SCORE-INFORMED TRANSCRIPTION FOR AUTOMATIC PIANO TUTORING

Emmanouil Benetos, Anssi Klapuri, and Simon Dixon

Centre for Digital Music, Queen Mary University of London, London E1 4NS, UK
{emmanouilb,anssik,simond}@eecs.qmul.ac.uk

ABSTRACT

In this paper, a score-informed transcription method for automatic piano tutoring is proposed. The method takes as input a recording made by a student which may contain mistakes, along with a reference score. The recording and the aligned synthesized score are automatically transcribed using the non-negative matrix factorization algorithm for multi-pitch estimation and hidden Markov models for note tracking. By comparing the two transcribed recordings, common errors occurring in transcription algorithms such as extra octave notes can be suppressed. The result is a piano-roll description which shows the mistakes made by the student along with the correctly played notes. Evaluation was performed on six pieces recorded using a Disklavier piano, using both manually-aligned and automatically-aligned scores as an input. Results comparing the system output with ground-truth annotation of the original recording reach a weighted F-measure of 93%, indicating that the proposed method can successfully analyze the student's performance.

Index Terms— Music signal analysis, score-informed transcription, NMF, HMMs

1. INTRODUCTION

Automatic music transcription is the process of converting an audio recording into some form of music notation. Although the field remains very active, results are still below human transcription performance. In contrast with unsupervised techniques, certain applications can also incorporate score information, such as the emerging field of informed source separation [1]. One application that can exploit score information is automatic piano tutoring, where a system evaluates a student's performance based on a reference score. Thus, the problem that needs to be addressed is score-informed piano transcription. Such systems can assist the student in eliminating basic mistakes during practice, thus giving the piano teacher the opportunity to focus on more advanced concepts during the lessons. In the past, the problem of informed transcription has received limited attention, with the most notable work done in automatic violin tutoring in [2], which fuses audio and video transcription with score information.

In this work, a method for score-informed music transcription is proposed which is applied to automatic piano tutoring. The algorithm takes as input a non-aligned reference MIDI score and a recording by a student which contains performance mistakes. The proposed method performs MIDI-to-audio alignment, MIDI synthesis, automatic transcription of both the recording and the synthesized MIDI, and combines all information in order to analyze the student's performance. For evaluation, six complete piano pieces recorded from a Yamaha Disklavier were tested. Experiments were performed using manually-aligned and automatically-aligned scores, where it is shown that the proposed system can successfully analyze the student's performance.

2. SCORE-INFORMED TRANSCRIPTION SYSTEM

The input of the score-informed transcription system is a recording made by a student which contains mistakes and a reference MIDI score, which is aligned and synthesized. Next, the recording made by the student is transcribed, along with the synthesized MIDI. The evaluation of the student's performance is made by comparing the two transcribed recordings with the aligned MIDI. In Fig. 1, the diagram for the proposed score-informed transcription system is depicted.

2.1. MIDI-to-audio Alignment and Synthesis

For automatically aligning the reference MIDI score with the recording made by the student, we employ the windowed time warping (WTW) alignment algorithm proposed in [3]. This algorithm is computationally inexpensive, and can be utilized in a real-time automatic piano tutoring application. In the experiments performed in [3], it was shown that the alignment algorithm can correctly align 97% of the audio note onsets in the test set employed, using a 2 sec tolerance.

The result is an aligned MIDI file, which afterwards is synthesized using the TiMidity synthesizer using the *Merlin Vienna* soundfont library. For comparative purposes, manually-aligned MIDI files are also produced and synthesized, which are described in Section 3.1.

We acknowledge the support of the MIReS project, supported by the European Commission, FP7, ICT-2011.1.5 Networked Media and Search Systems, grant agreement No 287711. E. Benetos is funded by a Westfield Trust Research Studentship (Queen Mary University of London).

