EXPLICIT DURATION HIDDEN MARKOV MODELS FOR MULTIPLE-INSTRUMENT POLYPHONIC MUSIC TRANSCRIPTION

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 paper, a method for multiple-instrument automatic music transcription is proposed that models the temporal evolution and duration of tones. The proposed model supports the use of spectral templates per pitch and instrument which correspond to sound states such as attack, sustain, and decay. Pitch-wise explicit duration hidden Markov models (EDHMMs) are integrated into a convolutive probabilistic framework for modelling the temporal evolution and duration of the sound states. A two-stage transcription procedure integrating note tracking information is performed in order to provide more robust pitch estimates. The proposed system is evaluated on multi-pitch detection and instrument assignment using various publicly available datasets. Results show that the proposed system outperforms a hidden Markov model-based transcription system using the same framework, as well as several state-of-the-art automatic music transcription systems.

1. INTRODUCTION

Automatic music transcription (AMT) is the process of converting an acoustic musical signal into some form of music notation [13]. In the music information retrieval literature, AMT typically involves the detection of multiple concurrent pitches (multi-pitch detection), the estimation of note onsets and offsets (note tracking) and the estimation of instrument identities (instrument identification/assignment). It is generally considered to be an open problem, especially for highly polyphonic music signals and multiple instruments. For a recent review of AMT systems, the reader is referred to [12].

A large part of AMT systems employ spectrogram factorization methods for multi-pitch detection. These systems attempt to decompose an input time-frequency representation as a series of spectral components and pitch activations, using a variety of constraints (regarding polyphony level, instrument identities, spectral envelopes, and temporal continuity among others); systems related to the proposed work will be presented below.

In [6], Dessein et al. propose an AMT system for piano music which uses non-negative matrix factorization (NMF) with beta-divergence and pre-extracted note templates, which is able to transcribe pieces in real-time. Vincent et al. [16] propose a harmonic variant of NMF for decomposing a spectrogram into a series of narrowband harmonic spectra, which are also smooth across frequency (also called the spectral smoothness assumption [13]). In [4], Carabias-Orti et al. propose a system for multi-pitch detection and instrument identification using NMF with source-filter model constraints. Grindlay and Ellis [11] utilize a probabilistic variant of NMF called probabilistic latent component analysis (PLCA) for decomposing a spectrogram into a series of *eigeninstrument* templates, pitch activations, and source contributions, and evaluate their method for multi-pitch detection and instrument assignment. Yoshii and Goto [17] proposed a non-parametric model for music signal analysis which decomposes an input spectrogram as a series of source-filter templates derived from an autoregressive model. Finally, in [2] Benetos and Dixon proposed a variant of convolutive PLCA for modelling the evolution of notes using sound state templates (such as attack, sustain, decay) with hidden Markov model-based constraints.

In this paper, we integrate explicit duration hidden Markov models (EDHMMs) [7,18] within the spectrogram factorization framework of [2], in order to model the duration of sound states within a note. Contrary to hidden Markov models (HMMs), where the state duration is (implicitly) geometrically distributed, EDHMMs form a specific case of hidden semi-Markov models [18], where each state has a variable duration. Alternatively, it can be viewed that an EDHMM can emit a sequence of observations instead of a single one. The additional information in EDHMMs is modelled through the use of a duration probability per state. EDHMMs have been shown to overcome the limitations posed by HMMs regarding state durations and have been successfully used in a variety of applications (see [18] for a review).

