STATISTICAL CORRECTION OF TRANSCRIBED MELODY NOTES BASED ON PROBABILISTIC INTEGRATION OF A MUSIC LANGUAGE MODEL AND A TRANSCRIPTION ERROR MODEL

Yuki Hiramatsu *Go Shibata* *Ryo Nishikimi* *Eita Nakamura* *Kazuyoshi Yoshii*

Graduate School of Informatics, Kyoto University, Japan

ABSTRACT

This paper describes a statistical post-processing method for automatic singing transcription that corrects pitch and rhythm errors included in a transcribed note sequence. Although the performance of frame-level pitch estimation has been improved drastically by deep learning techniques, note-level transcription of singing voice is still an open problem. Inspired by the standard framework of statistical machine translation, we formulate a hierarchical generative model of a transcribed note sequence that consists of a music language model describing the pitch and onset transitions of a true note sequence and a transcription error model describing the addition of deletion, insertion, and substitution errors to the true sequence. Because the length of the true sequence might be different from that of the observed transcribed sequence, the most likely sequences with possible different lengths are estimated with Viterbi decoding and the most likely length is then selected with a sophisticated language model based on a long short-term memory (LSTM) network. The experimental results show that the proposed method can correct musically unnatural transcription errors.

Index Terms— Singing transcription, music language models, statistical modeling, symbolic music processing

1. INTRODUCTION

Automatic singing transcription (AST) refers to estimating a symbolic musical score from singing voice and has been considered to be an important task from the technical and practical points of view [1–4]. In AST, pitch errors, rhythm errors, and extra/missing note errors are unavoidable because the singing voice has complicated pitch trajectories. In this study, we tackle a new research topic that aims to correct such errors included in musical scores estimated by an AST method.

To reduce transcription errors in AST, music language models that represent a probability distribution of musical scores have often been used with audio transcription models [1, 2, 5]. Given that a more typical musical score has a higher generative probability, AST methods using both language and transcription models are expected to improve the musical naturalness of estimated scores. Such language models, for example, include a Markov model representing the transitions of semitone-level pitches [1] and a metrical Markov model representing the transitions of metrical onset positions [2]. Nonetheless, the estimated score still includes a number of musically-unnatural errors because language models are considered to have limited impact when they are used in combination with transcription models. This calls for a post-processing step that performs error correction in the purely symbolic domain.

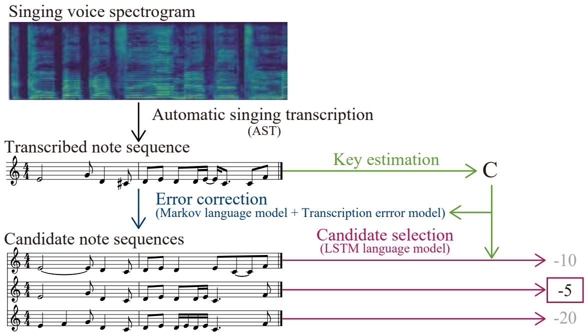
Singing voice spectrogram
Automatic singing transcription
(AST)
Transcribed note sequence
Key estimation
C
Error correction
(Markov language model + Transcription errror model)
Candidate note sequences
Candidate selection
(LSTM language model)
-10
-5
-20

Fig. 1. The overview of our statistical error correction method based on candidate estimation and selection.

In this paper, we propose a statistical error correction method that estimates multiple candidates of a true note sequence from a transcribed sequence including errors and then selects the most likely one (Fig. 1). More specifically, we formulate a hidden Markov model (HMM) that consists of a music language model representing the generative process of a true note sequence and a transcription error model representing that of a transcribed note sequence from the true sequence. Note that the lengths of the true and transcribed sequences might be different because the transcription error model represents the basic editing operations (*i.e.*, insertion, deletion, and substitution) for a true note sequence. Given a transcribed note sequence, the most likely key is estimated and the most likely true note sequences with specified lengths are then estimated with Viterbi decoding. Finally, the note sequence with the most likely length is selected with a sophisticated language model based on a long short-term memory (LSTM) network [6–8].

