From Audio to Symbolic Encoding

Shenli Yuan, Lingjie Kong, Jiushuang Guo
Stanford University

Problem

Speech recognition and automatic music transcription (AMT) share the very similar nature of translating certain audio signals to specific types of symbolic encoding. Speech recognition translates human spoken languages to word or phonetic transcriptions, while AMT transcribes music to symbolic music representations such as MIDI (Musical Instrument Digital Interface).

Inspired by the similarities between AMT and speech recognition, we developed an neural network architecture that could be trained to tackle both problems.

Symbolic encodings

Music	Speech
Onset	Start
Offset	End
Note	Phoneme
Velocity	

Data

We used two different datasets for AMT and speech recognition.

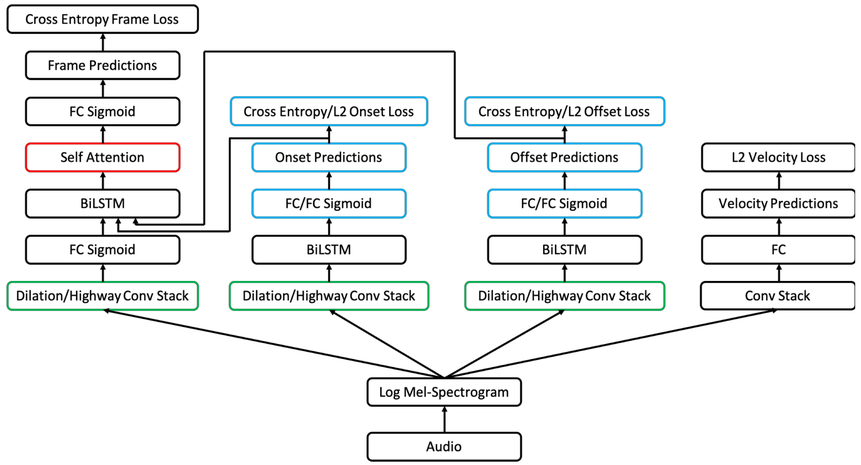
AMT
- **MAESTRO**[1] dataset
  - 1184 performances
  - ~430 compositions
  - 172.3 audio hours,
  - MIDI transcription

speech recognition
- **TIMIT**[2] dataset
  - 6300 sentences
  - 630 speakers
  - time-aligned orthographic, phonetic, word transcriptions

References

[1] 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.
[2] John S Garofolo. Timit acoustic phonetic continuous speech corpus. Linguistic Data Consortium, 1993,1993.
[3] 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:1710.11153, 2017.

Approach



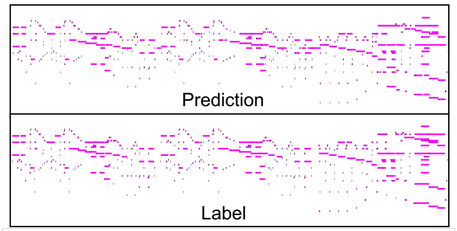
Our architecture is based on the Onsets and Frames [3] architecture. We implemented the following elements, and tested them individually and combined for the best outcome.

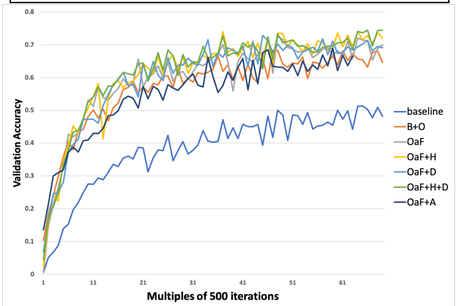
- **Baseline network**
  - Frame stack only, without onset stack, offset stack and velocity stack
- **Dilated convolutional stacks**
  - Robustness against scaling (especially in time domain), common in audio
- **Highway network**
  - Robustness against the depth of network
- **Self Attention**
  - Self attention of all notes among different frames
- **L2 loss for onset &offset time prediction**

Analysis

- Highway network and dilation conv layers improve the model; Highway network helps with deep network and dilation helps with data scaling.
- Self attention improves the model because it captures the relationship of information among different frames..
- L2 loss performs poorly, because most notes are off (0 for both onset and offset times) in most of the frames, which dominates the real onset and offset times.
- Speech recognition performs poorly, because 1) phonemes do not have distinct onset spikes, the amplitude remains relatively uniform; 2) phonetic frequency contents change significantly along with time, unlike those of music notes; 3) there is no post processing to remove the duplicated phonemes like CTC.

Results





Model	note	Note-w-o	Note-w-v	Note-w-ov	frame
baseline	0.698	0.523	0.683	0.514	0.514
OaF	0.898	0.708	0.872	0.693	0.870
OaF+H	**0.917**	**0.736**	**0.893**	0.721	**0.886**
OaF+D	0.904	0.714	0.879	0.699	0.865
OaF+H+D	0.915	**0.736**	0.891	**0.722**	0.884
OaF+A	0.911	0.732	0.878	0.714	0.881

For speech recognition, we used the phoneme error rate (PER) as the metric, but we were only achieve PER of ~0.9, which is much higher than other task-specific architectures

Conclusion

In this project, we developed a neural network architecture that was able to outperform the state-of-art architecture for AMT task. We investigated multiple variations of the architecture and compared their performances.
We were able to train models using the same architecture for the task of speech recognition, but the result were not ideal compared to the task specific models.