Audio Engineering Society

Conference Paper

Presented at the Conference on
Semantic Audio
2017 June 22 – 24, Erlangen, Germany

Polyphonic note and instrument tracking using linear dynamical systems

Emmanouil Benetos

Centre for Digital Music, Queen Mary University of London, UK

Correspondence should be addressed to Emmanouil Benetos (emmanouil.benetos@qmul.ac.uk)

ABSTRACT

In this paper, a system for automatic transcription of multiple-instrument polyphonic music is proposed, which supports tracking multiple concurrent notes using linear dynamical systems (LDS). The system is based on a spectrogram factorisation model which extends probabilistic latent component analysis (PLCA), and supports the detection of multiple pitches, instrument contributions, and pitch deviations. In order to jointly track multiple concurrent pitches, the use of LDS as prior to the PLCA model is proposed. LDS parameters are learned in a training stage using score-informed transcriptions; for LDS inference, online and offline variants are evaluated. The MAPS piano music dataset and the Bach10 multi-instrument dataset are used for note tracking experiments, with the latter dataset also being evaluated with respect to instrument assignment performance. Results show that the proposed LDS-based method can successfully track multiple concurrent notes, leading to an improvement of over 3% in terms of note-based F-measure for both datasets over benchmark note tracking approaches.

Introduction

Automatic music transcription (AMT) is a core problem in the area of music signal analysis, with several applications in music informatics, interactive music systems, and computational musicology [1]. The vast majority of AMT systems consist of a frame-based multi-pitch detection process, followed by a note tracking process in order to output note events with a start and end time. The most common note tracking approach is to perform temporal smoothing to the multi-pitch output at a postprocessing stage (e.g. using a median filter), followed by minimum duration pruning, where note events with small durations are deleted [2, 3, 4]. Another common approach is the use of 2-state on/off hidden Markov models (HMMs) [5], which consider each pitch independently. More recently, [6] proposed a method for note tracking at a postprocessing stage using maximum likelihood sampling.

In this work, the use of linear dynamical systems (LDS) [7, Ch. 18] is proposed for jointly tracking notes in polyphonic music. This approach, which is inspired by the use of LDS for environmental sound recognition in [8], assumes that the non-binary pitch activation output of an AMT system is the noisy observation in an LDS, with the latent states corresponding to the 'correct' and temporally smoothed pitch activation output. LDS parameters are learned through

EB is supported by a RAEng Research Fellowship (grant no. RF/128).

the use of fully observed data generated by score-informed transcriptions. The proposed method can be integrated with the multi-pitch estimation stage, by using the LDS posterior mean as prior in a spectrogram factorisation-based music transcription system [9]. Experiments on the MAPS piano database and the Bach10 multi-instrument dataset show that the proposed note tracking approach can lead to significant improvements in terms of multi-pitch detection and instrument assignment performance.

When compared with state-of-the-art note tracking methods presented in the above paragraphs, the proposed approach can both be used as a postprocessing step and also be integrated with multi-pitch detection. In addition, contrary to the use of pitch-wise independent HMMs, the proposed method jointly tracks all pitches, with an 88-dimensional latent space (corresponding to pitches A0-C8). It should be noted that it would be practically impossible to jointly model all possible note combinations using a single HMM, since the state space would be too large: with a maximum polyphony level of 6 when considering 60 possible notes, the number of possible note combinations would be $56 \cdot 10^6$ [10]; however this combinatorial problem is circumvented when using LDS. Finally, a model for non-negative LDS was proposed in [11] and applied to music signal analysis in [12]. However, the primary motivation behind the non-negative LDS model of [11] was to provide temporally smooth component activations while the observation model corresponds to standard non-negative matrix factorisation (NMF). Here, an added benefit of the proposed approach is that it provides a mapping between the observed and 'true' pitch activations, thus being able to correct common multi-pitch detection errors.

The outline of this paper is as follows. The next section ("Proposed System") presents motivation behind the proposed approach, as well as the multi-pitch detection and note tracking components of the system. Section "Evaluation" presents the datasets, evaluation metrics, comparative approaches and experimental results. The final section concludes the paper and outlines future work.

