DMRN+15: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2020

QUEEN MARY UNIVERSITY OF LONDON
TUE 15 DECEMBER 2020

Joint Piano-roll and Score Transcription for Polyphonic Piano Music

Lele Liu*, Veronica Morfi, and Emmanouil Benetos

Centre for Digital Music, Queen Mary University of London, UK, lele.liu@qmul.ac.uk

***Abstract*— We propose a method of joint multi-pitch detection and score transcription for polyphonic piano music. The outputs of our system include both a piano-roll representation (a descriptive transcription) and a symbolic musical notation (a prescriptive transcription). Instead of further converting MIDI transcriptions to scores, we use a multitask model combined with Convolutional Recurrent Neural Networks and Sequence-to-sequence models with attention mechanisms. We propose a reshaped score representation that outperforms a LilyPond representation both in prediction accuracy and time/memory resources, and compare different input audio spectrograms. The joint model outperforms a single task model in score transcription.**

I. INTRODUCTION

A large part of work in Automatic Music Transcription (AMT) falls under the tasks of multi-pitch detection and onset/offset detection. In this work, we discuss the problem of music audio-to-score transcription (A2S). Unlike in [1] which obtains a MIDI output in the beginning and transcribes music audio step by step, we use an end-to-end method that directly converts an audio input to a score format (see some early stage works in [2]).

In this work, we intend to extend the use of end-to-end A2S to a more general application scenario of polyphonic piano music with varying polyphony levels, as well as to support the estimation of music performance characteristics in a piano-roll format. We propose a multitask end-to-end model composed of convolutional layers, recurrent layers and sequence-to-sequence models with an attention mechanism for A2S, which is, to our knowledge, the first holistic model that transcribes polyphonic piano music into both a piano-roll format (corresponding to a descriptive notation of the music audio) and a score in Western staff notation (corresponding to a prescriptive notation of the musical audio). Additionally, we propose a new score representation for modelling polyphonic music that learns and predicts 7 times faster, uses less memory, and performs better than the LilyPond format score representation on this model. We also test the effect of using different input time-frequency representations, and the effect of combining multi-pitch detection and score transcription with a multitask model.

II. EXPERIMENTS

We carry out three experiments: 1) *comparison of time-frequency representations*, including Short-Time Fourier Transform (STFT), Mel Spectrogram, Constant-Q Transform (CQT), Harmonic Constant-Q Transform (HCQT), and Variable-Q Transform (VQT); 2) *comparison of score representations*, including a LilyPond format score representation and a Reshaped score representation (see in Figure 1); 3) *combination of piano-roll and symbolic score* in a multitask model. We use a joint model with shared convolutional layers, and separate recurrent layers/sequence-to-sequence networks for multi-pitch detection and score prediction.

*L. Liu is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by the China Scholarship Council and Queen Mary University of London.

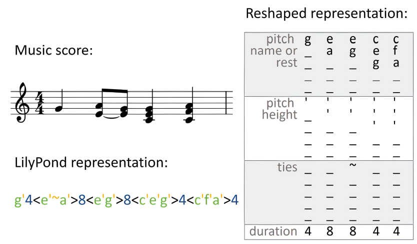

Figure 1. Example music score and corresponding LilyPond and Reshaped representation

We train and evaluate our system in a dataset with scores collected from the MusicScore website and audio recordings synthesized from the scores. Experimental results are shown in Tables 1 and 2. Among the five spectrogram types, VQT shows the best performance. The Reshaped representation runs around 7 times faster, uses around half the memory, and is slightly better than the LilyPond representation in terms of prediction accuracy. Overall, the joint model predicts better scores than a single task model.

Table 1. Benchmark F-measure of piano-roll prediction on different input representations and models.

Input representations/Models	F_f	F_{on}	F_{onoff}
STFT	89.5	81.0	61.7
Mel Spectrogram	89.0	82.1	63.0
CQT	**91.9**	85.4	67.4
HCQT	91.0	84.1	65.3
VQT	**91.9**	**85.7**	**68.5**
Piano-roll only	86.4	**67.6**	52.0
Joint	**88.0**	66.7	**53.6**

Table 2. Word error rates and MV2H [3] results in percentage for different models. LilyPond: Score-only model with LilyPond representation; Reshaped: Score-only model with Reshaped representation; Joint: Joint model with Reshaped representation.

WER	wer_{right}	wer_{left}	wer
LilyPond	38.0	39.0	38.5
Reshaped	37.8	**34.5**	**36.2**
Joint	**37.6**	35.3	36.5

MV2H	F_p	F_{voi}	F_{met}	F_{val}	F_{MV2H}
LilyPond	66.7	90.3	94.8	93.2	86.3
Reshaped	69.6	89.7	94.8	93.7	86.9
Joint	**71.1**	**90.8**	**94.9**	**94.4**	**87.8**

III. REFERENCES

[1] K. Shibata et al., “Non-local musical statistics as guides for audio-to-score piano transcription,” arXiv preprint arXiv:2008.12710, 2020.

[2] M. A. Román et al., “Data representations for autio-to-score monophonic music transcription,” Expert Systems with Applications, vol. 162, pp.113769, 2020.

[3] A. Mcleod and M. Steedman, “Evaluating automatic polyphonic music transcription,” in ISMIR, 2018, pp. 42-49.