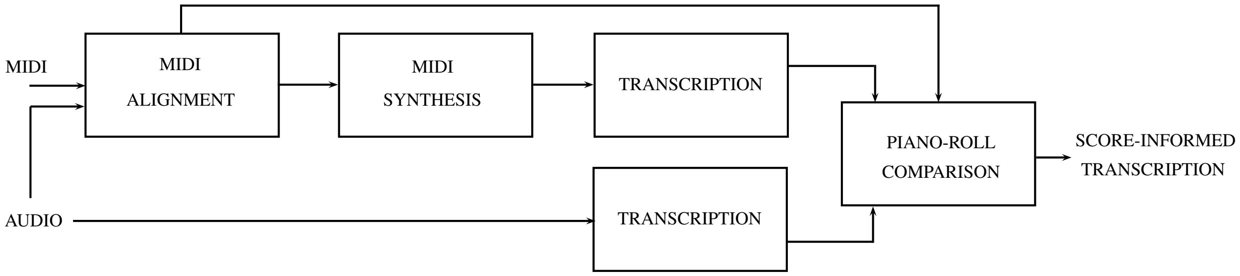

Fig. 1. Score-informed transcription system diagram.

2.2. Multi-pitch Detection

For transcribing the original and the synthesized recordings, we employ the non-negative matrix factorization (NMF) algorithm with β-divergence [4], using pre-extracted piano templates. The algorithm was utilized in [5] for real-time piano transcription. The NMF algorithm with β-divergence is computationally inexpensive and it has been shown to produce reliable results in piano-specific transcription [5].

Firstly, spectral templates for the complete piano note range were extracted, corresponding to notes from A0 to C8. We used recordings from 3 chromatic scales from a Yamaha U3 Disklavier, which was also used for the test recordings. In addition, we employed isolated note samples from 3 piano models from the MAPS database [6]. The fact that we are using training templates from the same piano source as in the test set is a reasonable assumption given the specific tutoring application, since the student can provide training examples in a setup stage. If templates from the same source are not available, general-purpose templates from e.g. the MAPS database can be used (related experiments shown in Section 3). For extracting the templates, the constant-Q transform (CQT) [7] was employed using a resolution of 120 bins/octave and lowest frequency at 27.5 Hz. Next, the NMF algorithm using a single component was employed for extracting the template from an isolated note recording. The single-component NMF model can be expressed by $\mathbf{V} \approx \mathbf{wh}$, where $\mathbf{V} \in \mathbb{R}^{f \times n}$ is the input CQT spectrogram, $\mathbf{w} \in \mathbb{R}^{f \times 1}$ is the computed spectral template, and $\mathbf{h} \in \mathbb{R}^{1 \times n}$ is the gain of the component [4].

For the multi-pitch detection step, the NMF model with β-divergence is employed, which is identical to the standard NMF model. The β-divergences are a parametric family of distortion functions which can be used in the NMF cost function. In essence, the choice of parameter $\beta \in \mathbb{R}$ controls the importance of high-energy and low-energy frequency components in the decomposition. For the present experiments, we used $\beta = 0.6$, which was shown to produce the best results for piano transcription in [5]. Since in our case the spectral template matrix is fixed, only the gain is iteratively updated (after random initialization) as:

$$\mathbf{h} \leftarrow \mathbf{h} \otimes \frac{\mathbf{W}^T((\mathbf{Wh})^{\beta-2} \otimes \mathbf{v})}{\mathbf{W}^T(\mathbf{Wh})^{\beta-1}} \qquad (1)$$

where $\mathbf{v} \in \mathbb{R}^{f \times 1}$ is a single frame from the test signal, $\mathbf{h}$ is the gain for the specific frame, $\otimes$ is the elementwise product and the fraction denotes elementwise division. Convergence is observed at 10-15 iterations.