Proposed System

Motivation

The aim of the proposed method is to jointly track multiple concurrent pitches in the context of automatic music transcription, through the use of linear dynamical systems (LDS). LDS, also called linear-Gaussian state space models or Kalman filters, are generalisations of HMMs where the state space is continuous [7]. In addition, the latent and observed variables are multivariate Gaussians whose means are linear functions of their parent states. A stationary LDS can be formulated as:

$$\begin{aligned} \mathbf{z}_t &= \mathbf{A}\mathbf{z}_{t-1} + \varepsilon_t \\ \mathbf{y}_t &= \mathbf{B}\mathbf{z}_t + \delta_t \\ \varepsilon_t &\sim \mathcal{N}(0, \mathbf{U}) \\ \delta_t &\sim \mathcal{N}(0, \mathbf{R}) \end{aligned} \quad (1)$$

where $\mathbf{z}_t$ is the latent state at time t, $\mathbf{A}$ is the transition model, ε_t is the system noise with covariance $\mathbf{U}$, $\mathbf{y}_t$ is the observation sequence, $\mathbf{B}$ is the observation model, and δ_t is the observation noise with covariance $\mathbf{R}$.

In this work, we assume that the noisy pitch activation output of a frame-based transcription system (presented in subsection 'Multi-Pitch Detection') corresponds to the observation sequence in an LDS, with the latent states corresponding to the correct pitch activations. Thus, the proposed LDS-based approach can both jointly track multiple pitches through its transition model, and at the same time correct any transcription errors through its observation model. The proposed approach does not impose any constraints on the level of polyphony. In addition, we propose the use of LDS as prior to the multi-pitch detection model.

Multi-Pitch Detection

The multi-pitch detection model is based on the system of [9], which extends probabilistic latent component analysis (PLCA), a spectrogram factorisation method. The system takes as input a log-frequency spectrogram $V_{\omega,t}$ (ω is the log-frequency index and t is the time index). Here, $V_{\omega,t}$ corresponds to the variable-Q transform (VQT) spectrogram of a music recording [13], with a log-frequency resolution of 60 bins/octave and a minimum frequency of 27.5Hz.

In the PLCA-based model, $V_{\omega,t}$ is approximated as a bivariate probability distribution $P(\omega,t)$, which is in turn decomposed into a dictionary of log-spectral templates, along with probability distributions for pitch,

instrument, tuning, and sound state[1] activation. The PLCA model is formulated as:

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

where q is the sound state index, p is the pitch index in semitone scale ($p \in \{1,\ldots,88\}$ corresponds to MIDI pitches A0 to C8), s is the instrument source index, and f denotes deviation from ideal tuning. $P(t)$ is defined as $\sum_\omega V_{\omega,t}$, which is a known quantity. Dictionary $P(\omega|q,p,f,s)$ is a 5-dimensional tensor of pre-extracted log-spectral templates per pitch p, instrument s, tuning f and sound state q. $P_t(f|p)$ denotes deviation from ideal tuning for a specific pitch across time, $P_t(s|p)$ denotes instrument contribution for a specific pitch across time, $P_t(p)$ is the pitch activation probability (the main output of the transcription system, used for multi-pitch detection evaluation), and $P_t(q|p)$ is the sound state activation probability per pitch across time. In the model of (2), $f \in [1,\ldots,5]$, with $f = 3$ corresponding to the ideal tuning position. Given that the employed time-frequency representation has a resolution of 5 bins/semitone, this means that f is able to capture tuning deviations of $\pm 20, \pm 40$ cent around the ideal tuning position. More information on the construction of dictionary $P(\omega|q,p,f,s)$ is given in subsection 'Datasets'.

Unknown parameters $P_t(f|p)$, $P_t(s|p)$, $P_t(p)$, and $P_t(q|p)$ can be iteratively estimated using the Expectation-Maximization (EM) algorithm [14]. The EM equations for the PLCA-based music transcription model can be found in [9]; in this work, 30 iterations are used for parameter estimation. The dictionary $P(\omega|q,p,f,s)$ is considered fixed and is not updated. Sparsity constraints are also incorporated on $P_t(p)$ and $P_t(s|p)$, as in [9]. This is done in order to control the polyphony level and the instrument contribution in the resulting transcription. The multi-pitch output of the PLCA model is given by $P(p,t) = P(t)P_t(p)$, i.e. the pitch activation probability weighted by the energy of the spectrogram. For instrument assignment evaluation, we can compute $P(s,p,t) = P(t)P_t(p)P_t(s|p)$, which is essentially an instrument-specific transcription.

[1] A sound state represents different stages in the temporal evolution of a musical note. For the case of piano, sound states can correspond to the attack, sustain, and decay.

