AN EFFICIENT TEMPORALLY-CONSTRAINED PROBABILISTIC MODEL FOR MULTIPLE-INSTRUMENT MUSIC TRANSCRIPTION

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 paper, an efficient, general-purpose model for multiple instrument polyphonic music transcription is proposed. The model is based on probabilistic latent component analysis and supports the use of *sound state* spectral templates, which represent the temporal evolution of each note (e.g. attack, sustain, decay). As input, a variable-Q transform (VQT) time-frequency representation is used. Computational efficiency is achieved by supporting the use of pre-extracted and pre-shifted sound state templates. Two variants are presented: without temporal constraints and with hidden Markov model-based constraints controlling the appearance of sound states. Experiments are performed on benchmark transcription datasets: MAPS, TRIOS, MIREX multiF0, and Bach10; results on multi-pitch detection and instrument assignment show that the proposed models outperform the state-of-the-art for multiple-instrument transcription and is more than 20 times faster compared to a previous sound state-based model. We finally show that a VQT representation can lead to improved multi-pitch detection performance compared with constant-Q representations.

1. INTRODUCTION

Automatic music transcription is defined as the process of converting an acoustic music signal into some form of musical notation [16] and is considered a fundamental problem in the fields of music information retrieval and music signal processing. The core problem of automatic music transcription is multi-pitch detection (i.e. the detection of multiple concurrent pitches), which despite recent advances is still considered an open problem, especially for a large polyphony level and multiple instruments.

A large subset of music transcription approaches use *spectrogram factorization* methods such as non-negative matrix factorization (NMF) and probabilistic latent component analysis (PLCA), which decompose an input time-frequency representation into a series of note templates and note activations. Several variants of the above methods propose more complex formulations compared to the original NMF/PLCA models, and also add musically- and acoustically-meaningful constraints. Such spectrogram factorization methods include amongst others [4, 8, 10, 13, 15, 18, 24]. Issues related to spectrogram factorization methods include: the choice of an input time-frequency representation, the ability to recognize instruments, the support of tunings beyond twelve-tone equal temperament, the presence or absence of a pre-extracted dictionary, the incorporation of any constraints, as well as computational efficiency (given ever-expanding collections and archives of music recordings).

In this paper, a model for multiple-instrument transcription is proposed, which uses a 5-dimensional dictionary of *sound state* spectral templates (sound states correspond to the various states in the evolution of a note, such as the attack, sustain, and decay states). The proposed model is based on PLCA and decomposes an input time frequency representation (in this case, a variable-Q transform spectrogram) into a series of probability distributions for pitch, instrument, tuning, and sound state activations. This model is inspired by a convolutive model presented in [4] that used a 4-dimensional dictionary and was able to transcribe a recording at 60 $\times$ real-time. This model uses pre-shifted spectral templates across log-frequency, thus introducing a new dimension in the dictionary and eliminating the need for convolutions. Thus, tuning deviations from equal temperament are supported and at the same time this model only uses linear operations that result in a system that is more than 20 times faster compared to the system of [4]. In addition, temporal constraints using pitch-wise hidden Markov models (HMMs) are incorporated, in order to model the evolution of a note as a sequence of sound states. Experiments are performed on several transcription datasets (MAPS, MIREX multiF0, Bach10, TRIOS) and experimental results for the multi-instrument datasets using the proposed system outperform the state-of-the-art. Finally, we show that a VQT representation leads to an improvement in transcription performance compared to the more common constant-Q transform (CQT) representation, especially on the detection of lower pitches. Code for the proposed model is also supplied (cf. Section 4).

The outline of this paper is as follows. The proposed system is presented in Section 2. The employed training and test datasets, evaluation metrics, and experimental re-

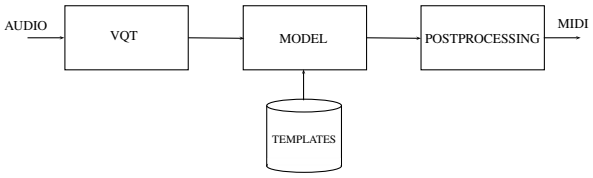

Figure 1. Diagram for the proposed system.

sults are shown in Section 3. Finally, a discussion on the proposed system followed by future directions is made in Section 4.

