MULTIPLE-INSTRUMENT POLYPHONIC MUSIC TRANSCRIPTION USING A CONVOLUTIVE PROBABILISTIC MODEL

Emmanouil Benetos and Simon Dixon
Centre for Digital Music, Queen Mary University of London, London E1 4NS, UK
{emmanouilb, simond}@eecs.qmul.ac.uk

ABSTRACT

In this paper, a method for automatic transcription of music signals using a convolutive probabilistic model is proposed, by extending the shift-invariant Probabilistic Latent Component Analysis method. Several note templates from multiple orchestral instruments are extracted from monophonic recordings and are used for training the transcription system. By incorporating shift-invariance into the model along with the constant-Q transform as a time-frequency representation, tuning changes and frequency modulations such as vibrato can be better supported. For postprocessing, Hidden Markov Models trained on MIDI data are employed, in order to favour temporal continuity. The system was tested on classical and jazz recordings from the RWC database, on recordings from a Disklavier piano, and a woodwind quintet recording. The proposed method, which can also be used for pitch content visualization, outperforms several state-of-the-art approaches for transcription, using a variety of error metrics.

1. INTRODUCTION

The goal of an automatic music transcription system is to convert an audio recording into a symbolic representation, such as a piano-roll, a MIDI file or a music sheet. The creation of a system able to transcribe music produced by multiple instruments with a high level of polyphony continues to be an open problem in the research community, although monophonic pitch transcription is largely considered solved. For a comprehensive overview on transcription approaches the reader is referred to [1].

Transcription or pitch tracking methods that employ probabilistic models related to the ones used in this work are detailed in Section 2. Other approaches related to this paper include the work by Poliner and Ellis [2], where piano note classification was performed using support vector machines (SVMs). In order to improve transcription performance, the classification output of the SVMs was fed as input to a hidden Markov model (HMM) [3] for postprocessing. The same note smoothing technique was also used in [4], where the main transcription algorithm consists of a spectral distance measure modeling polyphonic sounds as a weighted sum of Gaussian spectral models.

A signal processing-based multiple-F0 estimation was proposed by Saito et al. in [5], which uses the inverse Fourier transform of the linear power spectrum with log-scale frequency, called *specmurt* (an anagram of cepstrum). The input log-frequency spectrum is considered to be generated by a convolution of a single pitch template with a pitch indicator function. The deconvolution of the spectrum by the pitch template results in the estimated pitch indicator function. Previous work by the authors which is used for comparative purposes includes a signal processing-based polyphonic transcription system [6] which is based on joint multiple-F0 estimation using a feature-based score function and note onset and offset detection.

In this work, a system for automatic transcription of polyphonic music is introduced, which is based on a proposed extension of the shift-invariant probabilistic latent component analysis (PLCA) [7] model. Contrary to the models in [7, 8], which use a single spectral template for all pitches from the same instrument source, this model is able to support the use of multiple pitch templates extracted from multiple sources. Using a log-frequency representation and frequency shifting, detection of notes that are non-ideally tuned, or that are produced by instruments that exhibit frequency modulations is made possible. Sparsity is also enforced in the model, in order to further constrain the transcription result and the instrument contribution in the production of pitches. Also, an intermediate result of the proposed model is a time-pitch representation which can be used for pitch content visualization of polyphonic music. Finally, a hidden Markov model-based note tracking method is employed in order to provide a smooth piano-roll transcription. The system was tested on recordings from the RWC database [9], the Disklavier dataset in [2], as well as the MIREX multi-F0 woodwind quintet [10]. A comparison was performed with various transcription methods using error metrics found in the literature. It is shown that the proposed system outperforms several state-of-the-art approaches for the same experiment. Also, it is indicated that a shift-invariant model can improve the detection of non-ideally tuned notes.

The outline of the paper is as follows. In Section 2, the PLCA and shift-invariant PLCA methods are presented, along with their applications in music transcription and relative pitch tracking. The proposed polyphonic music transcription system is introduced in Section 3. Finally, the employed dataset, metrics and transcription experiments

performed are described in Section 4, while conclusions are drawn in Section 5.

