# GENERATING EMOTIONAL SPEECH IN CLONED VOICES

Wei Zhen Teoh
Industry Supervisor: Thibault de Boissière
Academic Supervisor: David Duvenaud

## Research Agenda

**In this project we work towards a Text-to-Speech (TTS) model for cloned voices with controllable prosody. Our goal is two-fold: to copy the vocal identity from a new speaker in a time-and-data efficient manner; and to generate artificial speech in the cloned voice with emotion that has not previously been observed in its speaker's recordings.**

**We divide this task into two stages. We train a multi-speaker neural network TTS model with controllable prosody style, which is capable of generating emotional speech for all seen speakers. We then fine-tune this model on a new unseen speaker's data.**

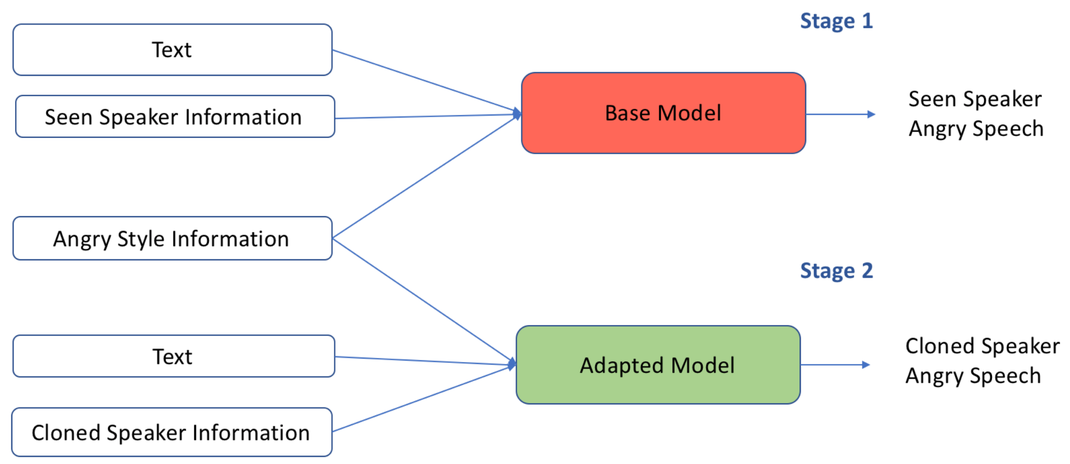


*Diagram 1*

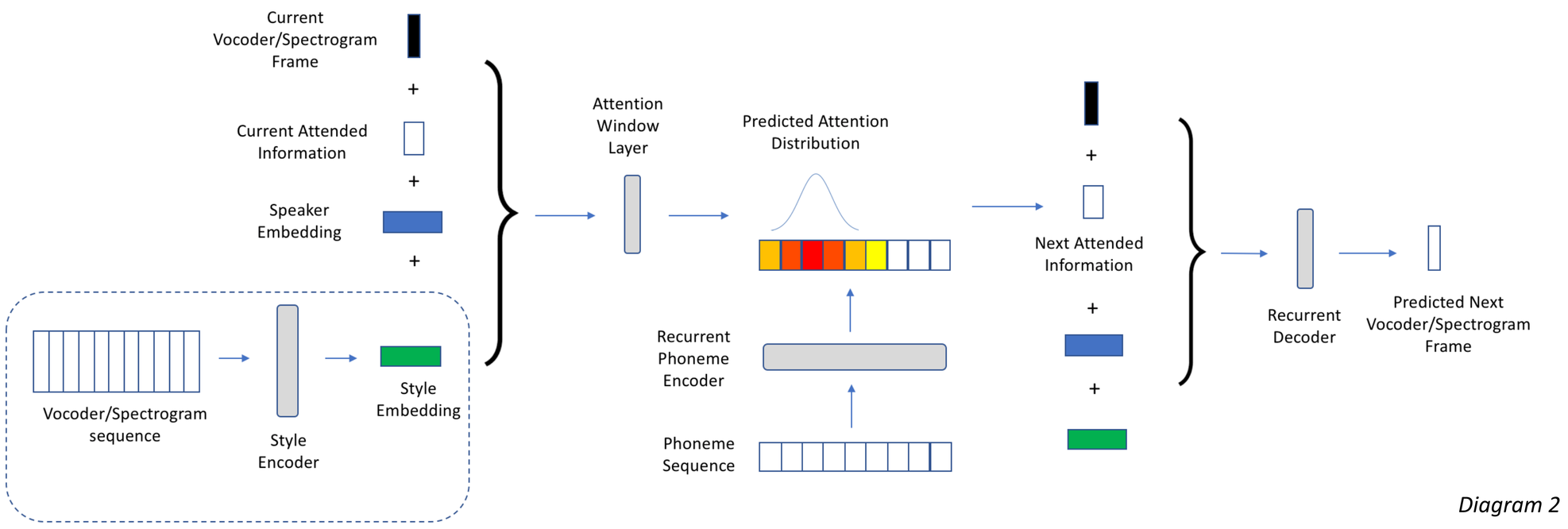


*Diagram 2*

## 1. Base Model

The backbone architecture underlying our base model is a sequence-to-sequence neural network model with attention. The model learns to generate intermediate audio representations such as WORLD[1] vocoder parameters or spectrogram, conditioned on phonetic information. Those audio representations can be decoded into audio waveform using signal processing techniques.