2. PROPOSED SYSTEM

2.1 Motivation

The overall aim of the proposed work is the creation of a system for automatic transcription of polyphonic music, that supports the identification of instruments along with multiple pitches, supports tunings beyond twelve-tone equal temperament along with frequency modulations, is able to model the evolution of each note (as a temporal succession of *sound states*), and is finally computationally efficient. The proposed system is based on work carried out in [4], which relied on a convolutive PLCA-based model and a 4-dimensional sound state dictionary. The aforementioned model was able to transcribe recordings at approximately 60 $\times$ real-time (i.e. for a 1min recording, transcription took 60min). This paper proposes an alternative linear model able to overcome the computational bottleneck of using a convolutive model, which is supported by the use of a 5-dimensional dictionary of pre-extracted and pre-shifted sound state spectral templates, at the same time providing the same benefits with the model of [4]. Finally, this paper proposes the use of a variable-Q transform (VQT) representation, in contrast with the more common constant-Q transform (CQT) or linear frequency representations (a detailed comparison is made in Section 3). On related work, a linear model that used a 4-dimensional dictionary which did not support sound state templates or temporal constraints was proposed in [3].

In Fig. 1, a diagram for the proposed system can be seen. As motivation on the use of sound state templates, two log-frequency representations for a G1 piano note are shown in Fig. 2; it is clear that the note evolves from an attack/transient state to a steady state, and finally to a decay state. Fig. 3 shows 3 spectral templates extracted for the same note, which correspond to the 3 sound states (the lower corresponds to the attack state, the middle to the steady state and the top to the decay state).

2.2 PLCA-based model

The first variant of the proposed system takes as input a normalised 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 $P(\omega, t)$. In this work, $V_{\omega,t}$ is a variable-Q time-frequency representation with a resolution of 60 bins/octave and minimum frequency

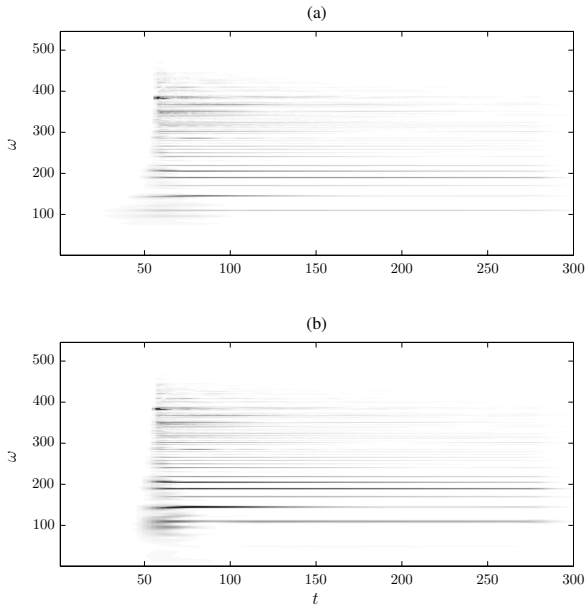

Figure 2. (a) The CQT spectrogram of a G1 piano note. (b) The VQT spectrogram for the same note.

