MIREX 2014 SUBMISSION: NOTE TRACKING BY MAXIMUM LIKELIHOOD SAMPLING

Zhiyao Duan
University of Rochester
Dept. Electrical and Computer Engineering
zhiyao.duan@rochester.edu

David Temperley
University of Rochester
Eastman School of Music
dtemperley@esm.rochester.edu

ABSTRACT

This document describes our three submissions to the Note Tracking subtask of the MIREX 2014 Multiple Fundamental Frequency Estimation & Tracking task. All submissions are built upon the frame-level Multi-pitch Estimation (MPE) results. Differently, the first submission "DT1" forms notes by just connecting pitch estimates that are close in both time and frequency, and then removes notes that are too short. Therefore, the formation of notes operates locally and does not consider interactions between different notes. The other two submissions "DT2" and "DT3" implement the idea presented in [6], which builds up a note sampling module based on results of "DT1" to sample subsets of notes as candidate transcriptions. It then builds a transcription evaluation module to select the best candidate as the final transcription. The evaluation module evaluates the likelihood of a transcription candidate in explaining the audio signal as a whole, considering the interactions between simultaneous notes. In "DT2", notes are sampled based on their salience and length; while in "DT3", the sampling also depends on the support of a note received from other notes in the transcription, which further considers interactions between notes.

1. INTRODUCTION

Automatic Music Transcription (AMT) is one of the fundamental problems in music information retrieval. On transcribing the pitch content, AMT can be performed at three levels from low to high: frame-level, note-level, and stream-level [4]. In this document, we describe our three submissions to the note tracking subtask of the MIREX 2014 Multiple Fundamental Frequency Estimation & Tracking task.

Most note tracking systems are built based on frame-level pitch estimates. The simplest way to convert frame-level pitch estimates to notes is to connect consecutive pitches into notes [2, 7, 9], followed by gap-filling [1, 4] and short-note-removal [1, 4]. This idea has also been implemented with more advanced techniques such as hidden Markov models [8] as well. This idea, albeit simple, forms notes independently without taking account the interaction with other simultaneous notes. Simultaneous notes that each is a good fit to the audio signal may not explain the signal very well as a whole, and vice versa.

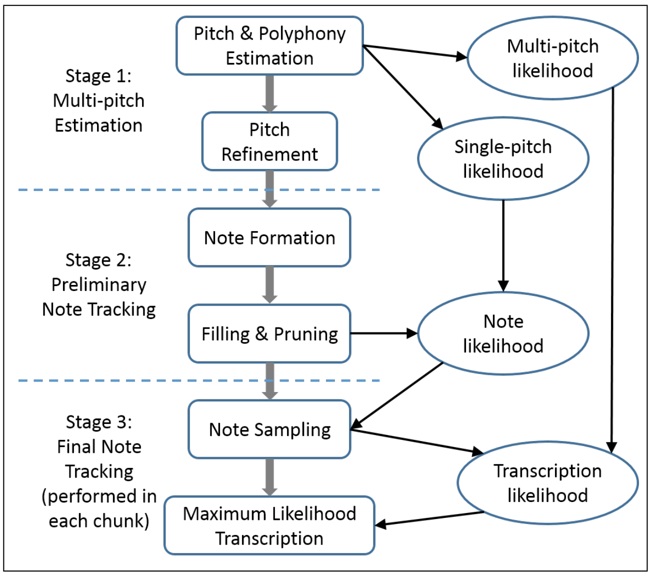

Figure 1. System overview of the proposed note-level transcription system in [6].

In [6], we proposed a new note-level music transcription system, whose architecture is shown in Figure 1. Similar to existing methods, it is built upon on frame-level pitch estimation (Stage 1) and adopts gap-filling and short-note-removal operations to generate preliminary note tracking results (Stage 2). However, a significant difference is in Stage 3. It contains two modules (note sampling, transcription evaluation) that together consider the interaction between simultaneous notes. The note sampling module samples subsets of notes of in the preliminary note tracking results, according to the note salience, length, and/or support from other notes. Each subset is treated as a transcription candidate. The transcription evaluation module then evaluates these candidates on how well they explains the audio signal as a whole, which considers the interaction between simultaneous notes in the candidate. Finally the best candidate is selected as the final note tracking result. This document describes our three submissions to the note tracking subtask.

