A Deterministic Annealing EM Algorithm for Automatic Music Transcription

Cheng, T; Dixon, S; Mauch, M

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. The method was tested on two standard multi-instrument transcription data sets (MIREX and Bach10). Experimental results show that the proposed method significantly outperforms a state-of-the-art reference method, according to both frame-based and note-based metrics. An additional analysis of instrument assignment results shows that instrument spectra are typically modelled as mixtures of templates from several instruments.

1. INTRODUCTION

Automatic music transcription is the process of transcribing audio into a symbolic music representation. To date, non-negative matrix factorisation (NMF) [15] and its probabilistic counterpart, probabilistic latent component analysis (PLCA) [17], 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). Although not yet providing the best transcription results, they provide a powerful mathematical model which can lead to a meaningful decomposition, using the constraints of non-negativity and sparsity. Another advantage of these methods is that they are easy to extend, by formulating a more complex model, adding variables or combining them with other models.

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 [16]. To address the local minimum problem of the EM algorithm, we make use of the deterministic annealing EM algorithm [18] by introducing a temperature parameter into an existing PLCA-based model [2]. The proposed method is tested on the Bach10 dataset [5] and the MIREX multi-F0 development dataset [1]. In comparison to the original PLCA-based model, the proposed method improves the results of multi-F0 estimation and note tracking, while the instrument assignment results vary for each individual instrument.

Although not much attention has been paid to the local minimum problem of automatic music transcription methods, there is still some related work. Bertin et al. [3] used a tempering scheme to favour the convergence of Itakura-Saito (IS) divergence to global minima. Experiments on music transcription show that IS-NMF can provide a good result by choosing a suitable temperature parameter. Hofmann [9] proposed a model based on the tempered EM algorithm to avoid overfitting in probabilistic latent semantic analysis. Kameoka et al. [11] introduced the DAEM algorithm into the harmonic-temporal-structured clustering (HTC) model for audio feature extraction. The HTC model is represented by a Gaussian kernel, and the DAEM algorithm is used to optimise the parameter convergence. Itaya et al. [10] used the DAEM algorithm to estimate the parameters of Gaussian mixture models (GMMs) and hidden

Markov models (HMMs). Experiments on speaker recognition and speech recognition show that DAEM is an effective method for GMM- and HMM-based acoustic modeling. Finally, Smaragdis et al. [16] stated that it is more likely to get "meaningful" decompositions and quick convergence by using "annealing" in PLCA.

The rest of this paper is organised as follows. In Section 2, we describe the PLCA model and the local minimum problem of this model. In Section 3, the update rules of a PLCA-based model are modified according to the DAEM algorithm. The results for three transcription subtasks are presented in Section 4. Finally discussion and conclusions are indicated in Section 5 and 6, respectively.

2. PLCA AND SHIFT-INVARIANT PLCA

Two basic PLCA models, PLCA and Shift-invariant PLCA, are presented in [17]. For automatic music transcription, the spectrogram is formulated by PLCA as:

$$V(\omega,t) \approx P(\omega,t) = P(t)\sum_{p} P(\omega|p)P(p|t) \qquad (1)$$

where $V(\omega,t)$ is the input spectrogram, $P(\omega,t)$ the approximated spectrogram, ω is the frequency bin, t the frame number. $P(t)$ is the energy of each time frame, $P(\omega|p)$ is the spectral bases corresponding to pitch p, and $P(p|t)$ the gain function.

To build a shift-invariant PLCA model, the spectrogram needs to be presented on a logarithmic frequency scale, such as the constant-Q transform. Assuming that the energy distributions of adjacent pitches are similar for any given instrument, the spectral basis can be shifted in frequency very easily, as the pattern of partial spacings is the same for all pitches, due to the logarithmic frequency axis. The spectrogram is formulated as:

$$\begin{aligned} V(\omega,t) \approx P(\omega,t) &= \sum_{z} P(z)P(\omega|z) *_{\omega} P(f,t|z) \\ &= \sum_{z} P(z)\sum_{f} P(\omega-f|z)P(f,t|z) \end{aligned} \qquad (2)$$

where $P(\omega|z)$ and $P(f,t|z)$ are the spectral templates and time-dependent shifted variant f of component z, and $P(z)$ is the prior distribution of the components.

In many recent systems the PLCA model is extended by introducing an instrument distribution, with templates trained per pitch per instrument. The spectrogram is formulated as:

$$V(\omega,t) \approx P(\omega,t) = P(t)\sum_{p,s} P(\omega|s,p)P(s|p,t)P(p|t) \qquad (3)$$

where $P(\omega|s,p)$ represents the spectral templates corresponding to each instrument s and pitch p, $P(s|p,t)$ the instrument contribution to each pitch in the t^{th} frame, and $P(p|t)$ the pitch probability distribution for each frame.