Note Tracking

The model of (2) does not contain any temporal constraints, which can lead to a fragmented pitch activation output. In order to track multiple pitches over time, we propose the use of LDS, where we assume that $P(p,t)$ corresponds to a 'noisy' observation $\mathbf{y}_t \in \mathbb{R}^{88}$ in an LDS ($t \in \{1,\ldots,T\}$). The aim of the proposed note tracking step is to estimate the latent state sequence $\mathbf{z}_t \in \mathbb{R}^{88}$ which would correspond to the desired clean and temporally smoothed output.

In order to estimate the LDS parameters $\mathbf{A},\mathbf{B},\mathbf{U}$, and $\mathbf{R}$, we can make use of *fully observed data* in a training step, following the process described in [7, Ch. 18]. In the case of multi-pitch detection, the latent states $\mathbf{z}_t$ (corresponding to the desired output) can be estimated by performing score-informed transcription in a training dataset. Given an input recording and aligned score, the pitch activation probability $P_t(p)$ can be initialised using a binary mask that corresponds to ground truth pitches. Following the PLCA iterations described in section 'Multi-Pitch Detection', the score-informed output (denoted as $P'(p,t)$) only has non-zero pitch activations in the time instants and pitch indices that correspond to the ground truth score.

Given the score-informed transcriptions $P'(p,t)$ and the automatic transcriptions $P(p,t)$, LDS parameters $\mathbf{A}$ and $\mathbf{B}$ are estimated by solving least squares problems for $\mathbf{z}_{t-1} \rightarrow \mathbf{z}_t$ and $\mathbf{z}_t \rightarrow \mathbf{y}_t$ [7]:

$$\begin{aligned} J(\mathbf{A}) &= \sum_t (\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1})^2 \\ J(\mathbf{B}) &= \sum_t (\mathbf{y}_t - \mathbf{B}\mathbf{z}_t)^2 \end{aligned} \quad (3)$$

where $\mathbf{y}_\tau = P(p,t=\tau)$ and $\mathbf{z}_\tau = P'(p,t=\tau)$. The system noise covariance matrix $\mathbf{U}$ is subsequently estimated from the residuals in predicting $\mathbf{z}_t$ from $\mathbf{z}_{t-1}$, and the observation noise covariance $\mathbf{R}$ is estimated from the residuals in predicting $\mathbf{y}_t$ from $\mathbf{z}_t$.

An example LDS observation matrix $\mathbf{B}$ can be seen in Fig. 1, learned from a fold from the MAPS piano database (see section 'Datasets' for details on the dataset). The strong diagonal shows a direct mapping between the observed and latent pitch; however the presence of 12-subdiagonal and 12-superdiagonal entries indicate the occurrence of possible octave errors (both subharmonic and superharmonic) in the transcription system, which are to be filtered out. For a

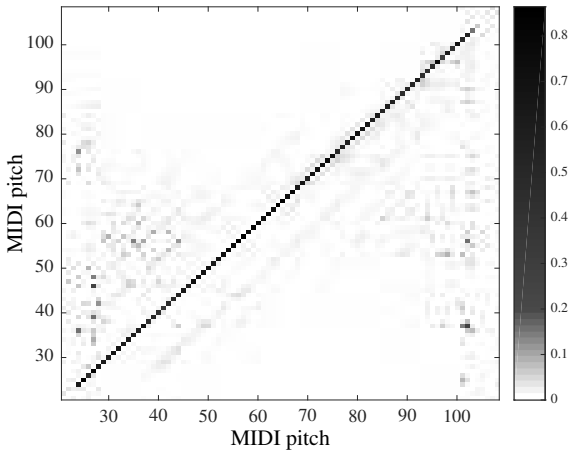

Fig. 1: The LDS matrix **B**, learned from fold 1 of the MAPS piano database. Rows correspond to the latent pitch dimension, columns to the observed pitch dimension.

certain pitch range, similar super- and sub-diagonal entries can be seen for one semitone intervals and 19-semitone intervals (the latter commonly occur as errors in transcription systems due to their correspondence to the 3rd harmonic of a given pitch).