2. LATENT VARIABLE METHODS

2.1 PLCA

Probabilistic latent component analysis (PLCA) is a model for acoustic analysis developed by Smaragdis et al. [11]. It provides a probabilistic framework that is extensible as well as easy to interpret. Considering the spectrogram as a probability distribution $P(\omega, t)$, the asymmetric PLCA model can be formulated as:

$$P(\omega, t) = P(t) \sum_{z} P(\omega|z)P(z|t) \quad (1)$$

where $P(\omega|z)$ are the spectral templates corresponding to component z, $P(z|t)$ are the component activations through time, and $P(t)$ is the energy distribution of the spectrogram. For estimating $P(\omega|z)$ and $P(z|t)$, iterative update rules are employed, which are based on the Expectation-Maximization (EM) algorithm.

In [12], an extension of the PLCA model was proposed for polyphonic music transcription, supporting multiple spectral templates for each pitch and multiple instruments. The concept of *eigeninstruments* was introduced, which models instruments as mixtures of basic models. Sparsity was enforced on the transcription matrix and the source contribution matrix of the model by a tempering-based approach. For experiments, stored pitch templates from various synthesized instrument sounds were used. Experiments were performed on instrument pairs taken from the multi-track woodwind recording used in the MIREX multi-F0 development set [10], as well as on three J.S. Bach duets.

2.2 Shift-invariant PLCA

An extension of the basic PLCA algorithm was proposed in [7], in order to extract shifted structures in non-negative data. The shift-invariant PLCA method can be used in conjunction with log-frequency spectrograms (e.g. the constant-Q transform) in order to extract pitch. This is feasible since in log-frequency spectra the inter-harmonic spacings are the same for any periodic sounds. The shift-invariant PLCA model is defined as:

$$\begin{aligned} P(\omega, t) &= \sum_{z} P(z)P(\omega|z) *_{\omega} P(f, t|z) \\ &= \sum_{z} P(z) \sum_{f} P(\omega - f|z)P(f, t|z) \quad (2) \end{aligned}$$

where the spectral template $P(\omega|z)$ corresponding to component z is convolved with the pitch impulse distribution $P(f, t|z)$. $P(z)$ is the prior distribution of the components. Again, in order to estimate $P(z)$, $P(\omega|z)$, and $P(f, t|z)$, the EM algorithm can be utilized.

The shift-invariant PLCA model was used in [8] for multiple-instrument relative pitch tracking, where one pitch template is attributed to each instrument source and is shifted across log-frequency. The constant-Q transform (CQT) was used as a time/frequency representation. Since the problem was unsupervised, additional constraints were imposed on eq. (2). Firstly, a sliding Gaussian Dirichlet prior distribution was used in the computation of $P(f, t|z)$ in order to eliminate any octave errors. In addition, in order to enforce temporal continuity, a Kalman filter type smoothing is applied to $P(f, t|z)$ at each iteration step. The method was tested on the MIREX [10] woodwind quintet using mixtures of two instruments at a time.

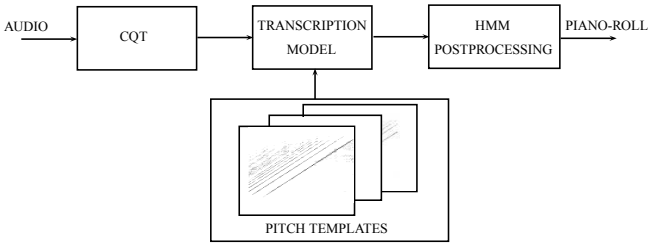

Figure 1. Diagram for the proposed polyphonic transcription system.

3. PROPOSED METHOD

The goal of the proposed transcription system is to provide a framework that supports multiple templates per pitch, in contrast to the relative pitch tracking method in [8], as well as multiple templates per musical instrument. In addition, the contribution of each instrument source is not constant for the whole recording as in [8], but is time-dependent. Also, its goal is to exploit the benefits given by a shift-invariant model coupled with a log-frequency representation, in contrast to the transcription method in [12], for detecting notes that exhibit frequency modulations and tuning changes.

In subsection 3.1, the extraction of pitch templates for various instruments is presented. The main transcription model is presented in subsection 3.2, while the HMM post-processing step is described in subsection 3.3. A diagram of the proposed transcription system is depicted in Fig. 1.

