Temporally-Constrained Convolutive Probabilistic Latent Component Analysis for Multi-pitch Detection

Emmanouil Benetos* and Simon Dixon

Centre for Digital Music, Queen Mary University of London
Mile End Road, London E1 4NS, UK
{emmanouilb,simond}@eecs.qmul.ac.uk

Abstract. In this paper, a method for multi-pitch detection which exploits the temporal evolution of musical sounds is presented. The proposed method extends the shift-invariant probabilistic latent component analysis algorithm by introducing temporal constraints using multiple Hidden Markov Models, while supporting multiple-instrument spectral templates. Thus, this model can support the representation of sound states such as attack, sustain, and decay, while the shift-invariance across log-frequency can be utilized for multi-pitch detection in music signals that contain frequency modulations or tuning changes. For note tracking, pitch-specific Hidden Markov Models are also employed in a post-processing step. The proposed system was tested on recordings from the RWC database, the MIREX multi-F0 dataset, and on recordings from a Disklavier piano. Experimental results using a variety of error metrics, show that the proposed system outperforms a non-temporally constrained model. The proposed system also outperforms state-of-the art transcription algorithms for the RWC and Disklavier datasets.

Keywords: Music signal analysis, probabilistic latent component analysis, hidden Markov models.

1 Introduction

Multi-pitch detection is one of the core problems of music signal analysis, having numerous applications in music information retrieval, computational musicology, and interactive music systems [4]. The creation of a robust multi-pitch detection system for multiple instrument sources is considered to be an open problem in the literature. The performance of multi-pitch estimation systems has not yet matched that of a human expert, which can be partly attributed to the non-stationary nature of musical sounds. A produced musical note can be expressed by a sound state sequence (e.g. attack, transient, decay, and sustain states) [1], and can also exhibit frequency modulations such as vibrato.

* E. Benetos is funded by a Westfield Trust research studentship (Queen Mary University of London).

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 364–371, 2012.

A method for modeling sound states in music signals was proposed by Nakano et al. in [8], combining the non-negative matrix factorization (NMF) algorithm with Markov-chained constraints. Smaragdis in [11] employed the shift-invariant probabilistic latent component analysis (PLCA) algorithm for pitch tracking, which can model frequency modulations. Mysore proposed a method for sound modeling which combined the PLCA method with temporal constraints using hidden Markov models (HMMs) [7]. In [3], the authors extended the shift-invariant PLCA model for multi-pitch detection, supporting multiple instrument and pitch templates, with time-dependent source contributions. Finally, the authors combined shift-invariant PLCA with HMMs using sound state templates for modeling the temporal evolution of monophonic recordings [2].

Here, we extend the single-instrument single-pitch model of [2] for multi-pitch detection of multiple-instrument recordings. This is accomplished by extracting sound state templates for the complete pitch range of multiple instruments, and utilizing multiple independent HMMs, one for each pitch, for modeling the temporal evolution of produced notes. Experiments performed on excerpts from the RWC database [6], Disklavier recordings [9], and the MIREX multi-F0 dataset showed that the proposed model outperforms the non-temporally constrained model of [3] and also provides accuracy rates that outperform state-of-the-art methods for automatic transcription.

2 Proposed Method

The motivation behind this model is to propose a multi-pitch detection algorithm which supports multiple instrument sources, can express the temporal evolution of a produced note (by modeling sound states), and can support frequency modulations (e.g. vibrati). Frequency modulations can be supported using a shift-invariant model and a log-frequency representation, while modeling the temporal evolution of a sound can be done by utilizing templates for different sound states and constraining the order of appearance of these states using HMMs. This would allow for a rich and informative representation of the music signal, addressing some drawbacks of current polyphonic transcription systems.

