Benetos, E. & Dixon, S. (2013). Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. The Journal of the Acoustical Society of America (JASA), 133(3), pp. 1727-1741. doi: 10.1121/1.4790351

Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model

Emmanouil Benetos[a] and Simon Dixon
Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom

(Received 16 August 2012; revised 17 January 2013; accepted 22 January 2013)

A method for automatic transcription of polyphonic music is proposed in this work that models the temporal evolution of musical tones. The model extends the shift-invariant probabilistic latent component analysis method by supporting the use of spectral templates that correspond to sound states such as attack, sustain, and decay. The order of these templates is controlled using hidden Markov model-based temporal constraints. In addition, the model can exploit multiple templates per pitch and instrument source. The shift-invariant aspect of the model makes it suitable for music signals that exhibit frequency modulations or tuning changes. Pitch-wise hidden Markov models are also utilized in a postprocessing step for note tracking. For training, sound state templates were extracted for various orchestral instruments using isolated note samples. The proposed transcription system was tested on multiple-instrument recordings from various datasets. Experimental results show that the proposed model is superior to a non-temporally constrained model and also outperforms various state-of-the-art transcription systems for the same experiment.
 [http://dx.doi.org/10.1121/1.4790351]

PACS number(s): 43.75.Zz, 43.75.Xz, 43.66.Mk [TRM] Pages: 1727–1741

I. INTRODUCTION

The automatic transcription task refers to the conversion of an audio recording into some form of music notation, usually a MIDI file or a music score. It is one of the fundamental problems of music information retrieval (MIR) and has additional applications in computational musicology and the creation of interactive music systems. The core problem of automatic transcription is multi-pitch detection, i.e., pitch estimation of several concurrent sounds over short frames of a recording. Additional subtasks of automatic transcription include onset/offset detection, instrument identification, and the extraction of rhythmic information (Klapuri and Davy, 2006). For an overview of the transcription and multi-pitch detection problem, the reader is referred to Klapuri and Davy (2006) and de Cheveigné (2006). Although for the single-pitch detection case the problem is generally considered to be solved, the multi-pitch case still remains open, especially in the case where the music signal is produced by multiple instruments.

Automatic transcription methods can be categorized according to the various techniques employed for multi-pitch detection. Several techniques employ audio features and music signal processing techniques (e.g., Klapuri and Davy, 2006; Pertusa and Iñesta, 2008; Yeh *et al.*, 2010; Benetos and Dixon, 2011a; Emiya *et al.*, 2010). A large subset of transcription systems (including the present work) employ methods stemming from spectrogram factorization techniques, which exploit the redundancies found in music spectrograms (e.g., Vincent *et al.*, 2010; Mysore and Smaragdis, 2009; Grindlay and Ellis, 2011; Carabias-Orti *et al.*, 2011). In Davy *et al.* (2006), a Bayesian framework for the estimation of pitch, dynamics, and instrument sources was proposed where the unknown parameters are estimated using a Markov chain Monte Carlo (MCMC) method. In Peeling *et al.* (2007) and Peeling and Godsill (2011), a generative model using a non-homogeneous Poisson process was proposed for multi-pitch detection. A machine learning-based transcription system was proposed in Poliner and Ellis (2007), while in Lee *et al.* (2011) sparse coding was used for piano-only transcription. In Duan *et al.* (2010), a maximum likelihood approach was proposed for multiple-F0 estimation by modeling spectral peaks and non-peak regions. Typically hidden Markov models (HMMs) are used in a postprocessing stage for note tracking due to the sequential structure offered by the models (e.g., Poliner and Ellis, 2007; Quesada *et al.*, 2010; Yeh *et al.*, 2010).

One of the drawbacks of current transcription systems is that in most cases, the non-stationarity of music sounds is not addressed. A note produced by a musical instrument can be expressed as a sequence of sound states, for example attack, transient, sustain, and decay parts (Bello *et al.*, 2005). One such example is given in Fig. 1, where the log-frequency spectrogram of a piano note can be seen, and various sound states are labeled. Additionally, depending on the instrument, frequency modulations such as vibrato and amplitude modulations such as tremolo might also take place. The problem of detecting frequency modulations using a single template for relative pitch tracking was addressed by Smaragdis (2009) using shift-invariant probabilistic latent component analysis (PLCA), which will be detailed in Sec. II. Also an algorithm that models the sound evolution in music signals was proposed in Nakano *et al.* (2010) where the non-negative matrix factorization algorithm is combined with HMMs. Finally, a non-parametric Bayesian extension

[a]Author to whom correspondence should be addressed. Electronic mail: emmanouilb@eecs.qmul.ac.uk

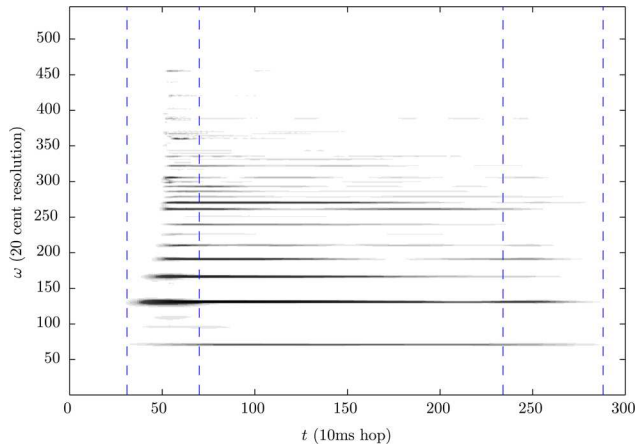

FIG. 1. (Color online) Log-frequency spectrogram of a B1 piano note (constant-Q transform with 60 bins/octave and lowest frequency at 27.5 Hz). The attack and release parts of the note can be seen in the marked areas around frames 50 and 250, respectively.

of the non-negative matrix factorization algorithm for analyzing music spectrograms was proposed by Nakano *et al.* (2011), where a model with infinite-state spectral bases was proposed.

The motivation for this work is to (1) propose a model that can deal with tuning and frequency modulations in music sounds by employing shift-invariance, (2) propose a model that will express the temporal evolution of sounds, and (3) integrate these aforementioned features in a multiple-source, multiple-pitch template model for automatic music transcription. Specifically, the model extends the shift-invariant PLCA technique by adding temporal constraints using multiple HMMs. Spectral templates that correspond to each sound state of a note are used across the complete pitch ranges of multiple orchestral instruments. The sequence of these sound states within a produced note is constrained via pitch-wise HMMs.

At the same time, the shift-invariance of the model supports the presence of tuning changes and frequency modulations within the music signal. A preliminary version of the proposed model was presented in Benetos and Dixon (2012). For the experiments reported in this paper, several sound state templates were extracted using an unsupervised single-pitch version of the proposed model, using the MAPS (Emiya *et al.*, 2010) and RWC (Goto *et al.*, 2003) databases. Experiments were performed using three widely used transcription datasets, and results are reported using several error metrics. It is shown that the proposed model outperforms a non-temporally constrained convolutive probabilistic model (Benetos and Dixon, 2011b) using the same time-frequency representation and note tracking steps. Also, the system is shown to outperform other state-of-the-art transcription systems for the same experiments. Finally, this model can also be applied for instrument identification in poly-phonic music. Instrument assignment experiments are made using the MIREX multi-F0 woodwind recording, where the proposed system produced promising results.

The outline of this paper is as follows. In Sec. II, the PLCA and shift-invariant PLCA methods are presented along with related applications of these methods to automatic music transcription and pitch tracking. The proposed temporally constrained convolutive model for single-pitch detection is detailed in Sec. III, while the multi-pitch model is described in Sec. IV. Section V presents the HMM-based postprocessing step for note tracking. The employed training and test datasets, error metrics, and experimental results on automatic music transcription using the proposed model are shown in Sec. VI. Finally, conclusions are drawn and future directions are indicated in Sec. VII.

II. RELATED WORK

A. PLCA

PLCA is a subspace analysis technique proposed in Smaragdis *et al.* (2006). It can be viewed as a probabilistic extension of the non-negative matrix factorization (NMF) algorithm (Lee and Seung, 1999) using the Kullback–Leibler cost function, providing a framework that is easy to generalize and interpret. PLCA can also offer a convenient way to incorporate priors over the parameters and control the resulting decomposition, for example using *entropic priors* (Shashanka *et al.*, 2008). In PLCA, the input spectrogram, which must be scaled to have integer entries, is modeled as the histogram of the draw of N independent random variables (ω_n, t_n), which are distributed according to $P(\omega,t)$ (ω denotes frequency, and t time) and a component activity matrix.

There are two ways of modeling $P(\omega,t)$, using symmetric or asymmetric factorizations. For the symmetric model, $P(\omega, t)$ is expressed as a mixture of two-dimensional latent factors with each factor being a product of one-dimensional marginal distributions (Shashanka *et al.*, 2008) and can be expressed as

$$P(\omega,t) = \sum_z P(z)P(\omega|z)P(t|z), \tag{1}$$

where z is the component index, $P(z)$ refers to the component weights, $P(\omega|z)$ is the spectral template that corresponds to the zth component, and $P(t|z)$ describes the time-varying energy of each component. In the context of music signal analysis, the components (or latent factors) typically refer to the constituent elements of a spectrogram (e.g., pitches or instrument sources).

The asymmetric factorization, which is called PLCA, treats ω and t differently and decomposes $P(\omega, t)$ as a product of a spectral basis matrix and a component activity matrix. It can be expressed as

$$P(\omega,t) = P(t)\sum_z P(\omega|z)P(z|t), \tag{2}$$

where z is the component index, $P(t)$ is the energy of the input spectrogram (known quantity), $P(\omega|z)$ is the spectral template that corresponds to the zth component, and $P(z|t)$ is the activation of the zth component. To estimate the unknown parameters $P(\omega|z)$ and $P(z|t)$, iterative update rules are applied, using the Expectation–Maximization (EM) algorithm (Dempster *et al.*, 1977). The derivation of the EM algorithm for PLCA

can be found in Smaragdis and Raj (2007). The update rules are guaranteed to converge to a local minimum.

Concerning PLCA-based work on music signal analysis, Grindlay and Ellis (2011) proposed an extension to the PLCA model for multiple-instrument transcription, supporting templates for multiple instrument sources. The notion of *eigeninstruments* was presented, by modeling the fixed spectral templates as a linear combination of basic instrument models in a training step. Sparsity was enforced on the pitch activity matrix and the source contribution matrix by modifying the model update equations. Experiments were performed on J. S. Bach duets and on pairs of tracks from the multi-track MIREX multi-F0 woodwind recording (MIREX, 2007), which is also used in this work.

