From Audio to Symbolic Encoding

Shenli Yuan
Department of Mechanical Engineering,
Center for Computer Research in Music and Acoustics (CCRMA)
Stanford University
Stanford, CA 94305
shenliy@stanford.edu

Lingjie Kong
Department of Computer Science
Stanford University
Stanford, CA 94305
ljkong@stanford.edu

Jiushuang Guo
Department of Statistics
Stanford University
Stanford, CA 94305
jguo18@stanford.edu

1 Key Information

- **Title:** From Audio to Symbolic Encoding[1]
- **Team members:**
  - Shenli Yuan: shenliy@stanford.edu
  - Lingjie Kong: ljkong@stanford.edu
  - Jiushuang Guo: jguo18@stanford.edu
- This is a **Custom Project**
- **Mentor:** (b) We would like to request Xiaoxue Zang or Annie Hu as our mentor.

2 Research paper summary

2.1 Paper information

- **Title:** Onsets and frames: Dual-objective piano transcription
- **Authors:** Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck.
- **Publication venue:** arXiv.org
- **publication year:** 2018
- **URL:** https://arxiv.org/abs/1710.11153

[1] We initially proposed an music information project (MIR) to Chris, and Chris suggested that the custom project, though open-ended, has to somehow involve human languages. He then proposed that we could combine audio-to-midi transcription with speech recognition, because they are both conversions of audio to symbolic encoding, and might share similar properties and neural models.

2.2 Task

The task of the paper is about automatic music transcription (AMT), which aims to convert raw audio to symbolic music representation, for example, Musical Instrument Digital Interface (MIDI) [1] - an industry standard music technology protocol widely used for music encoding. Instead of broadly focusing on automatic music transcription, this paper limits the scope to automatic piano transcription (APT), which means all the raw audios will be polyphonic piano recordings. As a fundamental problem of music information retrieval (MIR), AMT is considered a difficult task even for trained human experts. Polyphonic AMT is difficult because the concurrently active notes would create complex overlap of multiple harmonics in the acoustic signal.

2.3 Objective and approach

This paper presents a supervised neural network model for APT, more specifically, transcribing piano recordings to MIDI. The proposed model, using deep convolutional and recurrent neural network (RNN), is designed to transcribe polyphonic music without prior information about the recording environment.

There have been previous works using deep learning and neural networks for AMT [2, 3]. They were both inspired by models used for other tasks. For example, convolutional neural network(CNN) for image classification, or a combined CNN and RNN model commonly used for speech recognition.

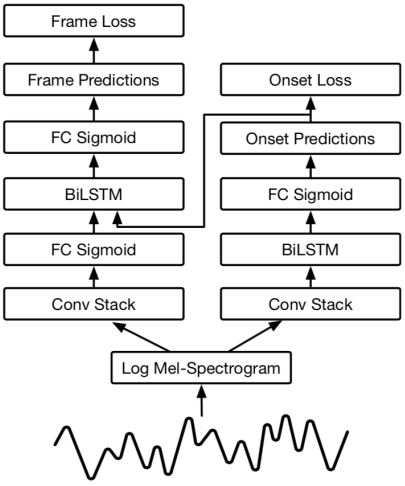

Figure 1: Onsets and Frames network diagram

Similarly as other deep learning and nerual networks based AMT models, this paper also converts the raw audio to a time-frequency representation (Log Mel-Spectrogram in this case), and use it as the input of neural networks. However, this paper improves the AMT model by emphasizing on the note onset (i.e. the very beginning of a note) detection. A dedicated note onset detector was trained and the output of the note onset detector is used as additional input for the framewise note activation detector. The reason behind such design is because the onset frame of a piano note is at the note's peak amplitude, followed by relatively sharp decay and therefore easier to identify.

The onset detector is built on top of the acoustic model presented in [2], followed by a bidirectional LSTM with 128 units in both directions, and then an 88-output sigmoid layer, the output of which represents the onset probability of each piano key.

The framewise note activation detector first have its input pass through a separate acoustic model and a fully connected sigmoid layer with 88 output. The output of this layer is concatenated with the output of the onset detector described previously. The concatenated output is then fed through a bidirectional LSTM with 128 units in both directions, followed by a sigmoid layer with 88 outputs.