The proposed model uses pitch-wise EDHMMs for constraining the order of the sound states, while also supporting the use of multiple templates per pitch and instrument, and also shift-invariance across log-frequency for supporting tuning changes and frequency modulations. In addi-

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

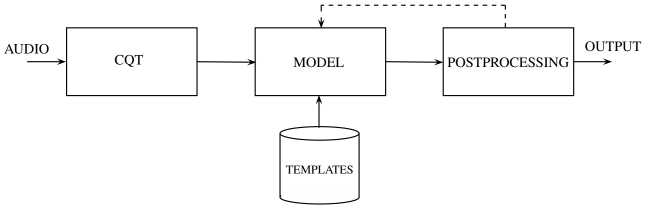

Figure 1. Proposed system diagram.

tion, we propose a two-stage transcription procedure in order to eliminate spurious pitch detections. The proposed model is trained on note samples from the RWC database [10] and is tested using recordings from the MAPS database [8], the TRIOS dataset [9], and the MIREX MultiF0 recording [1]. Multi-pitch detection and instrument assignment results show that the proposed EDHMM-based system is able to model the durations of sound states and the overall evolution of notes, and its temporal constraints lead to improved performance compared to hidden Markov models (HMMs) used in the same framework. Finally, the proposed system outperforms several AMT methods in the literature for the same experiments.

The outline of this paper is as follows. Section 2 presents the proposed EDHMM-based transcription model, along with the postprocessing steps. The datasets used for training and testing, as well as the evaluation metrics and experimental results, are presented in Section 3. Finally, conclusions are drawn and future directions are indicated in Section 4.

2. PROPOSED METHOD

In this section, the proposed EDHMM-constrained automatic transcription model is described, along with the update rules for estimating the various model parameters and the steps used for post-processing; the proposed system diagram can be seen in Fig. 1.

2.1 Model

The proposed model aims to express the evolution of notes in multiple-instrument polyphonic music as a succession of sound state templates, further constrained by the ordering and the expected duration of each sound state. These temporal constraints are incorporated into a model which supports multiple templates per pitch and instrument, and also supports shift-invariance across log-frequency in order to model tuning changes and frequency modulations. In order to achieve this, we integrate independent pitch-wise explicit duration hidden Markov models (EDHMMs) [18] into the HMM-constrained automatic transcription model of [2]. Thus, the proposed model can be called EDHMM-constrained shift-invariant PLCA.

More formally, the normalized magnitude log-frequency spectrogram $V_{\omega,t}$ (ω denotes log-frequency and t denotes time) which is used as input, is decomposed into a series of sound state spectral templates per instrument and pitch, a time-varying pitch shifting parameter, a time-varying instrument contribution per pitch, a pitch activation, and finally a sound state activation per pitch, which is controlled by its respective EDHMM. If we denote the collection of observations for all time frames as $\bar{\omega}$, the proposed model in terms of the observations is given by:

$$\begin{aligned}P(\bar{\omega}) = &\sum_{\bar{q}^{(1)},\cdots,\bar{q}^{(\mathcal{P})}} \sum_{\bar{d}^{(1)},\cdots,\bar{d}^{(\mathcal{P})}} P(q_1^{(1)})\cdots P(q_1^{(\mathcal{P})})\\ &P(d_1^{(1)})\cdots P(d_1^{(\mathcal{P})})\\ &\left(\prod_t P(q_t^{(1)}|q_{t-1}^{(1)},d_{t-1}^{(1)})P(d_t^{(1)}|q_t^{(1)},d_{t-1}^{(1)})\right)\cdots\\ &\left(\prod_t P(q_t^{(\mathcal{P})}|q_{t-1}^{(\mathcal{P})},d_{t-1}^{(\mathcal{P})})P(d_t^{(\mathcal{P})}|q_t^{(\mathcal{P})},d_{t-1}^{(\mathcal{P})})\right)\\ &\left(\prod_t P(\bar{\omega}_t|q_t^{(1)},\ldots,q_t^{(\mathcal{P})})\right) \qquad (1)\end{aligned}$$

where $p = 1,\cdots,\mathcal{P}$ denotes pitch, $q^{(p)}$ denotes the sound state for the p-th pitch, $d^{(p)}$ denotes the duration distribution for the p-th pitch, $P(q_1^{(p)})$ is the sound state prior for the p-th pitch, $P(d_1^{(p)})$ is the duration prior for the p-th pitch, $\bar{q}$ is the sequence of draws of q, $\bar{d}$ is the sequence of draws of d, and finally $P(\bar{\omega}_t|q_t^{(1)},\ldots,q_t^{(\mathcal{P})})$ is the observation probability for a given observation $\bar{\omega}_t$.