2.1 Model

The proposed model extends the single-pitch single-source algorithm proposed in [2], which incorporated temporal constraints into the single-component shift-invariant PLCA algorithm. Here, this method supports multiple concurrent pitches produced by multiple instrument sources, using as an input the log-frequency spectrogram $V_{\omega,t}$, where ω is the log-frequency index and t is the time index. The model approximates the input spectrogram as a probability distribution $P(\omega, t)$:

$$P(\omega,t) = P(t)\sum_{s,p} P_t(p)P_t(s|p)\sum_{q_t^{(p)}} P_t(q_t^{(p)}|p,\bar{\omega})P(\omega|s,p,q_t^{(p)}) *_\omega P_t(f|p) \quad (1)$$

where $p = 1, \ldots, 88$ is the pitch index, s denotes the instrument source, $q^{(p)}$ the sound state for each pitch, and f the pitch shifting. Thus, $P_t(p)$ expresses the piano-roll transcription, $P_t(q_t^{(p)}|p, \bar{\omega})$ is the sound state activation for the p-th pitch, $P_t(s|p)$ the s-th instrument source contribution, $P_t(f|p)$ the pitch impulse distribution, and $P(\omega|s, p, q_t^{(p)})$ the spectral template for the s-th source, p-th pitch, and $q^{(p)}$-th sound state. The convolution of $P(\omega|s, p, q_t^{(p)}) *_\omega P_t(f|p)$ takes place between ω and f using an area spanning one semitone around the ideal position of p, in order to constrain each template for the detection of the pitch it corresponds to. In addition, such formulation allows a greater control over the polyphony level of the signal, as explained in Section 2.2. It should also be noted that $P_t(f|p)$ is not dependent on the instrument source s for computational speed purposes. This design choice might have an effect in the rare case of two instruments producing the same note concurrently. Since 60 bins per octave are used in the input log-frequency spectrogram, f has a length of 5.

Since the sequence of each pitch-specific sound state is temporally constrained, the corresponding HMM for the p-th pitch is:

$$P(\bar{\omega}) = \sum_{\bar{q}^{(p)}} \sum_{\bar{s}} \sum_{\bar{p}} \sum_{\bar{f}} P(q_1^{(p)}) \prod_t P(q_{t+1}^{(p)}|q_t^{(p)}) \prod_t P_t(\omega_t|q_t^{(p)}) \tag{2}$$

where $\bar{\omega}$ refers to all observations, $P(q_1^{(p)})$ is the state prior distribution, $P(q_{t+1}^{(p)}|q_t^{(p)})$ is the transition probability, and $P_t(\omega_t|q_t^{(p)})$ is the observation probability for the pitch sound state. The observation probability is defined as:

$$P_t(\omega_t|q_t^{(p)}) = 1 - \frac{||P(\omega, t|q_t^{(p)}) - V_{\omega,t}||_2}{\sum_{q_t^{(p)}} ||P(\omega, t|q_t^{(p)}) - V_{\omega,t}||_2} \tag{3}$$

where $||\cdot||_2$ is the l^2 norm and

$$P(\omega, t|q_t^{(p)}) = P(t) \sum_s P_t(p) P_t(s|p) P_t(q_t^{(p)}|p, \bar{\omega}) \sum_f P(\omega - f|s, p, q_t^{(p)}) P_t(f|p) \tag{4}$$

is the spectrogram reconstruction for the p-th pitch and $q^{(p)}$-th sound state. Thus, for a specific pitch, a greater observation probability is given to the state spectrogram that better approximates the input spectrogram using the Euclidean distance. Again, for computational speed purposes, the HMMs are not dependent on s, which was done in order to avoid using $S \times 88$ HMMs.

2.2 Parameter Estimation

As in the single-pitch model from [2], the aforementioned parameters can be estimated using the Expectation-Maximization algorithm. For the *Expectation* step, the update equations are:

$$P_t(f_t, s, p, q_t^{(1)}, \ldots, q_t^{(88)}|\bar{\omega}) = P_t(q_t^{(1)}, \ldots, q_t^{(88)}|\bar{\omega}) P_t(f_t, s, p|q_t^{(1)}, \ldots, q_t^{(88)}, \omega_t) \tag{5}$$

$$P_t(q_t^{(1)}, \ldots, q_t^{(88)}|\bar{\omega}) = \prod_{p=1}^{88} P_t(q_t^{(p)}|\bar{\omega}) \tag{6}$$