The parameters of the PLCA models are estimated by iteratively decreasing the KL divergence of the input spectrogram $V(\omega,t)$ and the synthetic spectrogram $P(\omega,t)$ using the EM algorithm. The KL divergence is convex in one variable, but not convex in multiple variables [12]. In this case, the EM algorithm can only guarantee to find a local minimum for these parameters, so the results depend on the initialisation. The use of instrument templates is an effective way to deal with the initialisation sensitivity of the algorithm. Taking the model described in Eqn. (1) for example, if the templates are fixed as a constant, the gain function will be convex. This means that when we formulate the model as the product of the spectral bases and a gain function, we obtain a unique gain function corresponding to a fixed set of templates. On the one hand, the templates lead to a stable decomposition for automatic music transcription; on the other hand, the templates also limit the performance of the transcription. However, when encountering the extended model as described in Eqn. (3), the instrument contribution and the pitch contribution still face the risk of converging to local minima, even with fixed templates.

3. PROPOSED METHOD

To deal with the local minimum problem of PLCA models, we derive the update rules according to the deterministic annealing EM algorithm [18], 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.

3.1 The Baseline PLCA Model

Benetos and Dixon [2] proposed a model that adds an instrument distribution variable to shift-invariant PLCA. The time-frequency representation of the input signal was computed with the Constant-Q Transform [14] 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:

$$P(\omega,t) = P(t)\sum_{p,s} P(\omega|s,p) *_{\omega} P(f|p,t)P(s|p,t)P(p|t) \qquad (4)$$

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 [6] and RWC dataset [7].

The update rules are derived from the EM algorithm.

	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]

For the E-step, the posterior probability density is:

$$P(p,f,s|\omega,t) = \frac{P(\omega-f|s,p)P(f|p,t)P(s|p,t)P(p|t)}{\sum_{p,f,s} P(\omega-f|s,p)P(f|p,t)P(s|p,t)P(p|t)} \quad (5)$$

For the M-step, each parameter is estimated.

$$P(f|p,t) = \frac{\sum_{\omega,s} P(p,f,s|\omega,t)P(\omega,t)}{\sum_{f,\omega,s} P(p,f,s|\omega,t)P(\omega,t)} \quad (6)$$

$$P(s|p,t) = \frac{(\sum_{\omega,f} P(p,f,s|\omega,t)P(\omega,t))^{\alpha_1}}{\sum_s(\sum_{\omega,f} P(p,f,s|\omega,t)P(\omega,t))^{\alpha_1}} \quad (7)$$

$$P(p|t) = \frac{(\sum_{\omega,f,s} P(p,f,s|\omega,t)P(\omega,t))^{\alpha_2}}{\sum_p(\sum_{\omega,f,s} P(p,f,s|\omega,t)P(\omega,t))^{\alpha_2}} \quad (8)$$

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. (7) and (8) are used to enforce sparsity, where $\alpha_1, \alpha_2 > 1$. We set $\alpha_1 = 1.3$ and $\alpha_2 = 1.1$. The final piano-roll matrix $P(p,t)$ and the pitches assigned to each instrument $P(p,t,s)$ are given by:

$$P(p,t) = P(p|t)P(t) \quad (9)$$

$$P(p,t,s) = P(s|p,t)P(p|t)P(t) \quad (10)$$

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 $P(p,t,s)$.

3.2 The DAEM-based Model

To modify the update rules according to the DAEM algorithm, in the E-step, the posterior probability density in Eqn. (5) is modified by introducing a temperature parameter τ [1]:

$$P_\tau(p,f,s|\omega,t) = \frac{(P(\omega-f|s,p)P(f|p,t)P(s|p,t)P(p|t))^{1/\tau}}{\sum_{p,f,s}(P(\omega-f|s,p)P(f|p,t)P(s|p,t)P(p|t))^{1/\tau}} \quad (11)$$

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

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

  M-step: estimated $P(f|p,t)$, $P(s|p,t)$ and $P(p|t)$ by replacing $P(p,f,s|\omega,t)$ with $P_\tau(p,f,s|\omega,t)$.
- Decrease τ.
- 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:

$$P_\tau(p,f,s|\omega,t) = \frac{P(\omega-f|s,p)P(f|p,t)P(s|p,t)^{1/\tau}P(p|t)}{\sum_{p,f,s} P(\omega-f|s,p)P(f|p,t)P(s|p,t)^{1/\tau}P(p|t)} \quad (12)$$

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.

