Uncovering Latent Style Factors for Expressive Speech Synthesis

Yuxuan Wang, RJ Skerry-Ryan, Ying Xiao, Daisy Stanton, Joel Shor, Eric Battenberg,
Rob Clark, Rif A. Saurous

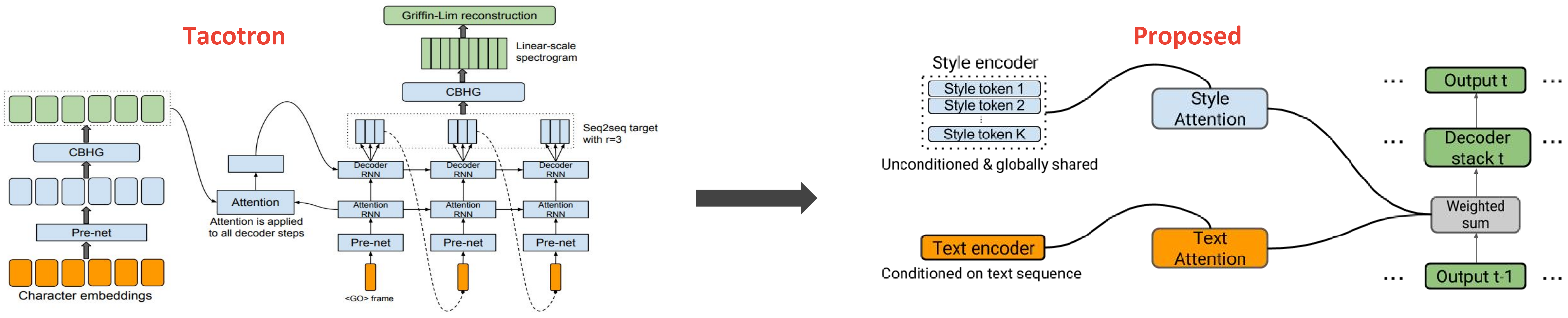
NIPS 2017, ML4Audio Workshop

ABSTRACT

Prosodic modeling is a core problem in speech synthesis. The key challenge is producing desirable prosody from textual input containing only phonetic information. In this preliminary study, we introduce the concept of “style tokens” in Tacotron, a recently proposed end-to-end neural speech synthesis model. Using style tokens, we aim to extract independent prosodic styles from training data. We show that without annotation data or an explicit supervision signal, our approach can automatically learn a variety of prosodic variations in a purely data-driven way. Importantly, each style token corresponds to a fixed style factor regardless of the given text sequence. As a result, we can control the prosodic style of synthetic speech in a somewhat predictable and globally consistent way.

MOTIVATION

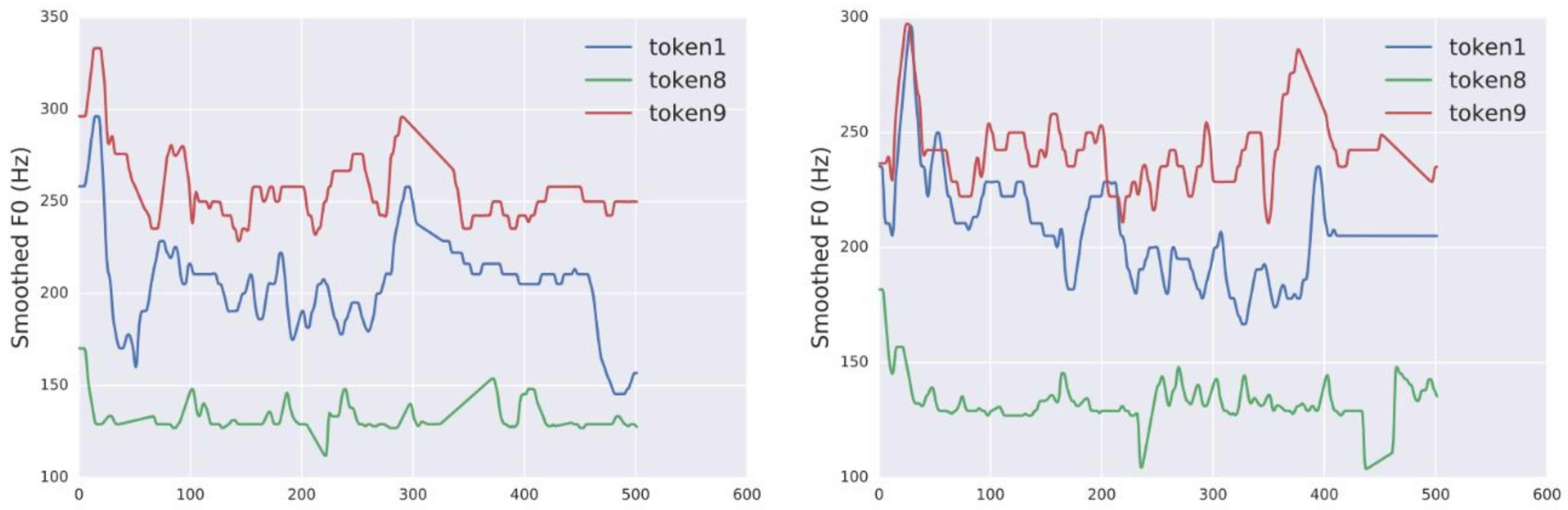
- Expressive speech synthesis
  - ○ Many applications in conversational assistants, long-form reading, etc.
- Unsupervised prosody modeling
  - ○ Prosody labels are either not given or hard to annotate. We aim to automatically learn prosody annotation/embedding
  - ○ Complements the scenario when prosody labels are given
- Enabling explicit prosody control
  - ○ Vailla end-to-end TTS lacks controllability
  - ○ Enabling style morphing, interpolation, etc.



