MIREX SUBMISSION: A DETERMINISTIC ANNEALING EM ALGORITHM FOR AUTOMATIC MUSIC TRANSCRIPTION

Tian Cheng, Simon Dixon and Matthias Mauch
Centre for Digital Music, Queen Mary University of London
{tian.cheng, simon.dixon, matthias.mauch} @eecs.qmul.ac.uk

ABSTRACT

In the past decade, non-negative matrix factorisation (NMF) and probabilistic latent component analysis (PLCA) have been used widely in automatic music transcription. Despite their successes, these methods only guarantee that the decomposition converges to a local minimum in the cost function. In order to find better local minima, we propose to extend an existing PLCA-based transcription method with the deterministic annealing EM (DAEM) algorithm. The PLCA update rules are modified by introducing a "temperature" parameter. At higher temperatures, general areas of the search space containing good solutions are found. As the temperature is gradually decreased, distinctions in the data are sharpened, resulting in a more fine-grained optimisation at each successive temperature. This process reduces the dependence on the initialisation, which is otherwise a limitation of NMF and PLCA approaches. There are three variants of the system submitted, two for multiple-F0 estimation task trained on different sets of instrument templates (with and without piano), and one trained on orchestral instruments and piano for note tracking task.

1. INTRODUCTION

Automatic music transcription is the process of transcribing audio into a symbolic music representation. To date, non-negative matrix factorisation (NMF) [7] and its probabilistic counterpart, probabilistic latent component analysis (PLCA) [9], have been used extensively for this task. These methods treat the spectrogram as a matrix, and decompose it into spectral bases, gain functions, and instrument distributions (when considering different instruments).

One obvious problem of non-negative matrix decomposition methods (such as NMF and PLCA) is that they are initialisation-sensitive and tend to converge to a local minimum. Training instrument templates is an effective way to initialise the spectral bases. By fixing the templates during the updating, we obtain a stable output for the gain function, independent of its initialisation. But when the model becomes more complicated, as by introducing an instrument variable into the model, which is used widely nowadays, it is not possible for us to find good initialisations for all variables.

In this paper, we tackle the local minimum problem by introducing an optimisation method. When using non-negative matrix decomposition methods, the transcription result is related to the cost function, the update rules and also the constraints. Here, We particularly focus on PLCA, which utilises the Kullback-Leibler (KL) divergence as the cost function and derives the update rules based on the EM algorithm [8]. To address the local minimum problem of the EM algorithm, we make use of the deterministic annealing EM algorithm [10] by introducing a temperature parameter into an existing PLCA-based model [2]. There are three variants of the system submitted, two for multiple-F0 estimation task trained on different sets of instrument templates (with and without piano), and one for note tracking task. Three variants of system are submitted for the MIREX 2013 evaluation, two for multiple-F0 estimation task trained on orchestral instruments only (CDM1) and on orchestral instruments and piano (CDM2), and one for note tracking task (CDM3) trained on orchestral instruments and piano. It should be noted that this paper is a shorter version of the ISMIR paper. If you'd like to cite this work, please use the ISMIR paper.

2. PROPOSED METHOD

To deal with the local minimum problem of PLCA models, we derive the update rules according to the deterministic annealing EM algorithm [10], which introduces a temperature parameter into the EM algorithm. The temperature parameter is employed on the posterior probability density in the E-step. Then by gradually reducing the temperature, the EM steps are iteratively executed until convergence at each temperature, leading the result to a global or better local minimum. We apply this method to a baseline PLCA-based model proposed in [2]. Since the templates are kept fixed, the temperature parameter is applied to the posterior probability density of the instrument distribution. In this way, we can enjoy the benefits of the DAEM algorithm and the templates.

2.1 The Baseline PLCA Model

Benetos and Dixon [2] proposed a model that adds an instrument distribution variable to shift-invariant PLCA. The

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2013 International Society for Music Information Retrieval.

	instrument	lowest note	highest note
1	Bassoon	34	72
2	Cello	26	81
3	Clarinet	50	89
4	Flute	60	96
5	Guitar	40	76
6	Horn	41	77
7	Oboe	58	91
8	Piano	21	108
9	Tenor Sax	44	75
10	Violin	55	100

Table 1: Instrument ranges, adapted from [1]

time-frequency representation of the input signal was computed with the Constant-Q Transform [6] using 120 bins per octave. Templates were trained for 10 instruments allowing shifts within a semitone range, in order to deal with arbitrary tuning and frequency modulation. The model is formulated as:

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

where $P(\omega,t)$ is the approximated spectrogram, $P(t)$ is the energy distribution of spectrogram. $P(\omega|s,p)$ are the templates of instrument s and pitch p, $P(f|p,t)$ is the shifted variant for each p, $P(s|p,t)$ is the instrument contribution for each pitch, and $P(p|t)$ is the pitch probability distribution for each time frame. The templates $P(\omega|s,p)$ are trained using the MAPS dataset [3] and RWC dataset [4].

The update rules are derived from the EM algorithm. For the E-step, the posterior probability density is:

$$P(p,f,s|\omega,t) = \frac{P(\omega-f|s,p)P(f|p,t)P(s|p,t)P(p|t)}{\sum_{p,f,s} P(\omega-f|s,p)P(f|p,t)P(s|p,t)P(p|t)} \tag{2}$$

For the M-step, each parameter is estimated.