In the experiment, the parameter τ took the values $10/i,\ i \in \{8,9,10\}$. When τ finally decreases to 1, the update rules agree with the original ones.

4. EVALUATION

4.1 Datasets

We used the Bach10 Dataset [5] and the MIREX Multi-F0 Development Dataset (MIREX dataset) [1] to test the performance of the proposed method. The Bach10 dataset consists of 10 quartet recordings performed on violin, clarinet, saxophone and bassoon. The MIREX dataset is an excerpt from a woodwind quintet recording, played on bassoon, clarinet, flute, horn, oboe.

4.2 Evaluation Metrics

The performance of the proposed system is evaluated on three subtasks of automatic music transcription. The first two, multiple F0 estimation and note tracking, are very commonly used. The third subtask, instrument assignment, evaluates the algorithms' ability to assign the notes to corresponding instruments.

[1] The parameter used in [18] is β, and the temperature is indicated by $1/\beta$. The reason of 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.

Dataset	Methods	P	R	F	Acc	$\boldsymbol{E_{tot}}$	E_{subs}	E_{miss}	E_{fa}
Bach10	BD(2012)	0.784	0.791	0.787	0.650	0.311	0.116	0.093	0.102
	Proposed	0.819	0.796	0.807	**0.677**	**0.282**	0.098	0.106	0.078
MIREX	BD(2012)	0.748	0.537	0.625	0.455	0.486	0.158	0.305	0.023
	Proposed	0.769	0.561	0.649	**0.480**	**0.461**	0.146	0.292	0.023
Both	BD(2012)	0.781	0.768	0.772	0.632	0.327	0.120	0.112	0.094
	Proposed	0.814	0.775	0.793	**0.659**	**0.299**	0.102	0.123	0.074

Table 2: Multiple F0 estimation results (see text for explanation of symbols).

In the multiple F0 estimation subtask, performance is evaluated frame by frame with an interval of $10ms$. The accuracy metrics are precision (P), recall (R), F-measure (F) [19] and the overall accuracy (*Acc*) [4], defined as follows:

$$P = \frac{N_{tp}}{N_{sys}},\ R = \frac{N_{tp}}{N_{ref}},\ F = \frac{2 \cdot R \cdot P}{R + P} \qquad (13)$$

$$Acc = \frac{N_{tp}}{N_{tp} + N_{fp} + N_{fn}} \qquad (14)$$

where N_{tp} is the number of true positives, N_{sys} and N_{ref} denote the number of the detected pitches and the ground-truth pitches, N_{fp} and N_{fn} are the number of false positives and false negatives respectively. The error metrics are the rates of total error (E_{tot}), substitution error (E_{subs}), missed detections (E_{miss}) and false alarms (E_{fa}). See the definitions in [13].

For the note tracking task, a note is considered correctly detected if the note is within the following ranges of ground truth.

pitch range $\pm 3\%$
onset range $\pm 50ms$
offset range $\pm$ max $\{20\%$ of the duration, $50ms\}$

The algorithms are evaluated in terms of onset-only and onset-offset accuracies, which are denoted by P_{on}, R_{on}, F_{on}, Acc_{on} and P_{off}, R_{off}, F_{off}, Acc_{off} respectively.

The instrument assignment task assesses whether the transcription not only identifies the correct pitch, but also the correct instrument. First, pitches are detected for each individual instrument. Then instruments actually occurring in the piece are evaluated according to the frame-based F-measure (13), whereas for the other instruments we calculate the false positive rate.

4.3 Results

We compare the performance of the proposed method to that of the baseline PLCA model introduced in Section 3.1 (mentioned as BD(2012) below). Here, we provide results for three subtasks on two different datasets.

4.3.1 Multiple F0 Estimation

Dataset	Methods	P_{on}	R_{on}	$\boldsymbol{F_{on}}$	Acc_{on}
Bach10	BD(2012)	0.319	0.339	0.328	0.197
	Proposed	0.399	0.354	**0.374**	0.231
MIREX	BD(2012)	0.628	0.420	0.503	0.336
	Proposed	0.690	0.459	**0.551**	0.380
Both	BD(2012)	0.347	0.346	0.344	0.209
	Proposed	0.427	0.364	**0.391**	0.245