For piano transcription, the spectral template matrix $\mathbf{W}$ was created from concatenating the spectral templates from either the 3 sets of the Disklavier or the MAPS templates:

$$\mathbf{W} = [\mathbf{W}^{(1)} \;\; \mathbf{W}^{(2)} \;\; \mathbf{W}^{(3)}] \qquad (2)$$

thus, $\mathbf{W} \in \mathbb{R}^{f \times 264}$. After the NMF update rule of (1) is applied to the input log-spectrogram $\mathbf{V}$, the pitch activation matrix is created by adding the component vectors from $\mathbf{H}$ that correspond to the same pitch:

$$\mathbf{H}' = \mathbf{H}_{1:88,:} + \mathbf{H}_{89:176,:} + \mathbf{H}_{177:264,:} \qquad (3)$$

where $\mathbf{H}' \in \mathbb{R}^{88 \times n}$.

2.3. Note Tracking

In [5], note activations are computed by simply thresholding the pitch activation matrix $\mathbf{H}'$. Here, we employ hidden Markov models (HMMs) for note tracking, similarly to [8]. Each pitch $p = 1, \ldots, 88$ is modeled using a 2-state HMM, which denotes pitch activity or inactivity. The pitch-wise hidden state sequence is given by $Q_p = \{q_p[n]\}$. For estimating the state priors $P(q_p[1])$ and state transitions $P(q_p[n]|q_p[n-1])$ for each pitch, MIDI files from the RWC database [9] are employed. The most likely state sequence for each pitch is:

$$\hat{Q}_p = \arg\max_{q_p[n]} \prod_n P(q_p[n]|q_p[n-1])P(o_p[n]|q_p[n]) \qquad (4)$$

where $P(o_p[n]|q_p[n])$ is the observation probability for frame n. The sequence $\hat{Q}_p$ is computed using the Viterbi algorithm.

For estimating the observation probability for an active pitch, we use a sigmoid curve which has as input the pitch activation $\mathbf{h}'_p = \mathbf{H}'_{p,n}$:

$$P(o_p[n]|q_p[n] = 1) = \frac{1}{1 + e^{-(\mathbf{h}'_p - \lambda)}} \qquad (5)$$

where λ is a parameter that controls the smoothing (a high value will discard pitches with low energy). The result of the postprocessing step is a binary piano-roll transcription.

In order to set the value of parameter λ for the transcribed recording and synthesized score, we used one piece from our dataset for training (detailed in Section 3.1). Also, we extract two additional piano-rolls from the transcribed recording using different values for λ, thus creating a 'strict' transcription (with high precision and low recall) and a 'relaxed' transcription (with high recall and low precision), which will be utilized in the output of the proposed system. The values of λ that were used for the normal, strict, and relaxed transcription, are respectively $\{1.3, 1.0, 2.1\}$.

Finally, the resulting piano-rolls are processed in order to detect any repeated notes which might appear in the final piano-roll as a continuous event (e.g. trills). For the piano, detecting note onsets can be achieved by simply detecting energy changes. Thus, peak detection is performed using the activation matrix for each detected note. If a peak is detected at least 200ms after the onset, then the note is split into two.

2.4. Piano-roll Comparison

In order to compare the performance of the student with the aligned score, we will utilize additional information using the transcribed synthesized score, as well as the strict and relaxed transcriptions of the recording. The motivation is that automatic transcription algorithms typically contain false alarms (such as octave errors) and missed detections (usually in the case of dense chords). However, the transcribed synthesized score might also contain these errors. Thus, it can assist in eliminating any errors caused by the transcription algorithm instead of attributing them to the student's performance.

Two *assumptions* are made in the algorithm: firstly, the recording does not contain any structural errors. Thus, only local errors can be detected, such as missed or extra notes played by the student. Secondly, evaluation is performed by only examining note onsets, thus discarding note durations.

The process comparing the piano-roll for the transcribed recording ($prStudent$), the synthesized MIDI ($prSynth$), and the aligned MIDI ($prGT$) is given in Algorithm 1. The tolerance for $onset(p, n)$ is set to $\pm 200ms$. In line 8, when an onset is present in the ground truth but is absent in both transcriptions, then we do not have enough knowledge to determine the existence of that note and it is set as correct.

