MULTIPLE-F0 ESTIMATION AND NOTE TRACKING FOR MIREX 2015 USING A SOUND STATE-BASED SPECTROGRAM FACTORIZATION MODEL

Emmanouil Benetos
Centre for Digital Music
Queen Mary University of London
emmanouil.benetos@qmul.ac.uk

Tillman Weyde
Department of Computer Science
City University London
t.e.weyde@city.ac.uk

ABSTRACT

In this submission for MIREX 2015 we utilize an efficient latent variable model for multiple-F0 estimation and note tracking, which uses an ERB-scale time-frequency representation as input. The transcription model is based on probabilistic latent component analysis and uses pre-extracted spectral templates corresponding to *sound states* for several instruments. Three system variants are submitted: one trained on orchestral instruments for multiple-F0 estimation, one trained on orchestral instruments 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 [8]. The problem is considered to be one of the most important ones in the fields of music information retrieval (MIR) and music signal processing, with applications in computational musicology, interactive music systems, and organisation of music collections. 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 proposed in [4]. In contrast to the aforementioned model, which utilised as input time-frequency representation the variable-Q transform (VQT) [9], in this submission we use an Equivalent Rectangular Bandwidth (ERB) scale time-frequency representation, which was also used for multi-pitch detection in [12]. Preliminary experiments showed that the ERB representation offers increased temporal resolution (which is particularly useful for the note tracking task), whilst offering a compact representation, suitable for an efficient system (250 frequency bins compared to the 545 log-frequency bins of the VQT). This is however at the cost of losing the shift-invariance abilities of [4], since the ERB scale is non-linear with respect to log-frequency.

The core model is based on probabilistic latent component analysis (PLCA) and supports the use of *sound state* spectral templates, which represent the temporal evolution of each note (e.g. attack, sustain, decay). It decomposes the input representation into a series of spectral templates per sound state, pitch, and instrument, as well as probability distributions for sound state, pitch, and instrument activations. As explained in [1], a sound state represents different segments in the temporal evolution of a note; e.g. for a piano, different sound states can correspond to the attack, sustain, and decay.

2. TRANSCRIPTION SYSTEM

2.1 Pitch template extraction

Pre-extracted sound state spectral templates are extracted for various instruments, namely alto saxophone, bass, bassoon, cello, clarinet, flute, guitar, horn, oboe, piano, and violin. For extracting the templates, we used isolated note samples from the RWC database [7] for all instruments apart from piano, where we used note samples from the MAPS database [6]. The complete note range of the instruments (given available data) is used.

As a time-frequency representation, we use the auditory-motivated Equivalent Rectangular Bandwidth (ERB) scale time-frequency representation that was used for multi-pitch detection in [12]. In short, the input signal is passed through a set of 250 filters, with frequencies linearly spaced between 5Hz and 10.8kHz on the ERB scale. Each subband is partitioned into disjoint 23ms time frames, and the rms magnitude $V_{\omega,t}$ is computed for each frame (ω is the frequency index, t is the time index). This leads to a compact representation of 250 frequency bins per frame (compared to the 545 bins of the VQT representation, or even higher numbers for STFT representations). For extracting the templates, we used the standard PLCA model [10] with one component.

E. Benetos is supported by a Royal Academy of Engineering Research Fellowship (grant no. RF/128).

2.2 Transcription model

The system takes as input the ERB representation of an audio recording ($V_{\omega,t}$) and approximates it as a bivariate probability distribution $P(\omega,t)$. In the model, $P(\omega,t)$ is decomposed into a series of spectral templates per sound state, pitch, and instrument, as well as probability distributions for sound state, pitch, and instrument. As explained in [1], a sound state represents different segments in the temporal evolution of a note; e.g. for a piano, different sound states can correspond to the attack, sustain, and decay.

The model is formulated as:

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

where q denotes the sound state, p denotes pitch, and s denotes instrument source. $P(t) = \sum_{\omega} V_{\omega,t}$, which is a known quantity. $P(\omega|q,p,s)$ is a 4-dimensional tensor that represents the pre-extracted spectral templates per sound state q, pitch p and instrument s. $P_t(s|p)$ is the instrument source contribution per pitch over time, $P_t(q|p)$ is the time-varying sound state activation per pitch, and finally $P_t(p)$ is the pitch activation, which is essentially the resulting multi-pitch detection output.