Having learned LDS parameters, note tracking is carried out by performing *LDS inference*, in other words estimating the posterior $P(\mathbf{z}_t|\mathbf{y}_{1:t})$ for an online LDS or $P(\mathbf{z}_t|\mathbf{y}_{1:T})$ in the offline case. Inference is achieved by applying the Kalman filter and Kalman smoother equations for the online and offline cases, respectively [7, Ch. 18]. For online inference, the posterior is represented as $P(\mathbf{z}_t|\mathbf{y}_{1:t}) = \mathcal{N}(\mathbf{z}_t|\mu_t, \Sigma_t)$. The output of the proposed online note tracking process is the LDS posterior mean $\mu_t \in \mathbb{R}^{88}$ (corresponding to 88 pitches in semitone scale from A0 to C8). For the offline model (i.e. having access to both past and future samples), the LDS posterior is $P(\mathbf{z}_t|\mathbf{y}_{1:T}) = \mathcal{N}(\mathbf{z}_t|\mu_{t|T}, \Sigma_{t|T})$, where the output of the note tracking process is the LDS posterior mean $\mu_{t|T}$.

In the present transcription system, the proposed LDS-based note tracking process can either be applied as a postprocessing step, or can be used as prior information in the PLCA updates. For the postprocessing use of the LDS-based note tracking method, the multi-pitch detection output of the AMT system $P(p,t)$ is given as input to an LDS, and the LDS mean μ_t or $\mu_{t|T}$ (for the online and offline model, respectively) is used as the postprocessed output of the transcription system. The second use of the proposed approach is to use the LDS-based tracking as prior in the PLCA model. In PLCA models, prior information can be incorporated in the form of Dirichlet priors [15]; we thus define the Dirichlet hyperparameter for the 'clean' pitch activation as:

$$\phi(p|t) \propto \mu_t \tag{4}$$

We then modify the *M-step* update rule for estimating $P_t(p)$ as to include a weighted prior with the LDS-based note tracking output (see [9] for the original update equation):

$$P_t(p) \propto (w-1) \cdot \Big(\sum_{\omega,f,s,q} P_t(q,p,f,s|\omega) V_{\omega,t} \Big) + w \cdot \phi(p|t) \tag{5}$$

where $w \in [0,1]$ is a weight indicating how much the prior should be imposed. The aim of using the LDS-based note tracking within the PLCA parameter estimation is to guide the PLCA convergence to a temporally smooth solution, whilst also providing a mapping between the observed and latent pitch activations. It should be noted that the LDS does not impose any constraints on the inputs being non-negative or summing to one. Thus, in order to ensure that the updated $P_t(p)$ remains non-negative in (5), only the non-negative values of μ_t (or $\mu_{t|T}$) are kept.

For instrument assignment evaluations (detecting multiple pitches and assigning each detected note to an instrument), we also propose the use of LDS tracking to the instrument contribution matrix. This essentially means performing LDS-based note tracking on the 2-dimensional instrument-specific pitch activation $P(s = i,p,t) = P_t(s = i|p)P_t(p)P(t)$, where $i \in \{1,\ldots,S\}$ corresponds to a specific instrument source. Instrument-specific LDS note tracking can be performed exactly in the same way as to that applied to the pitch activation $P_t(p)$ (i.e. using the same learned LDS parameters $\mathbf{A},\mathbf{B},\mathbf{U},\mathbf{R}$). The motivation behind this use is to temporally smooth the instrument contribution of each instrument over time, thus leading to a less temporally fragmented instrument assignment output.

Postprocessing

The output of the note tracking step is a smooth pitch activation μ_t or $\mu_{t|T}$, for online and offline LDS models, respectively. For multi-pitch detection, this non-binary time-pitch representation needs to be converted

into a list of detected pitches, with respective onsets and offsets. Firstly, the note tracking output is post-processed by first detecting the overall tuning level of the recording using information from $P_t(f|p)$ and compensating for any tuning deviations from 440 Hz tuning, following the method of [16]. The tuning-compensated LDS-smoothed pitch activation is then binarized by performing thresholding, followed by minimum duration pruning (i.e. removing detected pitches with a duration smaller than 20ms).

Evaluation

Datasets

For evaluation we used 2 benchmark music transcription datasets: the MAPS piano dataset [10] and the Bach10 multi-instrument dataset [17]. For the PLCA model, a dictionary of spectral templates $P(\omega|q,p,f,s)$ was created by using isolated note samples from all 9 pianos from the MAPS dataset. The complete piano note range was used; templates were pre-shifted across log-frequency in order to account for tuning deviations. The number of sound states is set to $Q = 3$. For multiple-instrument transcription using the Bach10 dataset, the PLCA dictionary contains spectral templates for bassoon, clarinet, alto sax and violin, all of which were extracted from isolated notes from the RWC database [18].