The main contribution of this paper is to build a statistical error correction framework based on a music language model and a transcription error model for AST in the same way as the statistical machine translation framework based on a target language model and a target-to-source back-translation model [9, 10]. Another noticeable contribution is to propose a hierarchical HMM as the transcription error model for dealing with the editing operations from the probabilistic generative point of view. Our model consists of an insertion-deletion model for alignment between true and transcribed note sequences with different lengths and a substitution model for modification of the pitches and onset positions of notes. Our method is thus more sophisticated than the basic HMM-based post-processing method that can correct only substitution errors.

2. RELATED WORK

This section reviews the use of language models for music and speech applications. Melody style conversion aims to change only the style

This work is supported in part by JST ACCEL No. JPM-JAC1602 and JSPS KAKENHI Nos. 16H01744, 19K20340, 19H04137, and 20K21813.

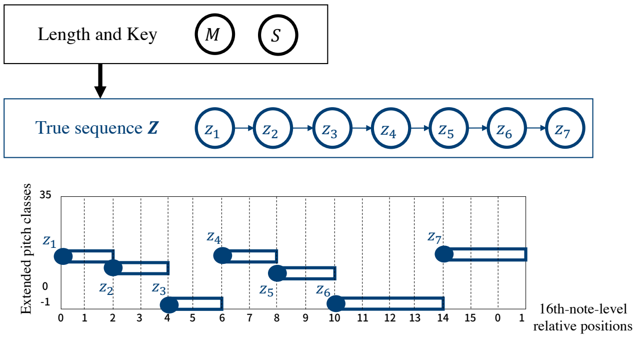

Fig. 2. The music language model that represents the generative probability of a true note sequence $\mathbf{Z}$ given a key S and a length M.

of a note sequence while preserving the original content. Inspired by the statistical machine translation framework [9], a statistical melody style conversion method was proposed by integrating music language models of individual styles and conversion models between different styles, where these models can be learned from existing melodies in an unsupervised manner [11]. The language model of a target style is based on a Markov model that represents the generative process of a target-style note sequence. The conversion model between target and source styles represents the back-translation process of a source-style note sequence from a target-style sequence. Given a source-style sequence, the most likely target-style sequence can be estimated with Viterbi decoding.

In this study, we take the same approach to error correction in AST. Specifically, we integrate a music language model representing the generative process of a true note sequence and a transcription error model representing the generative process of a transcribed sequence from a true sequence. While the conversion model proposed in [11] does not allow the number of notes to be changed, *i.e.*, deals with only substitutions, our transcription error model deals with insertions, deletions, and substitutions.

In automatic speech recognition (ASR), rescoring of N-best candidates estimated by an ASR method has often been used [12–14], where a BERT-based language model can be used for computing the generative probabilities of candidate word sequences [12]. Such a post-processing method based on a complicated yet powerful language model is useful, especially when the language model is hard to integrate with the inference process of an ASR system. In this study, we take the same approach to error correction in AST. Specifically, true sequence candidates with different lengths are estimated with a Markov language model and the best candidate is selected with an LSTM language model.

3. PROPOSED METHOD

We formulate a hierarchical generative model of a transcribed note sequence consisting of a Markov language model and a transcription error model. Given a transcribed sequence, we estimate a key with the language model and infer true sequence candidates with different lengths via Viterbi decoding. Finally, we select the best candidate by rescoring the candidates with the LSTM language model.

