MULTIPLE-F0 ESTIMATION AND NOTE TRACKING FOR MIREX 2012 USING A SHIFT-INVARIANT LATENT VARIABLE MODEL

Emmanouil Benetos and Simon Dixon
Centre for Digital Music, Queen Mary University of London
{emmanouilb,simond}@eecs.qmul.ac.uk

ABSTRACT

This submission for MIREX 2012 utilizes a shift-invariant latent variable model for multiple-F0 estimation and note tracking. It extends the shift-invariant probabilistic latent component analysis (SI-PLCA) method and employs several note templates from multiple orchestral instruments. By incorporating shift-invariance into the model along with the constant-Q transform as a time-frequency representation, tuning changes and frequency modulations such as vibrato can be better supported. Three variants of the system are utilized, one trained on orchestral instruments for multiple-F0 estimation, one trained on orchestral instruments plus piano for note tracking, and a final one trained on piano templates for piano-only note tracking.

1. INTRODUCTION

Automatic music transcription is the process of converting an audio recording into a symbolic representation, such as a piano-roll, a MIDI file or a music sheet. The creation of a system able to transcribe music produced by multiple instruments with a high level of polyphony continues to be an open problem in the literature, although monophonic pitch transcription is largely considered to be a solved problem. For a comprehensive overview on transcription approaches the reader is referred to [7], while a more recent overview of multiple-F0 estimation approaches can be found in [8].

Here, a system for automatic transcription of polyphonic music is utilized, which was first introduced in [2]. The system extends the shift-invariant probabilistic latent component analysis (SI-PLCA) method of [11]. This model is able to support the use of multiple pitch templates extracted from multiple sources. Using a log-frequency representation and frequency shifting, detection of notes that are non-ideally tuned, or that are produced by instruments that exhibit frequency modulations is made possible. Sparsity is also enforced in the model, in order to further constrain the transcription result and the instrument contribution in the production of pitches. It should be noted that this model was submitted by the authors for the MIREX 2011 evaluation [1], using a different set of instrument templates, model parameters, and a different note tracking procedure.

Emmanouil Benetos was supported by a Westfield Trust research studentship (Queen Mary University of London).

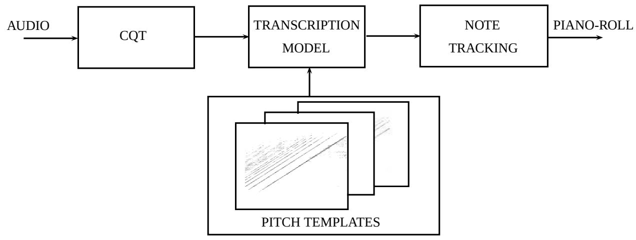

Figure 1. Diagram for the proposed polyphonic transcription system.

2. TRANSCRIPTION SYSTEM

The goal of the utilized transcription system is to provide a framework that supports multiple templates per pitch, in contrast to the relative pitch tracking method in [9], as well as multiple templates per musical instrument. In addition, the contribution of each instrument source is not constant for the whole recording as in [9], but is time-dependent. Also, its goal is to exploit the benefits given by a shift-invariant model coupled with a log-frequency representation, in contrast to the transcription method in [6], for detecting notes that exhibit frequency modulations and tuning changes.

In subsection 2.1, the extraction of pitch templates for various instruments is presented. The main transcription model is presented in subsection 2.2, while the system variants used for evaluation are discussed in subsection 2.3. A diagram of the proposed transcription system is depicted in Fig. 1.

2.1 Extracting Pitch Templates

Firstly, spectral templates are extracted for various instruments, for each note, using their whole note range. Isolated note samples from three different piano types were extracted from the MAPS dataset [4] and templates for other orchestral instruments were extracted from monophonic recordings from the RWC database [5]. For extracting the note templates, the constant-Q transform (CQT) was computed [10] with spectral resolution of 60 bins per octave. Afterwards, the standard PLCA model of [11] us-

Instrument	Lowest note	Highest note
Bassoon	34	72
Cello	26	81
Clarinet	50	89
Flute	60	96
Guitar	40	76
Horn	41	77
Oboe	58	91
Piano	21	108
Tenor Sax	44	75
Violin	55	100

Table 1. MIDI note range of the instrument templates used in the transcription system.

ing only one component z was employed in order to extract the spectral template $P(\omega|z)$, where ω is the log-frequency index. In Table 1, the pitch range of each instrument used for template extraction is shown.