$$P_t(q_t^{(p)}|\bar{\omega}) = \frac{\alpha_t(q_t^{(p)})\beta_t(q_t^{(p)})}{\sum_{q_t^{(p)}} \alpha_t(q_t^{(p)})\beta_t(q_t^{(p)})} \tag{7}$$

$$P_t(f_t, s, p|\omega_t, q_t^{(1)}, \ldots, q_t^{(88)}) = \frac{P_t(p)P(\omega_t - f_t|s, p, q_t^{(p)})P_t(f_t|p)P_t(s|p)}{\sum_p P_t(p) \sum_{s,f_t} P(\omega_t - f_t|s, p, q_t^{(p)})P_t(f_t|p)P_t(s|p)} \tag{8}$$

Equation (5) is the model posterior, for the source components, sound state activity, pitch impulse, and pitch activity. In (7), $\alpha_t(q_t)$ and $\beta_t(q_t)$ are the HMM forward and backward variables, respectively, which can be computed using the forward/backward procedure described in [10] and the observation probability from (3). Also, the posterior for the pitch-wise transition matrices is:

$$P(q_{t+1}^{(p)}, q_t^{(p)}|\bar{\omega}) = \frac{\alpha_t(q_t^{(p)})P(q_{t+1}^{(p)}|q_t^{(p)})\beta_{t+1}(q_{t+1}^{(p)})P_t(\omega_{t+1}|q_{t+1}^{(p)})}{\sum_{q_t^{(p)}} \sum_{q_{t+1}^{(p)}} \alpha_t(q_t^{(p)})P(q_{t+1}^{(p)}|q_t^{(p)})\beta_{t+1}(q_{t+1}^{(p)})P_t(\omega_{t+1}|q_{t+1}^{(p)})} \tag{9}$$

For the *Maximization* step, the update equations for the unknown parameters are:

$$P(\omega|s, p, q^{(p)}) = \frac{\sum_{f,s,t} \overline{\sum}_{q_t^{(p)}} V_{\omega+f,t} P_t(f, s, p, q^{(1)}, \ldots, q^{(88)}|\omega + f)}{\sum_{\omega,f,s,t} \overline{\sum}_{q_t^{(p)}} V_{\omega+f,t} P_t(f, s, p, q^{(1)}, \ldots, q^{(88)}|\omega + f)} \tag{10}$$

where $\overline{\sum}_{q_t^{(p)}} = \sum_{q_t^{(1)}} \cdots \sum_{q_t^{(p-1)}} \sum_{q_t^{(p+1)}} \cdots \sum_{q_t^{(88)}}$,

$$P_t(f_t|p) = \frac{\sum_{\omega,s} \sum_{q_t^{(1)}} \cdots \sum_{q_t^{(88)}} V_{\omega,t} P_t(f_t, s, p, q_t^{(1)}, \ldots, q_t^{(88)}|\omega_t)}{\sum_{f_t,\omega,s} \sum_{q_t^{(1)}} \cdots \sum_{q_t^{(88)}} V_{\omega,t} P_t(f_t, s, p, q_t^{(1)}, \ldots, q_t^{(88)}|\omega_t)} \tag{11}$$

$$P(q_{t+1}^{(p)}|q_t^{(p)}) = \frac{\sum_t P(q_t^{(p)}, q_{t+1}^{(p)}|\bar{\omega})}{\sum_{q_{t+1}^{(p)}} \sum_t P(q_t^{(p)}, q_{t+1}^{(p)}|\bar{\omega})} \tag{12}$$

$$P_t(s|p) = \frac{\sum_{\omega,f_t} \sum_{q_t^{(1)}} \cdots \sum_{q_t^{(88)}} V_{\omega,t} P_t(f_t, s, p, q_t^{(1)}, \ldots, q_t^{(88)}|\omega_t)}{\sum_{s,\omega,f_t} \sum_{q_t^{(1)}} \cdots \sum_{q_t^{(88)}} V_{\omega,t} P_t(f_t, s, p, q_t^{(1)}, \ldots, q_t^{(88)}|\omega_t)} \tag{13}$$

$$P_t(p) = \frac{\sum_{\omega,f_t,s} \sum_{q_t^{(1)}} \cdots \sum_{q_t^{(88)}} V_{\omega,t} P_t(f_t, s, p, q_t^{(1)}, \ldots, q_t^{(88)}|\omega_t)}{\sum_{p,\omega,f_t,s} \sum_{q_t^{(1)}} \cdots \sum_{q_t^{(88)}} V_{\omega,t} P_t(f_t, s, p, q_t^{(1)}, \ldots, q_t^{(88)}|\omega_t)} \tag{14}$$