3.1. Problem specification

Our goal is to estimate a true note sequence $\mathbf{Z}{=}\{\mathbf{z}_m{=}(z^p_m, z^o_m)\}_{m=1}^M$ from an erroneous transcribed sequence $\mathbf{X} = \{\mathbf{x}_n = (x^p_n, x^o_n)\}_{n=1}^N$, where M is the length of $\mathbf{Z}$, N is the length of $\mathbf{X}$, each note $\mathbf{x}_n$ is represented by a pair of $x^p_n \in \{-1, 0, \cdots, 35\}$ indicating an *extended* pitch class in three octaves (-1 indicates the rest) and $x^o_n \in \{0, \cdots, 15\}$ indicating a 16th-note-level relative position in each bar, and $\mathbf{z}_n$ is defined in the same way. In this paper, the input sequence is assumed to have the time signature of 4/4 and include no key changes. The extended pitch-class sequence $\{x^p_n\}_{n=1}^N$ is computed from a MIDI note number sequence $\{\tilde{x}^p_n\}_{n=1}^N$ as follows:

$$x^p_1 = \tilde{x}_1 \mod 36, \tag{1}$$

$$x^p_n = x^p_{n-1} + (\tilde{x}^p_n - \tilde{x}^p_{n-1}) \mod 36. \tag{2}$$

This representation is convenient to capture the relative pitch dynamics of a sung melody within three octaves.

3.2. Model formulation

We formulate a joint model of a transcribed note sequence $\mathbf{X}$ and a true note sequence $\mathbf{Z}$. Let M and S be the latent length and key of $\mathbf{Z}$, where a key $S \in \{0, \ldots, 11\}$ indicates $\{C, C\#, \ldots, B\}$. Let $\mathbf{Y} = \{y_n\}_{n=1}^N$ be a latent index sequence that aligns $\mathbf{Z}$ with $\mathbf{X}$, where $y_n \in \{1, \cdots, M\}$ indicates that $\mathbf{x}_n$ is derived from $\mathbf{z}_{y_n}$, *i.e.*, $\mathbf{Z}_\mathbf{Y} \triangleq \{\mathbf{z}_{y_n}\}_{n=1}^N$ corresponds to $\{\mathbf{x}_n\}_{n=1}^N$ one by one. Using these latent variables, the full probabilistic model is given by

$$p(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, S, M) = p(\mathbf{X}, \mathbf{Y}|\mathbf{Z}, M)p(\mathbf{Z}|S, M)p(S, M), \tag{3}$$

where $p(\mathbf{X}, \mathbf{Y}|\mathbf{Z}, M)$ and $p(\mathbf{Z}|S, M)$ are the transcription error model and the music language model, respectively, and $p(S, M)$ is a prior distribution on the key S and the length M. Note that we assume $p(\mathbf{X}, \mathbf{Y}|\mathbf{Z}, S, M) = p(\mathbf{X}, \mathbf{Y}|\mathbf{Z}, M)$ in (3).

3.2.1. *Music language model*

The music language model $p(\mathbf{Z}|S, M)$ in (3) gives the generative probability of a true note sequence $\mathbf{Z}$ given a key S and a length M (Fig. 2). It is based on a standard autoregressive model as follows:

$$p(\mathbf{Z}|S, M) = p(\mathbf{z}_1|S) \prod_{m=2}^{M} p(\mathbf{z}_m|\mathbf{z}_{1:m-1}, S), \tag{4}$$

where the notation $i{:}j$ indicates a set of indices from i to j. In this paper, (4) is implemented as a first-order Markov model or an LSTM model.

In the first-order Markov model, each term of (4) is represented with a categorical distribution as follows:

$$p(\mathbf{z}_1|S) = \text{Categorical}(\mathbf{z}_1|\boldsymbol{\pi}^S), \tag{5}$$

$$\begin{aligned} p(\mathbf{z}_m|\mathbf{z}_{1:m-1}, S) &= p(\mathbf{z}_m|\mathbf{z}_{m-1}, S) \\ &= \text{Categorical}(\mathbf{z}_m|\boldsymbol{\phi}^S_{\mathbf{z}_{m-1}}), \end{aligned} \tag{6}$$

