CS 6784 (Spring 2014): Project proposal

Charles Hermann CIH5
Jonathan Shi JS2845

1. Team

Charles Hermann (cih5) and Jonathan Shi (js2845)

2. Motivation

One of the major areas of research in MIR is music transcription, the process of creating a musical score from an audio recording. Though this can be done manually by experienced musicians, this process is difficult and time consuming. As a result, the ability to automatically transcribe music has numerous positive ramifications in the musical community and the MIR community. In addition, transcription presents an interesting and difficult challenge even for seemingly easily versions of the task such as transcribing piano solos. Several factors make this a difficult area of study including but not limited to: the high-dimensionality of the datasets, the large margin for error (both false positives and false negatives), and the lack of a field-accepted feature vector/technique for extracting the data.

Currently, the primary techniques used in transcription are HMMs applied to temporal windows and frequency bins (Raphael, 2002), neural networks (Bock & Schedl, 2012), and simple discriminative techniques to identify if a singular note is being played in a temporal window (Poliner & Ellis, 2006). Also feature sets have been obtained via deep learning (Nam et al., 2011) and sparse nonnegative matrix factorization (COSTANTINI et al., 2013), which have performed better in classification-based transcription than manually specified features. Finally, techniques have been tuned to detect note onsets and note offsets for the piano (Benetos & Dixon, 2011).

3. Statement of Task

Can we improve upon the state-of-the-art in transcription? Specifically, how well will techniques such as max-margin markov models (involving a multi-class SVM for note identification and a HMM on top of identified notes) work within this area?

4. General Approach

One approach we intend to try to use to improve upon existing automated transcription techniques is by using a maximum-margin Markov model (Taskar et al., 2003) which improves upon an HMM by using SVM techniques to find optimal separating hyperplanes that respect correlations between components of output labels. In addition, we expect the maximum-margin Markov framework introduced in (Anguelov et al., 2005) to be of use.

In addition we also want to test other techniques such as tiling the frequency windows, different feature vectors, and adding the derivatives of the frequency windows into the features. Tiling the frequency windows has been done by the discriminative approaches but would be an interesting addition to the PGM style. In addition, there should be some investigation into what tiling approach would work best. Current feature vectors include the eigenvalues of the STFT and the amplitude of the wavelets. An improved version may contain other smaller features that can be picked out. One of the most promising ideas that we have had is using the "derivative" of the frequency frames. So for a given frame, we would subtract out the frame that came before it (or something similar). This may allow us to improve recognition of note onset.

We will evaluate these methods according to the standard set in (Poliner & Ellis, 2006) and (Bock & Schedl, 2012). In both of these papers, they worked with a large library of mp3 files accompanied by midi files. Their software produced output which could be directly compared with the midi files and measured for: notes correctly predicted, notes failed to be predicted, and notes incorrectly predicted.

Possible extension given time: we would like to test a PGM style approach, which mimics the F0 style of signal subtraction but does so in a generative manner. We believe this will give several benefits: it will allow us to ground the very heuristic/hack style of an effective technique, it will allow us more flexibility in dealing with these algorithms, and it will apply knowledge from another field of study.

5. Resources

There are several datasets that should be applicable to our task on http://deeplearning.net/datasets/. In particular, MIDI sequences and associated synthesized audio files can be found on the web pages http://www.piano-midi.de/ and http://musedata.stanford.edu/. Researcher-generated datasets used in (Poliner & Ellis, 2006) and (Emiya et al., 2010) are available and used in other papers in the area.

User-created tabs of songs can be found on http://www.ultimate-guitar.com and 30-second samples of these can be found on http://us.7digital.com/.

We also expect all of our cited works to contribute valuable insight and techniques, whether in classification methods or in feature extraction.

6. Schedule

Our milestones for this project are:

16 Feb. Project proposal submitted.

15 Mar. Literature reviewed, datasets acquired.

5 Apr. Max-margin Markov model implemented for chords and single notes.

15 Apr. New features tested. Implement rudimentary PGM. Start HMM smoothing.

25 Apr. Try to wrap up the entire system - note transcription into HMM smoothing. Begin evaluation.

30 Apr. Class presentation.

12 May. Final project report submitted.

7. Literature review

We finished our literature review of the current methods used in piano music transcription. We found that the current state of the art methods used are ¡blah¿. In particular, there are a variety of signal processing techniques used to acquire feature sets are including:

Short-term Fourier transform (STFT). This involves taking a set of discrete time-localized windows of the original signal, and applying the Fourier transform to each window, so as to get an analysis of the frequencies roughly present at particular times (Bock & Schedl, 2012) (Poliner & Ellis, 2006).

Wavelet transform. This is similar to the STFT, but differs in that while the STFT decomposes time-local windows of the signal into frequency-local sinusoidal basis functions, the wavelet transform decomposes the whole signal into roughly time-and-frequency-local "wavelet" basis functions. The main effect of this is that it offers better time resolution at high frequencies (since high-frequency wavelets will tend to be more time-localized), and better frequency resolution at low frequencies. Because the shape of the wavelet functions can be customized, it is possible that wavelets could also be used to decompose a signal into more application-relevant bases.

Constant Q transform. The Constant Q transform is closely related to a particular type of wavelet transform: the Morlet wavelet transform. The Morlet wavelet is a wavelet consisting of a complex-exponential/sinusoidal wave attenuated by a Gaussian envelope. Similarly, the Constant Q transform takes a STFT, and instead of using discrete rectangular windows, uses Gaussian windows whose time-widths vary with the particular frequency bucket being analyzed. This transform is reportedly well suited for music-related applications, since the action of this transform mirrors that of human perception (Benetos & Dixon, 2011) (COSTANTINI et al., 2013).

