# TOWARDS COMPLETE POLYPHONIC MUSIC TRANSCRIPTION: INTEGRATING MULTI-PITCH DETECTION AND RHYTHM QUANTIZATION

*Eita Nakamura*[1]*, Emmanouil Benetos*[2]*, Kazuyoshi Yoshii*[1]*, Simon Dixon*[2]

[1]Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
[2]Centre for Digital Music, Queen Mary University of London, London E1 4NS, UK

## ABSTRACT

Most work on automatic transcription produces "piano roll" data with no musical interpretation of the rhythm or pitches. We present a polyphonic transcription method that converts a music audio signal into a human-readable musical score, by integrating multi-pitch detection and rhythm quantization methods. This integration is made difficult by the fact that the multi-pitch detection produces erroneous notes such as extra notes and introduces timing errors that are added to temporal deviations due to musical expression. Thus, we propose a rhythm quantization method that can remove extra notes by extending the metrical hidden Markov model and optimize the model parameters. We also improve the note-tracking process of multi-pitch detection by refining the treatment of repeated notes and adjustment of onset times. Finally, we propose evaluation measures for transcribed scores. Systematic evaluations on commonly used classical piano data show that these treatments improve the performance of transcription, which can be used as benchmarks for further studies.

***Index Terms*—** Automatic transcription; multi-pitch detection; rhythm quantization; music signal analysis; statistical modelling.

## 1. INTRODUCTION

Automatic music transcription, or conversion of music audio signals into musical scores, is a fundamental and challenging problem in music information processing [1, 2]. As musical notes in scores are described with a pitch quantized in semitones and onset and offset times quantized in musical units (*score times*), it is necessary to recognize this information from audio signals. In analogy with statistical speech recognition [3], one approach is to integrate a score model and an acoustic model [4]. However, due to the huge number of possible combinations of pitches in chords, this approach is currently infeasible for polyphonic music. A more popular approach is to separately carry out multi-pitch detection (quantization of pitch) and rhythm quantization (recognition of onset and offset score times).

Multi-pitch detection methods receive a polyphonic music audio signal and output a list of notes (called *note-track data*) represented by onset and offset times (in sec), pitch, and velocity, describing the configuration of pitches for each time frame. State-of-the-art approaches typically fall into two groups: spectrogram factorization or deep learning. Spectrogram factorization methods decompose an input spectrogram typically into a basis matrix (corresponding to spectral templates of individual pitches or harmonic components) and a component activation matrix (indicating active pitches over time). These include non-negative matrix factorization (NMF), probabilistic latent component analysis (PLCA), and sparse coding [5–7]. Deep learning approaches for multi-pitch detection have used feed-forward, recurrent, and convolutional neural networks [8, 9].

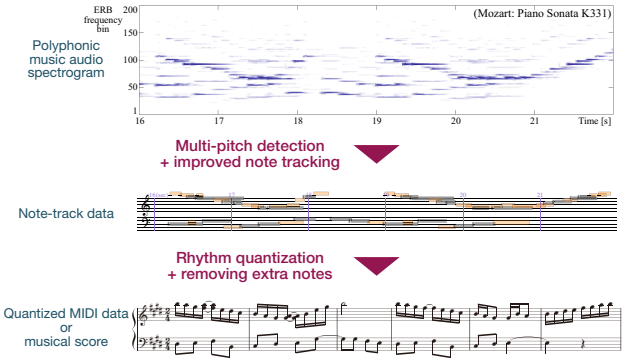


**Fig. 1**. Integration of multi-pitch detection and rhythm quantization for polyphonic transcription, with refinements on both parts.

Rhythm quantization methods receive note-track data or performed MIDI data (human performance recorded by a MIDI device) and output *quantized MIDI data* in which notes are associated with quantized onset and offset score times (in beats). Onset score times are usually estimated by removing temporal deviations in the input data, and approaches based on hand-crafted rules [10, 11], statistical models [12–18], and a connectionist approach [19] have been studied. A recent study [18] has shown that methods based on hidden Markov models (HMMs) are currently state of the art. Especially, the metrical HMM [13, 14] has the advantage of being able to estimate the metre and bar lines and avoid grammatically incorrect score representations (e.g. incomplete triplet notes). For recognition of offset score times or note values, a method using Markov random fields (MRFs) has achieved the current highest accuracy [20].