In Mysore (2010), temporal constraints were incorporated into the PLCA model for music signal analysis. This *non-negative hidden Markov model* expressed each note using a set of spectral templates linked to a hidden state in an HMM. Parameter estimation was achieved using the PLCA update rules combined with the HMM forward-backward procedure (Rabiner, 1989). An extension for two sources was also proposed by Mysore for source separation, which employed factorial HMMs (Ghahramani and Jordan, 1997). It should be noted that the model of Mysore (2010) cannot be used directly for automatic music transcription; the proposed approach extends the model of Mysore (2010) by incorporating shift-invariance across log-frequency and by introducing a sound state-pitch-instrument hierarchy instead of a component-source hierarchy.

B. Shift-invariant PLCA

Incorporating a shift-invariant model into the PLCA framework is practical because the sum of two random variables corresponds to a convolution of their distribution. Shift-invariant PLCA (Smaragdis *et al.*, 2008) was proposed for extracting shifted structures in non-negative data. It has been used in music signal processing applications using a normalized log-frequency spectrogram as an input because a shift over log-frequency corresponds to a pitch change. The shift-invariant PLCA (SI-PLCA) model can be defined as

$$P(\omega,t)=\sum_z P(z)\sum_f P(\omega-f|z)P(f,t|z), \tag{3}$$

where ω is the log-frequency index, z the component index, and f the shifting factor. $P(\omega - f|z)=P(\mu|z)$ denotes the spectral template for the zth component, $P(f,t|z)$ the time-varying pitch shifting, and $P(z)$ the component prior. Again the EM algorithm can be used for deriving update rules for the unknown parameters. An example of an SI-PLCA model is given in Fig. 2, where the input log-frequency spectrogram of a violin glissando is decomposed into a spectral template and a pitch impulse distribution.

In Smaragdis (2009), the SI-PLCA model was used for relative pitch tracking, where sparsity was enforced on the unknown matrices using an entropic prior. Mysore and Smaragdis (2009) used the SI-PLCA model for multiple-instrument relative pitch tracking, tested on the MIREX multi-F0 recording (MIREX, 2007). For eliminating octave errors, a sliding-Gaussian Dirichlet prior was used in the model, while a temporal continuity constraint using a Kalman filter type smoothing was applied to $P(f, t|z)$ to extract a smooth pitch track.

More recently, an extension of the SI-PLCA algorithm was proposed for harmonic signals by Fuentes *et al.* (2011). Each note is modeled as a weighted sum of narrowband log-spectra that are also shifted across log-frequency. This

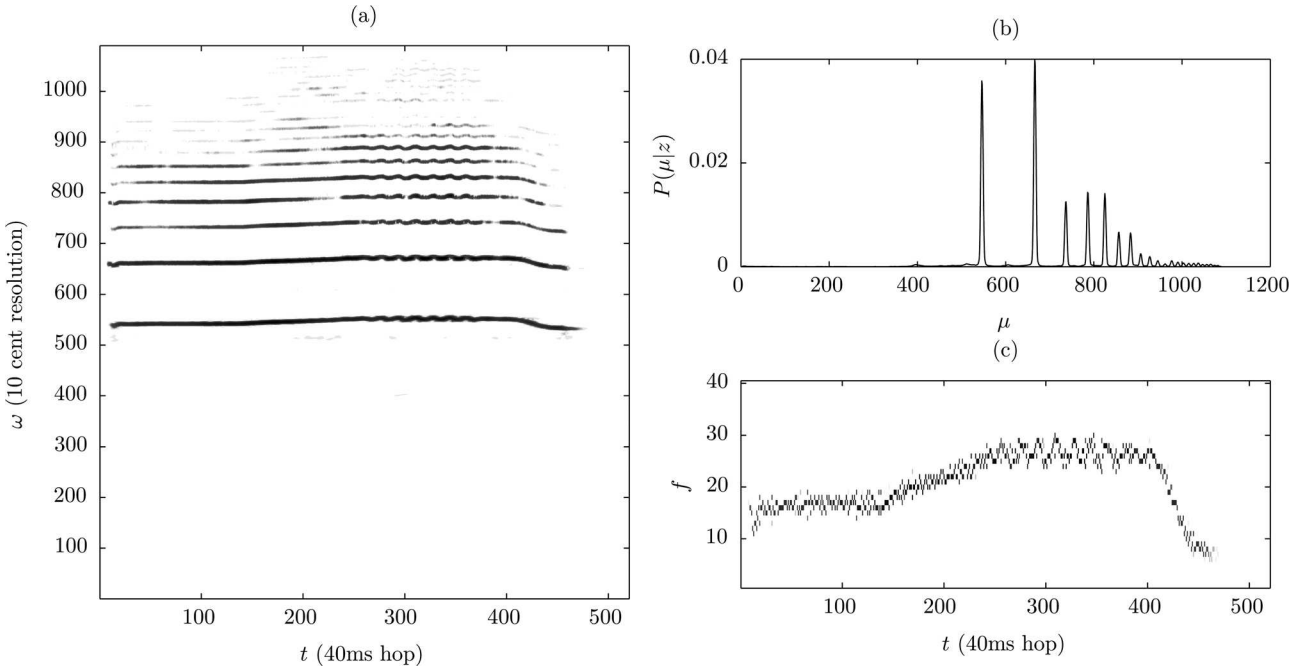

FIG. 2. A shift-invariant PLCA decomposition of a violin glissando (constant-Q transform with 120 bins/octave and lowest frequency at 27.5 Hz) with $z=1$. (a) Input log-frequency spectrogram, (b) spectral template $P(\mu|z)$, (c) pitch shift $P(f, t|z)$.

approach is a convolutive probabilistic formulation of the harmonic NMF algorithm proposed by Vincent (Vincent *et al.*, 2010) with added time dependence for the weights of the narrowband spectra. The harmonic SI-PLCA method was tested for single-pitch detection on isolated note samples, and a model was proposed for multi-pitch detection. An asymmetric minimum variance prior was also incorporated into the parameter update rules to eliminate any harmonic errors.

Finally, the authors proposed an extension of the shift-invariant PLCA model for automatic transcription in Benetos and Dixon (2011b). The model supported the use of one spectral template per pitch and instrument source. Each template could be shifted across log-frequency in a semitone range around the ideal pitch position. This system was also publicly evaluated in the MIREX 2011 competition (MIREX, 2007), where it ranked second in the note tracking task.

III. SINGLE-PITCH MODEL

In this section, a temporally constrained shift-invariant model first introduced by the authors in Benetos and Dixon (2011c) will be presented. The model expresses the evolution of monophonic music sounds as a sequence of sound state templates, shifted across log-frequency. The motivation behind it is to address drawbacks of current pitch detection approaches by (1) explicitly modeling sound states instead of using a constant spectral template for a complete note event, as in Smaragdis (2009), Mysore and Smaragdis (2009), and Grindlay and Ellis (2011) and (2) incorporating shift-invariance into the model to support the detection of notes that exhibit frequency modulations and tuning changes, extending the work done in Mysore (2010) and Nakano *et al.* (2010). Finally, compared to the NMF-based work in Nakano *et al.* (2010), the parameters for the temporal constraints are learned from a hidden Markov model instead of being pre-defined.

A. Formulation

The proposed method can be named as HMM-constrained SI-PLCA. The notion is that the input log-frequency spectrogram $V_{\omega,t}$ is decomposed as a sum of sound state spectral templates that are shifted across log-frequency, producing a pitch track. Each sound state q is constrained using an HMM. Here, $\omega \in [1, \Omega]$ is the log-frequency index and $t \in [1, T]$ the time index. The model in terms of the observations is defined as

$$P(\overline{\omega}) = \sum_{\overline{q}} \left(P(q_1) \prod_t P(q_{t+1}|q_t) \right) \left(\prod_t P(\overline{\omega}_t|q_t) \right), \quad (4)$$

where $\overline{\omega}$ is the complete sequence of draws for all time frames (observable via $V_{\omega,t}$), $\overline{q}$ is the sequence of draws of q, $P(q_1)$ is the sound state prior distribution, $P(q_{t+1}|q_t)$ is the state transition matrix, $P(\overline{\omega}_t|q_t)$ is the observation probability given a state, and $\overline{\omega}_t$ is the sequence of draws of ω at the tth frame.

The observation probability is calculated as

$$P(\overline{\omega}_t|q_t) = \prod_{\omega_t} P_t(\omega_t|q_t)^{V_{\omega,t}} \quad (5)$$

because $V_{\omega,t}$ represents the number of times ω has been drawn at time t. $P_t(\omega_t|q_t)$ is decomposed as

$$P_t(\omega_t|q_t) = \sum_{f_t} P(\omega_t - f_t|q_t) P_t(f_t|q_t). \quad (6)$$

Equation (6) denotes the spectrum reconstruction for a given state. $P(\omega - f|q) = P(\mu|q)$ are the sound state templates and $P_t(f|q)$ is the time-dependent pitch track for each state ($f \in [1, F]$). The subscript t in f_t, ω_t, q_t denotes the values of the random variables f, ω, q taken at frame t. It should also be noted that the observation probability of Eq. (5) is computed in the log-domain to avoid any underflow errors.

Thus the generative process for the proposed model is as follows:

(1) Choose an initial state according to $P(q_1)$.
(2) Set $t = 1$.
(3) Repeat the following steps V_t times ($V_t = \sum_\omega V_{\omega,t}$):
  (a) Choose μ according to $P(\mu_t|q_t)$.
  (b) Choose f according to $P_t(f_t|q_t)$.
  (c) Set $\omega_t = \mu_t + f_t$.
(4) Choose a new state q_{t+1} according to $P(q_{t+1}|q_t)$.
(5) Set $t = t + 1$ and go to step 3 if $t < T$.

B. Parameter estimation

The unknown parameters $P(\mu_t|q_t)$ and $P_t(f_t|q_t)$ can be estimated by maximizing the log-likelihood of the data, using the EM algorithm (Dempster *et al.*, 1977). The update equations are a combination of the SI-PLCA update rules and the HMM forward-backward algorithm (Rabiner, 1989). The posterior distribution of the model is given by $P(\overline{f}, \overline{q}|\overline{\omega})$, where $\overline{f}$ is the sequence of draws of f.

