City Research Online

City, University of London Institutional Repository

Citation: Benetos, E. and Dixon, S. (2011). A temporally-constrained convolutive probabilistic model for pitch detection. Paper presented at the Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on, 16 - 19 Oct 2011, New Paltz, NY, US.

This is the unspecified version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/2769/

Link to published version:

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

A TEMPORALLY-CONSTRAINED CONVOLUTIVE PROBABILISTIC MODEL FOR 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

A method for pitch detection which models the temporal evolution of musical sounds is presented in this paper. The proposed model is based on shift-invariant probabilistic latent component analysis, constrained by a hidden Markov model. The time-frequency representation of a produced musical note can be expressed by the model as a temporal sequence of spectral templates which can also be shifted over log-frequency. Thus, this approach can be effectively used for pitch detection in music signals that contain amplitude and frequency modulations. Experiments were performed using extracted sequences of spectral templates on monophonic music excerpts, where the proposed model outperforms a non-temporally constrained convolutive model for pitch detection. Finally, future directions are given for multipitch extensions of the proposed model.

Index Terms— Shift-invariant probabilistic latent component analysis, hidden Markov models, pitch detection

1. INTRODUCTION

Pitch estimation of music signals is the core problem in the development of automatic transcription systems, with numerous applications in the fields of music information retrieval, interactive computer systems, and automated musicological analysis [1]. While pitch estimation for monophonic music signals is considered to be a solved problem, estimating the pitch of multiple concurrent sources still remains open. One of the reasons why the performance of multipitch estimation systems has not yet matched that of a human expert lies in the non-stationarity of music sounds. A note produced by a musical instrument could be expressed as a sequence of sound states, namely the attack, transient, sustain, and decay parts [2]. Additionally, depending on the instrument, frequency modulations such as vibrato and amplitude modulations such as tremolo might also take place. In the past, models have been proposed which attempt to address these pitch detection issues, such as frequency modulations (e.g. [3]) or sound production states (e.g. [4]).

In this work, a method is proposed which attempts to model the temporal evolution of music sounds and also address frequency modulations. This work is based on the model in [5], which combined probabilistic latent component analysis (PLCA) [6] with hidden Markov models (HMMs) [7], where each hidden state corresponds to a musical note. Here, the proposed model constrains the shift-invariant PLCA model [3] with an HMM, where each hidden state corresponds to a temporal state of the produced sound. For experiments, spectral templates of sound states for a piano, cello, and an oboe were extracted. Using the proposed model, a supervised pitch detection method is proposed using the extracted templates. For comparison, the shift-invariant PLCA method [3] is also employed for pitch detection. Three monophonic excerpts were used for evaluation, where the proposed model is shown to outperform the shift-invariant PLCA model. Finally, a discussion is made on extending the proposed model for multiple pitch estimation using factorial HMMs.

The outline of the paper is as follows. Related work is presented in Section 2 and the proposed model is introduced in Section 3. Section 4 describes the pitch detection experiments that were performed. Finally, a discussion on extending the proposed model for multipitch estimation is made in Section 5 and conclusions are drawn in Section 6.

2. RELATED WORK

Related work to the proposed model is presented here. In [6], the probabilistic latent component analysis (PLCA) model is proposed, which is essentially a probabilistic version of the non-negative matrix factorization (NMF) method. In PLCA, the input spectrogram is considered to be a multivariate distribution $P(\omega, t)$, which can be expressed as a product of a spectral basis matrix $P(\omega|z)$ for each component z and a component gain matrix $P(z|t)$. For estimating these parameters, iterative update rules can be derived using the Expectation-Maximization (EM) algorithm [8]. An extension of the PLCA model was proposed in [9] for polyphonic music transcription, which supported multiple spectral templates for each pitch and multiple instruments.

In [3], a relative pitch tracking algorithm was proposed, which was based on a convolutive variant of PLCA (also called shift-invariant PLCA). The shift-invariant PLCA method can be used in conjunction with log-frequency spectrograms in order to extract pitch tracks. This is feasible since in log-frequency spectra the inter-harmonic spacings are the same for any periodic sounds. For a single source, the shift-invariant PLCA model is defined as:

$$P(\omega, t) = P(\omega) *_{\omega} P(f, t) \quad (1)$$

where a constant spectral template $P(\omega)$ is convolved with the pitch impulse distribution $P(f, t)$ over f in order to approximate the input spectrogram. The shift invariant PLCA model was also formulated for multiple instrument sources in [3], where a spectral template corresponds to each source. In addition, an extension of the shift-invariant PLCA model was proposed by the authors in [10]

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

for multiple pitch estimation, where the proposed method supports multiple instrument and pitch templates.