3.1 Extracting Pitch Templates

Firstly, spectral templates are extracted for various instruments, for each note, using their whole note range. Isolated note samples from three different piano types were extracted from the MAPS dataset [13] and templates from other orchestral instruments were extracted from monophonic recordings from the RWC database [9]. For extracting the note templates, the constant-Q transform (CQT) was computed [14] with spectral resolution of 120 bins per octave. Afterwards, the PLCA model of eq. (1) using only one component z was employed in order to extract the spectral template $P(\omega|z)$. In Table 1, the pitch range of each instrument used for template extraction is shown.

3.2 Transcription Model

Utilizing the extracted instrument templates and by extending the shift-invariant PLCA algorithm, a model is proposed which supports the use of multiple pitch and instrument templates in a convolutive framework, thus support-

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

Table 1. MIDI note range of the instrument templates used in the proposed transcription system.

ing tuning changes and frequency modulations. By considering the input CQT spectrum as a probability distribution $P(\omega, t)$, the proposed model can be formulated as:

$$P(\omega, t) = P(t) \sum_{p,s} P(\omega|s, p) *_\omega P(f|p, t)P(s|p, t)P(p|t) \tag{3}$$

where $P(\omega|s, p)$ is the spectral template that belongs to instrument s and MIDI pitch $p = 21, \ldots, 108$, $P(f|p, t)$ is the time-dependent impulse distribution that corresponds to pitch p, $P(s|p, t)$ is the instrument contribution for each pitch in a specific time frame, and $P(p|t)$ is the pitch probability distribution for each time frame.

By removing the convolution operator, the model of (3) can be expressed as:

$$P(\omega, t) = P(t) \sum_{p,f,s} P(\omega - f|s, p)P(f|p, t)P(s|p, t)P(p|t) \tag{4}$$

In order to only utilize each template $P(\omega|s, p)$ for detecting the specific pitch p, the convolution of $P(\omega|s, p) *_\omega P(f|p, t)$ takes place using an area spanning one semitone around the ideal position of p. Since 120 bins per octave are used in the CQT spectrogram, f has a length of 10.

The various parameters in (3) can be estimated using iterative update rules derived from the EM algorithm. For the expectation step the update rule is:

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

For the maximization step, the update equations for the proposed model are:

$$P(\omega|s, p) = \frac{\sum_{f,t} P(p, f, s|\omega + f, t)P(\omega + f, t)}{\sum_{\omega,t,f} P(p, f, s|\omega + f, t)P(\omega + f, t)} \tag{6}$$

$$P(f|p, t) = \frac{\sum_{\omega,s} P(p, f, s|\omega, t)P(\omega, t)}{\sum_{f,\omega,s} P(p, f, s|\omega, t)P(\omega, t)} \tag{7}$$

$$P(s|p, t) = \frac{\sum_{\omega,f} P(p, f, s|\omega, t)P(\omega, t)}{\sum_{s,\omega,f} P(p, f, s|\omega, t)P(\omega, t)} \tag{8}$$

$$P(p|t) = \frac{\sum_{\omega,f,s} P(p, f, s|\omega, t)P(\omega, t)}{\sum_{p,\omega,f,s} P(p, f, s|\omega, t)P(\omega, t)} \tag{9}$$

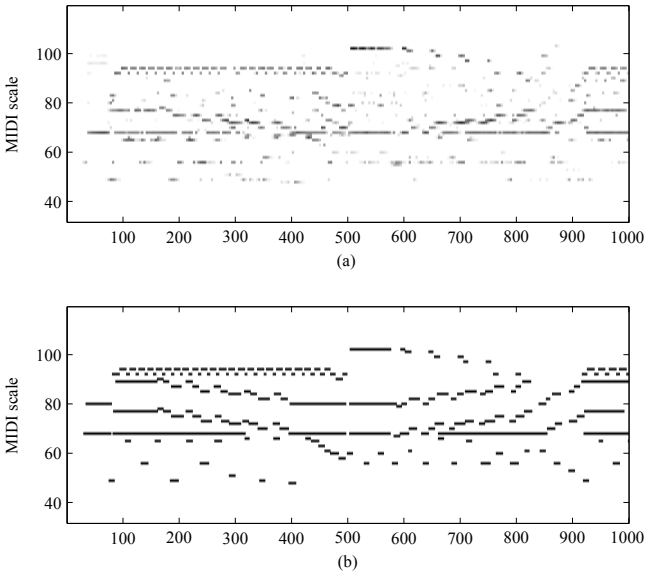

Figure 2. (a) The transcription matrix $P(p, t)$ of the first 10s of the MIREX woodwind quintet. (b) The pitch ground truth of the same recording. The abscissa corresponds to 10ms.