For the *Expectation* step, we compute the contribution of the latent variables f, q over the complete model reconstruction:

$$P_t(f_t, q_t|\overline{\omega}) = \frac{P_t(f_t|\overline{\omega}, q_t) P_t(\overline{\omega}, q_t)}{P(\overline{\omega})} P_t(f_t|\omega_t, q_t) P_t(q_t|\overline{\omega}), \quad (7)$$

where

$$P_t(f_t|\omega_t, q_t) = \frac{P(\omega_t - f_t|q_t) P_t(f_t|q_t)}{\sum_{f_t} P(\omega_t - f_t|q_t) P_t(f_t|q_t)}, \quad (8)$$

$$P_t(q_t|\overline{\omega}) = \frac{P_t(\overline{\omega}, q_t)}{\sum_{q_t} P_t(\overline{\omega}, q_t)} = \frac{\alpha_t(q_t)\beta_t(q_t)}{\sum_{q_t} \alpha_t(q_t)\beta_t(q_t)}. \quad (9)$$

Equation (7) is the posterior of the hidden variables over the observations and is computed using the fact that $P_t(f_t|\overline{\omega}, q_t) = P_t(f_t|\omega_t, q_t)$. Equation (8) is computed using Bayes' rule and the notion that $P(\omega_t|f_t, q_t) = P(\omega_t - f_t|q_t)$. Equation (9) is the time-varying contribution of each sound state and is derived from the following:

$$\begin{aligned} P_t(\overline{\omega}, q_t) &= P(\overline{\omega}_1, \overline{\omega}_2, \ldots, \overline{\omega}_t, q_t) P(\overline{\omega}_{t+1}, \overline{\omega}_{t+2}, \ldots, \overline{\omega}_T|q_t) \\ &= \alpha_t(q_t)\beta_t(q_t), \end{aligned} \quad (10)$$

where T is the total number of frames, and $\alpha_t(q_t)$ and $\beta_t(q_t)$ are the HMM forward and backward variables (Rabiner, 1989), respectively.

The forward variable $\alpha_t(q_t)$ can be computed recursively using the forward-backward algorithm as follows:

$$\alpha_1(q_1) = P(\overline{\omega}_1|q_1)P(q_1),$$
$$\alpha_{t+1}(q_{t+1}) = \left(\sum_{q_t} P(q_{t+1}|q_t)\alpha_t(q_t)\right) \cdot P(\overline{\omega}_{t+1}|q_{t+1}), \tag{11}$$

while the backward variable $\beta_t(q_t)$ can be computed as

$$\beta_T(q_T) = 1,$$
$$\beta_t(q_t) = \sum_{q_{t+1}} \beta_{t+1}(q_{t+1})P(q_{t+1}|q_t)P(\overline{\omega}_{t+1}|q_{t+1}). \tag{12}$$

The posterior for the sound state transition matrix is given by

$$P_t(q_t, q_{t+1}|\overline{\omega}) = \frac{P_t(\overline{\omega}, q_t, q_{t+1})}{\sum_{q_t}\sum_{q_{t+1}} P_t(\overline{\omega}, q_t, q_{t+1})} = \frac{\alpha_t(q_t)P(q_{t+1}|q_t)\beta_{t+1}(q_{t+1})P(\overline{\omega}_{t+1}|q_{t+1})}{\sum_{q_t,q_{t+1}} \alpha_t(q_t)P(q_{t+1}|q_t)\beta_{t+1}(q_{t+1})P(\overline{\omega}_{t+1}|q_{t+1})}. \tag{13}$$

For the *Maximization* step, we derive the update equations for the unknown parameters $P(\mu|q)$, $P_t(f_t|q_t)$, $P(q_{t+1}|q_t)$, and $P(q_1)$ using the computed posteriors:

$$P(\mu|q) = \frac{\sum_{f,t} V_{\omega,t}P_t(f, q|\overline{\omega})}{\sum_{\omega,f,t} V_{\omega,t}P_t(f, q|\overline{\omega})}, \tag{14}$$

$$P_t(f_t|q_t) = \frac{\sum_{\omega_t} V_{\omega,t}P_t(f_t, q_t|\overline{\omega})}{\sum_{f_t,\omega_t} V_{\omega,t}P_t(f_t, q_t|\overline{\omega})}, \tag{15}$$

$$P(q_{t+1}|q_t) = \frac{\sum_t P_t(q_t, q_{t+1}|\overline{\omega})}{\sum_{q_{t+1}}\sum_t P_t(q_t, q_{t+1}|\overline{\omega})}, \tag{16}$$

$$P(q_1) = P_1(q_1|\overline{\omega}). \tag{17}$$

After estimating the unknown parameters, the activation of each sound state is given by

$$P_t(q_t|\overline{\omega})\sum_{\omega} V_{\omega,t}. \tag{18}$$

An example of the single-source model is given in Fig. 3, where the 10-cent resolution log-frequency spectrogram of a B1 piano note from the MAPS database (Emiya *et al.*, 2010) is used as input. Here, a four-state left-to-right HMM is used. The temporal succession of spectral templates can be seen in Fig. 3(d).

IV. MULTI-PITCH MODEL

Here we will extend the single-source model of Sec. III for supporting multiple sources as well as multiple components per source. The goal is to create a multi-pitch detection system for multiple instruments, supporting also multiple sets of sound state tem-plates per source. At the same time, the model will be able to support tuning changes and frequency modulations using a shift-invariant formulation. For modeling, the temporal evolution of the sound state templates, one HMM will be linked with each pitch. Sparsity will also be enforced on certain distributions, as in Grindlay and Ellis (2011) and Benetos and Dixon (2011b) for further constraining the solution. All of the preceding features will allow for an informative representation of the input music signal, addressing some drawbacks of current multi-pitch detection systems.

A. Formulation

This model decomposes an input log-frequency spectrogram $V_{\omega,t}$ as a series of sound state templates per source and pitch, a shifting parameter per pitch, a pitch activation, a source activation, and a sound state activation. The sound state sequence for each pitch $p = 1,\ldots, 88$ (denoting notes A0 to C8) is constrained using a corresponding HMM. The proposed model can be given in terms of the observations as

$$P(\overline{\omega}) = \sum_{\overline{q}^{(1)}} \cdots \sum_{\overline{q}^{(88)}} P\left(q_1^{(1)}\right) \cdots P\left(q_1^{(88)}\right) \times \left(\prod_t P\left(q_{t+1}^{(1)}|q_t^{(1)}\right)\right) \cdots \left(\prod_t P\left(q_{t+1}^{(88)}|q_t^{(88)}\right)\right) \times \left(\prod_t P\left(\overline{\omega}_t|q_t^{(1)}, \ldots, q_t^{(88)}\right)\right), \tag{19}$$

where $\overline{q}^{(p)}$ refers to the state sequences for a given pitch, $P(q_1^{(p)})$ is the sound state prior distribution for pitch p, $P(q_{t+1}^{(p)}|q_t^{(p)})$ is the sound state transition matrix, and $P(\overline{\omega}_t|q_t^{(1)}, \ldots, q_t^{(88)})$ is the observation probability.

The observation probability is calculated as

$$P\left(\overline{\omega}_t|q_t^{(1)}, \ldots, q_t^{(88)}\right) = \prod_{\omega_t} P_t\left(\omega_t|q_t^{(1)}, \ldots, q_t^{(88)}\right)^{V_{\omega,t}}, \tag{20}$$

where

$$P_t\left(\omega_t|q_t^{(1)}, \ldots, q_t^{(88)}\right) = \sum_{s_t,p_t,f_t} P_t(p_t)P_t(s_t|p_t)P\left(\omega_t - f_t|s_t, p_t, q_t^{(p_t)}\right)P_t(f_t|p_t). \tag{21}$$

In Eq. (21), s denotes the instrument sources, f is the log-frequency pitch shifting parameter, and $q^{(p)}$ is the sound state sequence linked to pitch p. $P_t(p)$ is the pitch activity matrix (which is the output of the transcription system), and $P_t(s|p)$ is the contribution of each instrument source for each pitch across time. $P(\omega - f|s, p, q^{(p)}) = P(\mu|s, p, q^{(p)})$ denotes a spectral template for the qth sound state, pth pitch and sth source, and $P_t(f|p)$ is the time- and pitch-dependent log-frequency shifting distribution. For computing Eq. (21), we exploit the fact that $P(\omega_t - f_t|s_t, p_t, q_t^{(1)}, \ldots, q_t^{(88)}) = P(\omega_t - f_t|s_t, p_t, q_t^{(p_t)})$. To constrain the pitch shifting f so that each sound state template is associated with a single pitch, the shifting occurs in a semitone

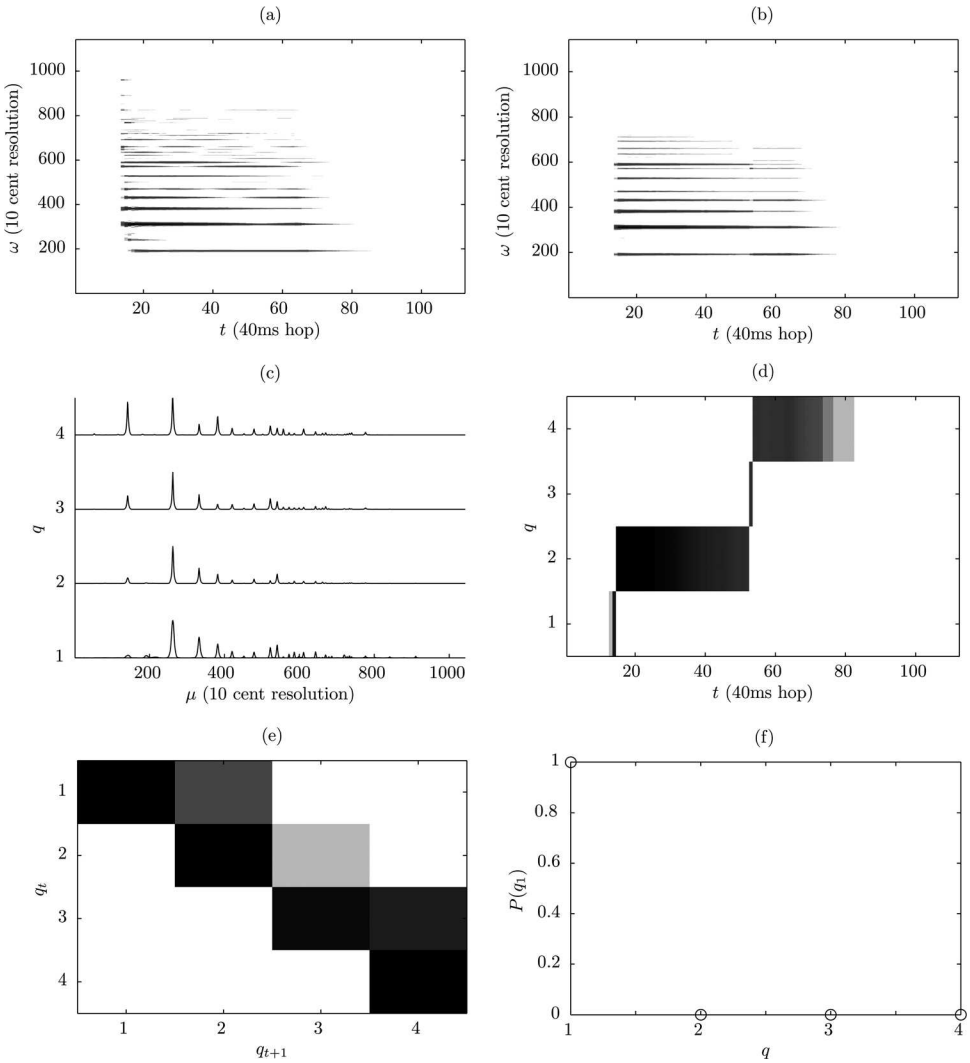

FIG. 3. (a) Log-frequency spectrogram $V_{\omega,t}$ of a B1 piano note. (b) Approximation of the spectrogram using estimated parameters from the single-source model. (c) Spectral templates $P(\mu|q)$. (d) Sound state activation $P_t(q_t|\overline{\omega})\sum_\omega V_{\omega,t}$. (e) Sound state transition matrix $P(q_{t+1}|q_t)$. (f) Sound state priors $P(q_1)$.

range around the ideal position of each pitch. Thus because we are using in this paper a log-frequency representation with a spectral resolution of 60 bins/octave, $f \in [-2, 2]$.