Given the recent progress of multi-pitch detection and rhythm quantization methods, we study their integration for a complete polyphonic transcription (Fig. 1). For this, we refine the frame-based multi-pitch detection part to provide a more musically meaningful output that is useful for subsequent rhythm quantization. Since note-track data typically contain erroneous notes, e.g. *extra notes* (false positives) that are not included in the ground-truth score, a rhythm quantization method that can reduce these errors is needed to avoid accumulating errors, as emphasized in [21]. Another issue is to adapt the parameters of rhythm quantization methods for note-track data that contain timing errors caused by the impreciseness of multi-pitch detection in addition to temporal deviations resulting from musical expression. Lastly, an evaluation methodology for the whole transcription process should be developed (see [22] for a recent attempt).

This work is supported by JSPS KAKENHI (Nos. 24220006, 26280089, 26700020, 15K16054, 16H01744, 16H02917, 16K00501, and 16J05486) and JST ACCEL No. JPMJAC1602. EN is supported by the JSPS Postdoctoral Research Fellowship and the long-term overseas research fund by the Telecommunications Advancement Foundation. EB is supported by a UK Royal Academy of Engineering Research Fellowship (grant no. RF/128).

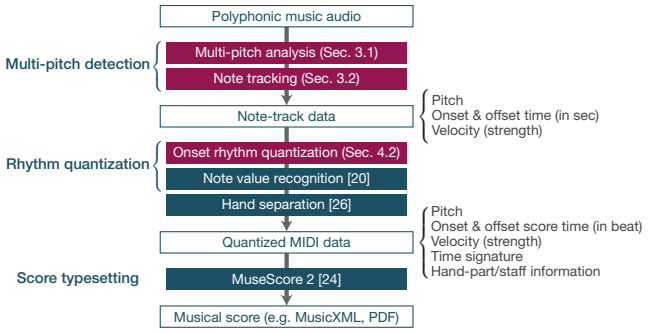


**Fig. 2**. Architecture of the proposed system.

The contributions of this study are as follows. First, we create a complete system for polyphonic transcription, from audio to rhythm-quantized musical score, which to our knowledge has not been attempted before in the literature. Second, we propose a novel method for rhythm quantization to reduce extra notes in note-track data. To incorporate top-down knowledge about musical notes like regularity in time, a generative model (named *noisy metrical HMM*) is constructed as a mixture process of a metrical HMM [13, 14] describing score-originated notes and a noise model describing the generation of extra notes. Third, we optimize the parameters for the rhythm quantization methods and examine the effect. Fourth, we refine a supervised multi-pitch detection method based on PLCA [7] by introducing processes for onset-time adjustment and repeated-note detection. Finally, we propose measures for evaluating estimated scores given ground-truth scores and report systematic evaluations on commonly used classical piano data [23], which can serve as benchmarks for further studies. We find that all of the above treatments contribute to improving accuracies (or reducing errors) and the best case significantly outperforms systems using commercial software (MuseScore 2 [24] or Finale 2014 [25]) for rhythm quantization.

## 2. SYSTEM ARCHITECTURE

The architecture of the proposed polyphonic music transcription system is illustrated in Fig. 2. Although the architecture is applicable to general polyphonic music, some components are adapted for piano transcription. The system has two main components: multi-pitch detection and rhythm transcription (see also Sec. 1).

The multi-pitch detection part (Sec. 3) consists of multi-pitch analysis (estimating multiple pitch activations for each time frame) and note tracking (detecting notes identified by onset and offset times, pitch, and velocity) and outputs note-track data. The rhythm quantization part consists of onset rhythm quantization (inferring the onset score times; Sec. 4) and note value recognition (inferring the offset score times). For note value recognition, we use the MRF method [20]. To include hand-part/staff information in quantized MIDI data, we apply the hand separation method in [26].

Finally, to obtain human/machine-readable score notation (e.g. MusicXML, PDF), we can apply the MIDI import function in score typesetting software. Specifically, we use MuseScore 2 [24], which has the ability to separate voices within each staff.