We augment the model with a **style encoder layer**, which **extracts the prosody information of each training utterance as a fixed length vector**. Along with speaker embedding, this information is used to **condition the decoding of phonetic information into vocoder sequence or spectrogram** in autoregressive mode. The general architecture of the model is illustrated in *Diagram 2*.

We train the model using a large dataset with many speakers. We only recorded emotional data for two of those speakers. We train the model in teacher-forcing mode to match each subsequent frame to ground-truth spectrogram or vocoder frame. The information of prosody style is therefore learned in an unsupervised manner.

## 2. Style Encoder

We adopt the work developed by Wang et al. [2] in designing the architecture of the style encoder. We use a memory augmented layer, whereby the **core prosody information components are stored as an external set of tokens**. These tokens are trained together with the base model. The full encoder architecture is illustrated in *Diagram 3*.

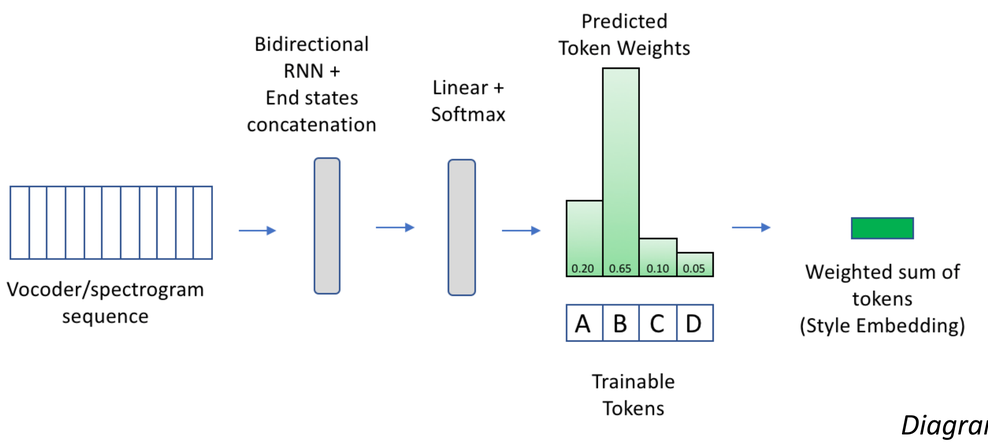


*Diagram 3*

## 3. Model Adaptation

We finetune the model with **less than 300 seconds of speech data** from a new speaker. We only require neutral emotion recordings from this speaker. A new speaker embedding is trained. We do not alter the style encoder weights at this stage**. This process takes 90 seconds of GPU time** on average.