An EDHMM has state transitions only at the end of a segment, and its duration distributions generate segment lengths only at every state switch [7]:

$$P(q_{t+1}^{(p)}|q_t^{(p)},d_t^{(p)}) = \begin{cases}\delta(q_{t+1}^{(p)},q_t^{(p)}), & d_t > 1\\ P(q_{t+1}^{(p)}|q_t^{(p)}), & \text{otherwise}\end{cases}$$

$$P(d_{t+1}^{(p)}|q_{t+1}^{(p)},d_t^{(p)}) = \begin{cases}\delta(d_{t+1}^{(p)},d_t^{(p)}-1), & d_t > 1\\ P(d_{t+1}^{(p)}|q_{t+1}^{(p)}), & \text{otherwise}\end{cases}$$

where $P(q_{t+1}^{(p)}|q_t^{(p)})$ is the pitch-wise sound state transition matrix, and $P(d_t^{(p)}|q_t^{(p)})$ is the pitch-wise sound state duration distribution. Also, $\delta(x,y)=1$ if $x=y$ and 0 otherwise.

Since in the PLCA-based models $V_{\omega,t}$ represents the number of times ω has been drawn at the t-th time frame, the observation probability is calculated as:

$$P(\bar{\omega}_t|q_t^{(1)},\ldots,q_t^{(\mathcal{P})}) = \prod_{\omega_t} P_t(\omega_t|q_t^{(1)},\ldots,q_t^{(\mathcal{P})})^{V_{\omega,t}} \qquad (2)$$

In the proposed model, $P_t(\omega_t|q_t^{(1)},\ldots,q_t^{(\mathcal{P})})$ is decomposed as:

$$\begin{aligned}&P_t(\omega_t|q_t^{(1)},\ldots,q_t^{(\mathcal{P})}) =\\ &\sum_{s_t,p_t,f_t} P_t(p_t)P_t(s_t|p_t)P(\omega_t-f_t|s_t,p_t,q_t^{(p_t)})P_t(f_t|p_t)\end{aligned} \qquad (3)$$

where s denotes the instrument source, f is the pitch shifting parameter, $P_t(p_t)$ is the pitch activation, $P_t(s_t|p_t)$ is the time-varying instrument contribution for each pitch, $P(\omega|s,p,q^{(p)})$ are the sound state spectral templates per source s, pitch p, and sound state $q^{(p)}$, and $P_t(f_t|p_t)$ is the log-frequency shifting distribution per pitch over time. The subscript t in f_t, ω_t, s_t, p_t denotes the values of variables f, ω, s, p taken at time t. The shifting parameter f is constrained to a semitone range around the ideal tuning position of each pitch. Since in the proposed system the time-frequency representation used is the constant-Q transform (CQT) with a log-frequency resolution of 60 bins/octave and a 40ms step [15], this implies that $f \in [1, 5]$. We also set a maximum duration for each sound state: $d \in [1, 20]$, which means that the maximum duration of each sound state is 800ms.

2.2 Parameter Estimation

The unknown model parameters of Section 2.1 can be estimated using the Expectation-Maximization (EM) algorithm [5]. For the E-step, the posterior for all hidden variables is:

$$P_t(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(\mathcal{P})}|\bar{\omega}) = P_t(q_t^{(1)}, \ldots, q_t^{(\mathcal{P})}|\bar{\omega}) P_t(f_t, s_t, p_t|\omega_t, q_t^{(1)}, \ldots, q_t^{(\mathcal{P})}) \quad (4)$$