Fine-tuned spectral features. Sometimes previous work has incorporated feature sets from spectrograph data that have been specially tuned for the transcription task. These include a pitch salience function (Benetos & Dixon, 2011), deeply learned feature sets (Nam et al., 2011), features acquired via sparse nonnegative matrix factorization (COSTANTINI et al., 2013), and features generated by a model of human inner ear hair cells (Marolt, 2004).

Using any of the above methods, we obtain spectograms, which are plots of which particular frequencies are present at any particular time within a signal. Observable features within a spectogram include:

Overtones. Each note, when played, consists of a fundamental frequency, which is the pitch of the note that we perceive, as well as a set of overtones at integer multiples of the original frequency.

Partials. A partial of a note is one of the theoretically pure sine wave constituents of that note. The fundamental frequency will be one of the partials, and each overtone will contribute another partial.

Rather than being pure sine waves, which would show up in the spectograph as straight think horizontal lines, in reality each partial will be a bit fuzzy and show up as blurred horizontal bars in the spectograph.

Missing fundamentals. In certain cases, a human will hear a note being played at a certain pitch even if the fundamental frequency is omitted from the signal. This is because the human audio processing system can infer the fundamental frequency from the present overtones.

Beat frequencies. When two distinct frequencies are played over each other, they give rise to a beat frequency equal to the difference of the two original frequencies. Human ears hear these as subjective tones, and some previous attempts have incorrectly detected beat frequencies as new note onsets (Marolt, 2004).

Difficult points for previous classifiers have also included octave errors, since notes that are an octave apart from each other are very similar in their partials (Bock & Schedl, 2012) (Poliner & Ellis, 2006). Previous methods also experienced difficulties in distinguishing which notes were onset and which were just held when some note onset event was detected (Marolt, 2004).

Previous work on this task have focused mostly around:

1. Markov networks, SVMs, and CRFs trained to detect particular notes or note events (Ryynanen & Klapuri, 2005) (Poliner & Ellis, 2006) (Gang et al., 2009).
2. Neural network approaches using recurrent networks (Marolt, 2004) (Bock & Schedl, 2012).
3. Estimation and subtraction, or expectation-maximization, of sets of overtones for detected fundamental frequencies (Benetos & Dixon, 2011).

Manual inspection of a spectrograph of a piano playing reveals that some partials can fade more quickly than others, and some partials can twinkle out of existence for brief moments before reappearing within the same note strike.

We were pointed to another paper (Gang et al., 2009) dealing with the use of max-margin methods to transcribe musical notes. Our approach differs substantially from theirs because they apply a max-margin markov network to classify sets of pre-segmented partials according to which instrument they belong to. Meanwhile we skip the analysis of partials and attempt to classify notes directly from the spectograms as features.

8. Refinement of ideas

We've narrowed our focus and elaborated our ideas since the project proposal. For instance, we've decided to focus on implementation of a max-margin markov network for transcribing particular notes given spectographic data as features. We understand the basic steps that need to be implemented (transforming the signal, possibily with downsampling, and feeding the transformed signal directly as features into a classifier). The structured output consists of a set of labels for which notes are present in each time window, exploiting correlations between notes both for different frequencies at the same time and for the same or similar frequencies at different times.

We've established a concrete plan to evaluate our results.

9. Data

We've obtained access to the datasets used in (Poliner & Ellis, 2006) and (Marolt, 2004). There are also a large number of piano mp3s and midi files available at `http:/www.piano-midi.de/`, though these require significant processing for use as experimental or training data.

References

Anguelov, Dragomir, Taskarf, B, Chatalbashev, Vassil, Koller, Daphne, Gupta, Dinkar, Heitz, Geremy, and Ng, Andrew. Discriminative learning of markov random fields for segmentation of 3d scan data. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pp. 169–176. IEEE, 2005.

Benetos, E. and Dixon, S. Polyphonic music transcription using note onset and offset detection. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 37–40, May 2011. doi: 10.1109/ICASSP.2011.5946322.

Bock, S and Schedl, Markus. Polyphonic piano note transcription with recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 121–124. IEEE, 2012.

COSTANTINI, GIOVANNI, TODISCO, MASSIMILIANO, and PERFETTI, RENZO. Nmf based dictionary learning for automatic transcription of polyphonic piano music. *WSEAS Transactions on Signal Processing*, 9(3), 2013.

Emiya, Valentin, Badeau, Roland, and David, Bertrand. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1643–1654, 2010.

Gang, Ren, Bocko, Mark F, Headlam, Dave, and Lundberg, Justin. Polyphonic music transcription employing max-margin classification of spectrograhic features. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pp. 57–60. IEEE, 2009.

Marolt, Matija. A connectionist approach to automatic transcription of polyphonic piano music. *Multimedia, IEEE Transactions on*, 6(3):439–449, 2004.

Nam, Juhan, Ngiam, Jiquan, Lee, Honglak, and Slaney, Malcolm. A classification-based polyphonic piano transcription approach using learned feature representations. In *ISMIR*, pp. 175–180, 2011.

Poliner, Graham E and Ellis, Daniel PW. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, 2007, 2006.

Raphael, Christopher. Automatic transcription of piano music. In *ISMIR*, 2002.

Ryynanen, Matti P and Klapuri, Anssi. Polyphonic music transcription using note event modeling. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pp. 319–322. IEEE, 2005.

Taskar, Ben, Guestrin, Carlos, and Koller, Daphne. Max-margin markov networks. In . MIT Press, 2003.