After Algorithm 1 is completed, the extra and missed notes present in $prResult$ are re-processed using the 'strict' piano-roll $prStrict$ and the 'relaxed' piano-roll $prRelaxed$, respectively. The notion is that if that same extra note is not present in $prStrict$, then it is simply caused by a deficiency in the transcription algorithm of the original recording. Likewise, if a missed note appears in $prRelaxed$, then it is taken that it was played but was not detected due to the transcription of the original recording.

The final output of the comparison step is the resulting piano-roll, which contains information on correct notes, missed notes, and extra played notes. In Fig. 2, the score-informed transcription of a piece can be seen, compared to the ground-truth of the student's performance.

Algorithm 1 Piano-roll comparison

```
Input: $prStudent, prSynth, prGT$
 1: for each onset($p, n$) $\in prGT$ do
 2:   if onset($p, n$) $\in prStudent$ then
 3:     $prResult(p, n) = correct\ note$
 4:   else
 5:     if onset($p, n$) $\in prSynth$ then
 6:       $prResult(p, n) = missed\ note$
 7:     else
 8:       $prResult(p, n) = correct\ note$
 9:     end if
10:   end if
11: end for
12: for each onset($p, n$) $\in prStudent$ do
13:   if onset($p, n$) $\notin prGT, prSynth$ then
14:     $prResult(p, n) = extra\ played\ note$
15:   end if
16: end for
17: return $prResult$
```

	Composer	Title
1	Josef Haydn	Andante from Symphony No. 94
2	James Hook	Gavotta, Op. 81
3	Pauline Hall	Tarantella
4	Felix Swinstead	A Tender Flower
5	Johann Krieger	Bourrée from Sechs musicalishe Partien
6	Johannes Brahms	The Sandman, WoO 31
7	Tim Richards (arr.)	Down by the Riverside

Table 1. The score-informed piano transcription dataset.

3. EVALUATION

3.1. Dataset

Since no dataset exists for score-informed piano transcription experiments, 7 recordings were made using a Yamaha U3 Disklavier. The piano was slightly untuned, making the recording conditions more realistic. The recordings were selected from the Associated Board of the Royal Schools of Music 2011/12 syllabus for grades 1 and 2. A list of the recorded pieces can be seen in Table 1. Each recording contains mistakes compared to the original score and MIDI ground-truth was created detailing those mistakes. The first recording was used for development, whereas the last six recordings were used for testing. The dataset is available online[1].

3.2. Metrics

Since the task of score-informed transcription is a relatively unexplored one, we will present a set of metrics for evalu-