For testing piano transcription, we carried out 4-fold cross validation on the 270 recordings of the MAPS dataset using the folds created in [19]. LDS parameters and weight w are learned in each fold using its respective training set. The aforementioned folds are music piece-independent, which means that any music compositions found in the training set are not present in the test set, thus allowing for a fair evaluation of the LDS (since they can be viewed as a music language model). For consistency with [19], only the first 30 sec of each recording are evaluated.

For multi-instrument transcription, we used the 10 complete recordings found in the Bach10 dataset [17], which are performed by violin, clarinet, saxophone, and bassoon. We use the Bach10 dataset both for multi-pitch detection and instrument assignment experiments using the final 4-voice mixes (not the individual instrument tracks). For note tracking using the Bach10 dataset we used an LDS trained on the complete set of piano recordings from the MAPS database; this was done in order to investigate the generalisation capabilities of the LDS to a different dataset and instrument set.

Metrics

For multi-pitch detection evaluation, the onset-based metric used in the MIREX Note Tracking task [20] is utilised. A detected note event is assumed to be correct if its pitch corresponds to the ground truth pitch and its onset is within a ± 50 ms range of the ground truth onset. Using the above rule, precision ($\mathscr{P}$), recall ($\mathscr{R}$), and F-measure ($\mathscr{F}$) metrics are defined. For instrument assignment evaluations with the Bach10 dataset, the pitch ground-truth of each instrument is used and compared against the instrument-specific output of the system. As with the multi-pitch metrics, we define the following note-based instrument assignment metrics: $\mathscr{F}_v$, $\mathscr{F}_c$, $\mathscr{F}_s$, $\mathscr{F}_b$, corresponding to violin, clarinet, saxophone, and bassoon, respectively. We also use an overall instrument assignment metric (averaged across all 4 instruments), denoted as $\mathscr{F}_{ins}$ (see [9] for more information on metrics).

Comparative Approaches

For comparison with the proposed LDS-based note tracking approach, firstly median filtering is applied on $P(p,t)$ followed by thresholding with minimum note duration pruning, as in [2]. The span of the filter is determined on each training fold.

A second comparative note tracking approach is employed using HMM-based postprocessing, following the method of [5]. Each pitch $p \in \{1,\ldots,88\}$ is modelled by a 2-state on/off HMM. Pitch-wise transition probabilities are estimated using ground truth MIDI for each training fold, while priors are assumed to be uniform. The observation probability of each pitch-wise HMM for an active pitch is defined as a sigmoid function [21]:

$$P(o_t^{(p)}|q_t^{(p)} = 1) = \frac{1}{1 + e^{-P(p,t)-\lambda}} \qquad (6)$$

where $o_t^{(p)}$ is the observation at time t for the p-th HMM, $q_t^{(p)}$ is the corresponding latent state (0 for an 'off' pitch, 1 for an 'on' pitch), and λ is a threshold that controls smoothing (computed for each training fold). The Viterbi algorithm [22] is used for inference. Both median filtering and HMM-based approaches also include tuning estimation and compensation, as described in subsection 'Postprocessing'.

System	$\mathscr{F}$	$\mathscr{P}$	$\mathscr{R}$
Median filtering [2]	69.6%	72.7%	68.5%
HMM postprocessing [5, 21]	68.4%	86.2%	58.1%
LDS postprocessing	70.0%	81.3%	62.8%
LDS prior	73.0%	83.0%	66.2%

Table 1: Multi-pitch detection results for the MAPS piano dataset, using the proposed LDS-based note tracking configurations and comparative approaches.

System	$\mathscr{F}$	$\mathscr{P}$	$\mathscr{R}$
Median filtering [2]	64.6%	57.5%	73.7%
HMM postprocessing [5, 21]	61.3%	79.6%	50.0%
LDS postprocessing	64.7%	59.3%	71.2%
LDS prior	66.5%	61.8%	72.2%
LDS prior 2	67.6%	63.1%	72.9%

Table 2: Multi-pitch detection results for the Bach10 multi-instrument dataset, using the proposed LDS-based note tracking configurations and comparative approaches.