## 3. MULTI-PITCH DETECTION

### 3.1. Multi-pitch analysis

Our acoustic model is based on the work of [7], which performs multi-pitch analysis through spectrogram factorization. The model extends PLCA [27] and takes as input an equivalent rectangular bandwidth (ERB) spectrogram denoted as $V_{\omega,t}$, where $\omega$ stands for the frequency index and $t$ stands for the time index. The spectrogram has $\Omega = 250$ filters, with frequencies linearly spaced between 5 Hz and 10.8 kHz on the ERB scale and has a 23 ms hop size (instead of a variable-Q transform spectrogram used in [7]).

In the acoustic model, the input ERB spectrogram is approximated as a bivariate probability $P(\omega, t)$. This is in turn decomposed into marginal probabilities for pitch, instrument source, and sound-state activations. The model is formulated as follows:

$$P(\omega, t) = P(t) \sum_{q,p,i} P(\omega|q, p, i) P_t(i|p) P_t(p) P_t(q|p), \quad (1)$$

where $p$ is the pitch index ($p \in \{1 = \mathrm{A0}, \ldots, 88 = \mathrm{C8}\}$); $q \in \{1, \ldots, Q\}$ is the sound-state index (with $Q = 3$, denoting attack, sustain, and release); and $i \in \{1, \ldots, I\}$ is the instrument-source index (with $I = 8$, here corresponding to 8 piano models). $P(t)$ corresponds to $\sum_\omega V_{\omega,t}$, a known quantity. $P(\omega|q, p, i)$ corresponds to a pre-learned 4-dimensional dictionary of spectral templates per instrument $i$, pitch $p$, and sound state $q$. $P_t(i|p)$ refers to the instrument-source contribution for a specific pitch over time, $P_t(p)$ is the pitch activation, and $P_t(q|p)$ is the sound-state activation per pitch over time.

Unknown parameters $P_t(i|p)$, $P_t(p)$, and $P_t(q|p)$ are iteratively estimated using the expectation-maximization algorithm [28]. The dictionary $P(\omega|q, p, i)$ is considered fixed and is not updated. Sparsity constraints are incorporated on $P_t(p)$ and $P_t(i|p)$, as in [7], to control the polyphony level and the instrument-source contribution in the resulting transcription. The output of the multi-pitch analysis is given by $P(p, t) = P(t)P_t(p)$, which is the pitch activation probability weighted by the magnitude of the spectrogram.

### 3.2. Note tracking

The note-tracking process converts the non-binary time-pitch representation of $P(p, t)$ into a list of detected pitches, with an onset and offset time. To do so, $P(p, t)$ is thresholded and note events with a duration less than 30 ms are removed. Following this, we introduce a repeated-note detection process. The process uses information from $V_{\omega,t}$ and detects magnitude peaks in time-frequency regions corresponding to detected notes. Any detected peaks in those regions indicate repeated notes, and the detected note is accordingly split into smaller segments. A final onset-time adjustment step slightly adjusts the position of detected onsets by looking at magnitude changes in the frequency bins of $V_{\omega,t}$ corresponding to each specific detected pitch, within a 50 ms window around the detected onset.

## 4. ONSET RHYTHM QUANTIZATION

### 4.1. Metrical HMM for onset rhythm quantization

We first review the metrical HMM [13, 14], which consists of a score model and a performance timing model. The score model generates the beat position (onset score time relative to bar lines) of the $n$ th note $b_n \in \{0, \ldots, B-1\}$ ($B$ is the length of a bar) from the first note ($n = 1$) to the last one ($n = N$). A binary variable (*chord variable*) $g_n$ is used to describe whether the $(n-1)$th and $n$ th notes are in a chord ($g_n = \mathrm{CH}$) or not ($g_n = \mathrm{NC}$). The $b_{1:N}$ and $g_{1:N}$ are generated with the initial probability $P(b_1, g_1)$ and transition probability $P(b_n, g_n|b_{n-1})$ with a constraint $b_n = b_{n-1}$ if $g_n = \mathrm{CH}$.