(a) onset-only accuracy

Dataset	Methods	P_{off}	R_{off}	$\boldsymbol{F_{off}}$	Acc_{off}
Bach10	BD(2012)	0.217	0.230	0.223	0.126
	Proposed	0.281	0.249	**0.263**	0.152
MIREX	BD(2012)	0.487	0.326	0.391	0.243
	Proposed	0.537	0.357	**0.429**	0.273
Both	BD(2012)	0.242	0.239	0.238	0.137
	Proposed	0.305	0.259	**0.279**	0.163

(b) onset and offset

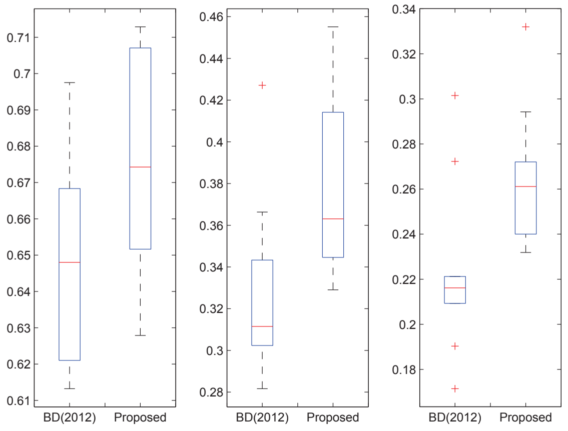
Table 3: Note-tracking results

The results for multiple F0 estimation using the Bach10 and MIREX datasets for two methods are shown in Table 2. It can be seen that the proposed method outperforms the BD(2012) method in terms of accuracy (*Acc*) on both individual datasets by at least 2.5 percentage points, leading to an increased overall accuracy of 0.659 (up 2.7 percentage points). The total error decreases by 2.8 percentage points. On the Bach10 dataset improvements are mainly due to a reduced false alarm rate (E_{fa}), which decreases from 10.2% to 7.8%. This is also reflected by increased precision (P) and stable recall (R). The improvement for the MIREX dataset mainly comes from reduction in both substitution error (E_{subs}) and missed detection error (E_{miss}) rates, leading to higher precision and recall.

In order to determine if the increase in accuracy (*Acc*) is significant we ran a Friedman test for this subtask. The resulting p-value of $0.0009 < 0.01$ indicates that the difference is highly significant. The distribution of *Acc* of the ten files in the Bach10 dataset is shown in Figure 1a.

4.3.2 Note Tracking

For the note tracking subtask, we found that the F-measure was improved by almost 5 percentage points for onset-only evaluation and around 4 percentage points for onset-offset evaluation for both datasets, as shown in Table 3. We ran a Friedman test with regard to the F-measures (F_{on} and F_{off}) for this subtask. For both onset-only and onset-offset metrics, the p-values are less than 0.01, showing that—here,



(a) *Acc* (b) F_{on} (c) F_{off}

Figure 1: Box-and-whisker plots of (a) accuracy; (b) onset-only F-measure; and (c) onset-offset F-measure; for the Bach10 dataset.

too—the differences are significant. The distributions of F_{on} and F_{off} for the Bach10 dataset are shown in Figures 1b and 1c.

The note tracking evaluation shows that both methods under consideration perform better on the MIREX dataset, whereas according to the frame-based evaluation (see Section 4.3.1) they perform better on the Bach10 dataset. This result is in line with the results from other methods on the same data, [3] and is likely to stem from the unusual co-occurrence of trills and legato notes that dominates the MIREX piece.

4.3.3 *Instrument Assignment*

The results for instrument assignment for the two datasets are shown in Table 4. In this subtask, we cannot identify a systematic advantage of either method, with the F-measure means over all instruments being very close (20.7% and 20.9% on the Bach10 dataset, and 35.1% and 34.3% on the MIREX dataset). Slight differences between the methods for particular instruments do not show a consistent advantage of one method either; we will therefore focus on the proposed method in the rest of the discussion. The most obvious differences in F-measure occur between instruments. For example, the results for the Bach10 dataset show that instrument assignment works better for the clarinet and bassoon than for the violin and saxophone. Also, since the note templates include instruments not present in the pieces, false positives occur for these instruments, with the largest ratio of false positives occurring for horn (18.6%) and piano (16.4%). The problem instrument in the MIREX dataset is the oboe, to which few notes are assigned, leading to a low F-measure of around 12-13%. Notes are detected in three instruments that do not feature in the music, with the largest ratio of false positives found in the piano (47.9%) and guitar (34.5%). No false positives were detected for saxophone or violin.