Thus the generative process for the multi-pitch model is as follows:

(1) Choose initial states for each p according to $P(q_1^{(p)})$.
(2) Set $t = 1$.
(3) Repeat the following steps V_t times ($V_t = \sum_\omega V_{\omega,t}$):
  (a) Choose p according to $P_t(p_t)$.
  (b) Choose s according to $P_t(s_t|p_t)$.
  (c) Choose f according to $P_t(f_t|p_t)$.
  (d) Choose μ according to $P(\mu_t|s_t, p_t, q_t^{(p_t)})$.
  (e) Set $\omega_t = \mu_t + f_t$.
(4) Choose new states $q_{t+1}^{(p)}$ for each p according to $P(q_{t+1}^{(p)}|q_t^{(p)})$.
(5) Set $t = t + 1$ and go to step 3 if $t < T$.

B. Parameter estimation

As in Sec. III, the unknown model parameters can be estimated using the EM algorithm (Dempster *et al.*, 1977). For the *Expectation* step, the posterior of all hidden variables is given by

$$P_t\left(f_t, s_t, p_t, q_t^{(1)}, \ldots q_t^{(88)}|\overline{\omega}\right) = P_t\left(q_t^{(1)}, \ldots q_t^{(88)}|\overline{\omega}\right) P_t\left(f_t, s_t, p_t|\omega_t, q_t^{(1)}, \ldots q_t^{(88)}\right). \tag{22}$$

Because independent HMMs are used, the joint probability of all pitch-wise sound states over the observations is given by

$$P_t\left(q_t^{(1)}, \ldots q_t^{(88)}|\overline{\omega}\right) = \prod_{p=1}^{88} P_t\left(q_t^{(p)}|\overline{\omega}\right), \tag{23}$$

where

$$P_t\left(q_t^{(p)}|\overline{\omega}\right) = \frac{P_t\left(\overline{\omega}, q_t^{(p)}\right)}{\sum_{q_t^{(p)}} P_t\left(\overline{\omega}, q_t^{(p)}\right)} = \frac{\alpha_t\left(q_t^{(p)}\right)\beta_t\left(q_t^{(p)}\right)}{\sum_{q_t^{(p)}} \alpha_t\left(q_t^{(p)}\right)\beta_t\left(q_t^{(p)}\right)} \tag{24}$$

and $\alpha_t(q_t^{(p)})$, $\beta_t(q_t^{(p)})$ are the forward and backward variables for the pth HMM (Rabiner, 1989), which can be computed recursively using Eqs. (11) and (12). The second term of Eq.

(22) can be computed using Bayes' theorem and the independence of the pitch-wise HMMs as

$$P_t\left(f_t, s_t, p_t|\omega_t, q_t^{(1)}, \ldots, q_t^{(88)}\right) = P_t\left(f_t, s_t, p_t|\omega_t, q_t^{(p_t)}\right) = \frac{P_t(p_t)P\left(\omega_t - f_t|s_t, p_t, q_t^{(p_t)}\right)P_t(f_t|p_t)P_t(s_t|p_t)}{\sum_{p_t} P_t(p_t)\sum_{s_t,f_t} P\left(\omega_t - f_t|s_t, p_t, q_t^{(p_t)}\right)P_t(f_t|p_t)P_t(s_t|p_t)}. \quad (25)$$

Finally, the posterior probability for the pth pitch transition matrix is given by

$$P_t\left(q_{t+1}^{(p)}, q_t^{(p)}|\overline{\omega}\right) = \frac{\alpha_t\left(q_t^{(p)}\right)P\left(q_{t+1}^{(p)}|q_t^{(p)}\right)\beta_{t+1}\left(q_{t+1}^{(p)}\right)P\left(\overline{\omega}_{t+1}|q_{t+1}^{(p)}\right)}{\sum_{q_t^{(p)}}\sum_{q_{t+1}^{(p)}} \alpha_t\left(q_t^{(p)}\right)P\left(q_{t+1}^{(p)}|q_t^{(p)}\right)\beta_{t+1}\left(q_{t+1}^{(p)}\right)P\left(\overline{\omega}_{t+1}|q_{t+1}^{(p)}\right)}, \quad (26)$$

where $P(\overline{\omega}_t|q_t^{(p)})$ is given from $\overline{\sum}_{q_t^{(p)}} P(\overline{\omega}|q_t^{(1)}, \ldots, q_t^{(88)}) \times P(q_t^{(1)}, \ldots, q_t^{(p-1)}, q_t^{(p+1)}, \ldots, q_t^{(88)})$, where $\overline{\sum}_{q_t^{(p)}} = \sum_{q_t^{(1)}} \ldots \sum_{q_t^{(p-1)}}\sum_{q_t^{(p+1)}} \ldots \overline{\sum}_{q_t^{(88)}}$.

For the *Maximization* step, the unknown parameters in the model can be computed using the following update equations:

$$P\left(\mu|s, p, q^{(p)}\right) = \frac{\sum_{f,s,t}\overline{\sum_{q_t^{(p)}}} V_{\omega,t}P_t\left(f, s, p, q^{(1)}, \ldots, q^{(88)}|\overline{\omega}\right)}{\sum_{\omega,f,s,t}\overline{\sum_{q_t^{(p)}}} V_{\omega,t}P_t\left(f, s, p, q^{(1)}, \ldots, q^{(88)}|\overline{\omega}\right)}, \quad (27)$$

$$P_t(f_t|p_t) = \frac{\sum_{\omega_t,s_t}\sum_{q_t^{(1)}}\ldots\sum_{q_t^{(88)}} V_{\omega,t}P_t\left(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(88)}|\overline{\omega}\right)}{\sum_{f_t,\omega_t,s_t}\sum_{q_t^{(1)}}\ldots\sum_{q_t^{(88)}} V_{\omega,t}P_t\left(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(88)}|\overline{\omega}\right)}, \quad (28)$$

$$P_t(s_t|p_t) = \frac{\sum_{\omega_t,f_t}\sum_{q_t^{(1)}}\ldots\sum_{q_t^{(88)}} V_{\omega,t}P_t\left(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(88)}|\overline{\omega}\right)}{\sum_{s_t,\omega_t,f_t}\sum_{q_t^{(1)}}\ldots\sum_{q_t^{(88)}} V_{\omega,t}P_t\left(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(88)}|\overline{\omega}\right)}, \quad (29)$$

$$P_t(p_t) = \frac{\sum_{\omega_t,f_t,s_t}\sum_{q_t^{(1)}}\ldots\sum_{q_t^{(88)}} V_{\omega,t}P_t\left(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(88)}|\overline{\omega}\right)}{\sum_{p_t,\omega_t,f_t,s_t}\sum_{q_t^{(1)}}\ldots\sum_{q_t^{(88)}} V_{\omega,t}P_t\left(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(88)}|\overline{\omega}\right)}, \quad (30)$$

$$P\left(q_{t+1}^{(p)}|q_t^{(p)}\right) = \frac{\sum_t P_t\left(q_t^{(p)}, q_{t+1}^{(p)}|\overline{\omega}\right)}{\sum_{q_{t+1}^{(p)}}\sum_t P_t\left(q_t^{(p)}, q_{t+1}^{(p)}|\overline{\omega}\right)}, \quad (31)$$

$$P\left(q_1^{(p)}\right) = P_1\left(q_1^{(p)}|\overline{\omega}\right). \quad (32)$$

It should be noted that the proposed multi-pitch transcription system uses pre-extracted sound state templates using the single-pitch model of Sec. III, thus the spectral template update rule of Eq. (27) is not utilized but is included here for completeness. After convergence using the update equations from the EM steps, the output of the system is a pitch activity matrix in MIDI scale and a pitch activity tensor in the resolution of the input time-frequency (T/F) representation, given by

$$P_t(p)\sum_\omega V_{\omega,t}, \quad P_t(p)P_t(f|p)\sum_\omega V_{\omega,t}. \quad (33)$$

A time-pitch representation can be created by stacking together matrix slices of tensor $P_t(p)P_t(f|p)\sum_\omega V_{\omega,t}$ for all pitch values. We will denote this time-pitch representation as $P(f', t)$, which can be used for pitch visualization purposes or for extracting tuning information. An example of the proposed model is given in Fig. 4, where the output time-pitch representation $P(f', t)$ and the MIDI ground-truth of a guitar recording can be seen.