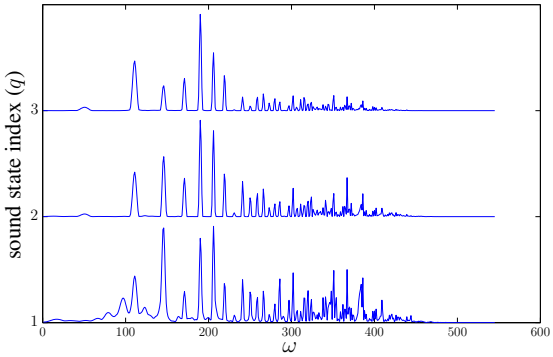

Figure 3. Sound state spectral templates for a G1 piano note (extracted using a VQT representation).

of 27.5Hz, computed using the method of [22]. As discussed in [22], a variable-Q representation offers increased temporal resolution in lower frequencies compared with a constant-Q representation. At the same time, a log-frequency transform represents pitch in a linear scale (where inter-harmonic spacings are constant for all pitches), thus allowing for pitch changes to be represented by shifts across the log-frequency axis.

In the model, $P(\omega, t)$ is decomposed into a series of log-frequency spectral templates per sound state, pitch, instrument, and log-frequency shifting (which indicates deviation with respect to equally tempered tuning), as well as probability distributions for sound state, pitch, instrument, and tuning activations. As explained in [4], 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:

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

where q denotes the sound state, p denotes pitch, s denotes instrument source, and f denotes log-frequency shifting. $P(t)$ is the energy of the log-spectrogram, which is a known quantity. $P(\omega|q,p,f,s)$ is a 5-dimensional tensor that represents the pre-extracted log-spectral templates per sound state q, pitch p and instrument s, which are also pre-shifted across log-frequency f. The proposed pre-shifting operation is made in order to account for pitch deviations, without needing to formulate a convolutive model across log-frequency, as in [4]. $P_t(f|p)$ is the time-varying log-frequency shifting distribution per pitch, $P_t(s|p)$ 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.

In the proposed model, $f \in [1,\ldots,5]$, where $f=3$ is the ideal tuning position for the template (using equal temperament). Given that the input time-frequency representation has a resolution of 5 bins per semitone, this means that all templates are pre-shifted across log-frequency on a ± 20 and ± 40 cent range around the ideal tuning position, thus accounting for small tuning deviations or frequency modulations. The proposed model also uses 3 sound states per pitch; more information on the extraction of the sound state spectral templates is given in subsection 3.1.

The unknown model parameters ($P_t(f|p)$, $P_t(s|p)$, $P_t(p)$, $P_t(q|p)$) can be iteratively estimated using the expectation-maximization (EM) algorithm [9]. For the *Expectation* step, the following posterior is computed:

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

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

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

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

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

$$P_t(q|p) = \frac{\sum_{\omega,f,s} P_t(q,p,f,s|\omega)V_{\omega,t}}{\sum_{q,\omega,f,s} P_t(q,p,f,s|\omega)V_{\omega,t}} \quad (6)$$

Eqs. (2)-(6) are iterated until convergence; typically 15-20 iterations are sufficient. No update rule for the sound state templates $P(\omega|q,p,f,s)$ is included, since they are considered fixed in the model. As in [4], 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 $P(p,t) = P(t)P_t(p)$, while a time-pitch representation $P(f',t)$ can also be derived from the model, as in [4] (this representation has the same pitch resolution as in the input representation, i.e. 20 cent resolution).

2.3 Temporally-constrained model

This model variant proposes a formulation that expresses the evolution of each note as a succession of sound states, following work carried out in [4]. These temporal constraints are modelled using pitch-wise hidden Markov models (HMMs). This also follows the work done by Mysore in [17] on the non-negative HMM (a spectrogram factorization framework where the appearance of each template is controlled by an HMM).

As discussed, one HMM is created per pitch p, which has as hidden states the sound states q (assuming 88 pitches that cover the entire note range of a piano, 88 HMMs are used). Thus, the basic elements of this pitch-wise HMM are: the sound state priors $P(q_1^{(p)})$, the sound state transitions $P(q_{t+1}^{(p)}|q_t^{(p)})$, and the observations $P(\bar{\omega}_t|q_t^{(p)})$. Following the notation of [17], $\bar{\omega}$ corresponds to the sequence of observed spectra from all time frames, and $\bar{\omega}_t$ is the observed spectrum at the t-th time frame. Also, $q_t^{(p)}$ is the value of the hidden sound state at the t-th frame for pitch p.