It should be noted that since the instrument-pitch templates have been extracted during the training stage, the update rule for the templates (6) is not used, but is included for the sake of completeness. Using these constant templates, convergence is quite fast, usually requiring 10-20 iterations. The resulting piano-roll transcription matrix and pitch matrix are respectively given by:

$$\begin{aligned} P(p, t) &= P(t)P(p|t) \\ P(f, p, t) &= P(t)P(p|t)P(f|p, t) \end{aligned} \tag{10}$$

By stacking together slices of the pitch matrix $P(f, p, t)$ for all pitch values: $P(f, t) = [P(f, 21, t) \cdots P(f, 108, t)]$ we can create a time-pitch representation which can be used for visualization purposes. In $P(f, t)$, f has a length of $88 \times 10 = 880$, thus representing pitch in a 10 cent resolution. In Fig. 2, the transcription matrix $P(p, t)$ for an excerpt of the MIREX multi-F0 woodwind quintet recording can be seen, along with the corresponding pitch ground truth. Also, in Fig. 3, the time-pitch representation of an excerpt of the 'RWC MDB-C-2001 No. 12' (string quartet) recording can be seen, where the frequency modulations caused by vibrato are visible.

In order for the algorithm to provide as meaningful solutions as possible, sparsity is encouraged on transcription matrix $P(p|t)$, expecting that only few notes are present at a given time frame. In addition, sparsity can be enforced to matrix $P(s|p, t)$, meaning that for each pitch at a given time frame, only a few instrument sources contributes to its production. The same technique used in [12] was employed for controlling sparsity, by modifying the update equations (8) and (9):

$$P(s|p, t) = \frac{\left(\sum_{\omega,f} P(p, f, s|\omega, t)P(\omega, t)\right)^\alpha}{\sum_s \left(\sum_{\omega,f} P(p, f, s|\omega, t)P(\omega, t)\right)^\alpha} \tag{11}$$

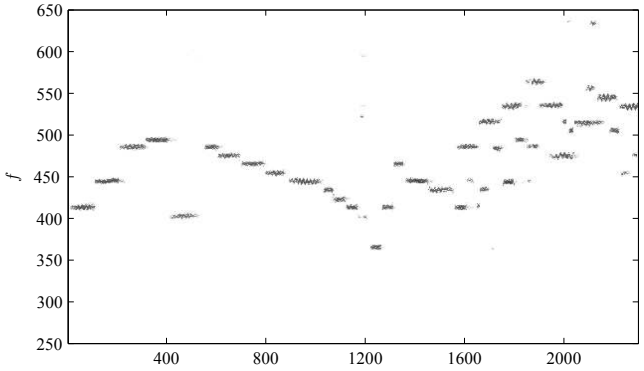

Figure 3. The time-pitch representation $P(f,t)$ of the first 23s of 'RWC MDB-C-2001 No. 12' (string quartet) in a 10 ms time scale.

$$P(p|t) = \frac{\left(\sum_{\omega,f,s} P(p,f,s|\omega,t)P(\omega,t)\right)^{\beta}}{\sum_{p}\left(\sum_{\omega,f,s} P(p,f,s|\omega,t)P(\omega,t)\right)^{\beta}} \quad (12)$$

By setting $\alpha, \beta > 1$, the entropy in matrices $P(s|p,t)$ and $P(p|t)$ is lowered and sparsity is enforced.

3.3 Postprocessing

Instead of simply thresholding $P(p,t)$ for extracting the piano-roll transcription as in [12], additional postprocessing is applied in order to perform note smoothing and tracking. Hidden Markov models (HMMs) [3] have been used in the past for note smoothing in signal processing-based transcription approaches (e.g. [2, 6]). Here, a similar approach to the HMM smoothing procedure employed in [2] is used, but modified for the probabilistic framework of the proposed transcription system.