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

$$P(s|p,t) = \frac{(\sum_{\omega,f} P(p,f,s|\omega,t)P(\omega,t))^{\alpha_1}}{\sum_s(\sum_{\omega,f} P(p,f,s|\omega,t)P(\omega,t))^{\alpha_1}} \tag{4}$$

$$P(p|t) = \frac{(\sum_{\omega,f,s} P(p,f,s|\omega,t)P(\omega,t))^{\alpha_2}}{\sum_p(\sum_{\omega,f,s} P(p,f,s|\omega,t)P(\omega,t))^{\alpha_2}} \tag{5}$$

The templates $P(\omega|s,p)$ are not updated as they are previously trained and kept fixed. The parameters α_1 and α_2 used in Eqn. (4) and (5) are used to enforce sparsity, where $\alpha_1, \alpha_2 > 1$. The final piano-roll matrix $P(p,t)$ and the pitches assigned to each instrument $P(p,t,s)$ are given by:

$$P(p,t) = P(p|t)P(t) \tag{6}$$

$$P(p,t,s) = P(s|p,t)P(p|t)P(t) \tag{7}$$

For post-processing, instead of using an HMM, the note events are extracted by performing thresholding on $P(p,t)$ and using minimum-length pruning (deleting notes shorter than $50ms$). The instrument-wise note events are detected in the same way using $P(p,t,s)$.

2.2 The DAEM-based Model

To modify the update rules according to the DAEM algorithm, in the E-step, the posterior probability density in Eqn. (2) is modified by introducing a temperature parameter τ [1]:

$$P_\tau(p,f,s|\omega,t) = \frac{(P(\omega-f|s,p)P(f|p,t)P(s|p,t)P(p|t))^{1/\tau}}{\sum_{p,f,s}(P(\omega-f|s,p)P(f|p,t)P(s|p,t)P(p|t))^{1/\tau}} \tag{8}$$

And the update rules are extended by adding a τ-loop:

1. Set $\tau \leftarrow \tau_{max}(\tau_{max} > 1)$.
2. Iterate the following EM-steps until convergence:
   E-step: calculate $P_\tau(p,f,s|\omega,t)$.

   M-step: estimate $P(f|p,t)$, $P(s|p,t)$ and $P(p|t)$ by replacing $P(p,f,s|\omega,t)$ with $P_\tau(p,f,s|\omega,t)$.
3. Decrease τ.
4. If $\tau \geq 1$, repeat from step 2; otherwise stop.

By gradually decreasing τ, the temperature is cooling down. At higher temperatures, the distributions are smoothed and general areas of the search space containing good solutions are found. As the temperature is gradually decreased, distinctions in the data are sharpened, resulting in a more fine-grained optimisation at each successive temperature.

Considering the properties of this particular model, we simplify the posterior probability density to:

$$P_\tau(p,f,s|\omega,t) = \frac{P(\omega-f|s,p)P(f|p,t)P(s|p,t)^{1/\tau}P(p|t)}{\sum_{p,f,s} P(\omega-f|s,p)P(f|p,t)P(s|p,t)^{1/\tau}P(p|t)} \tag{9}$$

The convolution of the templates and the pitch impulse distribution, giving the terms $P(\omega-f|s,p)P(f|p,t)$, works as the shift-invariant templates here. These are not modified by the temperature parameter, as the templates are fixed during the iterative process [2]. In addition, having observed that the pitch distribution $P(p|t)$ is dependent on the instrument distribution $P(s|p,t)$ in this model, we only need to modify $P(s|p,t)$ in the posterior probability density.

2.3 System Variants

There are three variants of system for the MIREX 2013, two for multiple-F0 estimation task (CDM1 and CDM2), and one for note tracking task (CDM3). As we try to find out whether templates trained on an extra instrument (piano) can improve the transcription result or not, we

[1] The parameter used in [10] is β, and the temperature is indicated by $1/\beta$. The reason for using τ here is because we want to indicate the temperature directly by τ and distinguish the proposed method from the β-divergence.

[2] This was also confirmed by test experiments where the power $1/\tau$ was also applied to the pitch impulse distribution $P(f|p,t)$, giving similar transcription results.

use two different sets of instrument templates for Task 1: one trained on orchestral instruments only (CDM1), the other one trained on orchestral instruments and piano (CDM2). Templates used for note tracking (CDM3) are trained on orchestral instruments and piano. In all case, we set $\alpha_1 = 1.3$ and $\alpha_2 = 1.1$; the parameter τ took the values $10/i,\ i \in \{8, 9, 10\}$. The number of iterations for each τ is set to 12.

3. REFERENCES

[1] Music Information Retrieval Evaluation eXchange (MIREX). `http://www.music-ir.org/mirex/wiki/MIREX_HOME`.

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

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

[4] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC Music Database: Music Genre Database and Musical Instrument Sound Database. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR03)*, pages 229–230, 2003.

[5] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562. MIT Press, 2001.

[6] C. Schoerkhuber and A. Klapuri. Constant-q transform toolbox for music processing. In *the 7th Sound and Music Computing Conference*, 2010.

[7] P. Smaragdis and J. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, 2003.

[8] P. Smaragdis and B. Raj. Shift-invariant probabilistic latent component analysis. Technical report, 2007.

[9] P. Smaragdis, B. Raj, and M. Shashanka. Sparse and shift-invariant feature extraction from non-negative data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP08)*, pages 2069–2072, Apr. 2008.

[10] N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11(2):271 – 282, 1998.