2. SUBMISSION DETAILS

2.1 DT1

This system only contains the first two stages in Figure 1. Stage 1 implements the system proposed in [5]. We use the frame size to 46 ms and hop size to 10 ms. We set the pitch range to C2-B6 and the maximum instantaneous polyphony to 7. For Stage 2, we first connect pitches whose frequency difference is less than 0.3 semitones and time difference is less than 50 ms. Each connected component is then viewed as a note. Then notes shorter than 50 ms are removed.

2.2 DT2

This system is built upon the results obtained by DT1. A salience/likelihood value is calculated for each pitch estimate in Stage 1. This salience is then averaged over all pitches within a note to calculate the note salience/likelihood. The note salience/likelihood is used together with note length in the sampling step in Stage 3. Longer notes with high likelihood values are more likelihood to be sampled into transcription candidates.

To avoid the combinatorial explosion problem of the sampling space, the sampling is performed in each chunk independently instead of the entire piece. A chunk size of 100 frames (hence 1 second long) is used. To generate each transcription candidate, notes are sampled one by one without replacement. With more notes sampled, the instantaneous polyphony of the transcription increases. The process stops when the maximum instantaneous polyphony exceeds 6. All intermediate subsets during the sampling process that contain at least half of all notes in the chunk are kept as a transcription candidate. This diversifies the candidates to have different polyphony. For each chunk, in total 100 transcription candidates are generated.

A transcription likelihood is calculated for each candidate. It is defined as the product of the multi-pitch likelihood (defined in [5]) of all time frames in the transcription. Since multi-pitch likelihood considers interactions between simultaneous pitches, the transcription likelihood also considers interactions between simultaneous notes. The candidate with the highest likelihood is returned as the final transcription of the chunk. Transcriptions of different chunks are combined together by merging duplicating or overlapping notes with the same pitch into the final transcription of the piece.

2.3 DT3

This system is similar to DT2. The only difference is at the note sampling module. Notes are sampled not only according to their likelihood and length, but also according to the support they receive from other notes in the entire piece (global support) and from already sampled notes in the transcription of the chunk (local support). Both global and local supports basically count the number of notes that have a close onset time or the same pitch. Notes with high support values are more likely to be sampled. The detailed implementation is out of the scope of this extended abstract, but the idea is that it further considers interactions between notes.

3. RESULTS

The evaluation results will be added here after the evaluation is performed.

4. REFERENCES

[1] J.P. Bello, L. Daudet, M.B., Sandler, "Automatic piano transcription using frequency and time-domain information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2242-2251, 2006.

[2] E. Benetos and S. Dixon, "A shift-invariant latent variable model for automatic music transcription," *Computer Music J.*, vol. 36, no. 4, pp. 81-94, 2012.

[3] A. Dessein, A. Cont, G. Lemaitre, "Real-time polyphonic music transcription with nonnegative matrix factorization and beta-divergence," in *Proc. ISMIR*, 2010, pp. 489-494.

[4] Z. Duan, J. Han, and B. Pardo, "Multi-pitch streaming of harmonic sound mixtures," *IEEE Trans. Audio Speech Language Processing*, vol. 22, no. 1, pp. 1-13, 2014.

[5] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Audio Speech Language Processing*, vol. 18, no. 8, pp. 2121-2133, 2010.

[6] Z. Duan, D. Temperley, "Note-level music transcription by maximum likelihood sampling," in *Proc. ISMIR*, 2014.

[7] G. Grindlay and D. Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1159-1169, 2011.

[8] G. Poliner, and D. Ellis, "A discriminative model for polyphonic piano transcription," in *EURASIP J. Advances in Signal Processing*, vol. 8, pp. 154-162, 2007.

[9] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528-537, 2010.