In this paper, the model formulation is the same as in (1), where the following assumption is made:

$$P_t(q|p=i) = P_t(q_t^{(p=i)}|\bar{\omega}) \quad (7)$$

which means that the sound state activations are assumed to be produced by the posteriors (also called *responsibilities*) of the HMM for pitch p. Following [17], the observation probability is calculated as:

$$P(\bar{\omega}_t|q_t^{(p)}) = \prod_{\omega_t} P(\omega_t|q_t^{(p)})^{V_{\omega,t}} \quad (8)$$

where $P(\omega_t|q_t^{(p)})$ is the approximated spectrum for a given sound state and pitch. The observation probability is calculated as above since in PLCA-based models, $V_{\omega,t}$ represents the number of times ω has been drawn at the t-th time frame [17].

In order to estimate the unknown parameters of this proposed temporally-constrained model, the EM algorithm is also used, which results in a series of iterative update rules that combine PLCA-based updates as well as the HMM forward-backward algorithm [20]. For the Expectation step, the HMM posterior per pitch is computed as:

$$P_t(q_t^{(p)}|\bar{\omega}) = \frac{P_t(\bar{\omega},q_t^{(p)})}{\sum_{q_t^{(p)}} P_t(\bar{\omega},q_t^{(p)})} = \frac{\alpha_t(q_t^{(p)})\beta_t(q_t^{(p)})}{\sum_{q_t^{(p)}} \alpha_t(q_t^{(p)})\beta_t(q_t^{(p)})} \quad (9)$$

where $\alpha_t(q_t^{(p)})$ and $\beta_t(q_t^{(p)})$ are the forward and backward variables for the p-th HMM, respectively, and can be computed using the forward-backward algorithm [20]. The posterior for the transition probabilities $P_t(q_{t+1}^{(p)}, q_t^{(p)}|\bar{\omega})$ is also computed as in [4]. Finally, the model posterior is computed using (2) and (7).

For the Maximization step, unknown parameters $P_t(f|p)$, $P_t(s|p)$, and $P_t(p)$ are computed using eqs. (3)-(5). Finally, the sound state priors and transitions per pitch p are estimated as:

$$P(q_1^{(p)}) = P_1(q_1^{(p)}|\bar{\omega}) \tag{10}$$

$$P(q_{t+1}^{(p)}|q_t^{(p)}) = \frac{\sum_t P_t(q_t^{(p)}, q_{t+1}^{(p)}|\bar{\omega})}{\sum_{q_{t+1}^{(p)}} \sum_t P_t(q_t^{(p)}, q_{t+1}^{(p)}|\bar{\omega})} \tag{11}$$

In our experiments, it was found that an initial estimation of the pitch and source activations using the PLCA-only updates in the Maximization step leads to a good initial solution. In the final iterations (set to 3 in this case), the HMM parameters are estimated as well, which leads to an estimate of the sound state activations, and an improved solution over the non-temporally constrained model of subsection 2.2.

2.4 Post-processing

For both the non-temporally constrained model of subsection 2.2 and the temporally-constrained model of subsection 2.3, the resulting pitch activation $P(p,t) = P(t)P_t(p)$ (which is used for multi-pitch detection evaluation) as well as the pitch activation for a specific instrument $P(s,p,t) = P(t)P_t(p)P_t(s|p)$ (which is used for instrument assignment evaluation) need to be converted into a binary representation such as a piano-roll or a MIDI file. As in the vast majority of spectrogram factorization-based music transcription systems (e.g. [10, 15]), thresholding is performed on the pitch and instrument activations, followed by a process for removing note events with a duration less than 80ms.