F-measure	Violin	Clarinet	Saxophone	Bassoon	Mean
BD(2012)	0.175	0.313	0.092	0.246	0.207
Proposed	0.190	0.275	0.127	0.243	0.209

(a) Bach10

F-measure	Bassoon	Clarinet	Flute	Horn	Oboe	Mean
BD(2012)	0.292	0.444	0.485	0.409	0.125	0.351
Proposed	0.294	0.420	0.489	0.385	0.129	0.343

(b) MIREX

Table 4: Instrument assignment results

The discrepancy between the satisfactory multiple F0 estimation results and the comparatively low results for instrument assignment is due to the fact that often the correct pitch is detected, but assigned to a wrong instrument or combination of instruments. That is, note templates from different instruments are combined to approximate the observed spectra. The proposed method provides a better reconstruction of the observed data using combinations of templates at the correct pitches, resulting in better performance for frame level and note tracking tasks.

5. DISCUSSION

The use of the temperature parameter τ that is central to the DAEM algorithm in Eqn. (11) is similar to the use of the sparsity parameters in Eqn. (7) and Eqn. (8). In fact, the sparsity method used here is related to the Tempered EM algorithm [8]. Both the DAEM and sparsity equations 'put an exponent on a distribution'. When the exponent is larger than one, the distribution becomes sharper and sparser; when the exponent is smaller than one, the distribution is smoothed, as in the case of high-temperature stages of DAEM.

So far we have used DAEM with only one configuration of three temperature steps. In the future, we would like to explore different configurations to see whether we can further improve the results of multiple F0 estimation and note tracking.

We have shown that DAEM can improve the performance of an EM-based model, but further investigations are needed to show how well this result generalises. For example, preliminary tests have shown that applying DAEM directly in the standard PLCA model in Eqn. (1) without templates, fails to provide better results.

We observe that the previously-trained templates are very important and work as an excellent initialisation for the spectral bases in the PLCA models. On the other hand, they also influence the result of the gain functions, which means that the transcription result will be poor if we use poor or inappropriate templates. The risk of updating the templates during the iteration is that an updated template might no longer accord with its labels (pitch, instrument). Due to the different ways a note can be played and differences in sound transmission, templates will never match observations precisely. Spectral decomposition algorithms compensate for this mismatch by finding mixtures of templates which provide a better approximation of the data

[3] as published on the MIREX website [1].

(see Section 4.3.3). In order to capture the variations of instrument sounds in a single model, we intend to explore physical modelling for time-varying templates in future work.

6. CONCLUSIONS

In this paper, we modified a baseline PLCA model for automatic music transcription. The model's update rules were changed according to the DAEM algorithm to tackle the local minimum problem. The DAEM algorithm introduces a temperature parameter to the update rules and leads the decomposition to converge to a global or better local minimum by gradually lowering the temperature. The proposed method was tested using two standard transcription datasets, the Bach10 dataset and the MIREX dataset. The results show that the proposed method significantly outperforms the baseline method in multiple F0 estimation (accuracy increases by 2.7 percentage points) and note tracking (F-measure increases by 4 percentage points). Although results on an additional instrument assignment task show no significant difference between the methods, they reveal that both methods use mixtures of instrument templates to approximate observed spectra in the test data. We noted several aspects that call for further study: DAEM temperature configurations, the extension of DAEM to more general PLCA models, and the use of physical modelling to generate more flexible instrument templates.

7. ACKNOWLEDGEMENTS

Tian Cheng is supported by China Scholarship Council (CSC)/Queen Mary Joint PhD scholarships. We would like to thank Emmanouil Benetos and Roland Badeau for their comments on this paper.

8. 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] N. Bertin, C. Févotte, and R. Badeau. A tempering approach for itakura-saito non-negative matrix factorization. with application to music transcription. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP09)*, pages 1545–1548, Apr. 2009.

[4] S. Dixon. On the computer recognition of solo piano music. In *Proceedings of Australasian Computer Music Conference*, pages 31–37, 2000.

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

[6] 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.

[7] 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.

[8] G. Grindlay and D. Ellis. A probabilistic subspace model for multi-instrument polyphonic transcription. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, August 9-13, 2010.

[9] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI99*, pages 289–296, 1999.

[10] Y. Itaya, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura. Deterministic annealing EM algorithm in acoustic modeling for speaker and speech recognition. *IEICE Transactions*, 88-D(3):425–431, 2005.

[11] H. Kameoka, T. Nishimoto, and S. Sagayama. Harmonic-temporal structured clustering via deterministic annealing EM algorithm for audio feature extraction. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR05)*, pages 115–122, Sep. 2005.

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

[13] G. Poliner and D. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, pages 154–162, 2007.

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

[15] 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.

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

[17] 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.

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

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