The difference between the $(n-1)$th and $n$ th score times is given as

$$[b_{n-1}, b_n, g_n] = \begin{cases} 0, & g_n = \mathrm{CH}; \\ b_n - b_{n-1}, & g_n = \mathrm{NC},\ b_n > b_{n-1}; \\ b_n - b_{n-1} + B, & g_n = \mathrm{NC},\ b_n \le b_{n-1}. \end{cases}$$

The performance timing model generates onset times denoted by $t_{1:N}$. To allow tempo variations, we introduce the local tempo variables $v_{1:N}$ that are assumed to obey a Gaussian-Markov model:

$$v_1 = \mathrm{Gauss}(v_{\mathrm{ini}}, \sigma^2_{\mathrm{ini}\,v}), \quad v_n = \mathrm{Gauss}(v_{n-1}, \sigma^2_v), \tag{2}$$

where $\mathrm{Gauss}(\mu, \Sigma)$ denotes the Gaussian distribution with mean $\mu$ and variance $\Sigma$, $v_{\mathrm{ini}}$ the initial (reference) tempo, $\sigma_{\mathrm{ini}\,v}$ the standard deviation describing the amount of global tempo variation, and $\sigma_v$ the standard deviation describing the amount of tempo changes. The onset time of the $n$ th note $t_n$ is determined stochastically by the previous onset time $t_{n-1}$ and variables $v_{n-1}$, $b_{n-1}$, $b_n$, $g_n$ as [18]:

$$t_n = \begin{cases} \mathrm{Gauss}(t_{n-1} + v_{n-1}[b_{n-1}, b_n, g_n], \sigma^2_t), & g_n = \mathrm{NC}; \\ \mathrm{Exp}(t_{n-1}, \lambda_t), & g_n = \mathrm{CH}, \end{cases} \tag{3}$$

where $\mathrm{Exp}(x, \lambda)$ denotes the exponential distribution with scale parameter $\lambda$ and support $[x, \infty)$. For onset rhythm quantization, we can infer $b_{1:N}$, $g_{1:N}$, and $v_{1:N}$ from given inputs $t_{1:N}$, with the Viterbi algorithm with discretization of the tempo variables.

### 4.2. Noisy metrical HMM

The noisy metrical HMM is constructed by combining the metrical HMM and a noise model. The noise model generates onset times as

$$P_*(t_n|t') = \mathrm{Gauss}(t_n; t', \sigma^2_*), \tag{4}$$

where $\sigma_*$ is a standard deviation that is supposed to be larger than $\sigma_t$. The reference time $t'$ will be set to $\tilde{t}_n$ introduced below. To construct a model combining the metrical HMM and the noise model, we introduce a binary variable $s_n \in \{\mathrm{S}, \mathrm{N}\}$ obeying a Bernoulli distribution: $P(s_n) = \alpha_{s_n}$ ($\alpha_{\mathrm{S}} + \alpha_{\mathrm{N}} = 1$). If $s_n = \mathrm{S}$, $t_n$ is generated according to the metrical HMM in Sec. 4.1; if $s_n = \mathrm{N}$, it is generated according to Eq. (4). This process is described as a merged-output HMM [18] with a state space indexed by $z_n = (s_n, b_n, g_n, v_n, \tilde{t}_n)$ and the following transition and output probabilities (Fig. 3):

$$\begin{aligned} P(z_n|z_{n-1}) = {} & \delta_{s_n\mathrm{N}}\, \alpha_{\mathrm{N}}\, \delta_{b_{n-1}b_n} \delta_{g_{n-1}g_n} \delta(v_n - v_{n-1}) \delta(\tilde{t}_n - \tilde{t}_{n-1}) \\ & + \delta_{s_n\mathrm{S}}\, \alpha_{\mathrm{S}}\, P(b_n, g_n|b_{n-1}) P(v_n|v_{n-1}) P(\tilde{t}_n|\tilde{t}_{n-1}), \end{aligned} \tag{5}$$

$$P(t_n|z_n) = \delta_{s_n\mathrm{S}}\, \delta(t_n - \tilde{t}_n) + \delta_{s_n\mathrm{N}} P_*(t_n|\tilde{t}_n), \tag{6}$$

where $\delta$ denotes Kronecker's delta for discrete arguments and Dirac's delta function for continuous arguments and $P(\tilde{t}_n|\tilde{t}_{n-1})$ is given in Eq. (3). The $\tilde{t}_n$ memorizes the previous onset time from the signal model: $\tilde{t}_n = t_{n'}$ for the largest $n' < n$ with $\alpha_{s_{n'}} = \mathrm{S}$.