At the last part of the model configuration, the proposed model was further extended by adding another stack to predict velocities for each onset.

In summary, figure 1 above sums up the onsets and frames networks we are describing.

2.4 Dataset and evaluation metrics

This paper uses the MAPS dataset [4] for training and testing. The MAPS dataset is composed of about 31 GB of CD-quality recordings in .wav format, as well as the corresponding ground truth in MIDI and text formats.

The metrics used to evaluate the model in this paper are precision, recall and F1 score at both frame level and note level. precision, recall and F1, as defined in [5], are metrics in statistical analysis commonly used to measure the test accuracy of binary classification.

2.5 Reason for choosing this paper

We are interested in applying deep learning techniques for transcribing audio (music, speech) to symbolic encoding (music language, human language). Potentially, we would like to explore the possibility of developing a generic model that solve this general problem of converting audio to symbolic encoding, or compare similarities and differences between music-specific and speech-specific tasks. The summarized paper, which fits into the scope of our interest, presents a state-of-art approach for automatic music transcription. The model proposed, though specifically tailored towards AMT tasks, can be easily modified for speech recognition tasks. In addition, the neural architecture proposed in this paper is very well presented with all the details; the code is also open-sourced in Tensorflow, which allows us to re-implement the model and make modifications quickly.

2.6 Other related papers

Before deep learning is widely studied to transcribe music audio to symbolic encoding, it has been studied and widely used for natural language processing already. Model like Connectionist Temporal Classification (CTC) uses a softmax layer to define a separate output distribution $Pr(k|t)$ at every step t along the input sequence [6]. This distribution covers the number of possible phonemes plus an extra blank symbol $\varnothing$. Different from the onsets and frames architecture [7], which utilizes an onset network to only look for note present or absent and then an separated frame network to inference the note, CTC for human language allows the network to decide whether or not to emit any label, or no label, at every time step. Unlike music which has the richest information at the very beginning and decays after the onsets, human language word has a relatively uniform distribution over time. Therefore, it further complicates the problem.

CTC defines a distribution over phoneme sequence that only depends on the acoustic input sequence. Therefore, it is an acoustic-only model. A recent update, known as RNN transducer [8], combines both a CTC network with an independent RNN that predict each phoneme given the previous predicted one. This overall yields a language and acoustic model.

With the advancement on enhancing RNN conditioned on input data through an attention mechanism in machine translation, image caption generation, and handwriting synthesis, attention has also been extended for speech recognition [9]

With inspiration from human language processing, we would like to propose attention-based onset and frame transducer architecture for music acoustic to symbolic encoding. We believe the transducer can yield both a music and acoustic model to help enhance the accuracy. Meanwhile, we would also like to apply the onset concept to human language processing to see whether it can help or not.

3 Project description

3.1 Main goals

The goals of this project is to tackle the generic problem of transcribing audio to symbolic encoding. We would like to explore both voice recognition and automatic piano transcription, their similarities and differences. Ideally, we would like to develop a generic model that would work on both tasks, or two models with minimal differences. The idea behind this is that we believe that there are intrinsic similarities and connections between these tasks, as there have been studies for human beings regarding the relationships between music and languages.

3.2 Datasets and evaluation metrics

3.2.1 Piano Transcription

For the task of piano transcription, we will be using the MAESTRO dataset [10], which contains 1184 performances, approximately 430 compositions with 172.3 hours of total audio hours. The dataset was just released in 2018, thus has not been extensively used for similar tasks, but it is an order of magnitude larger than other commonly used dataset (e.g. the MAPS dataset). In fact, followed by our summarized paper, the Onsets and frames model was later trained and tested on the MAESTRO dataset as well.

Similarly to the summarized paper, we are going to use frame-level and note-level precision, recall and F1 scores as our evaluation metrics. These metrics are commonly used for MIR tasks, and adopting them would allow us easily compare the results with previous works. The frame level precision, recall and F1 scores achieved by Onsets and Frames trained on MAESTRO dataset are 92.86, 78.46 and 84.91, respectively, while the note-level scores are 87.46, 85.58 and 86.44, respectively.

