TEMPLATE ADAPTATION FOR IMPROVING AUTOMATIC MUSIC TRANSCRIPTION

Emmanouil Benetos[†], Roland Badeau[‡], Tillman Weyde[†] and Gaël Richard[‡]
† Department of Computer Science, City University London, UK
{emmanouil.benetos.1, t.e.weyde}@city.ac.uk
‡ Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, France
{roland.badeau, gael.richard}@telecom-paristech.fr

ABSTRACT

In this work, we propose a system for automatic music transcription which adapts dictionary templates so that they closely match the spectral shape of the instrument sources present in each recording. Current dictionary-based automatic transcription systems keep the input dictionary fixed, thus the spectral shape of the dictionary components might not match the shape of the test instrument sources. By performing a *conservative* transcription pre-processing step, the spectral shape of detected notes can be extracted and utilized in order to adapt the template dictionary. We propose two variants for adaptive transcription, namely for single-instrument transcription and for multiple-instrument transcription. Experiments are carried out using the MAPS and Bach10 databases. Results in terms of multi-pitch detection and instrument assignment show that there is a clear and consistent improvement when adapting the dictionary in contrast with keeping the dictionary fixed.

1. INTRODUCTION

Automatic music transcription (AMT) is defined as the process of converting an acoustic music signal into some form of music notation [3]. Subtasks of AMT include multi-pitch detection, onset/offset detection, and instrument identification. Recently, the vast majority of transcription approaches use *spectrogram factorization* methods such as non-negative matrix factorization (NMF) and probabilistic latent component analysis (PLCA), which attempt to decompose an input non-negative spectrogram into spectral templates and note activations (e.g. [2, 10, 17]). The spectral templates can either be pre-extracted and stored in a template dictionary [2, 17] or can be estimated using parametric spectral models [10]. An open problem with dictionary-based methods is that the templates might not match the spectral shape of the input instrument sources. Also, unconstrained methods such as NMF and standard PLCA that jointly update the spectral templates and pitch activations can lead to the creation of non-informative bases, and thus, to poor transcription results. It has been shown (e.g. [3]) that the use of templates from the same instrument model or recording conditions can dramatically improve transcription performance.

Related work on automatically estimating or adapting templates for transcription includes [12], where the authors proposed a system for user-assisted (i.e. semi-automatic) music transcription in an NMF setting. The user can label a few notes in the recording; knowledge of the labelled notes can be used in order to create a dictionary that matches the input source. In addition, in [18], the authors propose a dictionary adaptation step within a sparse model that is suitable for single-instrument multi-pitch detection.

In this paper, we propose a method for template adaptation suitable for multiple-instrument polyphonic music transcription (supporting both multi-pitch detection and instrument assignment). The proposed method is based on a multiple-instrument transcription system using PLCA, and supporting tuning changes and frequency modulations. By performing a conservative transcription in a pre-processing step, notes are detected with a high degree of confidence and are used in order to expand the current template dictionary. An additional PLCA-based dictionary adaptation step can further refine the dictionary, so that it matches closely the input source(s). Two system variants are proposed, for single- and multiple-instrument transcription. Experiments using the MAPS [8] and Bach10 [7] databases show a consistent improvement in multi-pitch detection and instrument assignment performance when the proposed template adaptation method is used.

The outline of the paper is as follows. In Section 2, the proposed single-instrument transcription system is presented, with the multiple-instrument version presented in Section 3. The employed datasets, evaluation metrics, and results are detailed in Section 4. Finally, conclusions are drawn and future directions are indicated in Section 5.

2. SINGLE-INSTRUMENT SYSTEM

In the following, we describe a method for single-instrument polyphonic music transcription based on a dictionary of pre-extracted note templates, which is adapted in order to

EB is supported by a City University London Research Fellowship. This work was supported by a Télécom ParisTech sabbatical grant.

match the input instrument source. The proposed system contains a “conservative” transcription pre-processing step in order to detect notes with a high degree of confidence, a dictionary adaptation step, and a final transcription step. The diagram of the proposed system can be seen in Fig. 1.