3. EVALUATION

3.1 Training data

Sound state templates are extracted for several orchestral instruments, using isolated note samples from the RWC database [14]. Specifically, templates are extracted for bassoon, cello, clarinet, flute, guitar, harpsichord, oboe, piano, alto sax, and violin, using the variable-Q transform as a time-frequency representation [22]. The complete note range of the instruments (given available data) is used. The sound state templates are computed in an unsupervised manner, using a single-pitch and single-instrument variant of the model of (1), with the number of sound states set to 3.

3.2 Test data

Several benchmark and freely available transcription datasets are used for evaluation (all of them contain pitch ground truth). Firstly, thirty piano segments of 30s duration are used from the MAPS database using the 'ENSTDkCl' piano model. This test dataset has in the past been used for multi-pitch evaluation (e.g. [7, 18], the latter also citing results using the method of [24]).

System	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
§2.2	70.08%	76.78%	65.27%
§2.3	71.56%	77.95%	66.89%

Table 1. Multi-pitch detection results for the MAPS-ENSTDkCl dataset using the proposed models.

The second dataset consists of the woodwind quintet recording from the MIREX 2007 multiF0 development dataset [1]. The multi-track recording has been evaluated in the past either in its complete duration [4], or in shorter segments (e.g. [19, 24]).

Thirdly, we employ the Bach10 dataset [11], a multi-track collection of multiple-instrument polyphonic music, suitable for both multi-pitch detection and instrument assignment experiments. It consists of ten recordings of J.S. Bach chorales, performed by violin, clarinet, saxophone, and bassoon.

Finally, the TRIOS dataset [12] is also used, which includes five multi-track recordings of trio pieces of classical and jazz music. Instruments included in the dataset are: bassoon, cello, clarinet, horn, piano, saxophone, trumpet, viola, and violin.

3.3 Metrics

For assessing the performance of the proposed system in terms of multi-pitch detection we utilise the onset-based metric used in the MIREX note tracking evaluations [1]. A note event is assumed to be correct if its pitch corresponds to the ground truth pitch and its onset is within a $\pm$50 ms range of t ground truth onset. Using the above rule, precision ($\mathcal{P}$), recall ($\mathcal{R}$), and F-measure ($\mathcal{F}$) metrics can be defined:

$$\mathcal{P} = \frac{N_{tp}}{N_{sys}}, \quad \mathcal{R} = \frac{N_{tp}}{N_{ref}}, \quad \mathcal{F} = \frac{2 \cdot \mathcal{R} \cdot \mathcal{P}}{\mathcal{R} + \mathcal{P}} \tag{12}$$

where N_{tp} is the number of correctly detected pitches, N_{sys} is the number of detected pitches, and N_{ref} is the number of ground-truth pitches. For comparison with other state-of-the-art methods, we also use frame-based multiple-F0 estimation metrics, defined in [2], denoted as $\mathcal{P}_f, \mathcal{R}_f, \mathcal{F}_f$.

For the instrument assignment evaluations with the Bach-10 dataset, we use the pitch ground-truth of each instrument and compare it with the instrument-specific output of the system. As for the multi-pitch metrics, we define the following note-based instrument assignment metrics: $\mathcal{F}_v$, $\mathcal{F}_c$, $\mathcal{F}_s$, $\mathcal{F}_b$, corresponding to violin, clarinet, saxophone, and bassoon, respectively. We also use a mean instrument assignment metric, denoted as $\mathcal{F}_{ins}$.

3.4 Results

Experiments are performed using the two proposed model variants from Section 2: the non-temporally constrained version of subsection 2.2 and the HMM-constrained version of subsection 2.3. In both versions, the post-processing

System	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
§2.2	71.75%	68.78%	74.98%
§2.3	72.50%	73.31%	71.71%

Table 2. Multi-pitch detection results for the MIREX multiF0 recording using the proposed models.

System	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
§2.2	64.43%	56.99%	74.16%
§2.3	65.01%	57.35%	75.11%

Table 3. Multi-pitch detection results for the Bach10 dataset using the proposed models.

steps are the same. For the HMM-constrained model, the HMMs are initialized as ergodic, with uniform priors and state transition probabilities.