Finally, the pitch-wise initial state probabilities are: $P(q_1^{(p)}) = P_1(q_1^{(p)}|\bar{\omega})$. It should be noted that the spectral template update rule in (10) is not used in this system since we are utilizing pre-extracted templates, but is included for completeness.

Sparsity constraints were also incorporated, in order for the algorithm to provide as meaningful solutions as possible. Using the technique shown in [3], sparsity was enforced on the update rules for the pitch activity matrix $P_t(p)$ and the source contribution matrix $P_t(s|p)$. This means that we would like few notes active in a time frame, and that each note is produced by few instrument sources. The same sparsity parameters that were used in [3] were used. A pitch spectrogram can also be created using $P(f,p,t) = P(t)P_t(p)P_t(f|p)$ and stacking together slices of tensor $P(f,p,t)$ for all pitch values: $P(f,t) = [P(f,1,t) \cdots P(f,88,t)]$.

2.3 Postprocessing

For performing note smoothing and tracking, the resulting pitch activity matrix $P(p,t) = P(t)P_t(p)$ is postprocessed using pitch-wise HMMs, as in [9,3]. Each pitch p is modeled by a 2-state on/off HMM, while the hidden state sequence is $q'_p[t]$ and the observed sequence $o_p[t]$. MIDI files from the RWC database [6] were employed in order to estimate the pitch-wise state priors and the state transition matrices. For estimating the observation probability for each active pitch $P(o_p[t]|q'_p[t] = 1)$, we use a sigmoid curve which has $P(p,t)$ as input:

$$P(o_p[t]|q'_p[t] = 1) = \frac{1}{1 + e^{-P(p,t)}} \tag{15}$$

and use the Viterbi algorithm [10] for extracting the note tracking output for each pitch. The result of the HMM postprocessing step is a binary piano-roll transcription which can be used for evaluation.

3 Evaluation

3.1 Datasets

For training, the spectral templates $P(\omega|s,p,q^{(p)})$ were extracted for various instruments, over their complete pitch range, using $q = 3$ sound states. The extraction process was performed using the unsupervised single-source single-pitch model of [2] and the constant-Q transform with 60 bins/octave as input. Isolated note samples from 3 piano models were used from the MAPS database [5] and templates from cello, clarinet, flute, guitar, harpsichord, oboe, and violin were extracted from the RWC musical instrument sounds dataset [6]. An example of the sound state template extraction process is given in Fig. 1.

For evaluation, we employed 12 excerpts from the RWC classical and jazz datasets which are widely used for transcription (see [3] for comparative results). We also used the woodwind quintet recording from the MIREX multi-F0 development set[1]. Finally, 10 one-minute recordings taken from a Yamaha Disklavier piano which were presented in [9] were also utilized.

[1] `http://www.music-ir.org/mirex`

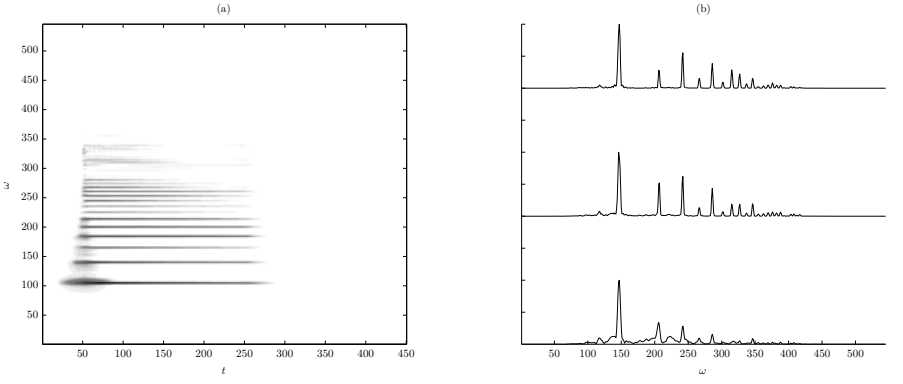

Fig. 1. (a) Spectrogram $V_{\omega,t}$ of a D3 piano note (b) Extracted spectral templates using the method in [2] corresponding to different sound states

