MULTIPLE-F0 ESTIMATION AND NOTE TRACKING FOR MIREX 2014 USING A VARIABLE-Q TRANSFORM

Emmanouil Benetos and Tillman Weyde
Music Informatics Research Group, Department of Computer Science, City University London
{emmanouil.benetos.1, t.e.weyde}@city.ac.uk

ABSTRACT

In this submission for MIREX 2014 we utilize an efficient latent variable model for multiple-F0 estimation and note tracking, which uses a variable-Q transform as a time-frequency representation. In contrast to the constant-Q transform of the 2013 submission, the variable-Q transform is able to provide a better temporal resolution in low frequencies with the same frequency resolution. The model is based on probabilistic latent component analysis and uses pre-extracted note templates from multiple instruments. The templates are also pre-shifted across log-frequency in order to support pitch deviations and frequency modulations. In contrast to typical shift-invariant models which need to perform convolutions for estimating model parameters, the present model avoids such computations by using the aforementioned pre-shifted templates. Three system variants are submitted: one trained on orchestral instruments for multiple-F0 estimation, one trained on orchestral instruments and 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 acoustic musical signal into some form of music notation [5]. The problem is considered to be one of the most important ones in the field of music information retrieval (MIR), with applications beyond the field, such as in computational musicology. However, the creation of an automated system able to transcribe multiple-instrument polyphonic music without any constraints on instrument identities or on the level of polyphony continues to be an open problem in the field [2].

In this MIREX submission for the Multiple-F0 Estimation and Note Tracking tasks, we utilise the polyphonic music transcription system that was first introduced in [1]. In contrast to last year's submission though, which utilised as input time-frequency representation the constant-Q transform [6], in this submission we use a variable-Q transform (VQT), as proposed in [7]. Compared to the constant-Q transform, the VQT is able to provide a better temporal resolution in low frequencies with the same frequency resolution. The model extends the probabilistic latent component analysis method [8] by supporting the use of pre-extracted and pre-shifted templates for multiple instruments. By using shift-invariance in the log-frequency domain, the system can support the detection of small pitch changes, tuning deviations, or frequency modulations. The employed model is also a variant of the shift-invariant probabilistic latent component analysis method cite Smaragdis09, where the convolution operations only occur in a training stage, thus making the model computationally efficient.

E. Benetos is supported by a City University London Research Fellowship.

2. TRANSCRIPTION SYSTEM

2.1 Pitch template extraction

Pre-extracted and pre-shifted spectral templates are extracted for various instruments, namely bassoon, clarinet, saxophone, violin, flute, horn, oboe, guitar, cello, and piano. For extracting the templates, we used isolated note samples from the RWC database [4]. As a time-frequency representation, we use the variable-Q transform (VQT) time-frequency representation proposed in [7], with a log-spectral resolution of 60 bins per octave and $\gamma = 30$. A comparison between the constant-Q and variable-Q representations can be seen in Figures 1 and 2, respectively, where the lower pitches played by the bassoon are more clearly located temporally for the VQT representation. For extracting the templates, we used the standard PLCA model [8] with one component. For pre-shifting the templates, we shift each note template -40, -20, 20, and 40 cent from the ideal tuning position (we also keep the original ideally tuned template).

2.2 Transcription model

The proposed model takes as input a log-frequency spectrogram $V_{\omega,t}$ (ω is the log-frequency index and t is the time index) and approximates it as a bivariate probability distribution $P(\omega, t)$. $P(\omega, t)$ is decomposed as:

$$P(\omega, t) = P(t) \sum_{p,f,s} P(\omega|s, p, f) P_t(f|p) P_t(s|p) P_t(p) \tag{1}$$

where p is the pitch index in semitone scale, s is the instrument source index, and f is the log-frequency shifting factor. $P(t)$ is the log-spectrogram energy, which is a known quantity. $P(\omega|s,p,f)$ are the pre-extracted and pre-shifted log-spectral templates for instrument s and pitch p. $P_t(f|p)$ is the time-varying log-frequency shifting factor for each pitch, which corresponds to one of the 5 shifts for each note template (-40,-20,0,20,and 40 cent centered at the ideal tuning position). $P_t(s|p)$ is the instrument contribution probability for each pitch at a given time frame, and finally $P_t(p)$ is the time-varying pitch activation, which is used for estimating the final transcription.

Unknown model parameters ($P_t(f|p)$, $P_t(s|p)$, $P_t(p)$) can be estimated in an iterative fashion using the expectation-maximization (EM) algorithm [3]. For the expectation step, the following posterior is computed:

$$P_t(p,f,s|\omega) = \frac{P(\omega|s,p,f)P_t(f|p)P_t(s|p)P_t(p)}{\sum_{p,f,s} P(\omega|s,p,f)P_t(f|p)P_t(s|p)P_t(p)} \tag{2}$$

For the maximization step, unknown parameters are updated using the posterior from (2):

$$P_t(f|p) = \frac{\sum_{\omega,s} P_t(p,f,s|\omega)V_{\omega,t}}{\sum_{f,\omega,s} P_t(p,f,s|\omega)V_{\omega,t}} \tag{3}$$

$$P_t(s|p) = \frac{\sum_{\omega,f} P_t(p,f,s|\omega)V_{\omega,t}}{\sum_{s,\omega,f} P_t(p,f,s|\omega)V_{\omega,t}} \tag{4}$$

$$P_t(p) = \frac{\sum_{\omega,f,s} P_t(p,f,s|\omega)V_{\omega,t}}{\sum_{p,\omega,f,s} P_t(p,f,s|\omega)V_{\omega,t}} \tag{5}$$