An algorithm called the non-negative hidden Markov model (N-HMM) was proposed in [5], which is able to combine the PLCA algorithm with HMMs, in order to model the pitch changes in a monophonic recording. Each hidden state corresponds to a single pitch component, and multiple templates per pitch are supported. Parameter estimation can be achieved using the EM algorithm, by combining the PLCA update steps with the HMM forward-backward procedure [7]. An extension for multiple sources was also proposed, which employed factorial HMMs.

Finally in [4], an extension of the NMF method with Markov-chained constraints is proposed for music spectrogram modeling. The non-stationarity of music sounds is addressed by learning the time-varying spectral patterns of musical instruments. Parameter estimation is achieved using the NMF update rules combined with the Viterbi algorithm (HMM transition probabilities are fixed).

3. PROPOSED MODEL

3.1. Motivation

The goal of the proposed model is to provide a framework for music signal analysis, where the produced notes can be represented as time-varying spectral templates that can be also shifted across frequency, in order to account for frequency modulations. This will allow for a much more accurate representation of the input spectrogram, and will attempt to address the drawbacks of current pitch estimation systems. In contrast to the Markov-chained approaches in [4, 5], the goal is to also exploit the benefits given by shift-invariance in the log-frequency spectrum for pitch detection.

3.2. Formulation

The proposed algorithm can be named as HMM-constrained shift-invariant PLCA. We approximate the input log-frequency spectrum $V_{\omega,t}$ (where ω is the log-frequency index and t the time index) as a multivariate probability distribution $P(\omega, t)$, which can be decomposed using a succession of spectral templates corresponding to each sound state q that can also be shifted across log-frequency. The model can be formulated as:

$$P(\omega, t) = P(t) \sum_{q_t} P_t(q_t|\bar{\omega}) P(\omega|q_t) *_\omega P_t(f|q_t) \quad (2)$$

where $P(\omega|q)$ is the spectral template for state q, $P(t)$ is the energy of each time frame, $P_t(q_t|\bar{\omega})$ is the contribution of each state at the current time frame ($\bar{\omega}$ represents all observed spectra), and $P_t(f|q_t)$ is the pitch impulse distribution for each state across time.

Since the succession of the sound states $P_t(q_t|\bar{\omega})$ is temporally-constrained, the corresponding HMM in terms of all observations $\bar{\omega}$ is:

$$P(\bar{\omega}) = \sum_{\bar{q}} \sum_{\bar{f}} P(q_1) \prod_t P(q_{t+1}|q_t) \prod_t P_t(\omega_t|q_t) \quad (3)$$

where $P(q_1)$ is the state prior distribution, $P(q_{t+1}|q_t)$ is the transition probability, and $P_t(\omega_t|q_t)$ is the time-dependent observation probability given a current state. It should be noted that ω_t corresponds to the observed spectrum at time t. Here, we define the observation probability as:

$$P_t(\omega_t|q_t) = 1 - \frac{||P(\omega, t|q_t) - V_{\omega,t}||_2}{\sum_{q_t} ||P(\omega, t|q_t) - V_{\omega,t}||_2} \quad (4)$$

where $||\cdot||_2$ is the l^2 norm and the spectrogram that corresponds to state q is given by:

$$P(\omega, t|q_t) = P(t) P_t(q_t|\bar{\omega}) P(\omega|q_t) *_\omega P_t(f|q_t) \quad (5)$$

Using (4), the state spectrogram that better approximates the input spectrogram using the Euclidean distance has a greater observation probability.

3.3. Parameter Estimation

The aforementioned parameters can be estimated by maximizing the log-likelihood of the data, using the EM algorithm [8]. Essentially, the update equations for each iteration are a combination of the shift-invariant PLCA rules [3] and the HMM forward-backward procedure [7].