## 4. Emotional Speech Generation

By selecting the input source for style encoder, we are able to control emotion of generated speech with cloned voices. *Diagram 4* shows the spectrograms of input sources and the corresponding generated speech audios with cloned voices. All the generated speech audios are conditioned on the same new text input. The input sources are speech audios from selected emotional speakers.

Notice that while the general signature of the speaker remains, the energy of the generated audio waveform concentrates at the lower frequency bands when conditioned on calm emotion style, and spreads over higher frequency bands when conditioned on angry emotion style.

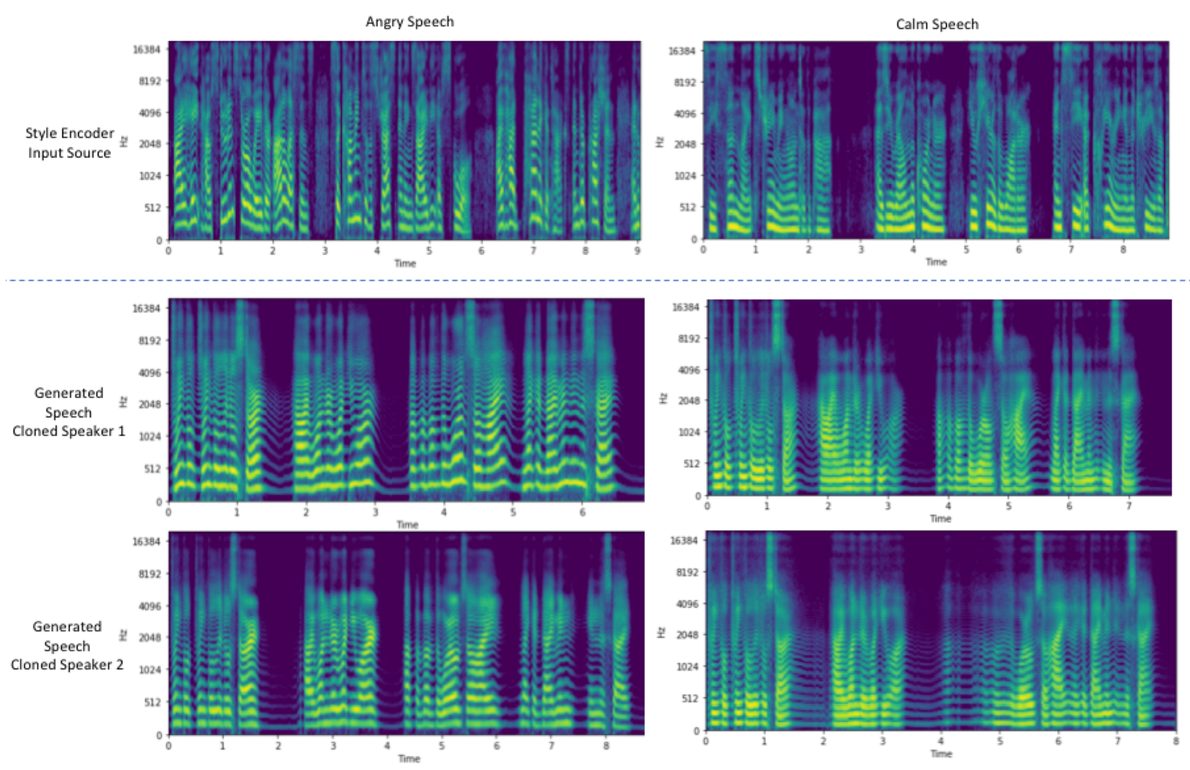


*Diagram 4*

References:
[1] M. Morise, F. Yokomori, and K. Ozawa: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877-1884, 2016
[2] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In International Conference on Machine Learning (ICML), pp. 5180–5189, 2018.