The information of duration and velocity in note-track data can be useful to identify extra notes since their distributions for extra notes have smaller means and variances compared to the case for score-originated notes. To utilize this information, we can extend the model to describe the generation of features $f_n$ for each note. (For notational simplicity, we use a unified notation $f_n$ to describe a general feature.) Their distribution is defined conditionally on $s_n$ as

$$P(f_n = f) = \delta_{s_n\mathrm{S}} P(f|\mathrm{S}) + \delta_{s_n\mathrm{N}} P(f|\mathrm{N}). \tag{7}$$

Because duration and velocity are defined for positive numbers, we here assume $P(f|s) = \mathrm{IG}(f; a_s, b_s)$, where $\mathrm{IG}(x; a, b) = b^a x^{-a-1} e^{-b/x}/\Gamma(a)$ denotes the inverse-gamma distribution with shape parameter $a$ and scale parameter $b$. (The formulation does not alter for the case of a more elaborate distribution.) The introduction of features can be seen as a modification to the probability $\alpha_{s_n}$:

$$\alpha_{s_n} \to \alpha'_{s_n} = \alpha_{s_n} \prod_{f:\ \mathrm{features}} P(f_n|s_n)^{w_f}, \tag{8}$$

where the normal model has $w_f = 1$. As the number of features we introduce is arbitrary, it is reasonable to consider $w_f$ as a variable that can be optimized by the maximum likelihood principle etc. In this study, we optimize $w_f$ according to the error rate of transcription (see Sec. 5). An inference algorithm for the noisy metrical HMM can be derived using a technique developed in [18].

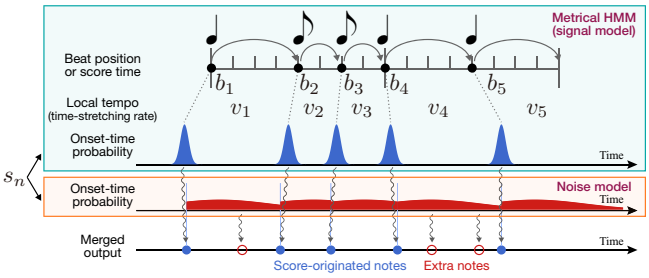


**Fig. 3**. Generation of onset times in the noisy metrical HMM.

## 5. EVALUATION

### 5.1. Evaluation measures

For evaluating the performance of the multi-pitch detection component of Sec. 3, we use the onset-based note-tracking metrics defined in [29], which are also used in the MIREX note-tracking public evaluations. These metrics assume that a note is correctly detected if its pitch is same as the ground-truth pitch and its onset time is within $\pm 50$ ms of the ground-truth onset time. Based on this rule, the precision $\mathcal{P}_{\mathrm{n}}$, recall $\mathcal{R}_{\mathrm{n}}$, and F-measure $\mathcal{F}_{\mathrm{n}}$ metrics are defined.

Measures for evaluating transcribed musical scores in comparison to the ground-truth scores have been proposed in the context of rhythm quantization [18, 20]. The rhythm correction cost (RCC) is defined as the minimum number of scale and shift operations for onset score times, which can be used for defining the onset-time error rate (ER) [18]. The offset-time ER can be defined by counting incorrect offset score times relatively to the adjacent onset score times [20]. To extend these ideas to the case with erroneous notes, we first align the estimated score to the ground-truth score using a state-of-the-art music alignment method that can also identify matched notes (i.e. correctly matched notes and notes with pitch errors), extra notes, and missing notes [30]. (A similar idea has been discussed in [22].) We notate the number of notes in the ground-truth score by $N_{\mathrm{GT}}$, that in the estimated score by $N_{\mathrm{est}}$, the number of notes with pitch errors by $N_{\mathrm{p}}$, that of extra notes by $N_{\mathrm{e}}$, and that of missing notes by $N_{\mathrm{m}}$, and define the number of matched notes as $N_{\mathrm{match}} = N_{\mathrm{GT}} - N_{\mathrm{m}} = N_{\mathrm{est}} - N_{\mathrm{e}}$. Then we define the pitch error rate $E_{\mathrm{p}} = N_{\mathrm{p}}/N_{\mathrm{GT}}$, extra note rate $E_{\mathrm{e}} = N_{\mathrm{e}}/N_{\mathrm{est}}$, missing note rate $E_{\mathrm{m}} = N_{\mathrm{m}}/N_{\mathrm{GT}}$, onset-time ER $E_{\mathrm{on}} = \mathrm{RCC}/N_{\mathrm{match}}$, and offset-time ER $E_{\mathrm{off}} = N_{\mathrm{o.e.}}/N_{\mathrm{match}}$, where the computation of RCC is explained in [18] and $N_{\mathrm{o.e.}}$ is the number of notes with an incorrect offset score time after normalization using the closest onset score time (similarly as in [20]). We define the mean of the five measures as the overall ER $E_{\mathrm{all}}$.

