An Efficient Shift-Invariant Model for Polyphonic Music Transcription

Emmanouil Benetos, Srikanth Cherla, and Tillman Weyde

Department of Computer Science, City University London, UK
{Emmanouil.Benetos.1,Srikanth.Cherla.1,t.e.weyde}@city.ac.uk

Abstract. In this paper, we propose an efficient model for automatic transcription of polyphonic music. The model extends the shift-invariant probabilistic latent component analysis method and uses pre-extracted and pre-shifted note templates from multiple instruments. Thus, the proposed system can efficiently transcribe polyphonic music, while taking into account tuning deviations and frequency modulations. Additional system improvements utilising massive parallel computations with GPUs result in a system performing much faster than real-time. Experimental results using several datasets show that the proposed system can successfully transcribe polyphonic music, outperforming several state-of-the-art approaches, and is over 140 times faster compared to a standard shift-invariant transcription model.

Keywords: Automatic music transcription; probabilistic latent component analysis; shift-invariance; GPU computing

1 Introduction

Automatic music transcription (AMT) is the process of converting an acoustic musical signal into some form of music notation [1]. Even though the problem of transcribing monophonic music is considered solved, multiple-instrument polyphonic transcription still remains an open problem. A large subset of current AMT systems use spectrogram factorization techniques such as non-negative matrix factorization (NMF) and probabilistic latent component analysis (PLCA) [2]. Convolutive extensions of such models like non-negative matrix deconvolution (NMD) and shift-invariant probabilistic latent component analysis (SI-PLCA) are able to take into account the constant inter-harmonic spacings in log-frequency representations in order to detect small pitch changes or frequency modulations [3] [4]. However, the convolutions in the aforementioned models make them computationally expensive.

In this work, we propose a spectrogram factorization-based model for AMT which uses pre-shifted note templates across log-frequency. This results in a linear model that is still able to support pitch deviations and frequency modulations. Additionally, the linear operations in the model can also benefit from GPU computing, which results in an AMT system which is able to run faster than real-time. Experiments show that the proposed model is equivalent to the shift-invariant model in [4], while being over 140 times faster.

2 Related Work

The SI-PLCA-based model of [4] takes as input a normalized log-frequency spectrogram $V_{\omega,t}$ and approximates it as a bivariate probability distribution $P(\omega, t)$ (where ω denotes frequency and t time). $P(\omega, t)$ is decomposed in the model as:

$$P(\omega, t) = P(t) \sum_{p,f,s} P(\omega - f|s, p)P_t(f|p)P_t(s|p)P_t(p) \tag{1}$$

where p, f, s denote pitch, log-frequency shifting, and instrument source, respectively. $P(t)$ is the spectrogram energy (known quantity), $P(\omega - f|s, p)$ are the pre-extracted spectral templates for pitch p and instrument s, shifted across log-frequency f, $P_t(f|p)$ is the time-varying log-frequency shifting for pitch p, $P_t(s|p)$ is the source contribution, and $P_t(p)$ is the pitch activation (i.e. the transcription). Parameter f is constrained to a semitone range. Unknown parameters can be iteratively estimated using the expectation-maximization (EM) algorithm.

3 Proposed Model

The motivation behind the proposed model is to create an efficient version of the model in [4] which would still incorporate shift-invariance. To that end, we create a model which uses pre-extracted note templates, which are also pre-shifted across log-frequency, thus avoiding the need for convolutions during parameter estimation. Thus the proposed model can be formulated as:

$$V_{\omega,t} \approx P(\omega, t) = P(t) \sum_{p,f,s} P(\omega|s, p, f)P_t(f|p)P_t(s|p)P_t(p) \tag{2}$$

where $P(\omega|s, p, f)$ are spectral templates for pitch p, instrument s, which are shifted according to parameter f. As a log-frequency representation we use the constant-Q transform [5] with 60 bins/octave, resulting in $f \in [1, \ldots, 5]$, where $f = 3$ is the ideal tuning position for the template (using equal temperament).

The EM formulations using the proposed model are as follows:

$$P_t(p, f, s|\omega) = \frac{P(\omega|s, p, f)P_t(f|p)P_t(s|p)P_t(p)}{\sum_{p,f,s} P(\omega|s, p, f)P_t(f|p)P_t(s|p)P_t(p)} \tag{3}$$

$$P_t(f|p) = \frac{\sum_{\omega,s} P_t(p, f, s|\omega)V_{\omega,t}}{\sum_{f,\omega,s} P_t(p, f, s|\omega)V_{\omega,t}} \tag{4}$$

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

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

Eqs. (3)-(6) are iterated until convergence; typically 15-20 iterations are sufficient. No update rule for the templates $P(\omega|s, p, f)$ is included, since they are

Method	[4]	Proposed	Proposed (GPU)
Runtime	2740sec	68sec	19sec

Table 1. Transcription runtimes for the MIREX multiF0 recording (duration: 55sec).

considered fixed in the model. As in [4], we also incorporated sparsity constraints on $P_t(p)$ and $P_t(s|p)$ in order to control the polyphony level and the instrument contribution in the resulting transcription (formulations are omitted due to lack of space). The resulting transcription is given by $P(p,t) = P(t)P_t(p)$, while a high-resolution time-pitch representation $P(f',t)$ can also be derived from the model, as in [4]. In order to extract note events, thresholding is performed on $P(p,t)$ followed by minimum note duration pruning set to 50ms.