For the Expectation step, the update equations are:

$$P_t(f, q_t|\bar{\omega}) = P_t(q_t|\bar{\omega}) P_t(f|\omega, q_t) \quad (6)$$

where

$$P_t(f|\omega, q_t) = \frac{P(\omega - f|q_t) P_t(f|q_t)}{\sum_f P(\omega - f|q_t) P_t(f|q_t)} \quad (7)$$

$$P_t(q_t|\bar{\omega}) = \frac{\alpha_t(q_t)\beta_t(q_t)}{\sum_{q_t} \alpha_t(q_t)\beta_t(q_t)} \quad (8)$$

Equation (6) is the model posterior, being the probability of the hidden variables given the observations. In (8), $\alpha_t(q_t)$ and $\beta_t(q_t)$ are the HMM forward and backward variables, respectively. The variables can be computed recursively by employing (4), using the forward/backward procedure described in [7]. Finally, the marginalized posterior for the transition matrix is given by:

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

For the Maximization step, the update equations are:

$$P(\omega|q) = \frac{\sum_{f,t} V_{\omega+f,t} P_t(f, q_t|\omega + f)}{\sum_{\omega,f,t} V_{\omega+f,t} P_t(f, q_t|\omega + f)} \quad (10)$$

$$P_t(f|q_t) = \frac{\sum_\omega V_{\omega,t} P_t(f, q_t|\omega)}{\sum_{f,\omega} V_{\omega,t} P_t(f, q_t|\omega)} \quad (11)$$

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

Finally, the state priors are computed using (8): $P(q_1) = P_1(q_1|\bar{\omega})$.

An example of the proposed model is given in Fig. 1, where the 10-cent resolution log-frequency spectrogram of a C4 piano note is used as input. The HMM topology for this example is a 4-state left-to-right model, and the pitch shifting span is one octave (thus, the length of f is 120). It can be seen from Fig. 1(d) that there is a clear succession of spectral templates across time.

4. EXPERIMENTS

The model presented in Section 3 was applied in a supervised manner for pitch detection in monophonic audio excerpts in order to demonstrate its superiority compared with standard approaches which do not take into account the temporal evolution of musical sounds. Three excerpts were utilized: a piano melody from

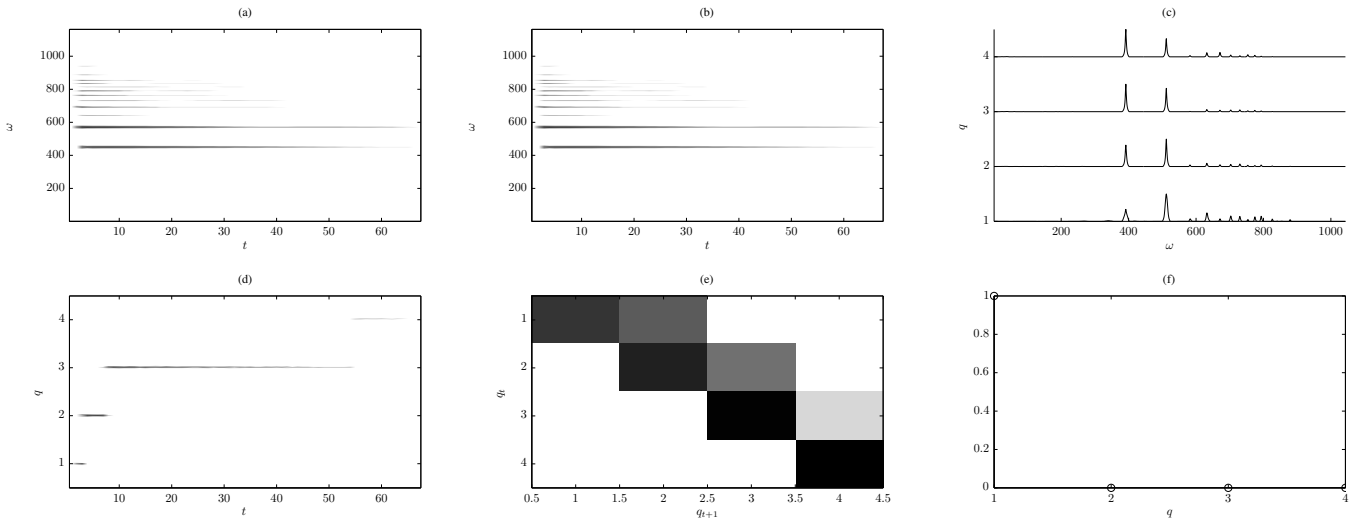

Figure 1: (a) Spectrogram $V_{\omega,t}$ of a C4 piano note (b) Approximation $P(\omega, t)$ of the spectrogram using estimated parameters (c) Spectral templates $P(\omega|q)$ (d) Stacked pitch distributions $P_t(f|q_t)$ (e) Sound state transition matrix $P(q_{t+1}|q_t)$ (f) Sound state priors $P(q_1)$

the beginning of J.S. Bach's Chromatic Fugue synthesized using the Native Instruments soundfonts[1], a cello melody from the RWC database [11] (RWC-MDB-C-2001 No. 12), and an oboe melody from the MIREX multi-F0 development set[2].