[1] `http://c4dm.eecs.qmul.ac.uk/rdr/`

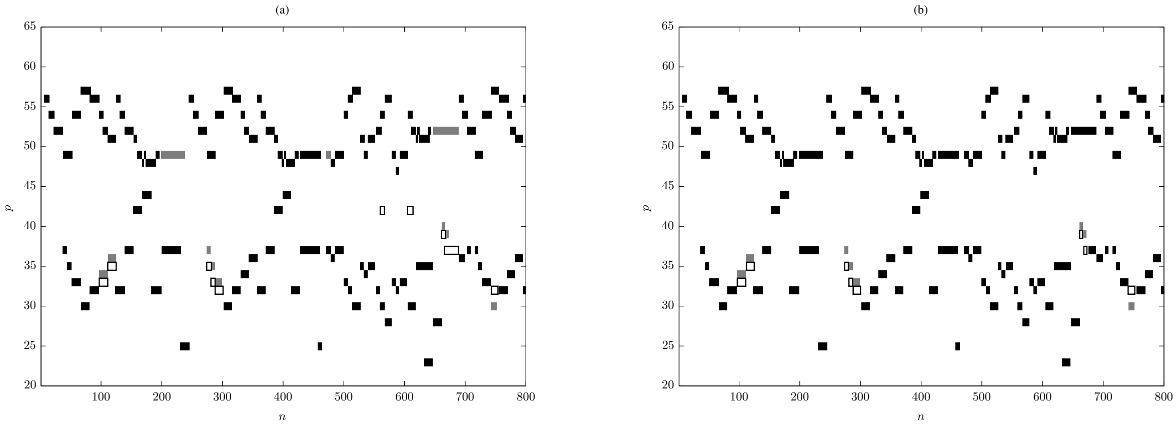

Fig. 2. (a) The score-informed transcription of a segment from Johann Krieger's Bourrée. (b) The performance ground-truth. Black corresponds to correct notes, gray to missed notes and empty rectangles to extra notes played by the performer.

ating the performance of the proposed method. Firstly, we will evaluate the method's efficiency for the task of automatic transcription by employing the onset-based note-level accuracy also used in [5]. This evaluation will be performed on the transcribed recording and synthesized score. A returned note event is assumed to be correct if its onset is within a +/-100ms range of a ground-truth onset. We define the number of correctly detected notes as N_{corr}, the number of false alarms as N_{fa} and the number of missed detections as N_{md}. The accuracy metric is defined as:

$$Acc = \frac{N_{corr}}{N_{corr} + N_{fa} + N_{md}} \tag{6}$$

In addition, the precision (Pre), recall (Rec), and F-measure ($\mathcal{F}$) are employed for evaluating the automatic transcription performance of the employed methods.

For the score-informed transcription experiments, each detected note from the student's recording can be classified as correct, missed, or extra. Thus, for each piece, three layers of ground-truth exist. Using (6) we will define Acc_{corr} as the algorithm's accuracy for the notes that were correctly played by the student. Likewise, Acc_{md} denotes the accuracy for the notes that the student omitted and Acc_{fa} the accuracy for the extra notes produced. Using the F-measure, a similar set of metrics is defined for the score-informed transcription evaluation: $\mathcal{F}_{corr}$, $\mathcal{F}_{md}$, $\mathcal{F}_{fa}$.

Finally, we will define weighted metrics joining all three layers of the ground-truth. Given that M_{corr} is the number of correctly played notes in the performance of the student, M_{md} is the number of the notes missed and M_{fa} is the number of extra notes, the weighted accuracy is defined as:

$$Acc_w = \frac{M_{corr}Acc_{corr} + M_{md}Acc_{md} + M_{fa}Acc_{fa}}{M_{corr} + M_{md} + M_{fa}} \tag{7}$$

A similar definition can be made for a weighted F-measure, denoted as $\mathcal{F}_w$.

	Acc	$\mathcal{F}$	Pre	Rec
Recording	83.88%	91.13%	93.34%	89.11%
Manual MIDI	84.73%	91.57%	93.56%	89.73%
Automatic MIDI	89.77%	94.55%	95.05%	94.09%

Table 2. Automatic transcription results.

3.3. Results

In Table 2, the automatic transcription results for the original recording and the synthesized MIDI (using manual and automatic alignment) are shown. In all cases the performance of the NMF-based transcription algorithm is quite high, with the $\mathcal{F}$ always surpassing 90%. The performance difference between the transcription of the manual and automatic MIDI is due to the fact that the note velocities (dynamics) are preserved in the synthesized manually-aligned MIDI. It should be stressed that when transcribing the synthesized MIDI, templates from the MAPS database [6] were used, whereas when transcribing the original recording, templates from the Disklavier were utilized. When using the MAPS templates for transcribing the recordings, $\mathcal{F}$ drops to 80.43%. When simple thresholding on $\mathbf{H}'$ is employed instead of the HMM-based note tracking procedure, the average $\mathcal{F}$ for the recordings drops to 84.92%.

