Q+2-0* does not use all cross-terms since that would lead to an unwieldy 702 dimensions, but instead uses the 324 features from *Q+1-0* and then adds just the 12 additional chroma features from 2 frames ahead without cross-terms.

The first five SVMstruct models do not show significant superior accuracy, but all the others do. In addition, *HMMg* shows small but statistically significant performance advantages over *HMMv*.

Results are almost uniformly better as the size of the training set grows, with the rates of improvement of the more complex SVMstruct models higher than those of the HMMs. Based on the modest gains from going from 60% to 90% training set size, however, there may not be much more to gain with more training data. Quadratic terms provide dramatic improvements, suggesting further gains may be achieved with non-linear kernels. Adding features from future frames also provides striking benefits, but only up to two frames ahead. Interestingly, adding features from past frames appears not to help. These observations may be relevant for other audio processing sequence labeling tasks, including melody or bass line transcription and perhaps speech recognition.