2.1 Pre-processing

As a pre-processing step, we perform an initial transcription which uses a fixed template dictionary (in which the templates might not be extracted from the same instrument source, model, or recording conditions). The main goal is to only detect notes for which we have a high degree of confidence; in order to achieve this, we perform a “conservative” transcription, as in [16], where the employed transcription system detects notes with high precision and low recall. In other words, the system returns few false alarms but might miss several notes present in the recording.

In order to perform the conservative transcription preprocessing step, we use the spectrogram factorization-based model of [2], which is based on probabilistic latent component analysis (PLCA) [14] and supports the use of a fixed template dictionary. It should be noted that the system in [2] ranked first in the MIREX transcription tasks [1]. The model of [2] takes as input a normalized log-frequency spectrogram $V_{\omega,t} \in \mathbb{R}^{\Omega \times T}$ (ω denotes frequency and t time) and approximates it as a bivariate probability distribution $P(\omega, t)$. $P(\omega, t)$ is in turn decomposed as:

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

where p, f, s denote pitch, log-frequency shifting, and instrument source (in the single-instrument case, s refers to instrument model), respectively. $P(t)$ is the spectrogram energy (known quantity) and $P(\omega|s, p, f)$ are pre-extracted spectral templates for pitch p, source/model s, which are also pre-shifted across log-frequency according to parameter 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. As a log-frequency representation we use the constant-Q transform [13] 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).

Using a fixed template dictionary, the parameters that need to be estimated are $P_t(f|p)$, $P_t(s|p)$, and $P_t(p)$. This can be achieved using the expectation-maximization (EM) algorithm [5], with 15-20 iterations being typically sufficient. The resulting multi-pitch output is given by $P(p, t) = P(t)P_t(p)$.

In order to extract note events in spectrogram factorization-based AMT algorithms, typically thresholding is performed on the pitch activations ($P(p, t)$ in this case). The value of the threshold θ controls the levels of precision/recall. A low threshold has a high recall and low precision; the opposite occurs with a high threshold. By selecting a high value of θ, in essence we perform a conservative transcription. The final output of the pre-processing step is a collection of pitches and time frames $\{p_1, t_1\}, \{p_2, t_2\}, ..., \{p_N, t_N\}$ which can be used in order to adapt the template dictionary.

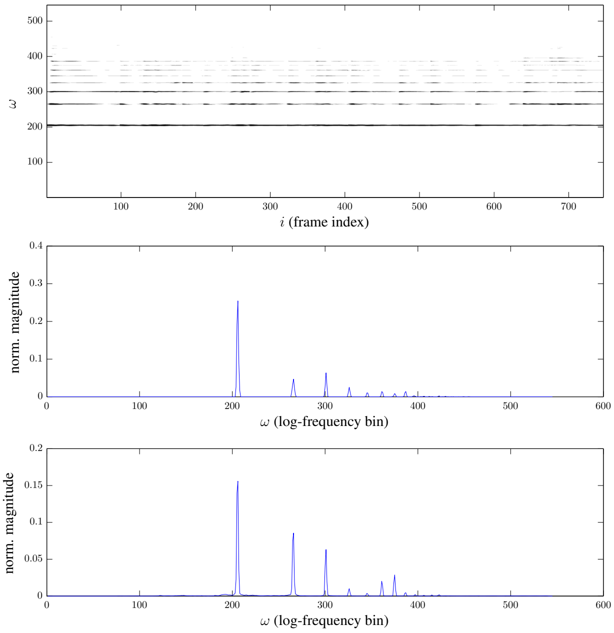

Figure 2. Top: a collection of spectra $\hat{V}^{(42)}$ (note D4) from piano recording ‘alb_se2’ taken from the MAPS database (piano model: ENSTDkCl). Middle: extracted normalised template $P(\omega|p = 42)$. Bottom: a D4 template from piano model AkPnBcht from the MAPS database.

2.2 Template Adaptation

Given a collection of detected pitches, the first step regarding template adaptation is to collect the spectra that correspond to the aforementioned pitches in the recording. Thus, for each pitch p all time frames t_{ip} that contain that pitch are collected (where $i = 1, ..., N_p$ and N_p is the number of frames containing p).