3.2 Results

For evaluation, the transcription metrics also used in [3] were utilized, namely the two accuracy measures (Acc_1, Acc_2), the total error (E_{tot}), the substitution error (E_{subs}), missed detection error (E_{fn}), and false alarm error (E_{fp}). Compared to Acc_1, accuracy Acc_2 also takes into account note substitutions. All evaluations take place by comparing the transcribed pitch output and the ground-truth MIDI files at a 10 ms scale.

For comparison, we employed the shift-invariant PLCA-based transcription model of [3] with the same CQT resolution as in the proposed model. The system used for comparison does not support any temporal constraints but uses the same formulation for source contribution, pitch impulse, pitch activity, as well as the same postprocessing step. Experiments were performed using ergodic HMMs (initialized with uniform transition probabilities), as they demonstrated superior performance compared to left-to-right HMMs for the single-pitch detection experiments in [2]. As explained in [2], although left-to-right HMMs might be more suitable for instruments exhibiting a clear temporal structure in note evolution (such as piano), in most instruments a fully connected HMM is more appropriate for expressing the temporal evolution of sound states. An example of the multi-pitch detection process can be seen in Fig. 2 where the pitch spectrogram of a guitar recording can be seen, along with the MIDI ground truth.

Results for the multi-pitch estimation experiments are presented in table 1, comparing the performance of the proposed method with the non-temporally constrained system of [3], over the three datasets. It can be seen that in all cases, the proposed method outperforms the shift-invariant PLCA-based model, with the smallest difference in terms of accuracy occurring for the MIREX recording. It should be noted that for the Disklavier dataset from [9], only piano templates were used in both systems. A common observation for all experiments is that the number of missed pitch detections is higher than the number of false positives.

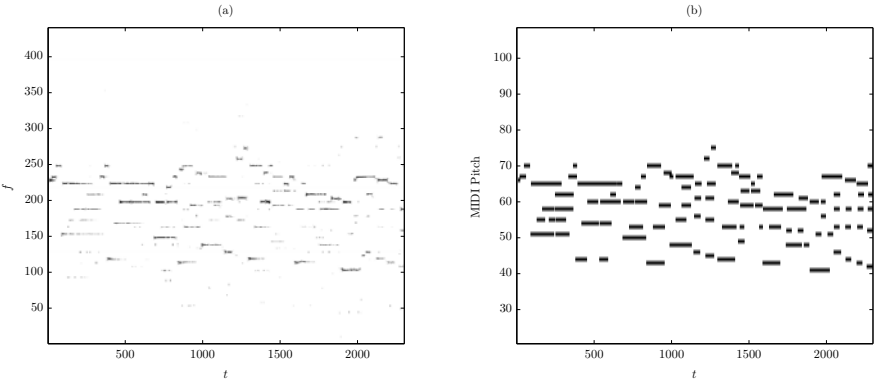

Fig. 2. (a) Pitch spectrogram $P(f, t)$ of an excerpt of "RWC-MDB-J-2001 No. 7" (guitar). (b) The pitch ground truth of the same recording. The abscissa corresponds to 10ms.