where $\boldsymbol{\pi}^S{\triangleq}\{\pi^S_{(z^p,z^o)}\}_{z^p=-1,z^o=0}^{35,15}$ is a set of the initial probabilities over the possible combinations of 37 pitches (including rest) and 16 onset positions under the key S, $\boldsymbol{\phi}^S_\mathbf{z}{\triangleq}\{\phi^S_{(z^p,z^o),(\hat{z}^p,\hat{z}^o)}\}_{\hat{z}^p=-1,\hat{z}^o=0}^{35,15}$ is a set of the transition probabilities from a note $\mathbf{z} = (z^p, z^o)$ under the key S. For standard notes $\mathbf{z} = (z^p, z^o)$ with $z^p \geq 0$, we assume the tonic invariance (transposition symmetry) as follows:

$$\pi^S_{(z^p,z^o)} = \pi^0_{(|z^p-S|_{36},z^o)}, \tag{7}$$

$$\phi^S_{(z^p,z^o),(\hat{z}^p,\hat{z}^o)} = \phi^0_{(|z^p-S|_{36},z^o),(|\hat{z}^p-S|_{36},\hat{z}^o)}, \tag{8}$$

where $|i - j|_k$ is the modulus of $|i - j|$ with respect to k.

In the LSTM model, each term of (4), *i.e.*, the categorical distribution of $\mathbf{z}_m$, is recursively predicted at each time step m by using an LSTM network. In this paper, the LSTM network is trained from existing melody note sequences that are transposed into the C major or A minor key. To evaluate the generative probability of a note sequence $\mathbf{Z}$ under an arbitrary key S, $\mathbf{Z}$ is thus transposed to the C major or A minor key and fed to the LSTM network.

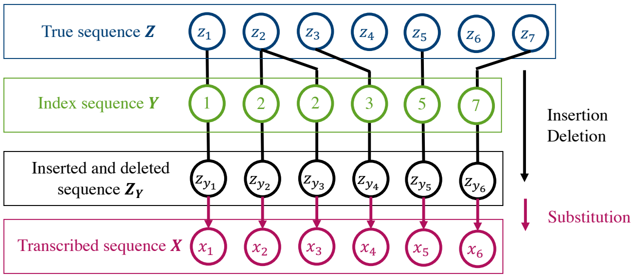

Fig. 3. The transcription error model consisting of the insertion-deletion model that aligns a true sequence $\mathbf{Z}$ with a transcribed sequence $\mathbf{X}$ through an index sequence $\mathbf{Y}$ and the substitution model that modifies an aligned sequence $\mathbf{Z_Y} = \{\mathbf{z}_{y_n}\}_{n=1}^{N}$ to $\mathbf{X}$.

3.2.2. Transcription error model

The transcription error model $p(\mathbf{X}, \mathbf{Y}|\mathbf{Z}, M)$ in (3) gives the generative probability of a transcribed note sequence $\mathbf{X}$ and an index sequence $\mathbf{Y}$ from a true sequence $\mathbf{Z}$ and a length M (Fig. 3). It is defined as a latent variable model given by

$$p(\mathbf{X}, \mathbf{Y}|\mathbf{Z}, M) = p(\mathbf{X}|\mathbf{Y}, \mathbf{Z})p(\mathbf{Y}|M), \tag{9}$$

where $p(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$ is the substitution model that represents the generative probability of a transcribed sequence $\mathbf{X}$ from a sequence of the same length $\mathbf{Z_Y}$ specified by a true sequence $\mathbf{Z}$ with an alignment $\mathbf{Y}$, and $p(\mathbf{Y}|M)$ is the insertion-deletion model that represents the probability distribution over all possible alignments between two sequences of lengths M and N ($|\mathbf{X}| = |\mathbf{Y}| = N$ and $|\mathbf{Z}| = M$).