- Add a parallel style attention pathway in addition to the text attention pathway. Use a “controller” to mix context vectors
- Token embeddings are randomly initialized and learned by backprop. Tokens are shared and unconditioned
- Style tokens are “control knobs” that we can tune to re-create prosody in different ways

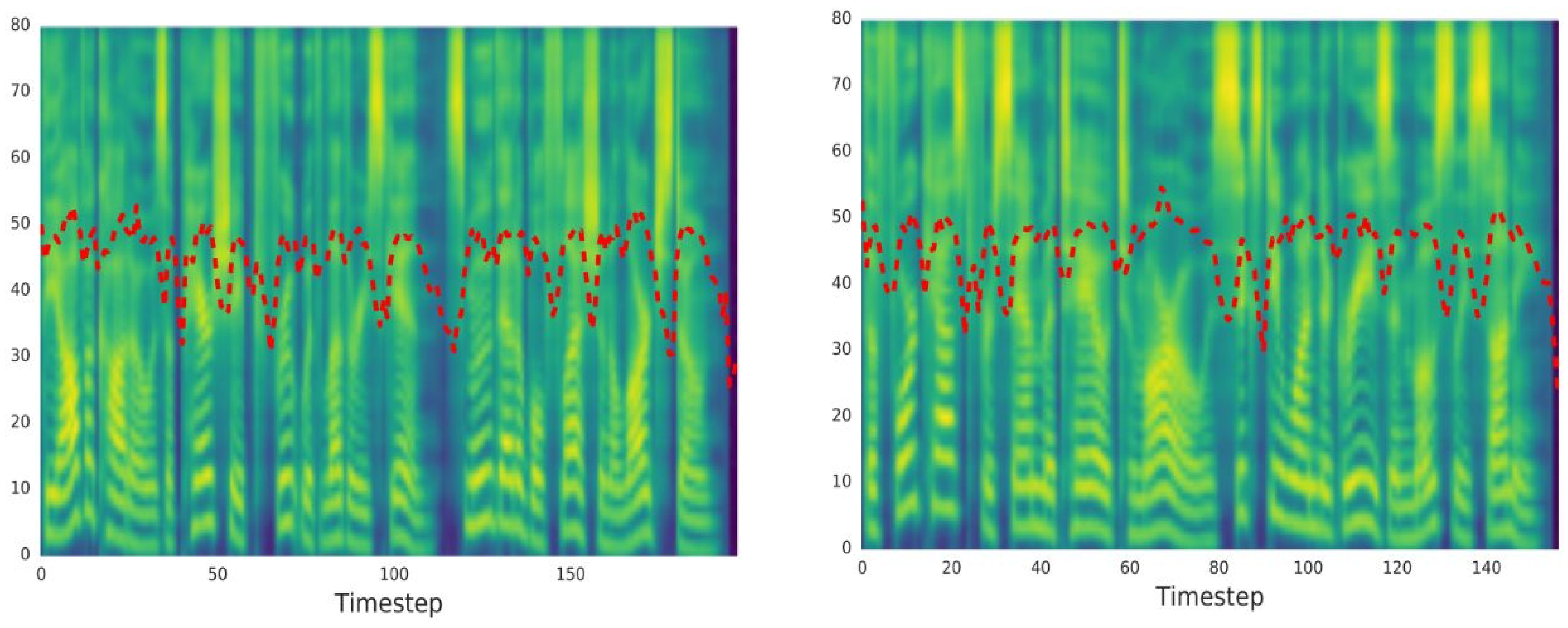
RESULTS

- Model:
  - ○ Based on an improved version of Tacotron (~4.0 MOS)
  - ○ 10 tokens, each embedding vector is 256-D
  - ○ Simple MLP controller
  - ○ Content-based style attention



Smoothed F0 trajectories of two different utterances, synthesized using 3 tokens.

- Data:
  - ○ Train on ~39 hours of data from a professional female speaker (US English)
  - ○ The speaker primarily speaks in a neutral prosody, but a small subset of the corpus uses more expression, including performing as a game show host, reading jokes and poems, etc.
  - ○ We aim to capture these variations, even though they appear in a minority of the training corpus.



Predicted mixing weights (dashed red lines) for the text attention overlaid on top of the predicted mel spectrograms.

Audio demos @ https://google.github.io/tacotron/