Table 1. Multi-pitch detection results using the proposed method compared to the one in [3] using three datasets

Dataset	**Method**	Acc_1	Acc_2	E_{tot}	E_{subs}	E_{fn}	E_{fp}
RWC	Proposed	61.6%	62.8%	37.2%	9.1%	18.3%	9.8%
	[3]	59.5%	60.3%	39.7%	9.2%	20.3%	10.2%
Disklavier	Proposed	58.6%	57.3%	42.7%	9.9%	16.3%	16.5%
	[3]	57.4%	55.5%	44.5%	10.8%	16.3%	17.4%
MIREX	Proposed	41.0%	47.0%	53.0%	25.4%	20.1%	7.5%
	[3]	40.5%	46.3%	53.8%	18.5%	32.3%	3.0%

Also, for the RWC and Disklavier datasets, results outperform state-of-the-art transcription algorithms (see [3] for transcription results using other methods in the literature). It should also be noted that most of the missed detections are located in the decay part of the produced notes. When no sparsity is used, the proposed method reports accuracy metrics $\{Acc_1, Acc_2\}$ of $\{56.3\%, 55.6\%\}$ for the RWC database, $\{56.8\%, 53.1\%\}$ for the Disklavier dataset, and $\{40.8\%, 46.9\%\}$ for the MIREX recording. Selected transcription examples are available online[2], along with the original recordings for comparison.

To the authors' knowledge, no statistical significance tests have been made for multi-pitch detection, apart from the piecewise Friedman tests in the MIREX task. However, given the fact that evaluations actually take place using 10 ms frames, even a small accuracy change can be shown to be statistically significant. Also, it should be noted that although using factorial HMMs (as in the source separation experiments of [7]) for the temporal constraints might in theory produce improved detection results, the model would be intractable, since it would need to compute 3^{88} sound state combinations.

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

4 Conclusions

In this work we proposed a model for multi-pitch detection that extends the shift-invariant PLCA algorithm by introducing temporal constraints using HMMs. The goal was to model the temporal evolution for each produced note using spectral templates for each sound state. Results indicate that the temporal constraints produce improved multi-pitch detection accuracy rates compared to the standard shift-invariant PLCA model. It is also seen that the proposed system outperforms the state-of-the-art methods for the RWC transcription dataset and the Disklavier [9] dataset.

In the future, the proposed model will be tested using different HMM topologies and by incorporating update scheduling procedures for the various parameters to be estimated. Finally, the proposed transcription system will be extended by including an instrument identification step and by jointly performing multi-pitch estimation with note tracking.

References

1. Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., Sandler, M.: A tutorial on onset detection of music signals. IEEE Trans. Audio, Speech, and Language Processing 13(5), 1035–1047 (2005)
2. Benetos, E., Dixon, S.: A temporally-constrained convolutive probabilistic model for pitch detection. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, pp. 133–136 (October 2011)
3. Benetos, E., Dixon, S.: Multiple-instrument polyphonic music transcription using a convolutive probabilistic model. In: 8th Sound and Music Computing Conf., Padova, Italy, pp. 19–24 (July 2011)
4. de Cheveigné, A.: Multiple F0 estimation. In: Wang, D.L., Brown, G.J. (eds.) Computational Auditory Scene Analysis, Algorithms and Applications, pp. 45–79. IEEE Press/Wiley (2006)
5. Emiya, V., Badeau, R., David, B.: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. IEEE Trans. Audio, Speech, and Language Processing 18(6), 1643–1654 (2010)
6. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: RWC music database: music genre database and musical instrument sound database. In: Int. Conf. Music Information Retrieval, Baltimore, USA (October 2003)
7. Mysore, G.: A non-negative framework for joint modeling of spectral structure and temporal dynamics in sound mixtures. Ph.D. thesis, Stanford University, USA (June 2010)
8. Nakano, M., Le Roux, J., Kameoka, H., Kitano, Y., Ono, N., Sagayama, S.: Nonnegative Matrix Factorization with Markov-Chained Bases for Modeling Time-Varying Patterns in Music Spectrograms. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 149–156. Springer, Heidelberg (2010)
9. Poliner, G., Ellis, D.: A discriminative model for polyphonic piano transcription. EURASIP J. Advances in Signal Processing (8), 154–162 (January 2007)
10. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE 77(2), 257–286 (1989)
11. Smaragdis, P.: Relative-pitch tracking of multiple arbitary sounds. J. Acoustical Society of America 125(5), 3406–3413 (2009)