In terms of multi-pitch detection evaluation, results for the MAPS, MIREX, Bach10, and TRIOS datasets are shown in Tables 1, 2, 3, and 4, respectively. In all cases, the HMM-constrained model outperforms the non-temporally constrained model. The difference over the two models in terms of F-measure is more prominent for the MAPS dataset (1.48%) and the TRIOS dataset (1.81%) compared to the MIREX (0.75%) and Bach10 (0.58%) datasets. This can be attributed to the presence of piano in the MAPS and TRIOS datasets, compared to the woodwind/string instruments present in the other two datasets; since the piano is a pitched percussive instrument with a clear attack and transient state, the incorporation of temporal constraints on sound state evolution can be considered more important compared to bowed string and woodwind instruments, that do not exhibit a clear decay state. As an example of the transcription performance of the proposed system, Fig. 4 shows the resulting pitch activation for the MIREX multiF0 recording along with the corresponding ground truth.

Instrument assignment results for the Bach10 dataset are presented in Table 5. As can be seen, the performance of the proposed system regarding instrument assignment is much lower compared to multi-pitch detection, which this can be attributed to the fact that instrument assignment is a much more challenging problem, since it not only requires a correct identification of a note, but also a correct classification of that detected note to a specific instrument. It is worth noting however that a clear improvement is reported when using the temporally-constrained model over the model of subsection 2.2. That improvement is consistent across all instruments.

3.4.1 Comparison with state-of-the-art

On comparison of the proposed system with other state-of-the art multi-pitch detection methods, for MAPS the proposed HMM-constrained method outperforms the spectrogram factorization transcription methods of [18] and [24] by 13.2% and 2.5% in terms of $\mathcal{F}$, respectively. It is however outperformed by the transcription system of [7] (4.9% difference); it should be noted that the system of [7] is

System	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
§2.2	57.55%	64.60%	54.04%
§2.3	59.36%	60.18%	59.45%

Table 4. Multi-pitch detection results for the TRIOS dataset using the proposed models.

System	F_v	F_c	F_s	F_b	F_{ins}
§2.2	10.55%	39.99%	33.87%	40.80%	31.30%
§2.3	12.28%	41.55%	34.53%	42.33%	32.67%

Table 5. Instrument assignment results for the Bach10 dataset using the proposed models.

developed specifically for piano, in contrast with the proposed multiple-instrument system.

Regarding comparison on the MIREX recording, the proposed method outperforms the method of [6] by 3.9% in terms of $\mathcal{F}$. In terms of $\mathcal{F}_f$, the first 30sec of the MIREX recording were evaluated using the systems of [24] and [19], leading to $\mathcal{F}_f = 62.5\%$ and $\mathcal{F}_f = 59.6\%$, respectively. The proposed HMM-constrained method reaches $\mathcal{F}_f = 70.35\%$, thus outperforming the aforementioned systems.

For the Bach10 dataset, a comparison is made using the accuracy metric defined in [11]. The proposed HMM-constrained method reaches an accuracy of 72.0%, whereas the method of [11] reaches 69.7% (the latter results are with unknown polyphony level, for direct comparison with the proposed method).

Finally, for the TRIOS dataset, multi-pitch detection results were reported in [6], with $\mathcal{F} = 57.6\%$. The proposed method reaches for the HMM-constrained case $\mathcal{F} = 59.3\%$, thus outperforming the system of [6].

3.4.2 Comparing time-frequency representations

In order to evaluate the use of the proposed input VQT time-frequency representation, a comparative experiment is made using the proposed system and having as input a constant-Q representation (using the method of [21], with a 60 bins/octave log-frequency resolution as with the VQT). For the comparative experiments, the MAPS-ENSTDkCl dataset is employed and both the non-temporally constrained and HMM-constrained models are evaluated. The post-processing steps are exactly the same as in the proposed method. Results show that when using the constant-Q representation $\mathcal{F} = 63.98\%$ for the non-temporally constrained model and $\mathcal{F} = 65.51\%$ for the temporally-constrained model, which are both significantly lower when compared to using a VQT representation as input (cf. Table 1).