2.2 Transcription Model

Utilizing the extracted instrument templates and by extending the shift-invariant PLCA algorithm, a model is proposed which supports the use of multiple pitch and instrument templates in a convolutive framework, thus supporting tuning changes and frequency modulations. By considering the input CQT spectrum as a probability distribution $P(\omega,t)$, the proposed model can be formulated as:

$$P(\omega,t) = P(t)\sum_{p,s} P(\omega|s,p) *_\omega P(f|p,t)P(s|p,t)P(p|t) \qquad (1)$$

where $P(\omega|s,p)$ is the spectral template that belongs to instrument s and MIDI pitch $p = 21,\ldots,108$, $P(f|p,t)$ is the time-dependent impulse distribution that corresponds to pitch p, $P(s|p,t)$ is the instrument contribution for each pitch in a specific time frame, and $P(p|t)$ is the pitch probability distribution for each time frame.

By removing the convolution operator, the model of (1) can be expressed as:

$$P(\omega,t) = P(t)\sum_{p,f,s} P(\omega-f|s,p)P(f|p,t)P(s|p,t)P(p|t) \qquad (2)$$

In order to only utilize each template $P(\omega|s,p)$ for detecting the specific pitch p, the convolution of $P(\omega|s,p) *_\omega P(f|p,t)$ takes place using an area spanning one semitone around the ideal position of p. Since 60 bins per octave are used in the CQT spectrogram, f has a length of 5.

The various parameters in (1) can be estimated using iterative update rules derived from the EM algorithm. For the expectation step the update rule is:

$$P(p,f,s|\omega,t) = \frac{P(\omega-f|s,p)P(f|p,t)P(s|p,t)P(p|t)}{\sum_{p,f,s} P(\omega-f|s,p)P(f|p,t)P(s|p,t)P(p|t)} \qquad (3)$$

For the maximization step, the update equations for the

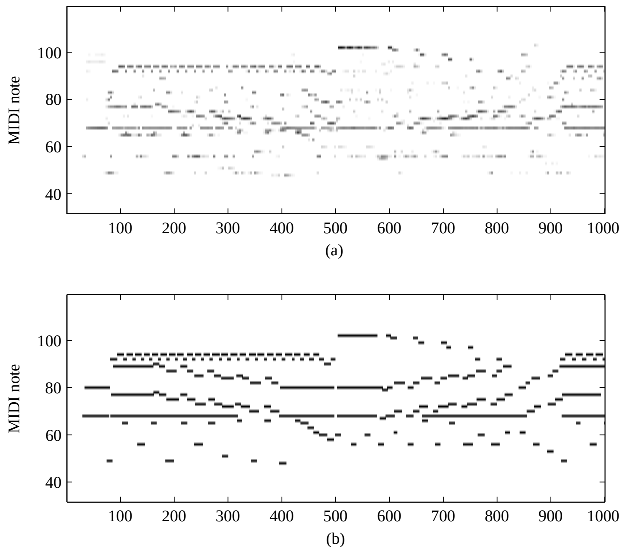

Figure 2. (a) The transcription matrix $P(p,t)$ of the first 10s of the MIREX woodwind quintet. (b) The pitch ground truth of the same recording. The abscissa corresponds to 10ms.

proposed model are:

$$P(\omega|s,p) = \frac{\sum_{f,t} P(p,f,s|\omega+f,t)P(\omega+f,t)}{\sum_{\omega,t,f} P(p,f,s|\omega+f,t)P(\omega+f,t)} \qquad (4)$$

$$P(f|p,t) = \frac{\sum_{\omega,s} P(p,f,s|\omega,t)P(\omega,t)}{\sum_{f,\omega,s} P(p,f,s|\omega,t)P(\omega,t)} \qquad (5)$$

$$P(s|p,t) = \frac{\sum_{\omega,f} P(p,f,s|\omega,t)P(\omega,t)}{\sum_{s,\omega,f} P(p,f,s|\omega,t)P(\omega,t)} \qquad (6)$$

$$P(p|t) = \frac{\sum_{\omega,f,s} P(p,f,s|\omega,t)P(\omega,t)}{\sum_{p,\omega,f,s} P(p,f,s|\omega,t)P(\omega,t)} \qquad (7)$$

It should be noted that since the instrument-pitch templates have been extracted during the training stage, the update rule for the templates (4) is not used, but is included for the sake of completeness. Using these constant templates, convergence is quite fast, usually requiring 10-20 iterations. The resulting piano-roll transcription matrix is given by:

$$P(p,t) \quad = \quad P(t)P(p|t) \qquad (8)$$

In Fig. 2, the transcription matrix $P(p,t)$ for an excerpt of the MIREX multi-F0 woodwind quintet recording can be seen, along with the corresponding pitch ground truth.