We assume that the pitch-wise EDHMMs are independent, thus the joint probability of all sound states is decomposed as:

$$P_t(q_t^{(1)}, \ldots, q_t^{(\mathcal{P})}|\bar{\omega}) = \prod_{p=1}^{\mathcal{P}} P_t(q_t^{(p)}|\bar{\omega}) \quad (5)$$

where:

$$P_t(q_t^{(p)}|\bar{\omega}) = \frac{\sum_{\tau<t} \left(\alpha_\tau^{*(p)}(q_{\tau+1}^{(p)})\beta_\tau^{*(p)}(q_{\tau+1}^{(p)}) - \alpha_\tau^{(p)}(q_\tau^{(p)})\beta_\tau^{(p)}(q_\tau^{(p)})\right)}{\sum_{q_t^{(p)},\tau<t} \left(\alpha_\tau^{*(p)}(q_{\tau+1}^{(p)})\beta_\tau^{*(p)}(q_{\tau+1}^{(p)}) - \alpha_\tau^{(p)}(q_\tau^{(p)})\beta_\tau^{(p)}(q_\tau^{(p)})\right)} \quad (6)$$

where $\alpha_\tau^*(q_{\tau+1})$, $\alpha_\tau(q_\tau)$ are the EDHMM forward variables and $\beta_\tau^*(q_{\tau+1})$, $\beta_\tau(q_\tau)$ are the EDHMM backward variables; all aforementioned forward-backward variables can be computed using recursive formulae [18].

The second term of (4) can be computed using Bayes' theorem and the notion that $P(\omega_t|f_t, s_t, p_t, q_t^{(p_t)}) = P(\omega_t - f_t|s_t, p_t, q_t^{(p_t)})$:

$$P_t(f_t, s_t, p_t|\omega_t, q_t^{(1)}, \ldots, q_t^{(\mathcal{P})}) = P_t(f_t, s_t, p_t|\omega_t, q_t^{(p_t)}) = \frac{P_t(p_t)P(\omega_t - f_t|s_t, p_t, q_t^{(p_t)})P_t(f_t|p_t)P_t(s_t|p_t)}{\sum_{p_t,s_t,f_t} P_t(p_t)P(\omega_t - f_t|s_t, p_t, q_t^{(p_t)})P_t(f_t|p_t)P_t(s_t|p_t)} \quad (7)$$

For the M-step, the update equations for the unknown parameters are as follows:

$$P_t(p_t) = \frac{\sum_{\omega_t,f_t,s_t,q_t^{(1)},\cdots,q_t^{(\mathcal{P})}} V_{\omega,t} P_t(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(\mathcal{P})}|\bar{\omega})}{\sum_{p_t,\omega_t,f_t,s_t,q_t^{(1)},\cdots,q_t^{(\mathcal{P})}} V_{\omega,t} P_t(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(\mathcal{P})}|\bar{\omega})} \quad (8)$$

$$P_t(s_t|p_t) = \frac{\sum_{\omega_t,f_t,q_t^{(1)},\cdots,q_t^{(\mathcal{P})}} V_{\omega,t} P_t(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(\mathcal{P})}|\bar{\omega})}{\sum_{s_t,\omega_t,f_t,q_t^{(1)},\cdots,q_t^{(\mathcal{P})}} V_{\omega,t} P_t(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(\mathcal{P})}|\bar{\omega})} \quad (9)$$

$$P_t(f_t|p_t) = \frac{\sum_{\omega_t,s_t,q_t^{(1)},\cdots,q_t^{(\mathcal{P})}} V_{\omega,t} P_t(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(\mathcal{P})}|\bar{\omega})}{\sum_{f_t,\omega_t,s_t,q_t^{(1)},\cdots,q_t^{(\mathcal{P})}} V_{\omega,t} P_t(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(\mathcal{P})}|\bar{\omega})} \quad (10)$$

$$P(q_1^{(p)}) = \frac{P_1(q_1^{(p)}|\bar{\omega})}{\sum_{q_1^{(p)}} P_1(q_1^{(p)}|\bar{\omega})} \quad (11)$$