The proposed model is already quite efficient, since linear operations (like tensor multiplications) can be used in the EM steps for estimating the unknown parameters, without needing any convolutions which considerably slow down SI-PLCA models. Such operations can also benefit from parallel computations; to that end, we exploit the massive parallelism offered on NVIDIA graphics processing units (GPUs), using the CUDA parallel computing framework which is supported in recent versions of Matlab[1]. To the authors' knowledge, this is the first AMT system which uses parallel computations. The Matlab code for the model, both in the normal and GPU-supported versions, can be found online[2].

4 Evaluation

Pre-extracted and pre-shifted spectral templates are extracted for bassoon, cello, clarinet, flute, guitar, harpsichord, horn, oboe, piano, tenor sax, and violin using the RWC database [6]. For testing, we used thirty 30sec piano segments from the MAPS-ENSTDkCl dataset [7], the MIREX multiF0 woodwind quintet [8] and the 5 complete recordings from the TRIOS dataset [9]. For evaluating the proposed system's performance we use the frame-based and onset-only note-based F-measure (F_f and F_n, respectively) [4].

Regarding runtimes, the original SI-PLCA-based model of [4] performed at about 50×real-time using a Sony VAIO S15 laptop. Using the proposed model, the runtime is close to 1×real-time, while the proposed model using GPU computing takes 0.3×real-time. As an example, the runtimes for the MIREX recording can be seen in Table 1.

In Table 2, averaged transcription results for the proposed model are shown. It should be noted that the performance of the shift-invariant system of [4] is roughly the same: the F_n is +0.9% for the MAPS database, -0.9% for the MIREX recording, and -0.1% for the TRIOS dataset. Performance comparisons for the MIREX recording can be seen in [4], where both systems outperform other state-of-the-art systems. For the MAPS recordings, the proposed model outperforms

[1] `http://www.mathworks.co.uk/discovery/matlab-gpu.html`
[2] `https://code.soundsoftware.ac.uk/projects/amt_mssiplca_fast`

Dataset	F_f	F_n
MAPS-ENSTDkCl	64.17%	63.14%
MIREX	67.19%	66.01%
TRIOS	66.46%	56.29%

Table 2. Transcription results using the proposed model.

the NMF-based method of [10] by 11% in terms of F_f. To the authors' knowledge, no transcription results have been reported for the TRIOS dataset.

5 Discussion

In this paper, we presented an efficient probabilistic model for AMT which relies on pre-shifted spectral templates. Using GPU computing, the system is able to run faster than real-time, while its performance outperforms several state-of-the-art systems. In the future, temporally-constrained spectrogram factorization models will be adapted in a similar way for more efficient performance.

Acknowledgments. Emmanouil Benetos is supported by a City University London Research Fellowship. The authors wish to thank Anssi Klapuri and Holger Kirchhoff for useful discussions on GPU computing technology.

References

1. Klapuri, A., Davy, M., eds.: Signal Processing Methods for Music Transcription. Springer-Verlag, New York (2006)
2. Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., Klapuri, A.: Automatic music transcription: breaking the glass ceiling. In: ISMIR. (October 2012) 379–384
3. Fuentes, B., Badeau, R., Richard, G.: Adaptive harmonic time-frequency decomposition of audio using shift-invariant PLCA. In: ICASSP. (May 2011) 401–404
4. Benetos, E., Dixon, S.: A shift-invariant latent variable model for automatic music transcription. Computer Music Journal **36**(4) (Winter 2012) 81–94
5. Schörkhuber, C., Klapuri, A.: Constant-Q transform toolbox for music processing. In: SMC. (July 2010)
6. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: RWC music database: music genre database and musical instrument sound database. In: ISMIR. (October 2003)
7. Emiya, V., Badeau, R., David, B.: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. IEEE Transactions on Audio, Speech, and Language Processing **18**(6) (August 2010) 1643–1654
8. MIREX: Music Information Retrieval Evaluation eXchange. `http://music-ir.org/mirexwiki/`
9. Fritsch, J.: High quality musical audio source separation. Master's thesis, UPMC / IRCAM / Telécom ParisTech (2012)
10. Carabias-Orti, J.J., Virtanen, T., Vera-Candeas, P., Ruiz-Reyes, N., Cañadas-Quesada, F.J.: Musical instrument sound multi-excitation model for non-negative spectrogram factorization. IEEE Journal of Selected Topics in Signal Processing **5**(6) (October 2011) 1144–1158