Results

For piano transcription experiments, 4 note tracking configurations are evaluated. The first configuration ("Median filtering") refers to using the frame-based PLCA model of subsection 'Multi-Pitch Detection' alone, with median filtering and minimum duration pruning applied in a postprocessing stage. The second configuration ("HMM postprocessing") corresponds to the pitch-wise 2-state HMMs described in the previous subsection. Configuration "LDS postprocessing" refers to using the proposed LDS-based method for postprocessing the PLCA pitch activation $P_t(p)$. Finally, the fourth configuration ("LDS prior") refers to the use of the LDS-based model as prior in the PLCA pitch activation probability. All LDS-based note tracking methods use the offline model (i.e. Kalman smoother) as the default option.

Table 1 presents multi-pitch detection results for the MAPS piano dataset, averaged across the 4 folds. Results show that the proposed LDS-based note tracking method when used as prior leads to an improvement of more than 3% in terms of $\mathscr{F}$, when compared with median filtering or HMM postprocessing. LDS-based note tracking used as postprocessing leads to a very small (0.4%) improvement over the use of LDS as prior. More insight is given by the precision and recall metrics: LDS tracking increases precision, whilst lowering recall at a smaller rate. HMM postprocessing leads to the best possible precision, at the expense of the lowest achieved recall. Essentially, the LDS method reduces the number of false alarms, to the expense of introducing additional missed detections.

In order to make a comparison between the online and offline variants of the note tracking model, experiments on the MAPS database were carried out using the online variant (corresponding to the Kalman filter) as prior to the pitch activation. Results using the same experimental setup show that the online model reaches $\mathscr{F} = 72.95\%$, thus the performance difference between the online and offline LDS variants for note tracking is negligible (this is attributed to the small size of time frames). With respect to comparison with state-of-the-art approaches, the proposed note tracking method using the LDS prior outperforms several neural network-based methods when using the MAPS database with the same experimental setup [19].

Results on multi-pitch detection using the multi-instrument Bach10 dataset are shown in Table 2. The LDS-based prior leads to an improvement of +1.9% in terms of $\mathscr{F}$ over median filtering, while HMM postprocessing suffers from low recall, as with the MAPS dataset. As with the piano transcription experiments, the proposed note tracking method leads to an increase in precision over median filtering. An additional experiment using the Bach10 dataset is also made using a fifth configuration ("LDS prior 2"), where the LDS prior is applied to both the pitch activation matrix $P_t(p)$ and the instrument contribution matrix $P_t(s|p)$ for each instrument separately (the LDS parameters were still trained from the complete set of piano recordings from the MAPS database). This combined pitch and instrument tracking outperforms both median filtering and HMM postprocessing, and shows that additional gains can be made by tracking each instrument source. Results using both LDS configurations also outperform the state-of-the-art wrt multi-pitch detection for the Bach10 dataset [9] which was at $\mathscr{F} = 65.0\%$.

An example transcription output is given in Fig. 2, for a segment from the Bach10 dataset. By comparing Fig 2(a) with 2(b), it can be seen that the LDS-based note tracking is able to suppress false alarms in the higher pitch region, as well as several semitone and octave errors occurring within the main 4 voices.

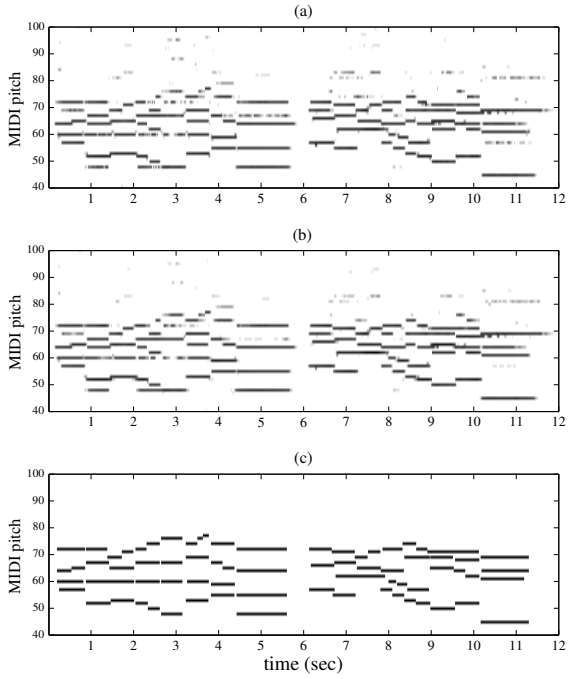

Fig. 2: (a) The pitch activation $P_t(p)$ of a segment from recording 'Ach Lieben Christen' from the Bach10 dataset using the PLCA model alone. (b) The pitch activation using the LDS-based prior. (c) The pitch ground truth.