$$P(q_{t+1}^{(p)}|q_t^{(p)}) = \frac{\sum_t \alpha_t^{(p)}(q_t^{(p)})P(q_{t+1}^{(p)}|q_t^{(p)})\beta_t^{*(p)}(q_{t+1}^{(p)})}{\sum_{q_{t+1}^{(p)},t} \alpha_t^{(p)}(q_t^{(p)})P(q_{t+1}^{(p)}|q_t^{(p)})\beta_t^{*(p)}(q_{t+1}^{(p)})} \quad (12)$$

$$P(d_t^{(p)}|q_t^{(p)}) = \frac{\sum_t \alpha_t^{*(p)}(q_{t+1}^{(p)})P(d_t^{(p)}|q_t^{(p)})\beta_{t+d}^{p}(q_{t+d}^{(p)})\prod_\tau P(\bar{\omega}_\tau|q_\tau^{(p)})}{\sum_{d_t^{(p)},t} \alpha_t^{*(p)}(q_{t+1}^{(p)})P(d_t^{(p)}|q_t^{(p)})\beta_{t+d}^{p}(q_{t+d}^{(p)})\prod_\tau P(\bar{\omega}_\tau|q_\tau^{(p)})} \quad (13)$$

where $\tau = t+1, \cdots, t+d$.

It should be noted that we consider the sound state templates to be fixed, so no update rule for $P(\omega|s, p, q^{(p)})$ exisits. Using fixed templates, 10-15 iterations using the update rules presented in the present section are sufficient for convergence. The output of the system is a pitch activation which is scaled by the energy of the log-spectrogram:

$$P_t(p) \sum_\omega V_{\omega,t} \quad (14)$$

In order to further constrain the model so that it reaches more meaningful solutions, sparsity is enforced in $P_t(p)$ and $P_t(s|p)$, by modifying the update rules in (8) and (9), where a power greater than 1 is applied to the numerators and denominators, which leads to sharpened distributions, thus encouraging sparsity [2]. Even though convergence is not guaranteed, it is observed in practice. This procedure implies that only few pitches need to be active at each time frame, and also that for a note at a given time frame, only few instruments are responsible for producing it.

As an example of the learned EDHMM parameters using the proposed system, Fig. 2 shows the learned duration distributions and sound state transitions for a D4 note, using a piano recording as input to the system. It can be seen that the duration distribution for the 1st sound state (which corresponds to an attack state) favors short durations, while the duration distribution for the 2nd state (which corresponds to the steady state) favors much longer durations. Also, the resulting transition matrix also shows the linear succession between the sound states.

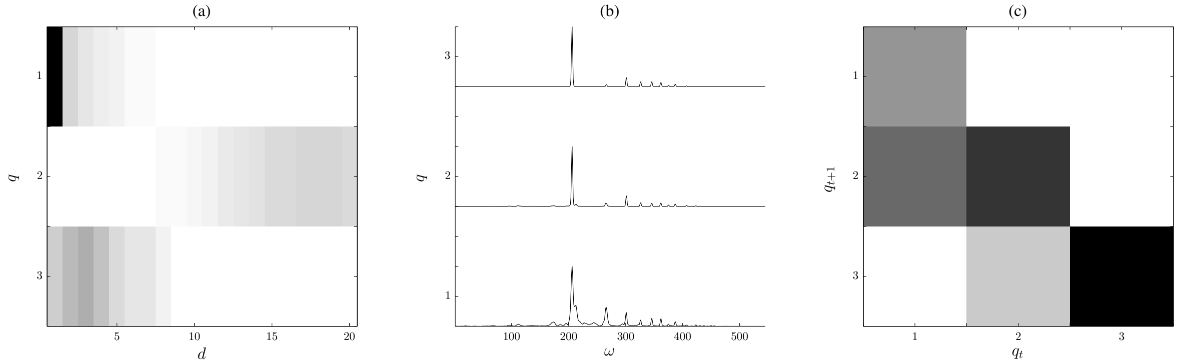

Figure 2. EDHMM parameters learned for note D4 using the 'MAPS_MUS-alb_se2_ENSTDkCl' piece from the MAPS database. (a) The duration distribution matrix $P(d|q)$. (b) Pre-computed piano sound state templates for note D4. (c) The sound state transition matrix $P(q_{t+1}|q_t)$.

