Voice Transmogrifier: Spoofing My Girlfriend's Voice

Ajay Shanker Tripathi | www.ajay.sexy | Stanford University | 12/12/2017

Goal

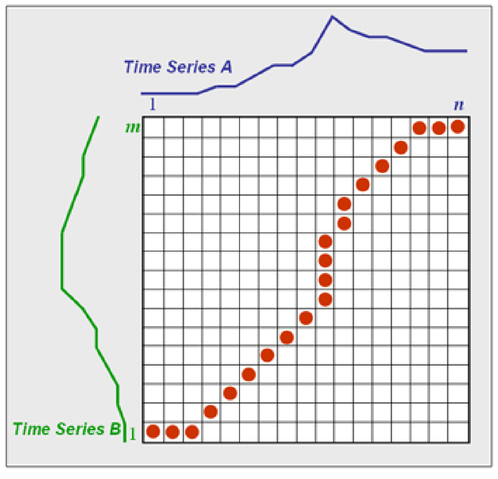
Input is my (Ajay's) voice, output is girlfriend's (Anh's) voice.

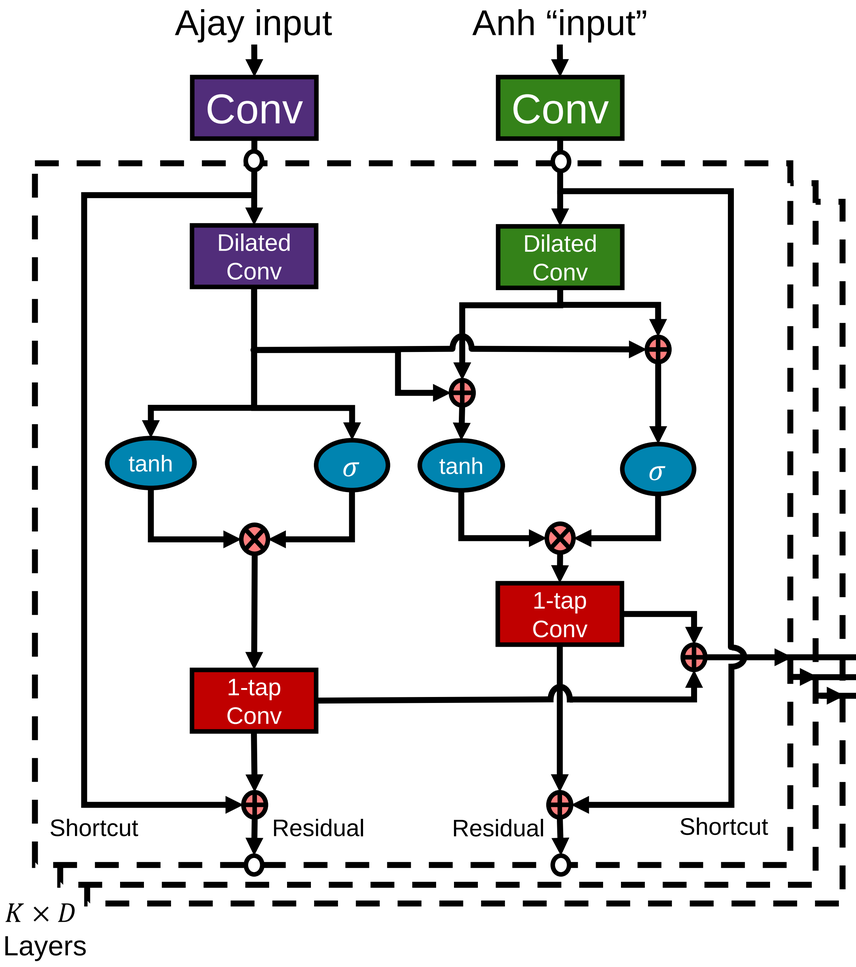
Big Challenges

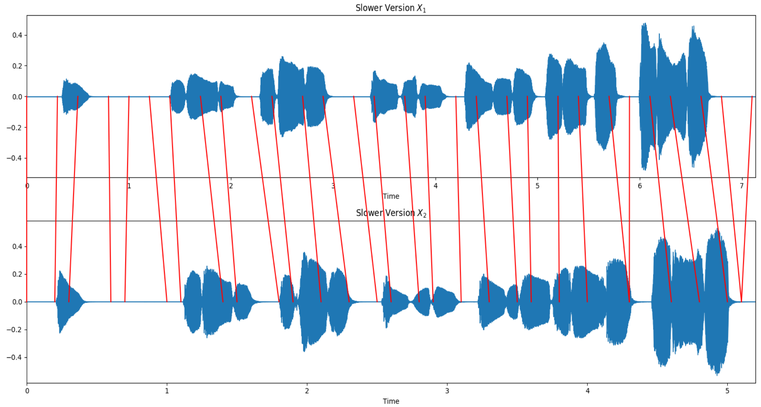
- For training data, need to align my voice input with her voice output.
- Girlfriend has a thick Vietnamese accent.
  - Non-standard pronunciation, emphasis, accent, etc.

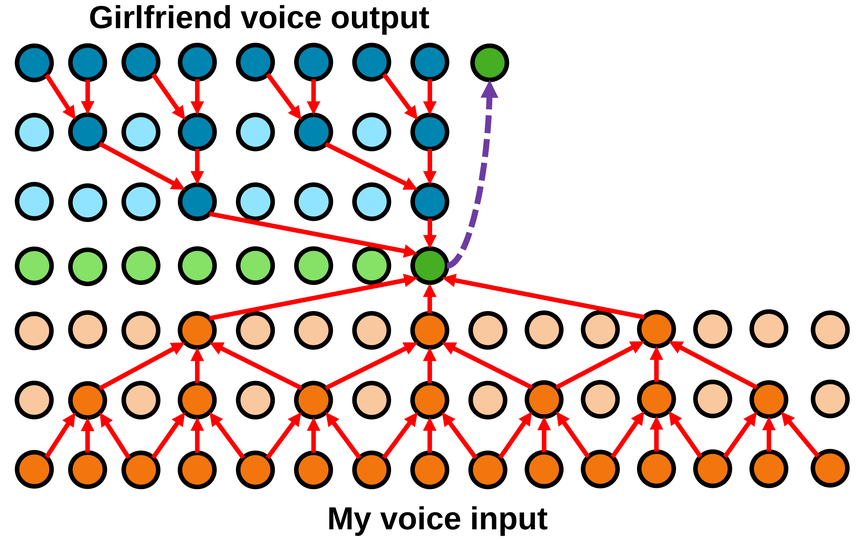
Aligning Voices

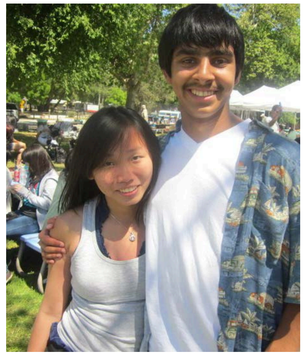
- Done by using DTW (Dynamic Time Warping).
- Audio broken into 20ms chunks, and edit-distance-like Dynamic Program is done. [2]
- Smooth monotonic interpolation done to prevent nasty jumps in speed











ANH-NET

- **A**uxiliary wave**N**et **H**armonizing neural **Net**work
- Wavenet by Deepmind is best known ANN model for audio [2]. Uses dilated convolution to achieve receptive field of 3000 samples (180 ms).
- Combine causal wavenet for girlfriend voice and non-causal wavenet for my voice.

Results

- Achieves prediction accuracy of >20%, where bins have a size $4.31 \cdot 10^{-5}$.
- Baseline scheme is LTI filter with 8 IIR taps and 17 FIR taps.
- Outperforms baseline by an order of magnitude.

[1] See librosa.github.io/librosa_gallery/auto_examples/plot_music_sync for example in Librosa python library.

[2] Original Wavenet Paper https://arxiv.org/abs/1609.03499