Subsequently, for each pitch p we create a collection of spectra where that pitch is observed:

$$\hat{V}^{(p)} = V_{\omega,t \in t_{ip}} \otimes \mathbf{h}_p \tag{2}$$

where $\mathbf{h}_p$ is a harmonic comb that serves as an indicator function (setting to zero all frequency bins not belonging to pitch p), and $\otimes$ denotes elementwise multiplication. In other words, $\hat{V}^{(p)} \in \mathbb{R}^{\Omega \times N_p}$ is a collection of the spectra corresponding to detected pitch p in the input recording.

Using information from $\hat{V}^{(p)}$, new spectral templates are created for each p that was detected in the conservative transcription step. In order to create the new templates, the standard PLCA algorithm is used with one component [14], with the input in each case being $\hat{V}^{(p)}$. The output for each p is a spectral template $\mathbf{w}^{(p)}$ which can be used in order to expand the present dictionary.

Given that the conservative transcription step might not have detected all possible pitches present in the recording, information from the extracted templates can be used in order to estimate missing templates. As in the user-assisted case of [12], we can derive templates at missing pitches by simply shifting existing templates across log-frequency. Given a missing pitch template, we consider a neighbor-

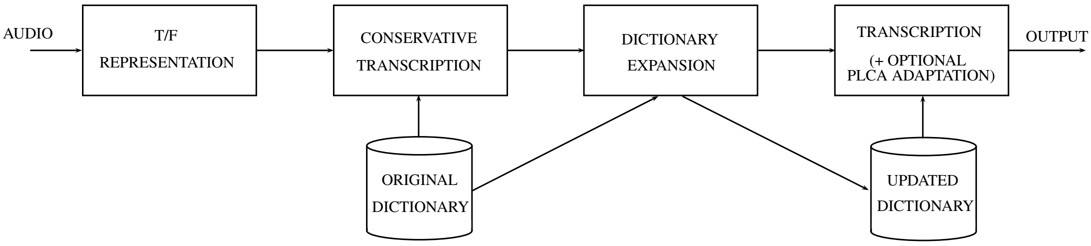

Figure 1. Proposed system diagram.

hood of up to 4 semitones; if a template exists in the neighborhood, it is shifted accordingly in order to estimate the missing template. Finally, the resulting template dictionary is pre-shifted across log-frequency over a semitone range in order to account for tuning deviations and frequency modulations. The output of the template adaptation step is normalized and denoted as $P(\omega|s = s_{new}, p, f)$, where s_{new} refers to the new instrument source that is added to the existing dictionary.

The template adaptation step is illustrated in Fig. 2, where a collection of extracted spectra for note D4 of a piano recording can be seen, along with the computed template, as well as with a template for the same note taken from a different piano model. By comparing the two piano spectra, the importance of adapting templates to the specific instrument source can be seen.

2.3 Transcription

Having created an expanded dictionary with a set of note templates taken from the instrument source present in the recording, the recording is re-transcribed using the new dictionary and the model of (1). In order to further adapt the extracted templates to the input source, an optional step is also applied on updating the new template set during the PLCA iterations. The modified iterative update rule is based on the work of [15] (which incorporated prior information on PLCA update rules) and is applied only for the new set of templates. It is formulated as:

$$\hat{P}(\omega|s_{new}, p, f) = \frac{\sum_t \alpha P_t(p, f, s_{new}|\omega) V_{\omega,t} + (1-\alpha) P(\omega|s_{new}, p, f)}{\sum_{\omega,t} \alpha P_t(p, f, s_{new}|\omega) V_{\omega,t} + (1-\alpha) P(\omega|s_{new}, p, f)} \quad (3)$$

where $P_t(p, f, s|\omega)$ is the posterior of the model (defined in [2]), and α is a parameter which controls the weight of the PLCA adaptation, with $(1-\alpha)$ giving weight to the set of extracted templates from the procedure of Section 2.2. In this work, α is set to 0.05, thus the PLCA template adaptation is only slightly changing the shape of the templates (given that the model is unconstrained, giving a large weight to the PLCA adaptation step would result in non-meaningful templates).