2.3 Post-processing

Since the resulting pitch activation from (14) is non-binary, a postprocessing procedure needs to take place in order to convert it to a binary piano-roll or a MIDI-like representation (this procedure is also called note tracking). As in the vast majority of spectrogram factorization-based automatic transcription systems (e.g. [6, 11]), we perform thresholding on the pitch activation, followed by a process for removing note events with a duration less than 80ms. We should note that the HMM-based postprocessing method of [2] was found not to perform well on pieces with fast tempo and rapid note changes. An example of the output of the post-processing step compared with a ground-truth transcription is given in Fig. 3, using a segment from a piano sonata.

A system variant is also proposed, where after detecting all active pitches in the final piano-roll, the update rules of subsection 2.2, are run again, but instead of setting p as to cover the entire pitch range, we only use the list of active pitches estimated in the first run. This two-stage process also helps in further constraining the solution by removing any pitches that might appear in $P_t(p)$ but are nevertheless removed in the postprocessing step.

3. EVALUATION

3.1 Training Data

Sound state templates are extracted for several orchestral instruments, using isolated note samples from the RWC database [10]. Specifically, we extract templates for bassoon, cello, clarinet, flute, guitar, harpsichord, oboe, organ, piano, tenor sax, and violin, using the CQT as a time-frequency representation [15]. The complete note range of the instruments is used, given the available training data. The sound state templates are computed in an unsupervised manner, using a single-pitch and single-instrument variant of the model of (3), where the number of sound states is set to $Q^{(p)} = 3$.

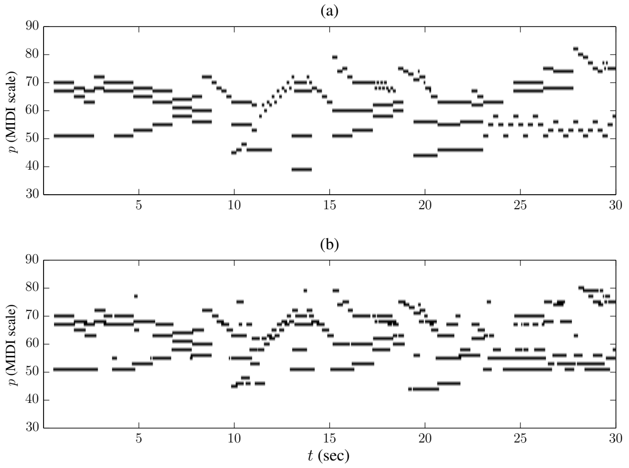

Figure 3. (a) The ground-truth piano-roll of the first 30sec of W.A. Mozart's Piano Sonata K.333, 2nd movement (from the MAPS database). (b) The piano-roll computed from the proposed transcription system.

3.2 Test Data

For testing, we use recordings from three publicly available transcription datasets. Firstly, we used thirty 30sec piano segments from the MAPS database [8], specifically from the 'ENSTDkCl' subset that has been used in the past for multi-pitch evaluation in [4, 17].

We also utilized the woodwind quintet recording used as a development set in the MIREX multiF0 and note tracking task [1]. Instruments present include bassoon, clarinet, flute, horn, and oboe, while manually-aligned ground truth for each instrument track is available online [1].

Finally, we used the TRIOS dataset [9], which includes five multitrack recordings of trio pieces of classical and jazz music. For the current experiments, we used the existing mixes of the multitracks. Instruments included in the dataset are: bassoon, cello, clarinet, horn, piano, saxophone, trumpet, viola, and violin. The dataset includes manually-aligned ground truth with instrument information per pitch. To the authors' knowledge, no transcription results have been reported for the TRIOS dataset.