In order to show the improved detection performance of a VQT representation with respect to lower pitches, the transcription performance for the MAPS dataset was computed when only taking into account notes below or above MIDI pitch 60 (middle C in the piano). Using the VQT, $\mathcal{F} = 65.18\%$ for the lower pitches and $\mathcal{F} = 74.98\%$ for the higher pitches. In contrast when using the CQT,

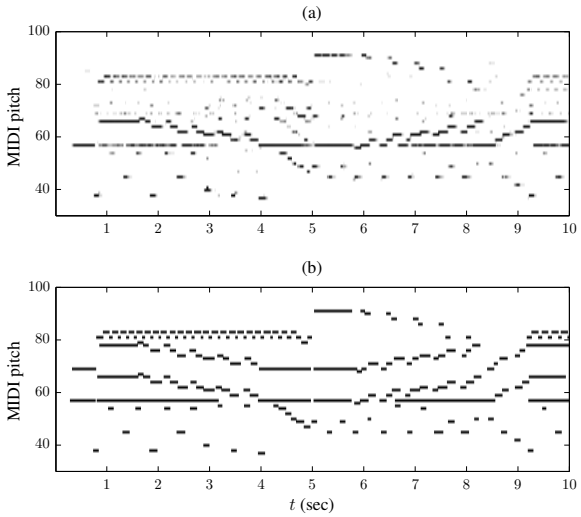

Figure 4. (a) The pitch activation output $P(p,t)$ for the first 10 sec of the MIREX multiF0 recording. (b) The corresponding pitch ground truth.

$\mathcal{F} = 51.17\%$ for the lower pitches and $\mathcal{F} = 74.58\%$ for the higher pitches. This result clearly demonstrates the benefit of using a VQT representation with respect to temporal resolution in lower frequencies, and by extension, to detecting lower pitches. As an example, Fig. 2 shows the CQT and VQT spectrograms for a G1 piano note, with the VQT exhibiting better temporal resolution in lower frequencies.

3.4.3 Sound state templates vs. note templates

Here, a comparison is performed between the use of the proposed 5-dimensional dictionary of sound state templates against the use of a 4-dimensional note template dictionary (which contains one template per pitch, instrument, and log-frequency shifting); the latter is supported by the method of [3]. In order to have a direct comparison, the method of [3] (for which the source code is publicly available) is modified as to use the same input VQT representation as well as post-processing steps with the proposed method, and is compared against the non-temporally constrained model of subsection 2.2.

When using a 4-dimensional dictionary, multi-pitch detection performance for the MAPS dataset reaches 64.65%, in contrast to 70.1% when using the 5-dimensional sound state dictionary. This shows the importance of using sound state templates, which are able to model the transient parts of the signal in contrast to simply using one (typically harmonic) note template for each pitch and instrument.

3.4.4 Runtimes

On computational efficiency, the proposed model requires linear operations like matrix/tensor multiplications in the EM steps; on the contrary, the previous model of [4] required the computation of convolutions which significantly slowed down computations. Regarding runtimes, the original HMM-constrained convolutive model of [4] runs at about 60 $\times$ real-time using a Sony VAIO S15 laptop. Using the proposed method, the runtime is approximately 1 $\times$ real-time for the non-temporally constrained model, and 2.5 $\times$ real-time for the HMM-constrained model (i.e. for a 1min recording, runtimes are 1min and 2.5min, respectively). Thus, the proposed system is significantly faster compared to the model of [4], making it suitable for large-scale MIR applications.