C. Sparsity

The multi-pitch model can be further constrained using sparsity restrictions. Sparsity was enforced in the shift-invariant models of Smaragdis (2009) and Mysore and Smaragdis (2009), using an entropic prior. However, those models were completely unconstrained because the spectral templates were not pre-extracted. Because we know that for a transcription problem few notes are active at a given time frame and that few instrument sources are responsible for creating a note event at a time frame, we impose sparsity on the pitch activity matrix $P_t(p_t)$ and the pitch-wise source contribution matrix $P_t(s_t|p_t)$. This is achieved in a similar way to the methods in Grindlay and Ellis (2011) and Benetos and Dixon (2011b), by modifying update Eqs. (29) and (30):

$$P_t(s_t|p_t) = \frac{\left(\sum_{\omega_t,f_t,q_t^{(1)},\ldots,q_t^{(88)}} V_{\omega,t}P_t\left(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(88)}|\overline{\omega}\right)^\kappa\right)}{\sum_{s_t}\left(\sum_{\omega_t,f_t,q_t^{(1)},\ldots,q_t^{(88)}} V_{\omega,t}P_t\left(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(88)}|\overline{\omega}\right)^\kappa\right)}, \quad (34)$$

$$P_t(p_t) = \frac{\left(\sum_{\omega_t,f_t,s_t,q_t^{(1)},\ldots,q_t^{(88)}} V_{\omega,t}P_t\left(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(88)}|\overline{\omega}\right)^\nu\right)}{\sum_{p_t}\left(\sum_{\omega_t,f_t,s_t,q_t^{(1)},\ldots,q_t^{(88)}} V_{\omega,t}P_t\left(f_t, s_t, p_t, q_t^{(1)}, \ldots, q_t^{(88)}|\overline{\omega}\right)^\nu\right)}. \quad (35)$$

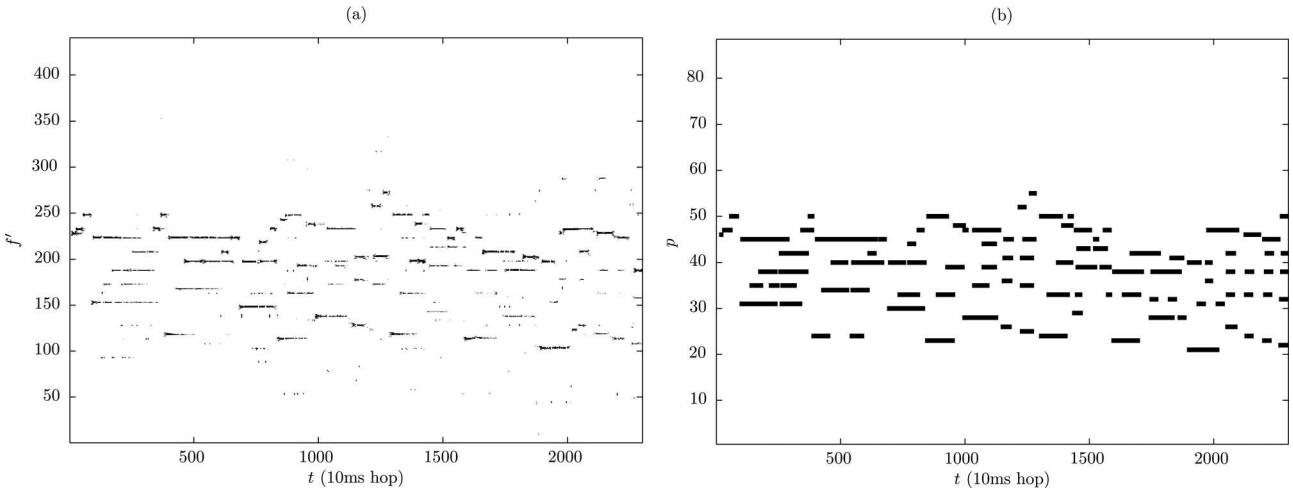

FIG. 4. (a) Time-pitch representation $P(f', t)$ of an excerpt of "RWC-MDB-J-2001 No. 7" (guitar). (b) The pitch ground truth of the same recording.

By setting $\kappa, \nu > 1$, the entropy in matrices $P_t(s|p)$ and $P_t(p)$ is lowered and sparsity is enforced (Grindlay and Ellis, 2011). It should be mentioned that this solution does not guarantee convergence, although it is observed in practice.

V. POSTPROCESSING

To estimate the note activations from the pitch activity matrix $P_t(p)\sum_\omega V_{\omega,t}$, a postprocessing step needs to take place. Most spectrogram factorization-based transcription or pitch tracking methods (Grindlay and Ellis, 2011; Mysore and Smaragdis, 2009; Dessein *et al.*, 2010) estimate the note activations by thresholding the pitch activity matrix. However, HMMs have been used in audio feature-based transcription approaches for note tracking, using the salience function of the system (Poliner and Ellis, 2007; Benetos and Dixon, 2011a; Ryynänen and Klapuri, 2005). Here we will employ pitch-wise HMMs for note tracking using $P_t(p)\sum_\omega V_{\omega,t}$, as in the non-temporally constrained single pitch template system of Benetos and Dixon (2011b).

We model each pitch using a two-state, on/off HMM, which denotes pitch activity/inactivity. The hidden state sequence for each pitch, which is the output of the note tracking step, is given by $Q'^{(p)} = \{q_t'^{(p)}\}$. For computing the note priors and transition matrices, we used 130 MIDI files from the classic and jazz genres from the RWC database (Goto *et al.*, 2003). The notes that are present in the training set fall within the A1–E6 range, which is representative for the RWC test recordings presented in Sec. VI B. The prior probability for an active note that is lower than A1 or higher than E6 is automatically set to 0.1. We denote the state priors for each pitch p as $P(q_1'^{(p)})$ and the corresponding transitions as $P(q_t'^{(p)}|q_{t-1}'^{(p)})$. The most likely state sequence for each pitch is given by

$$\hat{Q}'^{(p)} = \underset{q_t'^{(p)}}{\operatorname{argmax}} \prod_t P\left(q_t'^{(p)}|q_{t-1}'^{(p)}\right) P\left(o_t^{(p)}|q_t'^{(p)}\right), \tag{36}$$

where $P(o_t^{(p)}|q_t'^{(p)})$ is the observation probability for the p-HMM. Equation (36) can be estimated using the Viterbi algorithm (Rabiner, 1989). We define the observation probability for an active note event using $P(p, t)$ as

$$P\left(o_t^{(p)}|q_t'^{(p)} = 1\right) = \frac{1}{1 + e^{-P_t(p)\sum_\omega V_{\omega,t} - \lambda}}. \tag{37}$$

Equation (37) is a sigmoid curve with $P_t(p)\sum_\omega V_{\omega,t}$ as input. Parameter λ controls the smoothing (a high value will discard pitch candidates with low probability). Essentially, in a case of high values in the pitch activation for a given note, where a gap might occur due to an octave error, a high transition probability in an active state would help filling in that gap, thus performing note smoothing. The output of the postprocessing step is a piano-roll transcription, which can be used for evaluation. An example of the HMM-based note tracking step is given in Fig. 5, where the input pitch activity matrix and the output transcription piano roll of a string quartet recording can be seen.

VI. EXPERIMENTS

A. Training data

Sound state templates were extracted for various instruments, using their complete note range given the training data available. For extracting piano templates, the MAPS database was employed (Emiya *et al.*, 2010), where templates from three different piano models were extracted. Sound state templates for bassoon, cello, clarinet, flute, guitar, harpsichord, horn, oboe, organ, and violin were extracted using isolated notes from the RWC musical instrument samples database (Goto *et al.*, 2003). In total, source parameter s has a size of 13 (three sets of templates from the piano and 10 for the rest of the instruments). The note range of each instrument used for sound state template extraction can be seen in Table I. It should be noted that the proposed algorithm can support different note ranges for the existing instruments or can support additional instruments.

Ground-truth labels were given for each note and instrument type, but the sound state templates for each note segment were computed in an unsupervised manner, where the model learns the templates using the single-pitch model of Sec. III. Three sound states were set in the model of Eq. (6). As a time-frequency representation, the constant-Q transform (CQT) with 60 bins/octave was used (Brown, 1991).

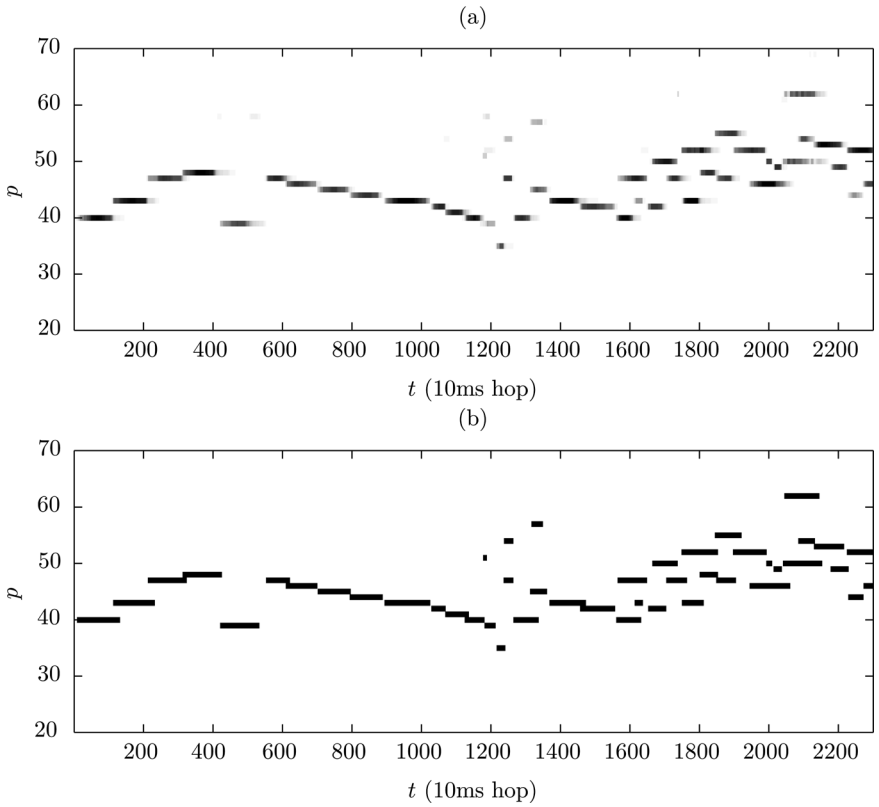

FIG. 5. (a) The pitch activity matrix $P_t(p)$ $\sum_\omega V_{\omega,t}$ of the first 23 s of "RWC-MDB-C-2001 No. 12" (string quartet). (b) The piano-roll transcription output of the note tracking step.

B. Test data

For testing, 12 excerpts from the RWC database (Goto *et al.*, 2003) were employed that have been used extensively for evaluating transcription systems (Kameoka *et al.*, 2007; Quesada *et al.*, 2010; Benetos and Dixon, 2011a). Details for the employed recordings can be seen, e.g., in Quesada *et al.* (2010). The excerpts belong to the classic and jazz genres from the RWC database and the duration of each excerpt is 23 s.

In addition, we employed the woodwind quintet recording from the MIREX 2007 multi-F0 development dataset (MIREX, 2007). The multi-track recording has been evaluated in the past in its complete duration (Benetos and Dixon, 2011a), in shorter segments (Vincent *et al.*, 2010; Peeling and Godsill, 2011; Grindlay and Ellis, 2011; Carabias-Orti *et al.*, 2011), or in pairs of tracks (Mysore and Smaragdis, 2009). Finally, we used the 10 Disklavier recordings developed in Poliner and Ellis (2007) that were additionally evaluated in Lee *et al.* (2011) and Benetos and Dixon (2011a). The Disklavier recordings are sampled at 8 kHz, while the RWC and MIREX recordings are sampled at 44.1 kHz.