Finally, the output of the transcription step is given by $P(p, t) = P(t)P_t(p)$. For converting the non-binary pitch activation into a binary piano-roll representation, as in [6] we perform thresholding on $P(p, t)$ followed by a process removing note events with a duration less than 80ms.

3. MULTIPLE-INSTRUMENT SYSTEM

In dictionary-based multiple-instrument transcription, the dictionary typically consists of one or more sets of templates per instrument. Thus, in order to update dictionary templates for multiple instruments, modifications need to be made from the system presented in Section 2.

Regarding the pre-processing step, we still use the model of (1), which supports multiple-instrument transcription. In this case, s denotes instrument source. An advantage of the model of (1) is that it can produce an instrument assignment output (i.e. each detected note is assigned to a specific instrument). Thus, having estimated the unknown model parameters, the instrument assignment output for instrument s_{ins} is given by the following time-pitch representation:

$$P(s = s_{ins}, p, t) = P_t(s = s_{ins}|p) P_t(p) P(t) \quad (4)$$

The representation $P(s, p, t)$ can be thresholded in the same way as the pitch activation in order to derive a binary piano-roll representation of the notes produced by a specific instrument. Here, we perform "conservative" thresholding (i.e. use a high θ value) for every instrument in $P(s, p, t)$ in order to create a collection of detected pitches and time frames per instrument:

$$\{s_1, p_1, t_1\}, \{s_2, p_2, t_2\}, ..., \{s_N, p_N, t_N\} \quad (5)$$

where $s \in 1, \ldots, S$, $p \in 1, \ldots, 88$, and $t \in 1, \ldots, T$.

For performing multi-instrument template adaptation, we collect all time frames t_{ips} that contain pitch p and instrument s. We create a collection of spectra $\hat{V}^{(p,s)}$ where a pitch is observed for a specific instrument, in the same way as in (2). Using information from $\hat{V}^{(p,s)}$, new spectral templates are created for specific cases of s and p using the single-component PLCA algorithm. As in Section 2.2, templates at missing pitches are derived by translating existing templates across log-frequency. The output of the template adaptation step is denoted as $P(\omega|s = \{s_{new1}, s_{new2}, ...\}, p, f)$ where $s_{new1}, s_{new2}, ...$ denote the new sets of templates for the existing instruments.

Finally, the input recording is re-transcribed using the model of (1), by utilizing the expanded dictionary. We also apply the same optional PLCA-based dictionary adaptation step shown in Section 2.3. The multiple-instrument transcription system has two sets of outputs: the pitch activation $P(p, t)$ (which is used for multi-pitch detection evaluation) and the instrument contribution $P(s, p, t)$ (which is used for instrument assignment evaluation).

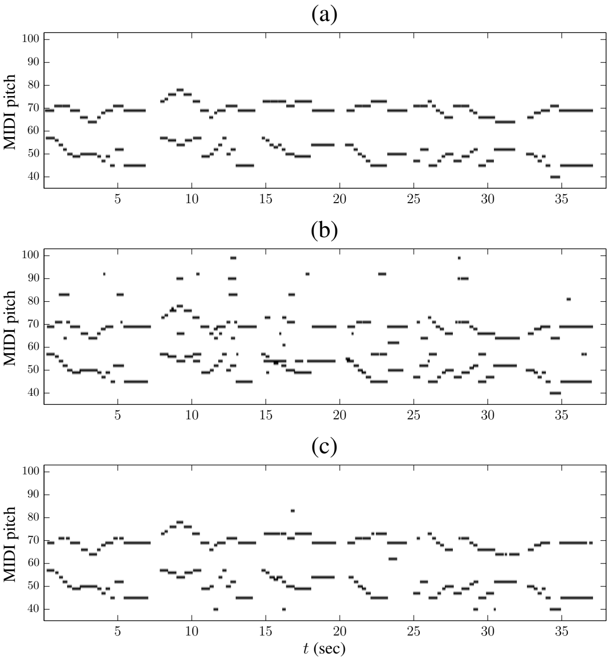

Figure 3. (a) The pitch ground truth for the bassoon-violin duet 'Nun bitten' from the Bach10 database. (b) The transcription piano-roll without template adaptation. (c) The transcription piano-roll with template adaptation.