In order for the algorithm to provide as meaningful solutions as possible, sparsity is encouraged on transcription matrix $P(p|t)$, expecting that only few notes are present at a given time frame. In addition, sparsity can be enforced to matrix $P(s|p,t)$, meaning that for each pitch at a given time frame, only a few instrument sources contributes to its production. The same technique used in [6] was employed for controlling sparsity, by modifying the update equations

(6) and (7):

$$P(s|p,t) = \frac{\left(\sum_{\omega,f} P(p,f,s|\omega,t)P(\omega,t)\right)^{\alpha}}{\sum_s \left(\sum_{\omega,f} P(p,f,s|\omega,t)P(\omega,t)\right)^{\alpha}} \quad (9)$$

$$P(p|t) = \frac{\left(\sum_{\omega,f,s} P(p,f,s|\omega,t)P(\omega,t)\right)^{\beta}}{\sum_p \left(\sum_{\omega,f,s} P(p,f,s|\omega,t)P(\omega,t)\right)^{\beta}} \quad (10)$$

By setting $\alpha, \beta > 1$, the entropy in matrices $P(s|p,t)$ and $P(p|t)$ is lowered and sparsity is enforced.

Finally, note events are extracted by performing thresholding on $P(p,t)$ followed by minimum duration pruning, set to 50ms as in [3].

2.3 System Variants

Three variants of the system are utilized for the MIREX 2012 evaluation; one trained on orchestral instruments only for the multiple-F0 estimation task (BD1), one trained on orchestral instruments plus piano for the note tracking task (BD2), and a system trained on the three sets of piano templates for the piano-only note tracking task (BD3). In all cases, $\beta = 1.1$, while $\alpha = 1.3$ for BD1 and BD2 and 1.0 for BD3. For computational speed purposes, the number of iterations in all variants was set to 12.

3. RESULTS

- For the Multiple Fundamental Frequency Estimation task, the submitted system (BD1) ranked 2nd out of 4 groups, reporting an accuracy of 57.9% and a chroma accuracy of 60.3%. Compared to the system submitted for the MIREX 2011 task [1] this system reports an accuracy increase of +0.5%.

- For the Note Tracking task, the submitted system (BD2) ranked 3rd out of 6 groups. Compared to last year's submission [1], there is an improvement of +2.6% in terms of onset-offset F-measure.

- For the Piano-only Note Tracking task, the submitted system (BD3) ranked 2nd out of 6 groups. Compared to last year's submission [1], there is an improvement of +7.2% in terms of onset-offset F-measure.

4. REFERENCES

[1] E. Benetos and S. Dixon. Multiple-F0 estimation and note tracking using a convolutive probabilistic model. In *Music Information Retrieval Evaluation eXchange*, Miami, Florida, USA, October 2011.

[2] E. Benetos and S. Dixon. Multiple-instrument polyphonic music transcription using a convolutive probabilistic model. In *8th Sound and Music Computing Conference*, pages 19–24, Padova, Italy, July 2011.

[3] A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *11th International Society for Music Information Retrieval Conference*, pages 489–494, Utrecht, Netherlands, August 2010.

[4] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1643–1654, August 2010.

[5] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: music genre database and musical instrument sound database. In *International Conference on Music Information Retrieval*, Baltimore, USA, October 2003.

[6] G. Grindlay and D. Ellis. A probabilistic subspace model for multi-instrument polyphonic transcription. In *11th International Society for Music Information Retrieval Conference*, pages 21–26, Utrecht, Netherlands, August 2010.

[7] A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription*. Springer-Verlag, New York, 2006.

[8] M. Müller, D.P.W. Ellis, A. Klapuri, and G. Richard. Signal processing for music analysis. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1088–1110, October 2011.

[9] G. Mysore and P. Smaragdis. Relative pitch estimation of multiple instruments. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 313–316, Taipei, Taiwan, April 2009.

[10] C. Schörkhuber and A. Klapuri. Constant-Q transform toolbox for music processing. In *7th Sound and Music Computing Conf.*, Barcelona, Spain, July 2010.

[11] P. Smaragdis, B. Raj, and M. Shashanka. Sparse and shift-invariant feature extraction from non-negative data. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2069–2072, Las Vegas, USA, April 2008.