In Table 3, score-informed transcription results are presented, using either manually-aligned or automatically-aligned MIDI. For the manually-aligned case, it can be seen that the method reaches very high accuracy for the correctly played notes by the student, while the detection performance for missed or extra notes is diminished. This can be attributed to errors in the two transcribed piano-rolls, where additional false alarms might be produced or notes might not be detected. However, the overall performance of the method in terms of $\mathcal{F}_w$ is quite high, reaching 96.76%. When

	$\mathcal{F}_w$	Acc_w	Acc_{corr}	Acc_{md}	Acc_{fa}
Manual MIDI	96.76%	94.38%	97.40%	70.63%	75.27%
Automatic MIDI	92.93%	88.20%	93.17%	49.16%	60.49%

Table 3. Score-informed transcription results.

automatically-aligned MIDI are used, the system performance is diminished, which is expected, as additional errors from imperfect alignment are introduced. The biggest decrease in performance can be observed for the missed notes by the student. This can be attributed to the fact that the alignment algorithm might place the non-played notes at different positions compared to the ground-truth. Still, the overall performance of the system using automatically-aligned MIDI files reaches an $\mathcal{F}_w$ of 92.93%.

In order to test the performance of different components of the proposed method, comparative experiments were performed by disabling the process for detecting repeated notes, using both manually-aligned and automatically-aligned MIDI. Using the manually-aligned score, $\mathcal{F}_w = 92.79\%$ while using the automatically-aligned score, $\mathcal{F}_w = 89.04\%$. Another experiment was performed using the templates from the MAPS dataset [6] for transcribing the recording. Using the manually-aligned MIDI, $\mathcal{F}_w = 90.75\%$ while using the automatically-aligned MIDI, $\mathcal{F}_w = 85.94\%$. Without processing $prResults$ with the 'strict' and 'relaxed' piano-rolls, the score-informed transcription results using manually-aligned scores reach $\mathcal{F}_w = 94.92\%$ and using automatically-aligned scores reach $\mathcal{F}_w = 90.82\%$. A final comparative experiment was performed by utilizing only the piano-roll of the aligned ground-truth for score information, instead of also using the piano-roll of the transcribed synthesized score. In this case, using the manually-aligned score $\mathcal{F}_w = 93.55\%$ and using the automatically-aligned score $\mathcal{F}_w = 89.47\%$, which demonstrates that transcribing the synthesized score can further assist in improving performance for a score-informed transcription system.

4. CONCLUSIONS

In this paper, a system for score-informed transcription is proposed for automatic piano tutoring, which takes as input an imperfect recording and a correct score and returns an analysis of the player's performance. Methods for automatic MIDI alignment, synthesis, multi-pitch detection, and note tracking were employed and an algorithm was proposed for producing a score-informed transcription. A dataset was created specifically for the task and metrics were proposed for evaluation. Results indicate that using manually-aligned scores, the proposed method reaches high accuracy, making it useful for real-life applications. Using automatically-aligned scores produces somewhat lower performance especially when the student deviates from the score.

Score-informed transcription is an unexplored research field and several of its sub-problems could be improved, for example creating robust source-specific transcription algorithms. Future work on the proposed system will focus on a MIDI-to-audio alignment algorithm specifically tailored for the piano alignment task, operating with higher precision as this was shown to be an important factor in the proposed method's performance. In addition, the detection of structural errors such as missed or replicated segments can be achieved through a more sophisticated alignment algorithm.

5. REFERENCES

[1] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *IEEE Int. Conf. Audio, Speech and Signal Processing*, May 2011, pp. 45–48.

[2] Y. Wang and B. Zhang, "Application-specific music transcription for tutoring," *IEEE MultiMedia*, vol. 15, no. 3, pp. 70–74, July 2008.

[3] R. Macrae and S. Dixon, "Accurate real-time windowed time warping," in *11th Int. Society for Music Information Retrieval Conf.*, Aug. 2010, pp. 423–428.

[4] R. Kompass, "A generalized divergence measure for non-negative matrix factorization," *Neural Computation*, vol. 19, no. 3, pp. 780–791, 2007.

[5] 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.*, Aug. 2010, pp. 489–494.

[6] 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*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

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

[8] 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.

[9] 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.