| Method | $\mathcal{P}_\mathrm{n}$ | $\mathcal{R}_\mathrm{n}$ | $\mathcal{F}_\mathrm{n}$ | p-val. |
|---|---|---|---|---|
| HNMF [5] | 62.3 | 76.9 | 67.9 | 0.0034 |
| PLCA-4D [7] | 79.4 | 66.0 | 71.7 | 0.080 |
| PLCA-4D-NT | 77.9 | 68.9 | 72.8 | — |

**Table 1**. Average accuracies (%) of multi-pitch detection on the MAPS-ENSTDkCl dataset, comparing acoustic models. The last column shows the p-values of $\mathcal{F}_\mathrm{n}$ with respect to PLCA-4D-NT.

| Method | $E_\mathrm{p}$ | $E_\mathrm{m}$ | $E_\mathrm{e}$ | $E_\mathrm{on}$ | $E_\mathrm{off}$ | $E_\mathrm{all}$ | p-val. |
|---|---|---|---|---|---|---|---|
| Finale 2014 | 5.6 | 24.2 | 18.3 | 53.3 | 54.0 | 31.1 | $< 10^{-5}$ |
| MuseScore 2 | 6.1 | 26.1 | 16.9 | 39.7 | 56.3 | 29.0 | $< 10^{-5}$ |
| MetHMM-def | 4.8 | 25.2 | 15.7 | 29.6 | 41.9 | 23.5 | 0.023 |
| MetHMM | 4.7 | 25.4 | 16.3 | 23.6 | 40.9 | 22.2 | 0.18 |
| NMetHMM | 4.4 | 28.6 | 13.3 | 21.6 | 39.3 | 21.4 | — |

**Table 2**. Average error rates (%) of the whole transcription systems on the MAPS-ENSTDkCl dataset, comparing rhythm quantization methods applied on the outputs of the PLCA-4D-NT method. The last column shows the p-values of $E_\mathrm{all}$ with respect to NMetHMM.

### 5.2. Experimental setup

For training the acoustic model in Sec. 3, we use a dictionary of spectral templates extracted from isolated note recordings in the MAPS database [23]. The dictionary contains sound-state templates for 8 piano models found in the database, apart from the 'ENSTDkCl' model, which is used for testing. The whole note range of the piano (A0 to C8) is used. Among the parameters of the symbolic model in Sec. 4, $P(b_1, g_1)$, $P(b_n, g_n|b_{n-1})$, $v_\mathrm{ini}$, $\sigma_{\mathrm{ini}\,v}$, and $\sigma_v$ are taken from a previous study [18] and $\alpha_s$, $a_s$, and $b_s$ are learned on the outputs of multi-pitch detection methods. The other parameters $\sigma_*$, $\sigma_t$, $\lambda_t$, and $w_f$ are optimized on the test data to maximize $E_\mathrm{all}$.

For testing the transcription system, we use 30 piano recordings in the 'ENSTDkCl' subset of the MAPS database [23], along with their corresponding ground-truth note-track data and MusicXML scores. For consistency with previous studies on multi-pitch detection, we only evaluate the first 30 s of each recording. For comparison, we also run the multi-pitch detection method based on harmonic NMF (HNMF) [5], which is based on adaptive NMF with pitch-specific spectra modelled as a weighted sum of narrowband spectra, and apply our rhythm quantization method on its outputs.

### 5.3. Results

Table 1 shows the accuracies of the multi-pitch detection methods. We refer to the original PLCA-based method of [7] as PLCA-4D and the note tracking additions of Sec. 3.2 as PLCA-4D-NT. The PLCA-4D-NT method slightly outperforms the PLCA-4D method by about 1% in terms of the note-based F-measure, with a lower precision and higher recall. The higher recall by the PLCA-4D-NT method is considered more useful for the noisy metrical HMM, which can reduce extra notes but cannot recover missing notes. The HNMF [5] method yields the highest recall but has the lowest F-measure.

