Musical Speech: A Transformer-based Composition Tool

Jason d'Eon[*]
Dalhousie University, Vector Institute

Sri Harsha Dumpala[*]
Dalhousie University, Vector Institute

Chandramouli Shama Sastry[*]
Dalhousie University, Vector Institute

Daniel Oore
IICSI, Memorial University of Newfoundland

Mengyu Yang
University of Toronto

Sageev Oore
Dalhousie University, Vector Institute

1 Overview

1.1 Technology

In this demo, we propose a compositional tool that generates musical lines (melody) based on the prosody of speech recorded by the user. The tool allows any user – regardless of musical training – to use their own speech to generate musical, satisfying melodies, while still being able to hear the direct connection between their recorded speech and the resulting music. This is achieved with a pipeline that combines speech-based signal processing [1, 2], musical heuristics, and a set of transformer models [3, 4] that we trained for new musical tasks. Importantly, the pipeline is designed to work with any kind of speech input and does not require a paired dataset for the training of the said transformer model.

Briefly, our approach consists of the following steps:

1. Estimate the F_0 values and loudness envelope of the speech signal.
2. Convert this into a sequence of musical constraints derived from the speech signal.
3. Apply one or more transformer models—each trained on different musical tasks or datasets—to this constraint sequence to produce musical sequences that follow or accompany the speech patterns in a variety of ways.

1.2 Audience Interaction

The demo is self-explanatory: The audience can interact with the system by either providing a live-recording using a web-based recording interface or by uploading a pre-recorded speech sample. The system then provides a visualization of the formant contours extracted from the provided speech sample, the set of note constraints obtained from the speech, and the sequence of musical notes as generated by the transformers. The audience can listen to—and interactively mix the levels (volume) of—the input speech sample, initial note sequences, and the musical sequences as generated by the transformer models.

References

[1] L Rabiner and BH Juang. Fundamentals of speech recognition. *Englewood Cliffs Publisher, New Jersey*, pages 200–232, 1993.

[*] Equal contribution

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

[2] Sri Harsha Dumpala, Jason d'Eon, and Sageev Oore. Sine-wave speech as pre-processing for downstream tasks. In *International Symposium on Frontiers of Research in Speech and Music*, 2020.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[4] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations*, 2018.