Finally, instrument assignment results using the Bach10 dataset are presented in Table 3, where the combined pitch and instrument tracking of configuration 5 outperforms both median filtering and HMM postprocessing (this is particularly attributed to an improvement in violin transcription/assignment performance for the LDS instrument prior). It is also worth noting that note tracking alone does not lead to an improvement in instrument assignment performance.

Conclusions

This paper proposed a method for note tracking in polyphonic music using linear dynamical systems. In the proposed approach, the pitch activation output of a transcription system is assumed to be the (noisy) observation in an LDS and the latent states correspond to clean pitch activations. LDS parameters are computed in a training step using score-informed transcriptions. When integrated as prior to a PLCA-based multiple-instrument transcription system, the proposed note tracking approach leads to an improvement in terms of both multi-pitch detection and instrument assignment performance, when compared to median filtering or HMM-based note tracking approaches. Future work includes tracking high-resolution pitch trajectories using LDS, by exploiting information from the pitch deviation probability $P_t(f|p)$. Finally, non-linear dynamical systems (such as extended and unscented Kalman filters [7]) will also be investigated for note and instrument tracking in polyphonic music.

References

[1] Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., and Klapuri, A., "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, 41(3), pp. 407–434, 2013.

[2] Dessein, A., Cont, A., and Lemaitre, G., "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *ISMIR*, pp. 489–494, 2010.

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

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

[5] Poliner, G. and Ellis, D., "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Advances in Signal Processing*, (8), pp. 154–162, 2007.

[6] Duan, Z. and Temperley, D., "Note-level music transcription by maximum likelihood sampling," in *ISMIR*, 2014.

[7] Murphy, K. P., *Machine Learning: A Probabilistic Perspective*, Adaptive computation and machine learning, MIT Press, 2012.

[8] Benetos, E., Lafay, G., Lagrange, M., and Plumbley, M. D., "Polyphonic sound event tracking using linear dynamical systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6), 2017, to appear.

System	$\mathscr{F}_v$	$\mathscr{F}_c$	$\mathscr{F}_s$	$\mathscr{F}_b$	$\mathscr{F}_{ins}$
Median filtering [2]	10.1%	40.4%	34.7%	41.6%	31.7%
HMM postprocessing [5, 21]	10.5%	39.9%	29.7%	50.1%	32.6%
LDS postprocessing	12.1%	40.9%	31.0%	44.6%	32.2%
LDS prior	13.0%	40.2%	30.1%	43.0%	31.6%
LDS prior 2	19.2%	41.8%	34.6%	47.0%	35.7%

Table 3: Instrument assignment results for the Bach10 dataset, using the proposed LDS-based note tracking configurations and comparative approaches.

[9] Benetos, E. and Weyde, T., "An efficient temporally-constrained probabilistic model for multiple-instrument music transcription," in *16th International Society for Music Information Retrieval Conference*, pp. 701–707, 2015.

[10] Emiya, V., Badeau, R., and David, B., "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6), pp. 1643–1654, 2010.

[11] Févotte, C., Roux, J. L., and Hershey, J. R., "Non-negative dynamical system with application to speech and audio," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3158–3162, 2013.

[12] Simsekli, U., Le Roux, J., and Hershey, J., "Hierarchical and coupled non-negative dynamical systems with application to audio modeling," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[13] Schörkhuber, C., Klapuri, A., Holighaus, N., and Dörfler, M., "A Matlab Toolbox for Efficient Perfect Reconstruction Time-Frequency Transforms with Log-Frequency Resolution," in *AES 53rd Conference on Semantic Audio*, 2014.

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

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

[16] Holzapfel, A. and Benetos, E., "The sousta corpus: beat-informed automatic transcription of traditional dance tunes," in *17th International Society for Music Information Retrieval Conference*, pp. 531–537, 2016.

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

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

[19] Sigtia, S., Benetos, E., and Dixon, S., "An End-to-End Neural Network for Polyphonic Piano Music Transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5), pp. 927–939, 2016.

[20] "Music Information Retrieval Evaluation eXchange (MIREX)," `http://music-ir.org/mirexwiki/`, Jan. 2017 (last accessed).

[21] Benetos, E. and Dixon, S., "A shift-invariant latent variable model for automatic music transcription," *Computer Music Journal*, 36(4), pp. 81–94, 2012.

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