An example of how template adaptation can improve transcription performance for a multiple-instrument recording (bassoon and violin) is given in Fig. 3, where the transcription output with template adaptation has significantly fewer false alarms compared with transcription without template adaptation (in which many extra detected notes can be seen in higher pitches).

4. EVALUATION

4.1 Datasets

For training the single-instrument system of Section 2, we used isolated note recordings from the 'AkPnBcht' and 'SptkBGCl' piano models of the MAPS database [8]. We used the standard PLCA algorithm with one component [14] in order to extract a single template per note, covering the complete piano note range. For testing the single-instrument system, we used thirty piano segments of 30s duration from the MAPS database from the 'ENSTDkCl' piano model; the test dataset has in the past been used for multi-pitch evaluation (e.g. [2,4,19]). For comparative purposes, we also extracted training templates from the same test source ('ENSTDkCl').

For training the multiple-instrument system of Section 3, we used isolated note samples of bassoon and violin from the RWC database [11], covering the complete note range of the instruments. For testing the multiple-instrument system, we created ten duets of bassoon-violin, mixed from single instrument tracks from the multi-track Bach10 dataset [7]. The duration of the recordings varies from 25-41sec. For comparative purposes, we also extracted dictionary templates for bassoon and violin from the single instrument tracks of the Bach10 database, in order to demonstrate the upper performance limit of the transcription system.

System	Pre_n	Rec_n	F_n
C1	66.41%	48.41%	55.33%
C2	68.07%	48.80%	56.26%
C3	67.84%	49.38%	56.56%
C4 (oracle)	70.43%	50.35%	58.17%

Table 1. Multi-pitch detection results for the single-instrument system using the MAPS database.

4.2 Metrics

For evaluating the performance of the proposed systems for multi-pitch detection, we employ onset-only note-based transcription metrics, which are used in the MIREX note tracking task [1]. A detected note is considered correct if its pitch matches a ground truth pitch and its onset is within a 50ms tolerance of a ground-truth onset. The resulting note-based precision, recall, and F-measure are defined as:

$$Pre_n = \frac{N_{tp}}{N_{sys}} \quad Rec_n = \frac{N_{tp}}{N_{ref}} \quad F_n = \frac{2Rec_n Pre_n}{Rec_n + Pre_n} \qquad (6)$$

where N_{tp} is the number of correctly detected pitches, N_{sys} is the number of pitches detected by the system, and N_{ref} is the number of reference pitches.

For the instrument assignment evaluations we use the pitch ground-truth of each instrument separately (compared with the instrument-specific piano-roll output of the system), and compute the F-measure metrics for bassoon (F_b) and violin (F_v).

4.3 Results - Single Instrument Evaluation

For single-instrument transcription evaluation using the 30 MAPS recordings, results are shown in Table 1 using four different system configurations. Configuration C1 corresponds to the system without template adaptation; C2 to the system with template adaptation; C3 to the system with template adaptation using both the creation of the new dictionary plus the PLCA update of the dictionary, as shown in Section 2.2. Finally, C4 refers to comparative experiments without template adaptation, but using templates from the same instrument source ('ENSTDkCl' model in the single-instrument case), which is meant to demonstrate the upper performance limit of the transcription system.

From the single-instrument multi-pitch detection results, it can be seen that an improvement of +0.9% in terms of F_n is reported when using the template adaptation procedure; the improvement rises to +1.2% when also using the PLCA dictionary adaptation updates. The performance difference between the original C1 system (without knowledge of the source templates) and the 'optimal' system (C4) which contains templates from the same test source is 2.8%; thus, the proposed template adaptation steps can help bridge the gap, without requiring any knowledge of

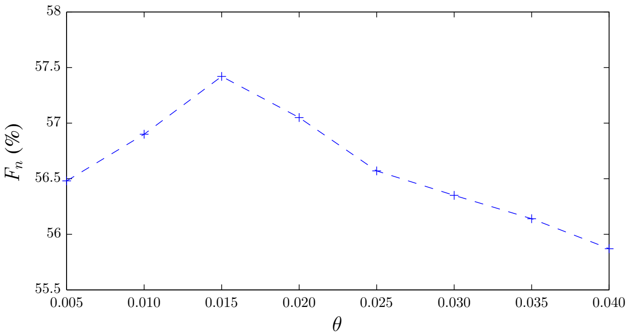

Figure 4. Multi-pitch detection results on the MAPS-ENSTDkCl set using different values of θ.