3.3 Metrics

For evaluating the performance of the proposed system for multi-pitch detection, we employ two sets of metrics: frame-based and note-based ones. For note-based evaluation, we use the onset-based transcription metrics which are used in the MIREX note tracking task [1]. A detected note is considered correct if its pitch matches a ground truth pitch and its onset is within a 50ms tolerance of a ground-truth onset. The resulting note-based precision, recall, and F-measure are defined as:

$$Pre_n = \frac{N_{tp}}{N_{sys}} \quad Rec_n = \frac{N_{tp}}{N_{ref}} \quad F_n = \frac{2Rec_n Pre_n}{Rec_n + Pre_n} \tag{15}$$

where N_{tp} is the number of correctly detected pitches, N_{sys} is the number of pitches detected by the system, and N_{ref} is the number of reference pitches.

For the frame-based metrics, evaluations are performed in a 10ms step as in the MIREX multiF0 evaluations [1], and we use the frame-based precision, recall, and F-measure, which are defined in a similar way to (15) and are denoted as Pre_f, Rec_f, and F_f, respectively.

3.4 Results

Experiments are performed using the proposed system of Section 2 using two variants; a one-stage version (using the update rules and the note tracking step) and the two-stage version presented in subsection 2.3. The proposed EDHMM-based system is compared with the HMM-based system of [2] using the same time-frequency representation and note tracking steps. In all cases, the Markov models were initialized as ergodic, with uniform priors and state transition probabilities.

In Table 1, multi-pitch detection results using the MAPS recordings are shown. It can be seen that using both sets of metrics, the EDHMM-based systems outperform the HMM-based one. It can also be seen that the two-stage version of the system makes a significant improvement in terms of performance. The differences in performance are not as clear using the frame-based metrics, but they are still evident. In all cases, the precision is higher compared to recall (e.g. for the two-stage EDHMM case, $Pre_n = 74.73\%$ and $Rec_n = 64.46\%$), which signifies that there is a larger number of missed detections compared to the number of false alarms. When comparing the reported results with other methods using the same dataset, it can be seen that the proposed system outperforms both the infinite composite autoregressive system of [17] (which reported $F_f = 48.4\%$) and the source-filter NMF model of [4], which reported $F_f = 52.4\%$ (where the best reported performance in [4] was reported using the using the SONIC algorithm [14], reaching $F_f = 58.0\%$). Finally, the note-based accuracy measure for the MAPS recordings is 53.42%; the accuracy reported in [3] for the MAPS-Disklavier dataset was 68.7%, although it should be stressed that in [3] the dataset was also used for training the system.

Results using the MIREX woodwind quintet recording are shown in Table 1; again it can be seen that the EDHMM-based system outperforms the HMM-based one. In the literature, an experiment using the first 30sec of the MIREX recording was made in [16], where $F_f = 62.5\%$. Using the first 30sec in the 2-stage EDHMM system, the frame-based F-measure reaches 66.95%.

Method/Instrument	HMM-based	EDHMM-based
Bassoon	47.72%	41.42%
Clarinet	64.33%	67.68%
Flute	51.18%	57.53%
Horn	39.86%	44.59%
Oboe	23.84%	22.17%
Mean	**45.39**%	**46.68**%

Table 2. Instrument assignment results (F_f) using the first 30sec of the MIREX MultiF0 recording.

Also, using the TRIOS dataset, similar results are reported, as can be seen in Table 1. It should be noted though that there is a large difference between the note-based metrics and the frame-based metrics, which can be attributed to the fact that the TRIOS dataset contains notes with long durations, which get oversegmented in the proposed system (where small gaps do not significantly affect the frame-based metrics).