Tables 2 and 3 show the results of evaluating the whole transcription method. For comparison, we run the metrical HMM with parameters taken from a previous study on rhythm quantization of performed MIDI data [18] (MetHMM-def) as well as the metrical HMM (MetHMM) and noisy metrical HMM (NMetHMM) with optimized parameters. We also compared MusicXML outputs converted from the note-track data with two commercial software for score typesetting (MuseScore 2 [24] and Finale 2014 [25]). For both outputs from

| Method | $E_\mathrm{p}$ | $E_\mathrm{m}$ | $E_\mathrm{e}$ | $E_\mathrm{on}$ | $E_\mathrm{off}$ | $E_\mathrm{all}$ | p-val. |
|---|---|---|---|---|---|---|---|
| Finale 2014 | 10.7 | 18.3 | 39.3 | 57.2 | 57.4 | 36.6 | $< 10^{-5}$ |
| MuseScore 2 | 12.3 | 19.9 | 34.4 | 49.7 | 62.6 | 35.8 | $< 10^{-5}$ |
| MetHMM-def | 10.5 | 18.6 | 33.2 | 36.5 | 44.1 | 28.6 | $< 10^{-5}$ |
| MetHMM | 9.6 | 17.5 | 33.0 | 25.5 | 42.1 | 25.5 | 0.00048 |
| NMetHMM | 7.2 | 20.8 | 19.8 | 24.1 | 41.2 | 22.6 | — |

**Table 3**. Same as Table 2 but for outputs of the HNMF method [5].

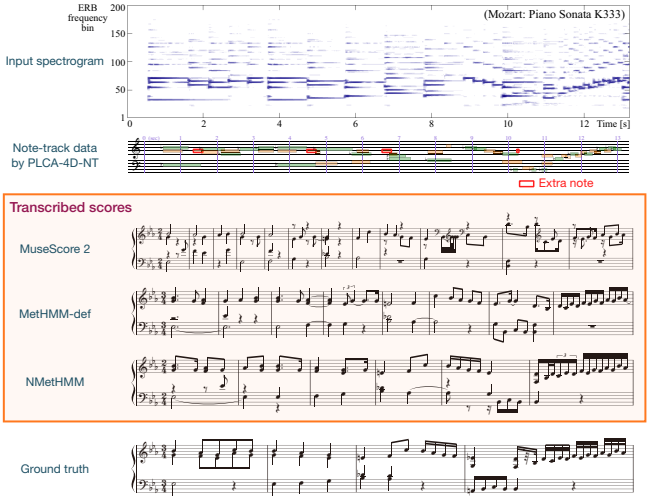


**Fig. 4**. Example transcription results (Mozart: Piano Sonata K333 in the MAPS-ENSTDkCl dataset).

the PLCA-4D-NT and HNMF methods, the NMetHMM yields the best average overall ER, which is significantly lower than the values for commercial software. We find that the optimization of the parameters of the MetHMM consistently reduces ERs. Compared to the MetHMM, the NMetHMM reduces all ERs except $E_\mathrm{m}$ and its effect is stronger for the higher-recall lower-precision outputs of the HNMF method. In Fig. 4, we find that the NMetHMM correctly removes one extra note (G4 at 10.23 s) and corrects a misalignment of chordal notes (E♭4 and G4) found in the fourth bar of the transcribed score by the MetHMM-def.

## 6. CONCLUSION

We have described integration of multi-pitch detection and rhythm quantization methods for polyphonic music transcription. We have improved the PLCA-based multi-pitch detection method by refining the note-tracking process and also proposed a rhythm quantization method based on the noisy metrical HMM aiming to remove extra notes in note-track data, both of which led to better performance of automatic transcription. We have also confirmed that the optimization of parameters of the metrical HMM describing temporal deviations reduces transcription errors.

Except for musically and acoustically simple cases, the transcribed scores obtained by our system contain musically incorrect configurations of pitches and unplayable notes and are still far from satisfactory. A limitation of the proposed noisy metrical HMM is that it does not describe the pitch information. By incorporating a pitch model, those notes with undesirable pitches are expected to be reduced. Correcting erroneous notes in note-track data other than extra notes, i.e. pitch errors and missing notes, is currently beyond the reach. Integration of a symbolic music language model with the acoustic model would be a necessary next step for this.

## 7. REFERENCES


[1] A. Klapuri and M. Davy (eds.), *Signal Processing Methods for Music Transcription*, Springer, 2006.