System	Pre_n	Rec_n	F_n	F_b	F_v
C1	64.79%	71.20%	67.72%	70.19%	42.10%
C2	69.71%	75.72%	72.51%	70.81%	45.98%
C3 (violin)	70.02%	75.41%	72.50%	70.54%	44.51%
C3 (bassoon)	72.49%	77.67%	74.90%	68.77%	45.87%
C3 (both)	71.30%	77.37%	74.11%	67.57%	44.08%
C4 (oracle)	74.90%	82.94%	78.64%	81.25%	62.05%

Table 2. Multi-pitch detection and instrument assignment results for the multiple-instrument system using the Bach10 dataset.

the test instrument source. Regarding precision and recall, in all cases it can be seen that the transcription system has fewer false alarms than missed note detections. The proposed template adaptation steps help in equally improving precision and recall.

In order to determine the value of the conservative transcription threshold θ, we used a training subset of 10 recordings from the MAPS 'SptkBGCl' models; the value of $\theta = 0.028$ was selected by maximising Pre_n. In Figure 4, transcription performance on the MAPS-ENSTDkCl set is reported by selecting various values for θ. It can be seen that the transcription performance can reach up to F_n=57.4% with $\theta = 0.015$, which enforces the argument that the proposed template adaptation method can successfully adapt dictionary templates so that they match the input instrument source.

Another comparison for the single-instrument system is made, where the dictionary derived from Section 2.2 replaces the dictionary of instrument 'SptkBGCl' (instead of expanding the original dictionary). The resulting F_n is 55.88%, indicating that expanding the dictionary leads to better results compared with replacing the dictionary. It should also be noted that the achieved transcription performance outperforms the system in [19] which reports a frame-based F-measure of 52.4%, whereas the template adaptation system reports a frame-based F of 59.73%. Finally, no rigorous figures for statistical significance of the results can be given since all signal frames cannot be considered as independent samples. However, the reported tests are run on several thousands of frames which leads, if the samples were independent, to a statistically significant difference of the order of 0.6% (with 95% confidence).

4.4 Results - Multiple Instrument Evaluation

For multiple-instrument evaluation, we also use the four different system configurations that were used for single-instrument transcription. For system configuration C3, we perform the PLCA dictionary update using 3 variants: by updating the bassoon only, by updating the violin only, or by updating both dictionaries. Transcription results for the multiple-instrument case are shown in Table 2.

It can be seen that without any template adaptation (C1), $F_n = 67.72\%$; by performing the proposed template adaptation step (C2), the multi-pitch detection F-measure improves by +4.8%.

By performing template adaptation with C3 which also includes the PLCA update rule of (3), although no performance gain is obtained over the C2 configuration for the violin updates, there is a +2.4% improvement over C2 when updating the bassoon dictionary only. Finally, when updating both dictionaries, there is a performance drop for F_b and F_v over the C2 configuration (but the system still outperforms the original C1 system). The performance of the PLCA-based dictionary updates can be explained by the fact that the update rule of (3) might combine the observed spectra from both instruments and produce dictionaries that might represent a combination of the two instruments. Finally, the C4 system represents the upper performance limit, which is +11.7% higher than when using a dictionary from a different instrument models or recording conditions. It can be seen that the proposed template adaptation methods help in bridging that performance gap, resulting in a dictionary that closely matches the test instrument sources.

Regarding instrument assignment performance, in all cases the bassoon note identification reports better results compared to violin note identification. It can be seen that with the proposed template adaptation, the bassoon identification remains relatively constant (a small improvement of +0.6% is reported when comparing C1 with C2). On the other hand, violin identification improves by +3.9%; this indicates that the RWC bassoon templates closely matched the Bach10 bassoon models, whereas the violin RWC templates could greatly benefit from template adaptation.