Assuming that pitch and onset substitution errors happen independently at any position of the aligned sequence $\mathbf{Z_Y}$, the substitution model $p(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$ in (9) is factorized as follows:

$$\begin{aligned} p(\mathbf{X}|\mathbf{Y}, \mathbf{Z}) &= \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_{y_n}) \\ &= \prod_{n=1}^{N} p(x_n^p|z_{y_n}^p)p(x_n^o|z_{y_n}^o), \end{aligned} \tag{10}$$

where $p(x_n^p|z_{y_n}^p)$ and $p(x_n^o|z_{y_n}^o)$ are the pitch and onset substitution probabilities, respectively. For $x^p, z^p \geq 0$, we assume that the pitch and onset substitution probabilities only depend on the differences of the pitches and the onset positions respectively as follows:

$$p(x_n^p = x^p|z_{y_n}^p = z^p) = \chi^p_{|x^p - z^p|_{16}}, \tag{11}$$

$$p(x_n^o = x^o|z_{y_n}^o = z^o) = \chi^o_{|x^o - z^o|_{36}}, \tag{12}$$

$$p(x_n^p = -1|z_{y_n}^p = -1) = \chi^p_{-1,-1}, \tag{13}$$

$$p(x_n^p = -1|z_{y_n}^p = z^p) = \chi^p_{z^p,-1}, \tag{14}$$

where $\boldsymbol{\chi}^p = \{\chi_{\Delta_p}\}_{\Delta_p=0}^{35}$ and $\boldsymbol{\chi}^o = \{\chi_{\Delta_o}\}_{\Delta_o=0}^{15}$ are the pitch and onset substitution probabilities, respectively, and $\chi^p_{-1,-1}$ and $\{\chi^p_{z^p,-1}\}_{z^p=0}^{35}$ are hyperparameters related to the rest note.

The insertion-deletion model $p(\mathbf{Y}|M)$ in (9) represents the conditional probability of an index sequence $\mathbf{Y}$ given the true length M. It is defined as a first-order left-to-right Markov model as follows:

$$p(\mathbf{Y}|M) = p(y_1|M) \prod_{n=2}^{N} p(y_n|y_{n-1}, M). \tag{15}$$

This model is parameterized by an insertion probability η_{ins} and a deletion probability η_{del}. For $m \leq M$, the initial and transition probabilities are defined as follows:

$$p(y_1 = 2|M) = \eta_{del}, \tag{16}$$

$$p(y_n = m|y_{n-1} = m, M) = \eta_{ins}, \tag{17}$$

$$p(y_n = m|y_{n-1} = m-2, M) = \eta_{del}. \tag{18}$$

When a decent AST method is used, it is natural to enforce the sequential alignment (prohibit the cross alignment) between $\mathbf{Z}$ and $\mathbf{X}$ and more than two successive notes are unlikely to be deleted at once. In addition to the left-to-right property of this Markov model, we thus consider the following constraints:

$$y_1 \in \{1, 2\}, \tag{19}$$

$$y_n - y_{n-1} \in \{0, 1, 2\}, \tag{20}$$

$$y_N \in \{M-1, M\}. \tag{21}$$

3.3. Model training

The music language models are trained using existing melody note sequences including no errors with key annotations. These sequences can be transposed into C major or A minor. The Markov language model is trained by maximum likelihood estimation and the LSTM language model is trained such that the sequential predictive probability of the next note given the history of notes is maximized.

The probabilities of the transcription error model $\boldsymbol{\chi}^p$, $\boldsymbol{\chi}^o$, η_{ins}, and η_{del} are calculated from the alignment obtained by the method of [15] between transcribed and true note sequences. We set the rest-to-note substitution probability $1 - \chi^p_{-1,-1}$ and the note-to-rest substitution probabilities $\{\chi^p_{z^p,-1}\}_{z^p=0}^{35}$ to zero because these probabilities cannot be calculated from the alignment where the rests are ignored and these substitutions can be represented as the combinations of the deletion, insertion, and note-to-note substitutions.