Each pitch p is modeled by a two-state HMM, denoting pitch activity/inactivity. The hidden state sequence for each pitch is given by $Q_p = \{q_p[t]\}$. MIDI files from the RWC database [9] from the classic and jazz subgenres were employed in order to estimate the state priors $P(q_p[1])$ and the state transition matrix $P(q_p[t]|q_p[t-1])$ for each pitch p. For each pitch, the most likely state sequence is given by:

$$\hat{Q}_p = \arg\max_{q_p[t]} \prod_t P(q_p[t]|q_p[t-1])P(o_p[t]|q_p[t]) \quad (13)$$

which can be computed using the Viterbi algorithm [3]. For estimating the observation probability for each active pitch $P(o_p[t]|q_p[t]=1)$, we use a sigmoid curve which has as input the transcription piano-roll $P(p,t)$ from the output of the transcription model:

$$P(o_p[t]|q_p[t]=1) = \frac{1}{1+e^{-P(p,t)}} \quad (14)$$

The result of the HMM postprocessing step is a binary piano-roll transcription which can be used for evaluation. An example of the HMM postprocessing step is given in Fig. 4, where the transcription matrix $P(p,t)$ of a piano recording from [2] is seen along with the output of the HMM smoothing.

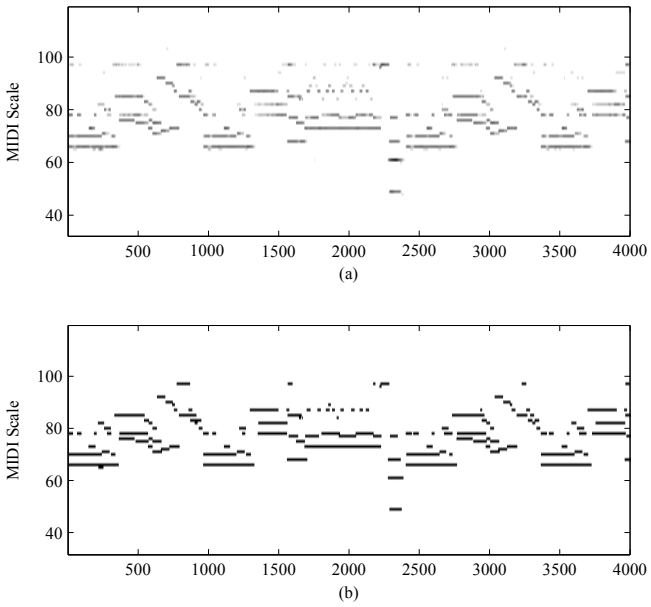

Figure 4. (a) The transcription matrix $P(p,t)$ of the first 40s of the J. Haydn Piano Sonata No.54 from the Disklavier dataset of [2] (b) The output of the HMM postprocessing step (the abscissa corresponds to 10 ms).

4. EVALUATION

4.1 Datasets

For the transcription experiments, we used recordings from three different sources. Firstly, 12 excerpts from the RWC database [9] were employed, which have been widely used for evaluating transcription systems (e.g. [15, 5, 4]). The dataset contains classical and jazz music produced by piano, guitar, flute, and bowed strings, with the majority being piano. Aligned ground-truth MIDI data was created using the original non-aligned MIDI reference for the first 23 sec of each recording, using Sonic Visualiser [1] .

In addition, the test dataset developed by Poliner and Ellis [2] was also used for transcription experiments. It contains 10 one-minute classical recordings from a Yamaha Disklavier grand piano, sampled at 8 kHz along with aligned MIDI ground truth. Finally, the full woodwind quintet recording from the MIREX multi-F0 development set [10] was also used for transcription experiments.

4.2 Evaluation Metrics

Several evaluation metrics are employed for the recordings used for the transcription experiments. All evaluations take place by comparing the transcribed output and the ground-truth MIDI files using a 10 ms scale, as in the MIREX multiple-F0 estimation task [10]. The first metric that is used is the overall accuracy (Acc_1) used in [2]. Also, an additional set of metrics is employed, namely the alternative accuracy measure (Acc_2), the total error (E_{tot}), the substitution error (E_{subs}), missed detection error (E_{fn}), and false alarm error (E_{fp}). Definitions for the aforementioned set of metrics can be found in [15, 5, 4].

[1] http://www.sonicvisualiser.org/

4.3 Results

Transcription experiments using the 12 excerpts from the RWC database were performed using only piano templates, or using the full list of instrument templates shown in Table 1. Results are presented in Table 2, comparing the performance of the system with other state-of-the-art methods [6, 4, 5, 15], while in Table 3 additional metrics are used in order to compare the performance of the proposed system with the method in [6]. It can be seen that when using the piano templates, the proposed method outperforms other systems with respect to the accuracy measure Acc_2. Also, most of the errors of the system consist of missed detections, while relatively few false alarms are detected. Concerning the signal processing-based method in [6], the improvement using Acc_2 is 0.5%, which rises to 1.0% when Acc_1 is utilized.