3.2.2 Speech recognition

For speech recognition, we would like to use the TIMIT dataset, which contains recordings of 630 speakers each reading ten phonetically rich sentences. The dataset includes time-aligned orthographic, phonetic and word transcriptions in addition to the raw audio recordings (16 bits and 16 kHz).

Like speech recognition with deep recurrent neural networks paper [6], we are going to use phoneme error rate (PER) and word error rate (WER) on the core test set to compare our model with the existed ones. These error rates are calculated by using Levenshtein algorithm to find the edit distance between reference sequence and hypothesis dividing the number of elements in reference.

3.3 Task and Methods

The task of our project is to provide a generic model to transcribe audio to symbolic encoding. We would like to divide our task into four detailed ones with proposed methods as below.

First, for automatic music transcription (AMT), we would like to build on top of the current state-of-the-art onsets and frames architecture [7] by introducing both transducer [8] and attention [11] inspired by natural language processing architecture. We believe that by using transducer, it can yield both a music and acoustic model to further increase the model accuracy. Meanwhile, we also believe that by using attention, it can better capture information about the source so far to resolve the information bottleneck problem. We would like to compare our introduced attention-based onsets and frames transducer architecture with classical music models through precision, recall, and F1 score .

Second, For end-to-end speech recognition, we would like to introduce an onset network on top of current CTC transducer [8] inspired by AMT model. Different from the music model which accurate onset identification are crucial because note identification are further complicated by the way that note energy decays after an onset, human language words are more uniformly distributed. However, we still believe that an independent onset network will help improve the current model in the following three reasons. First, by introducing an independent onset network, it separates the on/off prediction from CTC which predicts the number of possible phonemes plus an extra blank

symbol all together. Second, it should also speed up both training and inference to let the light onset network first decide whether there is a word embedding or not before running more intense network to decide what specific word embedding it should be. Third, it should also help resolve the post-processing problem that CTC might have in its decoding for duplicated words embedding.

Third, we would like to take advantages from both the state-of-the-art music and speech models to introduce our own model to solve generic audio to symbolic encoding problem.

Last, for most audio to symbolic encoding, a fixed window size is used to run a FFT to convert time series data into frequency domain before feeding the data into the neural network. Therefore, there is a trade off in how fine the window size should be to discretizing the input audio. Specific domain knowledge is needed to determine what might be the best window size for music as well as human language. Therefore, we would like to add another layer to let the network learn how fine the window size should be instead of treating it as a hyperparameter. This on-the-fly FFT concept will slow down the network training because we cannot preprocess the time serial data. However, we believe this will increase model accuracy.

3.4 Baselines

We would like to implement a simple RNN model as the first baseline for both APT and speech recognition tasks. In addition to that, we would like to reimplement the CNN model introduced in [2] as another baseline for APT, and the Hidden Markov Model (HMM) introduced in [12] as another baseline for speech recognition.

References

[1] Robert A Moog. Midi: musical instrument digital interface. *Journal of the Audio Engineering Society*, 34(5):394–404, 1986.

[2] Rainer Kelz, Matthias Dorfer, Filip Korzeniowski, Sebastian Böck, Andreas Arzt, and Gerhard Widmer. On the potential of simple framewise approaches to piano transcription. *arXiv preprint arXiv:1612.05153*, 2016.

[3] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon. An end-to-end neural network for polyphonic piano music transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5):927–939, 2016.

[4] Valentin Emiya, Roland Badeau, and Bertrand David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1643–1654, 2010.

[5] Tsung-Ping Chen and Li Su. Discovery of repeated themes and sections with pattern clustering. *Music Information Retrieval Evaluation eXchange (MIREX 2017)*, 2017.

[6] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. pages 6645–6649, 2013.

[7] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck. Onsets and frames: Dual-objective piano transcription. *arXiv preprint arXiv:1710.11153*, 2017.

[8] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

[9] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 577–585. Curran Associates, Inc., 2015.

[10] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the maestro dataset, 2018.

[11] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognitidon. In *Advances in neural information processing systems*, pages 577–585, 2015.

[12] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.