Finally, we perform experiments on instrument identification using information from matrix $P_t(s|p)$. In the instrument assignment task [11], a detected pitch is considered to be correct if, in addition to pitch and timing constraints, it is assigned to a correct instrument source. We performed experiments using the MIREX woodwind quintet, using a system variant which utilizes templates found in the recording (bassoon, clarinet, flute, horn, oboe). The output (for instrument s) is given by $P_t(p)P_t(s|p)\sum_\omega V_{\omega,t}$. For comparative purposes, we evaluated the first 30sec of the MIREX recording, as in [4], using the frame-based F-measure. Instrument assignment results are shown in Table 2, where it can be seen that the proposed EDHMM-based method performs better compared to the HMM-based one. It can be seen that the best performance is reported for clarinet, which has a relatively different spectral shape compared to the other instruments. It should be noted though that the HMM-based method performs better for bassoon and oboe, while the EDHMM-based method performs better for clarinet, flute, and horn. The reported F_f for the method in [4] is 37.0%, which indicates that the proposed method (which uses pre-extracted spectral templates instead of source-filter models within a spectrogram factorization framework) is more appropriate for the task.

4. CONCLUSIONS

In this paper, we proposed a model for automatic music transcription which models the temporal evolution of notes using pitch-wise explicit duration hidden Markov models, within a spectrogram factorization framework supporting multiple pitch and instrument templates, as well as shift-invariance across log-frequency. It was shown that the tem-

Dataset	MAPS 'ENSTDkCl'		MIREX		TRIOS	
Method / Metric	F_n	F_f	F_n	F_f	F_n	F_f
HMM-based	65.93%	66.41%	63.64%	66.01%	55.94%	67.76%
EDHMM-based	67.12%	66.82%	65.14%	66.42%	56.95%	69.54%
EDHMM-based (2-stage)	68.61%	67.99%	66.60%	66.98%	57.66%	71.17%

Table 1. Multi-pitch detection results (in F_n and F_f) using the three employed datasets.

poral constraints posed by the EDHMMs resulted in improved multi-pitch detection and instrument identification performance when compared to HMM-based constraints. Evaluation results outperformed state-of-the-art multi-pitch detection methods using the MAPS and MIREX datasets. Finally, a proposed two-stage transcription procedure helps in further eliminating transcription errors.

One of the main drawbacks of the proposed method is its computational complexity. Even with independent EDHMMs, the proposed method performs about 60×real-time, which is prohibitive for large-scale experiments or real-time applications. In the future, we will attempt to create computationally-efficient versions of the proposed system using more compact time-frequency representations and by replacing the expensive expectation-maximization algorithm with variational Bayesian methods. Finally, we will expand the existing spectrogram factorization framework in order to introduce additional constraints via musicological models, for example integrating information from chord and key detection for improving multi-pitch detection performance.

5. REFERENCES

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

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

[3] S. Böck and M. Schedl. Polyphonic piano note transcription with recurrent neural networks. In *IEEE Int. Conf. Audio, Speech and Signal Processing*, pages 121–124, March 2012.

[4] J. J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F. J. Cañadas-Quesada. Musical instrument sound multi-excitation model for non-negative spectrogram factorization. *IEEE J. Selected Topics in Signal Processing*, 5(6):1144–1158, October 2011.

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

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

[7] M. Dewar, C. Wiggins, and F. Wood. Inference in hidden Markov models with explicit state duration distributions. *IEEE Signal Processing Letters*, 19(4):235–238, 2012.

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

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

[10] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: music genre database and musical instrument sound database. In *Int. Conf. Music Information Retrieval*, October 2003.

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

[12] P. Grosche, B. Schuller, M. Müller, and G. Rigoll. Automatic transcription of recorded music. *Acta Acustica united with Acustica*, 98(2):199–215, March 2012.

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

[14] M. Marolt. A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Trans. Multimedia*, 6(3):439–449, June 2004.

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

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

[17] K. Yoshii and M. Goto. Infinite composite autoregressive models for music signal analysis. In *13th Int. Society for Music Information Retrieval Conf.*, pages 79–84, October 2012.

[18] S.-Z. Yu. Hidden semi-Markov models. *Artificial Intelligence*, 174(2):215–243, 2010.