When using the proposed system with all instrument templates, performance is significantly lowered, although it should be stressed that the majority of the RWC recordings are produced by piano. This also indicates that having a knowledge of the instruments present can significantly improve the performance of the proposed system. It is notable that when RWC recording 10 -a string quartet- is transcribed using the all-instruments model, its Acc_2 is 82.7%, far surpassing all other methods. Concerning sparsity parameters, after experimentation, no sparsity was added to the instrument contribution matrix ($\alpha = 1$) and sparsity was only enforced on the transcription matrix ($\beta = 1.5$). It is worth mentioning that with $\beta = 1$, performance using the piano templates drops to 58.6% for Acc_2. Also, in order to evaluate the contribution of the shift-invariant model, experiments were also performed by disabling convolution, resulting in a PLCA-based model similar to the one in [12]. In terms of Acc_2, performance for the RWC recordings was 60.1%, which indicates that using a shift-invariant model for transcription can improve performance when non-ideally tuned recordings or when frequency modulations are considered. To the authors' knowledge, no statistical significance tests have been made for transcription, apart from the piecewise tests in the MIREX task [10]. However, given the fact that transcription evaluations actually take place using 10 ms frames, even a small accuracy change can be shown to be statistically significant, using a method like [16].

Results using the 10 piano recordings from [2] are shown in Table 4, compared with results from other approaches reported in [2] and the method in [6]. For this experiment, only piano templates were used in the proposed system. Again, it is shown that the proposed system outperforms all other methods - compared to the one in [2], improvement is 1.1% with respect to Acc_1. It should be stressed also that the training set for the method of [2] used data from the same source as the test set. When compared with [6], the performance improvement is much larger compared to the RWC recordings (about 10.6%). This can be attributed to the much faster tempo of the pieces in [2], since the method in [6] is more suited to slower tempo due to the onset/offset detections performed, tending to accumulate transcription errors in cases of rapidly changing notes. Additional metrics for the recordings of [2] are included in Table 5.

Data	Proposed (piano)	Proposed (all)	[6]	[4]	[5]	[15]
1	64.3%	58.3%	60.0%	63.5%	59.0%	64.2%
2	70.5%	61.4%	73.6%	72.1%	63.9%	62.2%
3	70.3%	53.8%	62.5%	58.6%	51.3%	63.8%
4	67.0%	63.4%	65.2%	79.4%	68.1%	77.9%
5	66.9%	55.4%	53.4%	55.6%	67.0%	75.2%
6	71.7%	73.3%	76.1%	70.3%	77.5%	81.2%
7	67.0%	55.9%	68.5%	49.3%	57.0%	70.9%
8	67.7%	51.4%	60.1%	64.3%	63.6%	63.2%
9	51.9%	48.8%	50.3%	50.6%	44.9%	43.2%
10	55.3%	82.7%	72.4%	55.9%	48.9%	48.1%
11	57.1%	54.2%	56.2%	51.1%	37.0%	37.6%
12	30.4%	26.8%	36.6%	38.0%	35.8%	27.5%
Mean	**61.7%**	**57.1%**	**61.2%**	**59.1%**	**56.2%**	**59.6%**

Table 2. Transcription results (Acc_2) for the 12 RWC recordings compared with other approaches.

Method	Acc_1	Acc_2	E_{tot}	E_{subs}	E_{fn}	E_{fp}
Proposed (p)	60.8%	61.7%	38.3%	8.9%	19.6%	9.8%
Proposed (a)	54.8%	57.1%	42.9%	11.5%	24.4%	7.0%
[6]	59.8%	61.2%	38.8%	7.3%	24.8%	6.7%

Table 3. Transcription error metrics for the 12 RWC recordings using piano only (p) or all templates (a), compared with the approach in [6].

Finally, results using the proposed system are shown using the MIREX multi-F0 woodwind quintet [10] in Table 6. The MIREX recording is available in 5 instrument tracks; although results using pairs of these tracks have been reported in [8, 12], to the authors' knowledge no results using the complete 5-instrument recording have been published. Using the full instrument templates matrix, performance of the proposed system is 48.1% using Acc_2, while it is 39.0% using the system in [6]. Again, the reduced performance of [6] can be attributed to the fast tempo of the recording (a part of which is depicted in Fig. 2).