3.4. Error correction as Viterbi decoding

We first estimate the key S assuming that the key of $\mathbf{X}$ is the same as the key of $\mathbf{Z}$ as follows:

$$S^* = \arg\max_{S} p_{Markov}(\mathbf{Z} = \mathbf{X}|S, M = |\mathbf{X}|). \tag{22}$$

Given the key S^*, we then estimate the index sequence $\mathbf{Y}_M^*$ and the true sequence $\mathbf{Z}_M^*$ for all possible M as follows:

$$\mathbf{Y}_M^*, \mathbf{Z}_M^* = \arg\max_{\mathbf{Y}, \mathbf{Z}} p(\mathbf{X}, \mathbf{Y}, \mathbf{Z}|S^*, M). \tag{23}$$

We finally select the most likely true note sequence $\mathbf{Z}_{M^*}^*$ as follows:

$$M^* = \arg\max_{M} \frac{1}{M} \log p_{LSTM}(\mathbf{Z}_M^*|S^*, M), \tag{24}$$

where we use the LSTM language model.

To solve (23), we interpret $p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ as an ordinary HMM of length N consisting of $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N}$ and $\mathbf{Z_Y} = \{\mathbf{z}_{y_n}\}_{n=1}^{N}$ as observed and latent variables, respectively, as follows:

$$p(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = p(\mathbf{X}|\mathbf{Y}, \mathbf{Z})p(\mathbf{Y}, \mathbf{Z}), \tag{25}$$

where $p(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$ is given by (10) and $p(\mathbf{Y}, \mathbf{Z})$ is factorized as

$$p(\mathbf{z}_{y_1}, y_1) = \begin{cases} p(\mathbf{z}_1)p(y_1) & (y_1 = 1), \\ \max\limits_{\mathbf{z}_1}\{p(\mathbf{z}_1)p(\mathbf{z}_2|\mathbf{z}_1)\}p(y_1) & (y_1 = 2), \end{cases}$$

$$p(\mathbf{z}_{y_{n+1}}, y_{n+1}|\mathbf{z}_{y_n}, y_n) =$$

$$\begin{cases} \delta_{\mathbf{z}_{y_{n+1}}, \mathbf{z}_{y_n}} p(y_{n+1}|y_n) & (y_{n+1} = y_n), \\ p(\mathbf{z}_{y_{n+1}}|\mathbf{z}_{y_n})p(y_{n+1}|y_n) & (y_{n+1} = y_n + 1), \\ \max\limits_{\mathbf{z}_{y_n+1}}\{p(\mathbf{z}_{y_{n+1}}|\mathbf{z}_{y_n+1})p(\mathbf{z}_{y_n+1}|\mathbf{z}_{y_n})\}p(y_{n+1}|y_n) & \\ & (y_{n+1} = y_n + 2). \end{cases}$$

The most likely sequence $\mathbf{Z}_{\mathbf{Y}}^*$ can be estimated efficiently with Viterbi decoding.

Table 1. Evaluation results [%] (lower E_* and higher R_{dn} are better).

	E_p	E_e	E_m	E_{on}	E_{off}	E_{all}	R_{dn}
Transcribed	8.48	17.4	10.9	39.0	31.4	21.4	95.9
Corrected (LSTM)	9.43	15.5	15.0	41.7	31.7	22.7	97.5
Corrected (Markov)	9.95	17.7	12.7	42.0	31.8	22.8	97.3
Corrected (oracle)	9.15	13.3	14.6	39.8	29.4	21.2	97.4
Ground truth	-	-	-	-	-	-	97.1

Table 2. Cross entropies [bits/note] (lower is better).

	Markov	LSTM
Transcribed	5.89	5.62
Corrected (LSTM)	4.50	4.51
Ground truth	4.18	4.63

4. EVALUATION

We report experiments conducted to evaluate transcription accuracy and musical naturalness of the corrected sequences.