TABLE I. MIDI note range of the instruments employed for sound state template extraction.

Instrument	Lowest note	Highest note
Bassoon	34	72
Cello	26	81
Clarinet	50	89
Flute	60	96
Guitar	40	76
Harpsichord	28	88
Horn	41	77
Oboe	58	91
Organ	36	91
Piano	21	108
Violin	55	100

C. Metrics

For assessing the performance of the proposed system, we employ several metrics from the automatic transcription literature. Frame-based evaluations are made by comparing the transcribed output and the MIDI ground-truth frame by frame using a 10 ms scale as in the MIREX multiple-F0 estimation task (MIREX, 2007). The first employed metric is the overall accuracy, defined in Dixon (2000):

$$Acc_1 = \frac{\sum_n N_{tp}[n]}{\sum_n N_{fp}[n] + N_{fn}[n] + N_{tp}[n]}, \tag{38}$$

where $N_{tp}[n]$ is the number of correctly detected pitches at frame n, $N_{fn}[n]$ denotes the number of false negatives, and $N_{fp}[n]$ the number of false positives.

A second accuracy metric is also used, proposed in Kameoka *et al.* (2007), that also takes into account pitch substitutions:

$$Acc_2 = \frac{\sum_n N_{ref}[n] - N_{fn}[n] - N_{fp}[n] + N_{subs}[n]}{\sum_n N_{ref}[n]}, \tag{39}$$

where $N_{ref}[n]$ is the number of ground-truth pitches at frame n and $N_{subs}[n]$ is the number of pitch substitutions, given by $N_{subs}[n] = \min(N_{fn}[n], N_{fp}[n])$. We also employ the error

metrics defined in Poliner and Ellis (2007) that measure the substitution errors (E_{subs}), missed detection errors (E_{fn}), false alarm errors (E_{fp}), and the total error (E_{tot}).

We also used the frame-wise precision, recall, and F-measure metric for comparing the transcription performance of the MIREX recording with other methods in the literature, defined in Vincent *et al.* (2010) as

$$\mathcal{P}=\frac{\sum_n N_{tp}[n]}{\sum_n N_{sys}[n]},\quad \mathcal{R}=\frac{\sum_n N_{tp}[n]}{\sum_n N_{ref}[n]},\quad \mathcal{F}=\frac{2\mathcal{R}\mathcal{P}}{\mathcal{R}+\mathcal{P}}, \tag{40}$$

where $N_{sys}[n]$ is the number of detected pitches.

Finally, for note-based evaluation, we utilized the onset-based metric defined in Bay *et al.* (2009) that is also used in the MIREX note tracking task (MIREX, 2007). A note event is assumed to be correct if its onset is within a ± 50 ms range of a ground-truth onset. For this case, metrics are defined in a similar way to Eq. (40), resulting in the note-based precision, recall, and F-measure, denoted as $\mathcal{P}_n$, $\mathcal{R}_n$, and $\mathcal{F}_n$, respectively.

D. Results

Experiments were performed using the multi-pitch model of Sec. IV with the postprocessing method of Sec. V. For comparison, we utilized the non-temporally constrained shift-invariant PLCA transcription model of Benetos and Dixon (2011b) using the same time-frequency representation as an input (CQT with 60 bins/octave) and the same post-processing step. Experiments were performed using ergodic (fully connected) HMMs, which were initialized with uniform priors and transition probabilities. As in Grindlay and Ellis (2011), Dessein *et al.* (2010), and Benetos and Dixon (2011b), results are presented by selecting the parameter value (in this case λ) that maximizes the average accuracy in a dataset. For each dataset, results using state-of-the-art transcription methods published in the literature for the same experiment are reported for comparison.

Regarding runtimes, the computational time for extracting the sound state templates is negligible. In contrast, the multi-pitch estimation stage has a heavy computational burden, mostly due to the convolutions computed in the E-step of Eq. (25), and the M-step in Eqs. (28)–(30). In practice, the algorithm converges at about 10–15 iterations; 15 iterations are chosen for the present experiments. Using 32-bit MATLAB with a 1.5 GHz processor, the computation time is approximately 50×real time. The note tracking step takes about 1×real time. The computation time for the method in Benetos and Dixon (2011b) is approximately 30×real time and for the PLCA method is approximately 4×real time.

1. RWC dataset

Transcription results using the 12 excerpts from the RWC database (Goto *et al.*, 2003) and the complete set of instrument templates are shown in terms of Acc_2 in Table II. Comparisons are made with the non-temporally constrained SI-PLCA method of Benetos and Dixon (2011b) as well as the GMM-based method of Quesada *et al.* (2010) and the HTC method of Kameoka *et al.* (2007). It is clearly seen that the proposed method outperforms other transcription approaches for the same experiment. In terms of specific recordings, the lowest performance of the system is reported for recording 12, which is a piano-tenor duet. The lower performance can be attributed to the fact that the current system does not support any spectral templates for singing voice and does not track vibrati that span more than one semitone in range. On the other hand, the best system performance is reported for recording 10, which was performed by a string quartet. This demonstrates that the proposed method can accurately transcribe recordings of non-ideally tuned instruments that also exhibit vibrati, contrary to state-of-the-art audio feature-based methods.

Concerning the statistical significance of the accuracy improvement of the proposed system compared to the other reported systems from the literature, it should be noted that because transcription evaluations take place using 10 ms frames, even a small accuracy change can be shown to be statistically significant (Benetos and Dixon, 2011a). In particular, using the recognizer comparison technique of Guyon *et al.* (1998) for the experiments using the RWC

TABLE II. Transcription results (Acc_2) for the 12 RWC recordings compared with other approaches.

Data	Proposed	Benetos and Dixon (2011b)	Quesada *et al.* (2010)	Kameoka *et al.* (2007)
1 (%)	65.1	61.3	63.5	64.2
2 (%)	65.0	68.6	72.1	62.2
3 (%)	65.3	61.7	58.6	63.8
4 (%)	66.8	61.3	79.4	77.9
5 (%)	57.1	66.0	55.6	75.2
6 (%)	76.6	75.7	70.3	81.2
7 (%)	67.0	59.7	49.3	70.9
8 (%)	67.9	65.5	64.3	63.2
9 (%)	50.4	52.0	50.6	43.2
10 (%)	80.7	68.1	55.9	48.1
11 (%)	57.6	52.8	51.1	37.6
12 (%)	34.0	29.6	38.0	27.5
Mean (%)	**62.8**	**60.2**	**59.1**	**59.6**
Std. (%)	12.1	11.8	11.5	16.9

TABLE III. Transcription error metrics for the 12 RWC recordings.

Method	$\mathcal{F}_n$ (%)	Acc_1 (%)	Acc_2 (%)	E_{tot} (%)	E_{subs} (%)	E_{fn} (%)	E_{fp} (%)
Proposed	**44.3**	**61.8**	**62.8**	**37.2**	**8.7**	19.0	**9.5**
Benetos and Dixon (2011b)	42.0	59.6	60.2	39.8	9.7	18.7	11.4
PLCA (%)	38.8	58.5	58.6	41.4	10.6	**17.5**	13.4

transcription dataset, the significance threshold with 95% confidence is 1.1% in terms of Acc_2, which makes the improvement significant. Thus the differences reported between our current work and previously published results are significant.

Additional transcription metrics for the RWC dataset using the proposed method, the non-temporally constrained one in Benetos and Dixon (2011b), and a standard PLCA-based method using one template per instrument and pitch can be seen in Table III. The average note-based precision and recall for the proposed system are 51.2% and 40.4%, respectively. The most common errors occurring in the system are missed detections, usually occurring in dense chords, where only the root note is detected and the higher notes are considered as harmonics. Another source of missed detections in the frame-based evaluation also occurs when the decay part of a note is not recognized due to low energy. Given the fact that estimating note durations is a challenging task even for a human annotator, missed detections due to different note durations is not considered as important as, e.g., having octave errors. Note substitutions can also be octave errors when the lower note is missing or can be semitone errors when an instrument might be severely untuned or might momentarily change pitch. False alarms also occur that are usually octave errors taking place in the attack part of a note. When comparing the proposed system with the non-temporally constrained one, it is apparent that the proposed method outperforms the non-temporally constrained method of Benetos and Dixon (2011b) in terms of the lower false alarms produced as well as on note substitutions. The number of false alarms is diminished in the proposed system due to the fact that attack states have been modeled. Also octave errors counting as note substitutions have been diminished due to modeling the decay state of produced notes, where in some cases the fundamental might be suppressed (e.g., piano).

It can also be seen that the shift-invariant model of Benetos and Dixon (2011b) outperforms the standard PLCA-based transcription model. Most of the additional errors introduced by PLCA are false alarms, which are commonly extra notes one octave higher than the expected pitch. Note substitution errors also increased with the majority being semitone errors due to the inability of the PLCA-based model to estimate fine tuning or frequency modulations. It should be noted though that the improvement of a SI-PLCA model over a PLCA one is also dependent on the overall tuning of a dataset, it is expected that transcribing an untuned dataset will cause additional errors in a PLCA-based transcription model. It should be noted that the proposed model (which can extract pitch in high-frequency resolution) can also be useful for tuning and temperament estimation of music recordings.

To demonstrate the effect of the HMM-based postprocessing procedure of Sec. V, we perform a comparative experiment on the 12 RWC recordings using the proposed method with simple thresholding on $P_t(p)$ $\sum_\omega V_{\omega,t}$. In that case, $Acc_1 = 61.4\%$, $Acc_2 = 61.9\%$, and $\mathcal{F}_n = 42.1\%$. Thus the HMM-based postprocessing helps achieve improved performance, especially for the note tracking task.