Eqs. (2)-(5) are iterated until convergence; for the submitted system we set the number of iterations to 30. As in [1], we also enforced sparsity constraints on $P_t(p)$ and $P_t(s|p)$ in order to control the polyphony level and the number of instruments contributing to produced notes in the resulting transcription. The resulting transcription is given by $P(p,t) = P(t)P_t(p)$. An example for the pitch activation can be seen in Fig. 3. After performing 7-sample median filtering for note smoothing, thresholding is performed on $P(p,t)$ followed by minimum note duration pruning set to 40ms in order to convert $P(p,t)$ into a binary piano-roll representation. As an example, the $P(p,t)$ is depicted for the first 10sec of the MIREX multiF0 woodwind quintet. The flute trills in the upper register are particularly evident.

The system is quite efficient computationally, being able to produce a transcription in about 1.5 × real-time in a Sony VAIO S15 laptop (e.g. for a 30sec recording it requires 45sec). The code for the transcription model (using the CQT as input) is available online [1] , both in a CPU-based version as well as in a GPU-based version, which is about 3 times faster.

2.3 System variants

Three variants of the system are utilized for the MIREX 2014 evaluation; one trained on the instruments listed in subsection 2.1 minus piano for the multiple-F0 estimation task (BW1), one trained on the complete instrument set

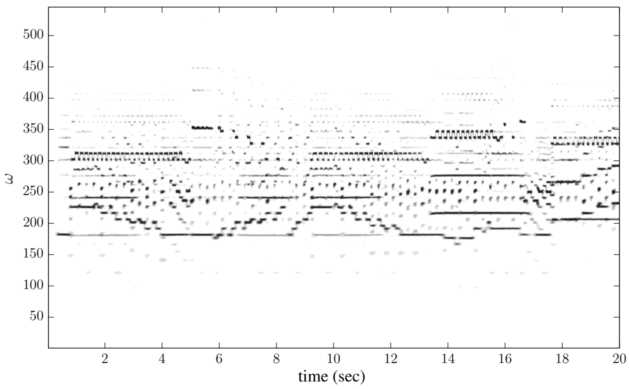

Figure 1. The constant-Q transform spectrogram for the first 20sec of the MIREX multiF0 development recording.

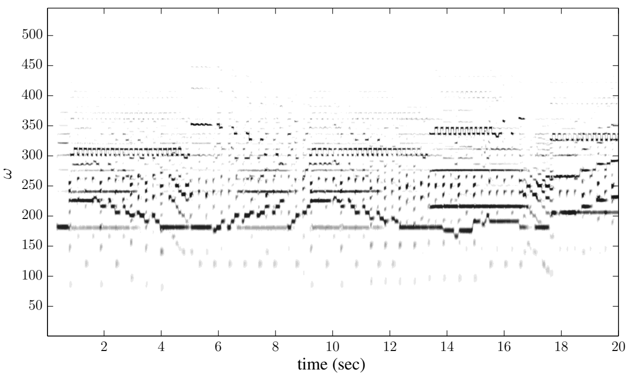

Figure 2. The variable-Q transform spectrogram for the first 20sec of the MIREX multiF0 development recording.

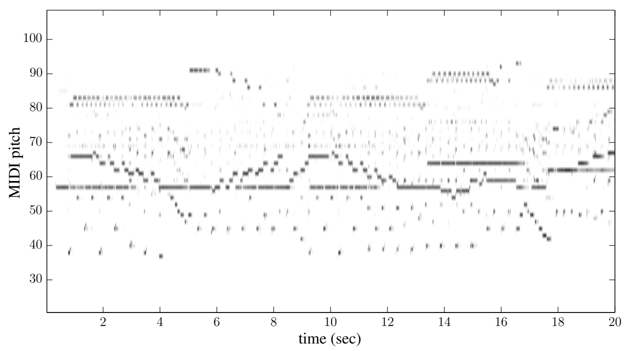

Figure 3. The pitch activation $P(\omega,t)$ for the first 20sec of the MIREX multiF0 development recording.

[1] https://code.soundsoftware.ac.uk/projects/amt_mssiplca_fast

for the note tracking task (BW2), and a system trained on piano templates only for the piano-only note tracking task (BW3).

3. RESULTS

The BW1 system ranked 2nd (out of 5 teams) for the Multiple-F0 Estimation task. The BW2 system ranked 3rd (out of 8 teams) for the multi-instrument note tracking task. The BW3 system ranked 2nd (out of 9 teams) for the piano note tracking task. Compared to the submitted system by the same team for MIREX 2013 (where the CQT was used as T/F representation), an improvement of +3.61% in terms of onset-offset F-measure is reported for the multi-instrument Note Tracking task, and an improvement of +1.1% is reported for the piano Note Tracking task. These results indicate the increased temporal precision of the VQT representation over the CQT.

4. REFERENCES

[1] E. Benetos, S. Cherla, and T. Weyde. An effcient shift-invariant model for polyphonic music transcription. In *6th International Workshop on Machine Learning and Music*, September 2013.

[2] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 2013. accepted.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[4] 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.

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

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

[7] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler. A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution. In *AES 53rd Conference on Semantic Audio*, page 8 pages, London, UK, January 2014.

[8] P. Smaragdis, B. Raj, and Ma. Shashanka. A probabilistic latent variable model for acoustic modeling. In *Neural Information Processing Systems Workshop*, Whistler, Canada, December 2006.