5. CONCLUSIONS

In this work, a system for automatic music transcription using a model based on shift-invariant probabilistic latent component analysis techniques was proposed. The main contribution of the paper is a transcription model that is able to support multiple instrument and pitch templates and is able to detect notes produced without ideal tuning or exhibiting frequency modulations. The system was tested on recordings from several sources, where it was shown to outperform other state-of-the-art transcription techniques using several error metrics. The system architecture makes it suitable for instrument-specific transcription applications. Also, a by-product of the system is a time-pitch representation that can also be used for pitch content visualization. Selected transcription examples are available online [2], along with the original excerpts for comparison.

[2] http://www.eecs.qmul.ac.uk/~emmanouilb/transcription.html

Method	Proposed	[6]	[2]	[17]
Acc_1	57.6%	47.0%	56.5%	41.2%

Table 4. Mean transcription results (Acc_1) for the piano recordings in [2] compared with other approaches.

Method	Acc_1	Acc_2	E_{tot}	E_{subs}	E_{fn}	E_{fp}
Proposed	57.6%	56.7%	43.3%	10.9%	16.9%	15.5%
[6]	47.0%	47.2%	52.8%	10.7%	33.6%	8.5%

Table 5. Transcription error metrics for the piano recordings in [2] compared with the approach in [6].

Since it was indicated that system performance can be improved by utilizing knowledge of the instruments present in the recording, instrument identification techniques will be incorporated in future versions of the system. Finally, future research will focus on producing templates for the attack, transient, sustain, and release states of the produced notes of each instrument and incorporate such formulation into the proposed model, in an effort to further reduce the number of missed detections.

Acknowledgments

E. Benetos is supported by a Westfield Trust PhD Studentship (QMUL). We would like to thank the late Graham Grindlay from Columbia University for providing part of the MIREX recording annotation.

6. REFERENCES

[1] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*, 2nd ed. New York: Springer-Verlag, 2006.

[2] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Advances in Signal Processing*, no. 8, pp. 154–162, Jan. 2007.

[3] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[4] F. Cañadas-Quesada, N. Ruiz-Reyes, P. V. Candeas, J. J. Carabias-Orti, and S. Maldonado, "A multiple-F0 estimation approach based on Gaussian spectral modelling for polyphonic music transcription," *J. New Music Research*, vol. 39, no. 1, pp. 93–107, Apr. 2010.

[5] S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama, "Specmurt analysis of polyphonic music signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 639–650, Mar. 2008.

[6] E. Benetos and S. Dixon, "Polyphonic music transcription using note onset and offset detection," in *IEEE Int. Conf. Audio, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

Method	Acc_1	Acc_2	E_{tot}	E_{subs}	E_{fn}	E_{fp}
Proposed	41.9%	48.1%	51.9%	23.6%	21.9%	6.4%
[6]	33.8%	39.0%	61.0%	28.1%	26.4%	6.5%

Table 6. Transcription error metrics for the MIREX woodwind quintet compared with the approach in [6].

[7] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Las Vegas, USA, Apr. 2008, pp. 2069–2072.

[8] G. Mysore and P. Smaragdis, "Relative pitch estimation of multiple instruments," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 313–316.

[9] 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*, Oct. 2003.

[10] "Music Information Retrieval Evaluation eXchange (MIREX)." [Online]. Available: http://music-ir.org/mirexwiki/

[11] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Neural Information Processing Systems Workshop*, Whistler, Canada, Dec. 2006.

[12] G. Grindlay and D. Ellis, "A probabilistic subspace model for multi-instrument polyphonic transcription," in *11th Int. Society for Music Information Retrieval Conf.*, Utrecht, Netherlands, Aug. 2010, pp. 21–26.

[13] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

[14] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *7th Sound and Music Computing Conf.*, Barcelona, Spain, Jul. 2010.

[15] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 982–994, Mar. 2007.

[16] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, "What size test set gives good error estimates?" *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 52–64, Jan. 1998.

[17] M. Ryynänen and A. Klapuri, "Polyphonic music transciption using note event modeling," in *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, Oct. 2005, pp. 319–322.