Spectral templates were extracted for the three aforementioned instruments, using samples for note C4 from the RWC Musical Instrument Sound database [11]. The time-frequency representation that was employed for analysis was the resonator time-frequency image (RTFI), which is a first-order complex resonator filter bank, having been used in the past for transcription experiments [12]. The reason the RTFI was selected instead of the more common constant-Q transform (CQT) is because it provides a more accurate temporal resolution in lower frequencies, which is attributed to the use of an exponential decay factor in the filterbank analysis. Here, a constant-Q RTFI with 120 bins per octave was selected, with a frequency range from 27.5 Hz (A0) to 12.5 kHz. For extracting the templates, the model in (2) was employed, using left-to-right HMMs with $Q = 4$ hidden sound states.

For the pitch detection experiments, the update rules in (6) - (12) were used, excluding the update rule for the spectral templates in (10), since the patterns for each sound state were considered fixed. The detected pitch for the recordings is summed from the pitch distribution for each sound state:

$$P(f,t) = P(t) \sum_{q_t} P_t(q_t|\bar{\omega}) P_t(f|q_t) \tag{13}$$

Using $P(f, t)$, a piano-roll representation was created by summing every 10 pitch bins (which make for one semitone). The output piano-roll representation was compared against existing MIDI ground truth for the employed recordings. In Fig. 2, an excerpt of the employed piano melody can be seen along with the weighted sound state transitions using the employed model with a left-to-right HMM. For each produced note, the transition from the attack state to two sustain states, followed by a brief decay state can clearly be seen. For evaluation, the transcription metrics also used in [10] were utilized, namely the overall accuracy (Acc), the total error (E_{tot}), the substitution error (E_{subs}), missed detection error (E_{fn}), and false alarm error (E_{fp}). It should also be noted that all evaluations take place by comparing the transcribed pitch output and the ground-truth MIDI files at a 10 ms scale. For comparative purposes, the shift-invariant PLCA method in [3] was also employed for transcription. In this case, one spectral template per source is employed, using the same training data as in the proposed method.

Pitch detection results using the proposed model are displayed for each recording in Table 1. Experiments using the proposed method were performed using left-to-right and ergodic HMMs (where all possible transitions between states were allowed). Although the use of an ergodic model might not be ideal in cases where the sound evolves clearly between the attack, transient, sustain, and decay states, it might be useful for instruments where different sustain states alternate (e.g. tremolo). It can be seen that in all cases, the proposed HMM-constrained shift-invariant PLCA methods outperform the shift-invariant PLCA method in terms of overall transcription accuracy. Also, the accuracy is relatively high for the piano and cello recordings, but significantly lower for the oboe recording. This can be attributed to the fact that the spectral pattern of oboe notes is not constant for all pitches, but in fact changes drastically. Most of the missed detections are located in the decay states of produced notes, whereas most false alarms are octave errors occurring in the attack part of notes. Finally, when comparing the HMM topologies, it can be seen that the ergodic model slightly outperforms the left-to-right one.

5. MODEL EXTENSIONS

The proposed model is a first attempt in introducing temporal constraints in a convolutive model. By using only one time-varying set of pitch templates per source with log-frequency shifting, transcription errors may occur. One solution would be to employ one set of spectral templates per pitch for each instrument source (as an extension to the works in [9, 10]), which would allow for a much more informative decomposition.