Regarding sparsity parameters κ and ν, the accuracy rates for the RWC dataset using different sparsity values for the two parameters are presented in Fig. 6, where the other sparsity parameter is set to 1.0. It can be seen that with increased source contribution sparsity the accuracy of the system diminishes, while enforcing sparsity on the pitch activation leads to

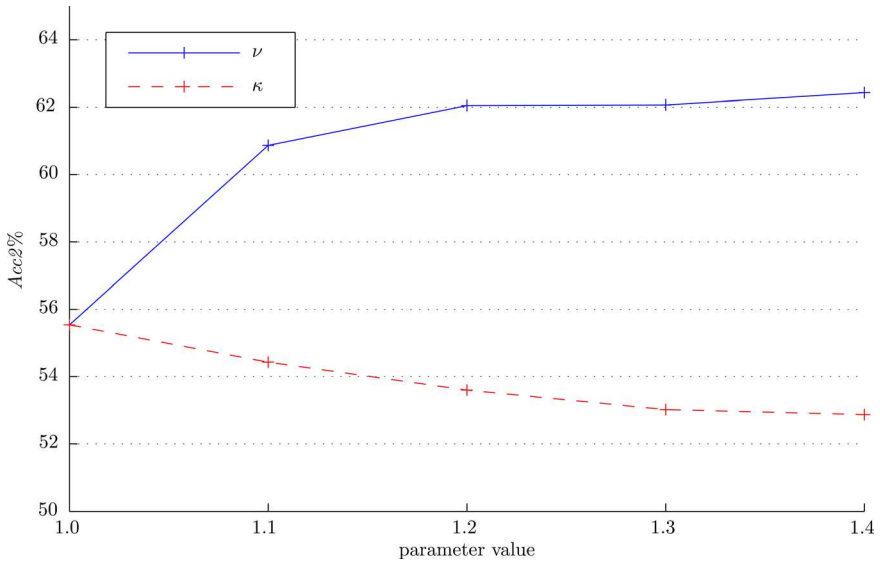

FIG. 6. (Color online) Transcription results (Acc_2) for the RWC dataset using various sparsity parameters (while the other parameter is set to 1.0).

TABLE IV. Frame-based $\mathcal{F}$ for the first 30 s of the MIREX recording compared with other approaches.

Method	$\mathcal{F}$ (%)
Proposed	65.9
Benetos and Dixon (2011b)	63.7
Peeling and Godsill (2011)	59.6
Vincent *et al.* (2010)	62.5

a significant improvement. However, the optimal combination of sparsity parameters was found to be $\kappa = 1.1$ and $\nu = 1.3$, after experimentation. This indicates that applying a sparsity constraint on the pitch activation results in a cleaner pitch track with less octave errors even though the resulting spectrogram approximation might deviate further compared to the original spectrogram. It should be noted, however, that for a monophonic piece, a greater value of ν might be needed compared to a polyphonic piece with a high polyphony level, where a lower ν value is more appropriate. It was also shown that by slightly applying sparsity to the source contribution helps in assigning produced notes to single instruments instead of instrument combinations.

2. MIREX recording

Results using the MIREX 2007 woodwind quintet recording are shown in Tables IV and V. We use the complete set of instrument templates for transcription. In Table IV, results using the first 30 s of the recording are reported using the F-measure and compared with the method of Benetos and Dixon (2011b), the harmonic NMF method of Vincent *et al.* (2010), and the likelihood search method using a Poisson process in Peeling and Godsill (2011) (the aforementioned methods used only the first 30 s of the MIREX recording). Again, the proposed method clearly outperforms other methods in the literature. It should be noted that the corresponding precision and recall for the proposed method are $\mathcal{P} = 63.7\%$ and $\mathcal{R} = 68.7\%$.

Additional transcription metrics using the complete 54 s recording are shown in Table V, compared with the method in Benetos and Dixon (2011b). A similar trend with the RWC dataset can be seen where the number of missed detections is significantly greater than the number of false alarms. The note-based precision and recall for the proposed system are 55.0% and 62.2%, respectively. In addition, the first 30 s of the piece were also utilized in Carabias-Orti *et al.* (2011), resulting in $\mathcal{F}_n = 66.9\%$. However, in the case of Carabias-Orti *et al.* (2011), the number of instruments present in the signal is known in advance, making again the experimental procedure not directly comparable with the present one. It should be noted that $\mathcal{F}_n$ is quite higher compared to the frame-based accuracy measure, while the opposite occurs for the RWC database. This can be attributed to the fact that the majority of the produced notes in the MIREX recording are flute trills (with extremely short duration) that are successfully detected by the system.

As far as the choice of templates is concerned, we also transcribe the MIREX recording by only using woodwind templates. The frame-based F-measure reaches 65.2%, which is about 1% lower compared to the full set of templates. This indicates that having a large set of templates that might include instruments not present in the recording does in fact improve transcription accuracy because the combination of different instrument templates might better approximate the spectra of the produced notes.

3. Disklavier dataset

Transcription results using the Disklavier dataset from Poliner and Ellis (2007) are presented in Table VI. For that case, the proposed system and the system of Benetos and Dixon (2011b) utilized only the sets of piano templates extracted from the MAPS database (Emiya *et al.*, 2010). Using Acc_1, it can be seen that the proposed system outperforms the non-temporally constrained system of Benetos and Dixon (2011b), the SVM classifier of Poliner and Ellis (2007), and the iterative spectral subtraction system with note tracking from Ryynänen and Klapuri (2005). Additional metrics for the Disklavier dataset are presented in Table VII. For the proposed method, $\mathcal{P}_n$ and $\mathcal{R}_n$ are 58.8% and 53.0%, respectively. Additional experiments using the Disklavier dataset were performed in the sparse coding system of Lee *et al.* (2011) using the frame-based F-measure as a metric. In that case, the reported $\mathcal{F}$ from Lee *et al.* (2011) was 70.2%, while the proposed system reaches $\mathcal{F} = 73.1\%$. For the Disklavier dataset (Poliner and Ellis, 2007), the statistical significance threshold with 95% confidence is 0.44% in terms of Acc_1, which makes the performance difference significant. As far as the choice of templates is concerned, comparative experiments were made using the full template set for the Disklavier recordings. The full set produced $Acc_1 = 59.4\%$ and $Acc_2 = 57.8\%$, which outperform the results using only the piano templates.

4. Instrument assignment

Finally, an evaluation on the performance of the proposed system for instrument identification in polyphonic music is also performed, using the first 30 s of the MIREX woodwind quintet recording. In this *instrument assignment* task, a pitch is only considered correct if it occurs at the correct time and is assigned to the proper instrument source (Grindlay and Ellis, 2011). Two variants of the system are utilized, one using only templates from the instruments that are present in the signal (bassoon, clarinet, flute, horn, and

TABLE V. Transcription error metrics for the complete MIREX woodwind quintet compared with the approach in Benetos and Dixon (2011b).

Method	$\mathcal{F}_n$ (%)	Acc_1 (%)	Acc_2 (%)	E_{tot} (%)	E_{subs} (%)	E_{fn} (%)	E_{fp} (%)
Proposed	**58.4**	**47.8**	**51.5**	**48.5**	23.7	**12.7**	12.2
Benetos and Dixon (2011b)	57.3	45.2	50.9	49.2	**18.5**	25.7	**5.0**

TABLE VI. Mean transcription results (Acc_1) for the piano recordings in Poliner and Ellis (2007) compared with other approaches.

Method	Acc_1 (%)
Proposed	58.2
Benetos and Dixon (2011b)	57.4
Poliner and Ellis (2007)	56.5
Ryynänen and Klapuri (2005)	41.2

oboe) and another using the complete set of instrument templates. The instrument-specific output is given by $P(s=i, p, t) = P_t(p)P_t(s=i|p) \sum_\omega V_{\omega,t}$, where i is the index for the selected instrument. Postprocessing using the method of Sec. V is applied to each instrument-pitch activation to produce a binary piano-roll, which is compared to the MIDI ground truth of the specific instrument track.

Results are presented in Fig. 7, where it can be seen that the system using the complete set of templates has a higher instrument identification accuracy compared to the system that uses only woodwind templates (a similar trend was reported in Grindlay and Ellis, 2011). This can be attributed to the fact that combining several instrument templates can help in better approximating produced notes. However, we can note that identification accuracy for bassoon and oboe was better when woodwind templates were used. Clarinet and flute are more accurately transcribed compared to the rest of the instruments; this might be attributed to the spectral shape of the clarinet templates and the pitch range of the flute (where the specific flute notes in the recording were mostly outside the pitch range of the other woodwind instruments). The same segment was also evaluated in Carabias-Orti *et al.* (2011) where $\mathcal{F} = 37.0\%$ in the case where the instrument sources are known. A 22 s segment of the same recording was also evaluated in Grindlay and Ellis (2011), where the reported F-measure for the complete set of templates was 40.0% and the performance for the instrument-specific transcription case interestingly drops at 35.0%. Using the same 22 s segment, the F-measure of the proposed system using the woodwind templates is 43.85% and rises to 45.49% for the complete template set. Thus the proposed system shows promising results for instrument assignment in polyphonic music.

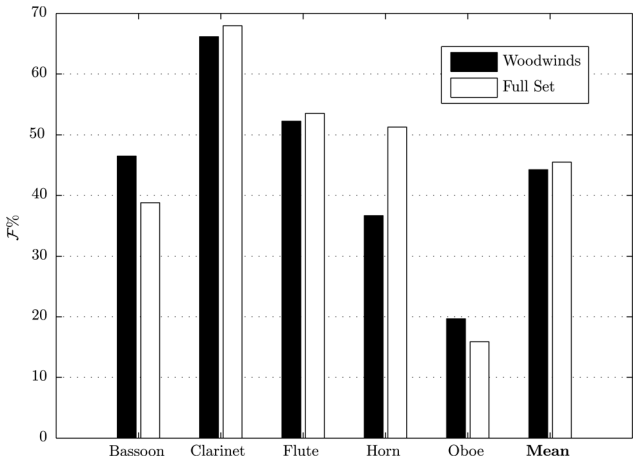

FIG. 7. Instrument assignment results ($\mathcal{F}$) using the first 30 s of the MIREX woodwind quintet.

VII. CONCLUSIONS

In this paper, we presented a polyphonic transcription system that supported the modeling of the temporal evolution of notes produced by multiple instruments. We presented a model that extracted sound state templates from monophonic recordings that was used for creating a multi-pitch multi-instrument template set. Also proposed was a model for multi-pitch detection that extended shift-invariant PLCA by including temporal constraints using multiple HMMs. The system was tested on three datasets that are widely used in the transcription literature, and results were reported using various error metrics. It was shown that the proposed model clearly outperforms a non-temporally constrained shift-invariant PLCA-based model previously proposed in Benetos and Dixon (2011b) as well as a standard PLCA-based model, using the same T/F representation and note tracking steps. Also the proposed transcription system outperformed several state-of-the-art multi-pitch detection and transcription systems, and the accuracy improvement achieved was shown to be statistically significant. All in all, the accuracy improvement of the model is attributed to the fact that notes are treated as a sequence of sound state templates that can also exhibit tuning changes and frequency modulations. We also showed that this model is useful for instrument identification in polyphonic music.