By comparing the MAPS and Bach10 evaluations, an observation can be made that the performance improvement using the proposed template adaptation method depends on the mismatch between the original dictionary and the spectral shape of the instruments present in the recordings. Thus, the 11.7% performance gap for the Bach10 dataset led to a greater improvement for the template adaptation method compared to the 2.8% performance gap reported for the MAPS dataset (which led to a smaller, yet consistent improvement when using the proposed template adaptation method).

5. CONCLUSIONS

In this paper, we proposed a novel method for template adaptation for automatic music transcription, that can be used in dictionary-based systems. We utilized a multiple-instrument transcription system based on probabilistic latent component analysis, and performed a conservative transcription pre-processing step in order to detect notes with a high confidence. Based on the initial transcription, the spectra of the detected notes are collected, processed, and are used in order to create a new dictionary that closely matches the spectral characteristics of the input instrument source(s). Both single-instrument and multi-instrument variants of the proposed method are presented and evaluated, in terms of multi-pitch detection and instrument assignment. Experimental results using the MAPS and Bach10 datasets show that there is a clear and consistent performance improvement when using the proposed template adaptation method, especially when there is a large discrepancy between the original dictionary and the spectral characteristics of the test instrument sources.

In the future, we will evaluate the proposed system using multiple-instrument recordings with more than two instruments. Parametric models (such as source-filter models) will also be investigated for updating the note templates, along with adaptive methods for deriving the conservative transcription threshold. We also plan to evaluate the proposed system in the next MIREX evaluations [1]. Finally, the proposed template adaptation steps will also be evaluated in the context of score-informed source separation using spectrogram factorization models [9].

6. REFERENCES

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

[2] E. Benetos, S. Cherla, and T. Weyde. An efficient shift-invariant model for polyphonic music transcription. In *6th Int. Workshop on Machine Learning and Music*, Prague, Czech Republic, September 2013.

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

[4] J. J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F. J. Cañadas-Quesada. Musical instrument sound multi-excitation model for non-negative spectrogram factorization. *IEEE J. Selected Topics in Signal Processing*, 5(6):1144–1158, 2011.

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

[6] A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *11th Int. Society for Music Information Retrieval Conf.*, pages 489–494, 2010.

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

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

[9] S. Ewert, B.Pardo, M. Müller, and M. D. Plumbley. Score-informed source separation for musical audio recordings. *IEEE Signal Processing Magazine*, 31(3):116–124, May 2014.

[10] B. Fuentes, R. Badeau, and G. Richard. Harmonic adaptive latent component analysis of audio and application to music transcription. *IEEE Trans. Audio, Speech, and Language Processing*, 21(9):1854–1866, 2013.

[11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: music genre database and musical instrument sound database. In *International Conference on Music Information Retrieval*, October 2003.

[12] H. Kirchhoff, S. Dixon, and Anssi Klapuri. Missing template estimation for user-assisted music transcription. In *IEEE Int. Conf. Audio, Speech and Signal Processing*, pages 26–30, 2013.

[13] C. Schörkhuber and A. Klapuri. Constant-Q transform toolbox for music processing. In *7th Sound and Music Computing Conf.*, Barcelona, Spain, July 2010.

[14] M. Shashanka, B. Raj, and P. Smaragdis. Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience*, 2008. Article ID 947438.

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

[16] D. Tidhar, M. Mauch, and S. Dixon. High precision frequency estimation for harpsichord tuning classification. In *IEEE Int. Conf. Audio, Speech and Signal Processing*, pages 61–64, Dallas, USA, March 2010.

[17] F. Weninger, C. Kirst, B. Schuller, and H.-J. Bungartz. A discriminative approach to polyphonic piano note transcription using supervised non-negative matrix factorization. In *IEEE Int. Conf. Audio, Speech and Signal Processing*, pages 6–10, May 2013.

[18] T. B. Yakar, P. Sprechmann, R. Litman, A. M. Bronstein, and G. Sapiro. Bilevel sparse models for polyphonic music transcription. In *14th Int. Society for Music Information Retrieval Conf.*, pages 65–70, 2013.

[19] K. Yoshii and M. Goto. Infinite composite autoregressive models for music signal analysis. In *13th Int. Society for Music Information Retrieval Conf.*, pages 79–84, October 2012.