[1] Available at: http://www.eecs.qmul.ac.uk/~emmanouilb/WASPAA.html
[2] http://www.music-ir.org/mirex/wiki/MIREX_HOME

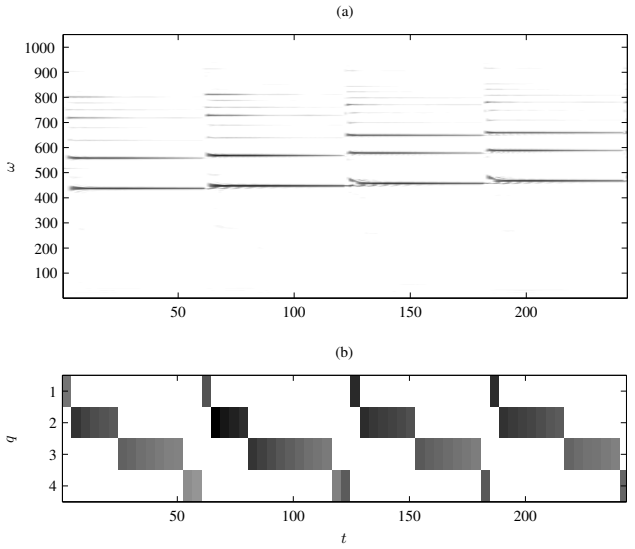

Figure 2: (a) Log-frequency spectrogram of a segment of the piano melody employed for experiments (b) Weighted state transitions $P(q_t, t|\bar{\omega}) = P(t)P_t(q_t|\bar{\omega})$.

Method	Instrument	Acc	E_{tot}	E_{subs}	E_{fn}	E_{fp}
Proposed (LtR)	Piano	81.5%	17.8%	2.2%	9.8%	5.8%
	Cello	80.3%	22.1%	8.3%	5.6%	15.7%
	Oboe	55.0%	39.1%	13.3%	22.6%	3.2%
Proposed (ergodic)	Piano	82.2%	16.9%	2.2%	9.5%	5.2%
	Cello	80.5%	22.2%	5.6%	5.4%	16.2%
	Oboe	55.6%	37.5%	14.9%	19.3%	3.2%
SIPLCA	Piano	80.1%	20.2%	1.6%	10.7%	7.9%
	Cello	75.0%	28.5%	1.2%	9.2%	18.0%
	Oboe	54.1%	41.9%	13.7%	20.5%	7.7%

Table 1: Pitch detection results using the proposed method with left-to-right and ergodic HMMs, compared with the shift-invariant PLCA method.

For multipitch estimation, the proposed model can be extended by utilizing multiple HMMs (one per pitch). Either independent HMMs could be employed, or factorial HMMs as in the non-negative HMM formulation in [5], which however would lead to greater computational complexity. For transcription of polyphonic music, the set of pitch templates could be shifted in a semitone span as in [10], which would allow the creation of a pitch spectrogram. The final goal would be a system which is able to exploit information from multiple instrument sources, multiple pitches, and multiple sound states per pitch, which could allow for a rich representation of the evolution of sound in polyphonic music. In order to further constrain the model, sparseness could also be enforced in the pitch impulse distribution as in [3], or in the source contribution as in [9, 10].

6. CONCLUSIONS

In this proof-of-concept work, we proposed an HMM-constrained convolutive model for pitch detection. The goal was to model the temporal evolution of each produced note and utilize the extra information for reducing pitch detection errors. A supervised variant of the proposed model was utilized for pitch detection on monophonic excerpts from a piano, cello, and oboe and was compared against a convolutive probabilistic model. Results showed that the proposed model can capture the temporal evolution of musical sounds and outperforms the single pitch template approach. Finally, future directions on extending the proposed model for temporally-constrained multipitch estimation are given.

7. ACKNOWLEDGMENT

The authors would like to thank Anssi Klapuri for his valuable feedback on this work.

8. REFERENCES

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

[2] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection of music signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 13, no. 5, pp. 1035–1047, Sept. 2005.

[3] P. Smaragdis, "Relative-pitch tracking of multiple arbitary sounds," *J. Acoustical Society of America*, vol. 125, no. 5, pp. 3406–3413, May 2009.

[4] M. Nakano, J. L. Roux, H. Kameoka, Y. Kitano, N. Ono, and S. Sagayama, "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms," in *Int. Conf. Latent Variable Analysis and Signal Separation*, Sept. 2010, pp. 149–156.

[5] G. Mysore, "A non-negative framework for joint modeling of spectral structure and temporal dynamics in sound mixtures," Ph.D. dissertation, Stanford University, USA, June 2010.

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

[7] 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.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

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

[10] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a convolutive probabilistic model," in *8th Sound and Music Computing Conf.*, July 2011, pp. 19–24.

[11] 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.

[12] R. Zhou, "Feature extraction of musical content for automatic music transcription," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Oct. 2006.