[2] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Challenges and future directions," *J. Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[3] S. Levinson, L. Rabiner, and M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *The Bell Sys. Tech. J.*, vol. 62, no. 4, pp. 1035–1074, 1983.

[4] C. Raphael, "A graphical model for recognizing sung melodies," in *Proc. ISMIR*, 2005, pp. 658–663.

[5] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE TASLP*, vol. 18, no. 3, pp. 528–537, 2010.

[6] K. O'Hanlon and M. D. Plumbley, "Polyphonic piano transcription using non-negative matrix factorisation with group sparsity," in *Proc. ICASSP*, 2014, pp. 3112–3116.

[7] E. Benetos and T. Weyde, "An efficient temporally-constrained probabilistic model for multiple-instrument music transcription," in *Proc. ISMIR*, 2015, pp. 701–707.

[8] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM TASLP*, vol. 24, no. 5, pp. 927–939, 2016.

[9] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, "On the potential of simple framewise approaches to piano transcription," in *Proc. ISMIR*, 2016, pp. 475–481.

[10] H. Longuet-Higgins, *Mental Processes: Studies in Cognitive Science*, MIT Press, 1987.

[11] D. Temperley and D. Sleator, "Modeling meter and harmony: A preference-rule approach," *Comp. Mus. J.*, vol. 23, no. 1, pp. 10–27, 1999.

[12] A. T. Cemgil, P. Desain, and B. Kappen, "Rhythm quantization for transcription," *Comp. Mus. J.*, vol. 24, no. 2, pp. 60–76, 2000.

[13] C. Raphael, "A hybrid graphical model for rhythmic parsing," *Artificial Intelligence*, vol. 137, pp. 217–238, 2002.

[14] M. Hamanaka, M. Goto, H. Asoh, and N. Otsu, "A learning-based quantization: Unsupervised estimation of the model parameters," in *Proc. ICMC*, 2003, pp. 369–372.

[15] H. Takeda, T. Otsuki, N. Saito, M. Nakai, H. Shimodaira, and S. Sagayama, "Hidden Markov model for automatic transcription of MIDI signals," in *Proc. MMSP*, 2002, pp. 428–431.

[16] D. Temperley, "A unified probabilistic model for polyphonic music analysis," *J. New Music Res.*, vol. 38, no. 1, pp. 3–18, 2009.

[17] A. Cogliati, D. Temperley, and Z. Duan, "Transcribing human piano performances into music notation," in *Proc. ISMIR*, 2016, pp. 758–764.

[18] E. Nakamura, K. Yoshii, and S. Sagayama, "Rhythm transcription of polyphonic piano music based on merged-output HMM for multiple voices," *IEEE/ACM TASLP*, vol. 25, no. 4, pp. 794–806, 2017.

[19] P. Desain and H. Honing, "The quantization of musical time: A connectionist approach," *Comp. Mus. J.*, vol. 13, no. 3, pp. 56–66, 1989.

[20] E. Nakamura, K. Yoshii, and S. Dixon, "Note value recognition for piano transcription using Markov random fields," *IEEE/ACM TASLP*, vol. 25, no. 9, pp. 1542–1554, 2017.

[21] E. Kapanci and A. Pfeffer, "Signal-to-score music transcription using graphical models," in *Proc. IJCAI*, 2005, pp. 758–765.

[22] A. Cogliati and Z. Duan, "A metric for music notation transcription accuracy," in *Proc. ISMIR*, 2017, pp. 407–413.

[23] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE TASLP*, vol. 18, no. 6, pp. 1643–1654, 2010.

[24] MuseScore, "MuseScore 2," https://musescore.org/en [online], accessed on: Oct. 11, 2017.

[25] MakeMusic, "Finale 2014," https://www.finalemusic.com [online], accessed on: Oct. 11, 2017.

[26] E. Nakamura, N. Ono, and S. Sagayama, "Merged-output HMM for piano fingering of both hands," in *Proc. ISMIR*, 2014, pp. 531–536.

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

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[29] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-F0 estimation and tracking systems," in *Proc. ISMIR*, 2009, pp. 315–320.

[30] E. Nakamura, K. Yoshii, and H. Katayose, "Performance error detection and post-processing for fast and accurate symbolic music alignment," in *Proc. ISMIR*, 2017, pp. 347–353.