Although the proposed model is very rich, it makes an assumption regarding instruments playing the same note at the same time. The pitch shifting tensor $P_t(f|p)$ is only dependent on the pitch and not on the source s. This was done for computational speed purposes to avoid using a fourth-order tensor in the form of $P_t(f|s,p)$. Likewise, the sound state sequence is not source-dependent, to avoid using $88 \times s$ HMMs in the formulation. Thus when two instruments play the same note at the same time, a single sound state is active for a given time frame. However, the proposed model can still detect the same note played concurrently by different instruments by using $P_t(s|p)$, which can be active for multiple sources, even though the pitch shifting and sound state information might not be accurate, thus leading to some expense in accuracy for this rare case.

In the future, the present system will be evaluated in the forthcoming MIREX multi-F0 and note tracking contest (MIREX, 2007) as was done with the non-temporally constrained system previously developed by the authors (Benetos

TABLE VII. Transcription error metrics for the piano recordings in Poliner and Ellis (2007) compared with the approach in Benetos and Dixon (2011b).

Method	$\mathcal{F}_n$ (%)	Acc_1 (%)	Acc_2 (%)	E_{tot} (%)	E_{subs} (%)	E_{fn} (%)	E_{fp} (%)
Proposed	**55.5**	**58.2**	**57.7**	**42.3**	**9.8**	18.6	**13.9**
Benetos and Dixon (2011b)	51.9	57.4	55.5	44.5	10.8	**16.3**	17.4

and Dixon, 2011b). Regarding the postprocessing step, a key induction procedure would assist in assigning priors and transition probabilities using training data in the same key. The number of sound states can also become instrument-dependent by performing slight modifications to the model. To that end, an analysis on the number of sound states needed to approximate each instrument source is needed. As far as instrument identification is concerned, although results outperformed the state-of-the-art for the same experiment, additional work needs to be done to improve the current instrument recognition performance of the system. This can be achieved by utilizing the information provided by the source contribution matrix $P_t(s|p)$, combined with features for characterizing music timbre (Peeters, 2004). Finally, the present model can be further extended by incorporating a musicological model of note transitions in the top level (Ryynänen and Klapuri, 2005), which can be done using HMMs or Bayesian networks.

ACKNOWLEDGMENTS

E.B. was funded by a Westfield Trust Research Studentship (Queen Mary University of London). We acknowledge the support of the MIReS project, supported by the European Commission, FP7, ICT-2011.1.5 Networked Media and Search Systems, Grant agreement No. 287711.

Bay, M., Ehmann, A. F., and Downie, J. S. (**2009**). "Evaluation of multiple-F0 estimation and tracking systems," in *10th International Society of Music Information Retrieval Conference*, Kobe, Japan, pp. 315–320.

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M. (**2005**). "A tutorial on onset detection of music signals," IEEE Trans. Audio, Speech Lang. Proc. **13**, 1035–1047.

Benetos, E., and Dixon, S. (**2011a**). "Joint multi-pitch detection using harmonic envelope estimation for polyphonic music transcription," IEEE J. Sel. Top. Signal Proc. **5**, 1111–1123.

Benetos, E., and Dixon, S. (**2011b**). "Multiple-instrument polyphonic music transcription using a convolutive probabilistic model," in *8th Sound and Music Computing Conference*, Padova, Italy, pp. 19–24.

Benetos, E., and Dixon, S. (**2011c**). "A temporally-constrained convolutive probabilistic model for pitch detection," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, pp. 133–136.

Benetos, E., and Dixon, S. (**2012**). "Temporally-constrained convolutive probabilistic latent component analysis for multi-pitch detection," in *International Conference on Latent Variable Analysis and Signal Separation*, Tel-Aviv, Israel, pp. 364–371.

Brown, J. C. (**1991**). "Calculation of a constant Q spectral transform," J. Acoust. Soc. Am. **89**, 425–434.

Carabias-Orti, J. J., Virtanen, T., Vera-Candeas, P., Ruiz-Reyes, N., and Canadas-Quesada, F. J. (**2011**). "Musical instrument sound multi-excitation model for non-negative spectro-gram factorization," IEEE J. Sel. Topics Signal Proc. **5**, 1144–1158.

Davy, M., Godsill, S., and Idier, J. (**2006**). "Bayesian analysis of western tonal music," J. Acoust. Soc. Am. **119**, 2498–2517.

de Cheveigné, A. (**2006**). "Multiple F0 estimation," in *Computational Auditory Scene Analysis, Algorithms and Applications*, edited by D. L. Wang and G. J. Brown (IEEE Press/Wiley, New York), pp. 45–79.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (**1977**). "Maximum likelihood from in-complete data via the EM algorithm," J. Royal Stat. Soc. **39**, 1–38.

Dessein, A., Cont, A., and Lemaitre, G. (**2010**). "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *11th International Society on Music Information Retrieval Conference*, Utrecht, The Netherlands, pp. 489–494.

Dixon, S. (**2000**). "On the computer recognition of solo piano music," in *2000 Australasian Computer Music Conference*, Brisbane, Australia, pp. 31–37.

Duan, Z., Pardo, B., and Zhang, C. (**2010**). "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," IEEE Trans. Audio, Speech Lang. Proc. **18**, 2121–2133.

Emiya, V., Badeau, R., and David, B. (**2010**). "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," IEEE Trans. Audio, Speech Lang. Proc. **18**, 1643–1654.

Fuentes, B., Badeau, R., and Richard, G. (**2011**). "Adaptive harmonic time-frequency decomposition of audio using shift-invariant PLCA," in *International Conference on Acoustical Speech and Signal Processing*, Prague, Czech Republic, pp. 401–404.

Ghahramani, Z., and Jordan, M. (**1997**). "Factorial hidden Markov models," Mach. Learn. **29**, 245–273.

Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (**2003**). "RWC music database: Music genre database and musical instrument sound database," in *International Conference on Music Information Retrieval*, Baltimore, MD.

Grindlay, G., and Ellis, D. (**2011**). "Transcribing multi-instrument polyphonic music with hierarchical eigen instruments," IEEE J. Sel. Top. Signal Proc. **5**, 1159–1169.

Guyon, I., Makhoul, J., Schwartz, R., and Vapnik, V. (**1998**). "What size test set gives good error estimates?" IEEE Trans. Pattern Anal. Mach. Intell. **20**, 52–64.

Kameoka, H., Nishimoto, T., and Sagayama, S. (**2007**). "A multipitch analyzer based on harmonic temporal structured clustering," IEEE Trans. Audio, Speech Lang. Proc. **15**, 982–994.

Klapuri, A., and Davy, M., editors (**2006**). *Signal Processing Methods for Music Transcription* (Springer-Verlag, New York), pp. 440.

Lee, C.-T., Yang, Y.-H., and Chen, H. (**2011**). "Automatic transcription of piano music by sparse representation of magnitude spectra," in *IEEE International Conference on Multimedia and Expo*, Barcelona, Spain, pp. 1–6.

Lee, D., and Seung, H. (**1999**). "Learning the parts of objects by non-negative matrix factorization," Nature **401**, 788–791.

MIREX (**2007**). "Music Information Retrieval Evaluation eXchange (MIREX)" available at http://music-ir.org/mirexwiki/ (Last viewed August 19, 2012).

Mysore, G. (**2010**). "A non-negative framework for joint modeling of spectral structure and temporal dynamics in sound mixtures," Ph.D. thesis, Stanford University, CA, pp. 143.

Mysore, G., and Smaragdis, P. (**2009**). "Relative pitch estimation of multiple instruments," in *International Conference on Acoustical Speech and Signal Processing*, Taipei, Taiwan, pp. 313–316.

Nakano, M., Roux, J. L., Kameoka, H., Kitano, Y., Ono, N., and Sagayama, S. (**2010**). "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms," in *9th International Conference on Latent Variable Analysis and Signal Separation*, St. Malo, France, pp. 149–156.

Nakano, M., Roux, J. L., Kameoka, H., Ono, N., and Sagayama, S. (**2011**). "Infinite-state spectrum model for music signal analysis," in *International Conference on Acoustical Speech and Signal Processing*, Prague, Czech Republic, pp. 1972–1975.

Peeling, P., and Godsill, S. (**2011**). "Multiple pitch estimation using non-homogeneous Poisson processes," IEEE J. Sel. Top. Signal Proc. **5**, 1133–1143.

Peeling, P., Li, C., and Godsill, S. (**2007**). "Poisson point process modeling for polyphonic music transcription," J. Acoust. Soc. Am. **121**, 168–175.

Peeters, G. (**2004**). "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Technical Report No. CUIDADO I.S.T. Project.

Pertusa, A., and Iñesta, J. M. (**2008**). "Multiple fundamental frequency estimation using Gaussian smoothness," in *International Conference on Acoustical Speech and Signal Processing*, Las Vegas, NV, pp. 105–108.

Poliner, G., and Ellis, D. (**2007**). "A discriminative model for polyphonic piano transcription," EURASIP J. Adv. Signal Process. 154–162.

Quesada, F. C., Ruiz-Reyes, N., Candeas, P. V., Carabias-Orti, J. J., and Maldonado, S. (**2010**). "A multiple-F0 estimation approach based on Gaussian spectral modeling for polyphonic music transcription," J. New Mus. Res. **39**, 93–107.

Rabiner, L. R. (**1989**). "A tutorial on hidden Markov models and selected applications in speech recognition," Proc, IEEE **77**, 257–286.

Ryynänen, M., and Klapuri, A. (**2005**). "Polyphonic music transciption using note event modeling," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, pp. 319–322.

Shashanka, M., Raj, B., and Smaragdis, P. (**2008**). "Probabilistic latent variable models as nonnegative factorizations," Comput. Intell. Neurosci. **2008**, 947438.

Smaragdis, P. (**2009**). "Relative-pitch tracking of multiple arbitary sounds," J. Acoust. Soc. Am. **125**, 3406–3413.

Smaragdis, P., and Raj, B. (**2007**). "Shift-invariant probabilistic latent component analysis," Technical Report No. TR2007-009, Mitsubishi Electric Research Laboratories.

Smaragdis, P., Raj, B., and Shashanka, M. (**2006**). "A probabilistic latent variable model for acoustic modeling," in *Neural Information Processing Systems Workshop*, Whistler, BC, Canada.
Smaragdis, P., Raj, B., and Shashanka, M. (**2008**). "Sparse and shift-invariant feature extraction from non-negative data," in *International Conference Acoustical Speech and Signal Processing*, Las Vegas, NV, pp. 2069–2072.
Vincent, E., Bertin, N., and Badeau, R. (**2010**). "Adaptive harmonic spectral decomposition for multiple pitch estimation," IEEE Trans. Audio, Speech Lang. Proc. **18**, 528–537.
Yeh, C., Röbel, A., and Rodet, X. (**2010**). "Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals," IEEE Trans. Audio, Speech Lang. Proc. **18**, 1116–1126.