The unknown model parameters ($P_t(s|p)$, $P_t(p)$, $P_t(q|p)$) can be iteratively estimated using the expectation-maximization (EM) algorithm [5]. For the *Expectation* step, the following posterior is computed:

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

For the *Maximization* step, unknown model parameters are updated using the posterior from (2):

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

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

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

Eqs. (2)-(5) are iterated until convergence; 30 iterations are set. No update rule for the sound state templates $P(\omega|q,p,s)$ is included, since they are considered fixed in the model. As in [1], we also incorporated sparsity constraints on $P_t(p)$ and $P_t(s|p)$ in order to control the polyphony level and the instrument contribution in the resulting transcription. The resulting multi-pitch detection output is given by $P(p,t) = P(t)P_t(p)$. After performing 5-sample median filtering for note smoothing, thresholding is performed on $P(p,t)$ followed by minimum note duration pruning set to 20ms in order to convert $P(p,t)$ into a binary piano-roll representation.

An example, the ERB representation for the MIREX multiF0 development woodwind quintet recording can be seen in Fig. 1; the corresponding pitch activation $P(p,t)$ can be seen in Fig. 2. The flute trills in the upper register are particularly evident.

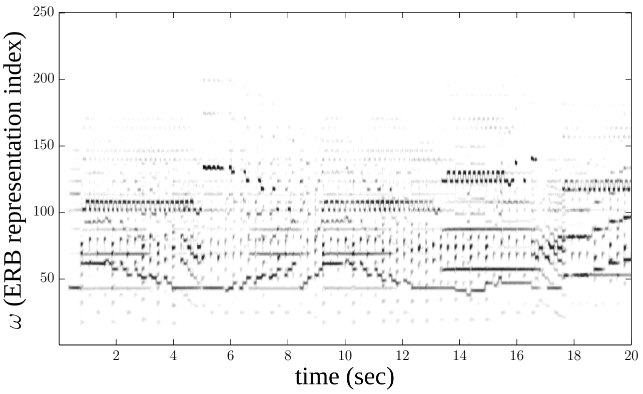

Figure 1. The ERB T/F representation for the first 20sec of the MIREX multiF0 development recording.

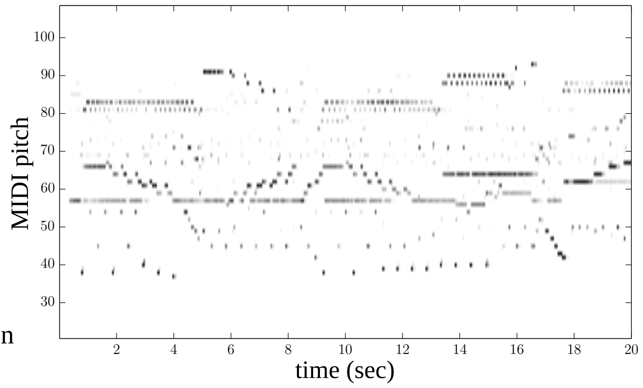

Figure 2. The pitch activation $P(\omega,t)$ for the first 20sec of the MIREX multiF0 development recording.

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 VQT representation as input) is available online [1].

2.3 System variants

Three variants of the system are utilized for the MIREX 2015 evaluation; one trained on the complete instruments set listed in subsection 2.1 for the multiple-F0 estimation task (BW1), one trained on the complete instrument set for the note tracking task (BW2), and a system trained on piano templates only for the piano-only note tracking task (BW3).

3. RESULTS

This year, evaluation was performed on two datasets: the MIREX dataset (used in previous years) and the Su dataset [11].

On the MIREX dataset, the BW1 system ranked 1st for the Multiple-F0 Estimation task (Task 1); the BW2 system

[1] `https://code.soundsoftware.ac.uk/projects/amt_plca_5d`

ranked 1st for the (multi-instrument) Note Tracking task (Task 2); and the BW3 system ranked 1st for the Piano-only Note Tracking task (Task 3). Compared to last year's submission by the same team of Benetos & Weyde [3], the current system had an improvement of +1.7% for Task 2 and an improvement of +6.6% for Task 3 (both in terms of onset-based F-measure).

On the Su dataset [11], our submission ranked 2nd for the Multiple-F0 Estimation task (Task 1); it ranked 1st for the (multi-instrument) Note Tracking task (Task 2); and ranked 1st for the Piano-only Note Tracking task (Task 3).

4. REFERENCES

[1] E. Benetos and S. Dixon. Multiple-instrument polyphonic music transcription using a temporally-constrained shift-invariant model. *Journal of the Acoustical Society of America*, 133(3):1727–1741, March 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*, 41(3):407–434, December 2013.

[3] E. Benetos and T. Weyde. Multiple-F0 estimation and note tracking for MIREX 2014 using a variable-Q transform. In *Music Information Retrieval Evaluation eXchange*, Taipei, Taiwan, October 2014.

[4] E. Benetos and T. Weyde. An efficient temporally-constrained probabilistic model for multiple-instrument music transcription. In *16th International Society for Music Information Retrieval Conference*, Malaga, Spain, October 2015.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] L. Su and Y.-H. Yang. Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription. In *International Symposium on Computer Music Multidisciplinary Research*, June 2015.

[12] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):528–537, March 2010.