4. CONCLUSIONS

In this paper, we proposed a computationally efficient system for multiple-instrument automatic music transcription, based on probabilistic latent component analysis. The proposed model employs a 5-dimensional dictionary of sound state templates, covering different pitches, instruments, and tunings. Two model variants were presented: a PLCA-only method and a temporally constrained model that uses pitch-wise HMMs in order to control the order of the sound states. Experiments were performed on several transcription datasets; results show that the temporally-constrained model outperforms the PLCA-based variant. In addition, the proposed system outperforms several state-of-the-art multiple-instrument transcription systems using the MIREX multiF0, Bach10, and TRIOS datasets. We also showed that a VQT representation can yield improved results compared to a CQT representation. Finally, the non-temporally constrained variant of the model is able to transcribe a recording at 1 $\times$ real-time, thus making this method useful for large-scale applications. The Matlab code for the HMM-constrained model can be found online [1] in the hope that this model can serve as a framework for creating transcription systems useful to the MIR community.

This system can also be extended beyond the proposed formulations, by exploiting recent developments in spectrogram factorization-based approaches for music and audio signal analysis. Thus, the proposed model can also incorporate prior information in various forms (e.g. instrument identities, key information, music language models), following the PLCA-based approach of [23]. It can also use alternate EM update rules to guide convergence [8] or can use additional temporal continuity and sparsity constraints [13]. Drum transcription can also be incorporated into the system, in the same way as in [5]. In the future, we will also incorporate temporal constraints on note transitions and polyphony level estimation and will continue work on instrument assignment by combining timbral features with PLCA-based models.

5. ACKNOWLEDGEMENT

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

6. REFERENCES

[1] Music Information Retrieval Evaluation eXchange (MIREX). http://music-ir.org/mirexwiki/.

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

[2] M. Bay, A. F. Ehmann, and J. S. Downie. Evaluation of multiple-F0 estimation and tracking systems. In *10th International Society for Music Information Retrieval Conference*, pages 315–320, Kobe, Japan, October 2009.

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

[4] 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.

[5] E. Benetos, S. Ewert, and T. Weyde. Automatic transcription of pitched and unpitched sounds from polyphonic music. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3131–3135, Florence, Italy, May 2014.

[6] E. Benetos and T. Weyde. Explicit duration hidden Markov models for multiple-instrument polyphonic music transcription. In *14th International Society for Music Information Retrieval Conference*, pages 269–274, Curitiba, Brazil, November 2013.

[7] T. Berg-Kirkpatrick, J. Andreas, and D. Klein. Unsupervised transcription of piano music. In *Advances in Neural Information Processing Systems*, pages 1538–1546, 2014.

[8] T. Cheng, S. Dixon, and M. Mauch. A deterministic annealing em algorithm for automatic music transcription. In *14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil.

[9] 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.

[10] 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.

[11] Z. Duan, B. Pardo, and C. Zhang. Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2121–2133, November 2010.

[12] J. Fritsch. High quality musical audio source separation. Master's thesis, UPMC / IRCAM / Telécom ParisTech, 2012.

[13] B. Fuentes, R. Badeau, and G. Richard. Harmonic adaptive latent component analysis of audio and application to music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1854–1866, September 2013.

[14] 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.

[15] G. Grindlay and D. Ellis. Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1159–1169, October 2011.

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

[17] G. Mysore. *A non-negative framework for joint modeling of spectral structure and temporal dynamics in sound mixtures*. PhD thesis, Stanford University, USA, June 2010.

[18] K. O'Hanlon and M.D. Plumbley. Polyphonic piano transcription using non-negative matrix factorisation with group sparsity. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3112–3116, May 2014.

[19] P.H. Peeling and S.J. Godsill. Multiple pitch estimation using non-homogeneous poisson processes. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1133–1143, October 2011.

[20] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

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

[22] 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.

[23] P. Smaragdis and G. Mysore. Separation by "humming": user-guided sound extraction from monophonic mixtures. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 69–72, New Paltz, USA, October 2009.

[24] 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.