4.1. Experimental conditions

For evaluation, we conducted 5-fold cross validation using 60 songs with the time signature of 4/4 and no key transposition taken from the RWC Music Database [16]. The note sequences were estimated by a convolutional neural network followed by an LSTM network. For each transcribed sequence $\mathbf{X}$ of length N, $\{[rN] : r \in \{\bar{r} - 0.1, \ldots, \bar{r} - 0.02, \bar{r}, \bar{r} + 0.02, \ldots, \bar{r} + 0.1\}\}$ were considered as the candidates of the true length M, where $\bar{r}$ was the average of ratios of true lengths to transcribed lengths. The optimal length was selected by using the LSTM language model or the Markov language model. We trained these music language models by using the melody scores of 206 Beatles songs and 328 J-pop songs.

To evaluate the proposed method, we calculated the pitch error rate E_p, the extra note rate E_e, the missing note rate E_m, the onset-time error rate E_{on}, the offset-time error rate E_{off}, and the overall error rate E_{all} [17] by comparing transcribed and corrected sequences with the ground-truth sequences. The musical naturalness was evaluated in terms of the rate of diatonic notes R_{dn} because the majority of notes should be on a scale. We considered C major scale {C, D, E, F, G, A, B}, C harmonic minor scale {C, D, Eb, F, G, Ab, B}, and the other 22 transposed scales. Because detailed key annotations were unavailable, we used as R_{dn} the maximum of the diatonic note rates computed for all scales.

4.2. Experimental results

The experimental results are listed in Table 1. Among the 60 songs, the proposed method reduced the overall error rate E_{all} for 15 songs. The overall error rate E_{all} of the proposed method with the LSTM language model was lower than that with the Markov language model only. This indicates the effectiveness of the LSTM language model in the rescoring step. If the length M was selected for each song so that the overall error rate of the song was minimized (oracle condition), E_{all} of the corrected sequences was reduced to 21.2%, which was lower than that of the transcribed sequences 21.4%. This showed the potential of the proposed method for improving the transcription accuracy. The diatonic note rate R_{dn} of the corrected sequences was higher than that of the transcribed sequences and was closer to that of the ground-truth sequences.

The cross-entropies per note obtained by the Markov and LSTM language models were shown in Table 2. The cross-entropy per note of the corrected sequences was lower than that of the transcribed sequences. The higher R_{dn} and the lower cross-entropy of corrected sequences indicate that the proposed method successfully improved the musically naturalness in exchange for the slight decrease of the transcription accuracy (trade-off problem). Three examples of error correction with the proposed method are shown in Figs. 4, 5, and 6. In Fig. 4, the proposed method corrected a pitch error and an extra note error. In Fig. 5, the proposed method deleted extra successive notes of the same pitch and the corrected sequence has a musically natural rhythm. In Fig. 6, the pitches of the corrected sequence were on a scale and were considered to be musically coherent although the proposed method made a pitch error.

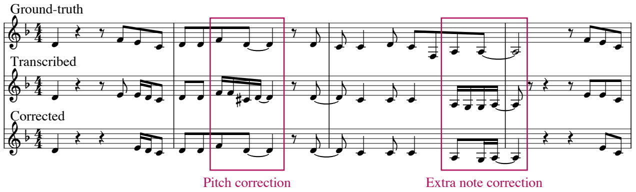

Fig. 4. The proposed method corrected a pitch error and an extra note error (RWC-MDB-P-2001 No.11).

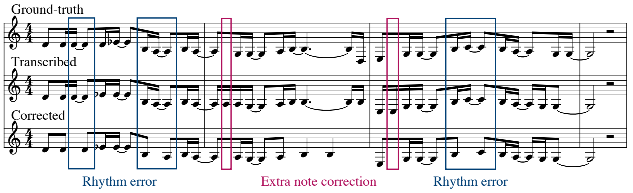

Fig. 5. The proposed method corrected extra note errors, but added rhythm errors (RWC-MDB-P-2001 No.74).

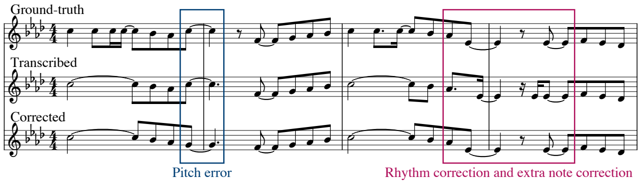

Fig. 6. The proposed method corrected a rhythm error and an extra note error, but added a pitch error (RWC-MDB-P-2001 No.80).

5. CONCLUSION

This paper presented a statistical error correction method as a post-processing step of AST. Our method is based on an HMM that consists of a Markov language model that generates a true sequence and a transcription error model that generates an erroneous transcribed sequence from a true sequence. Given a transcribed sequence, we estimate candidate note sequences via Viterbi decoding for all possible lengths and choose the optimal length based on an LSTM language model. We found that the corrected sequences were more musically natural than the transcribed sequences and the proposed method had potential for improving the transcription accuracy.

In future work, we plan to integrate the LSTM language model with the transcription error model. We believe that a post-processing correction method based on a powerful music language model is effective for AST because transcribed sequences are expected to be musically natural from the practical points of view[1].

[1] The demo page: `https://music-lsmtse.github.io`

6. REFERENCES

[1] M. Ryynänen and A. Klapuri, "Transcription of the singing melody in polyphonic music," in *ISMIR*, 2006, pp. 222–227.

[2] R. Nishikimi, E. Nakamura, M. Goto, K. Itoyama, and K. Yoshii, "Scale- and rhythm-aware musical note estimation for vocal F0 trajectories based on a semi-tatum-synchronous hierarchical hidden semi-markov model," in *ISMIR*, 2017, pp. 376–382.

[3] A. Laaksonen, "Automatic melody transcription based on chord transcription," in *ISMIR*, 2014, pp. 119–124.

[4] F. Rigaud and M. Radenen, "Singing voice melody transcription using deep neural networks," in *ISMIR*, 2016, pp. 737–743.

[5] S. Sigtia, E. Benetos, S. Cherla, T. Weyde, A. Garcez, and S. Dixon, "An RNN-based music language model for improving automatic music transcription," in *ISMIR*, 2014, pp. 53–58.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *INTERSPEECH*, 2012, pp. 194–197.

[8] A. Ycart and E. Benetos, "A study on LSTM networks for polyphonic music sequence modelling," in *ISMIR*, 2017, pp. 421–427.

[9] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[10] E. K. Ringger and J. F. Allen, "A fertility channel model for post-correction of continuous speech recognition," in *ICSLP*. IEEE, 1996, vol. 2, pp. 897–900.

[11] E. Nakamura, K. Shibata, R. Nishikimi, and K. Yoshii, "Unsupervised melody style conversion," in *ICASSP*. IEEE, 2019, pp. 196–200.

[12] J. Shin, Y. Lee, and K. Jung, "Effective sentence scoring method using BERT for speech recognition," in *ACML*, 2019, pp. 1081–1093.

[13] Y. Si, Q. Zhang, T. Li, J. Pan, and Y. Yan, "Prefix tree based N-best list re-scoring for recurrent neural network language model used in speech recognition system.," in *INTERSPEECH*, 2013, pp. 3419–3423.

[14] E. Arisoy, A. Sethy, B. Ramabhadran, and S. Chen, "Bidirectional recurrent neural network language models for automatic speech recognition," in *ICASSP*, 2015, pp. 5421–5425.

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

[16] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases," in *ISMIR*, 2002, pp. 287–288.

[17] E. Nakamura, E. Benetos, K. Yoshii, and S. Dixon, "Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization," in *ICASSP*, 2018, pp. 101–105.