Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron

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

Google AI

Prosody in Speech

What's Prosody?

- Intonation, rhythm, pitch, stress, loudness.
- Conveys emotion, emphasis, and additional meaning.
- Aids understanding.

How should we say this text? It depends.

- The **cat sat** on the **mat**.
- End-to-end TTS sounds **pretty good**.

Our working definition:

> **Definition.** *Prosody is the variation in speech signals that remains after accounting for variation due to phonetics, speaker identity, and channel effects (i.e. the recording environment).*

Prosody Transfer

How to Control Prosody:

- Prosody annotations (e.g., ToBI)
- Phoneme-wise pitch, energy, duration.
- **"Say it like this" (prosody transfer)**

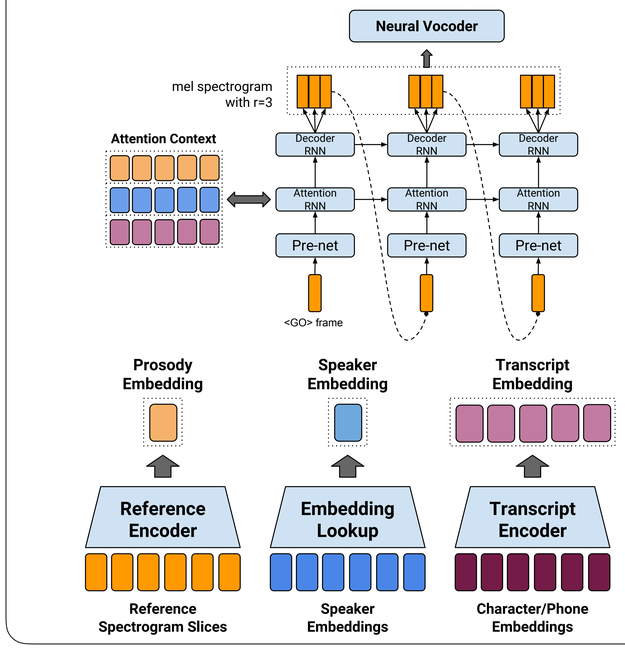
Prosody Transfer Desiderata:

- **Pitch relative** transfer (output within a speaker's natural range).
- **Robust to text modifications** (makes it scalable).

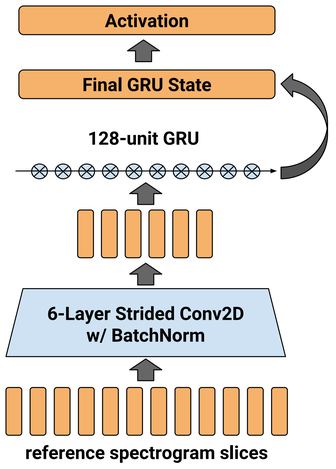
Tacotron Architecture

Specs: Phoneme inputs, GMM attention, CBHG transcript encoder, Griffin-Lim or WaveNet sample generation.

Multi-Speaker Tacotron with Prosody Embeddings



Reference Encoder



Quantitative / Subjective Results

Quantitative Similarity Measures

- **Mel Cepstral Distortion (MCD_K):** Mean squared error over first K MFCCs.

$$\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\sum_{k=1}^{K}\left(c_{t,k}-c'_{t,k}\right)^2}$$

- **F0 Frame Error (FFE):** Percentage of frames with either a >20% pitch error or a voicing decision error.

$$\frac{\sum_{t=0}^{T-1} 1\left[|p_t - p'_t| > 0.2p_t\right] 1[v_t]1[v'_t] + 1[v_t \neq v'_t]}{T}$$

Subjective Similarity Measure

- **Subjective:** Anchored side-by-side prosody similarity comparisons on a scale of [-3 to 3]

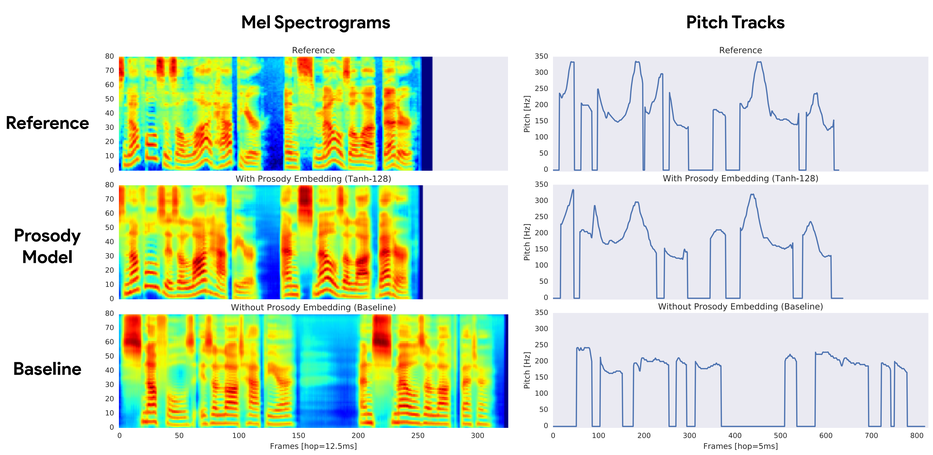
Results Table

Voice	Model	Reference	MCD_{13}	FFE	Subjective
Single-speaker	Baseline	Same speaker	10.63	53.2%	
Single-speaker	Tanh-128	Same speaker	**7.92**	**28.1%**	$\mathbf{1.611 \pm 0.164}$
Single-speaker	Baseline	Unseen speaker	11.22	59.6%	
Single-speaker	Tanh-128	Unseen speaker	**8.89**	**38.0%**	$\mathbf{1.465 \pm 0.132}$
Multi-speaker	Baseline	Same speaker	9.93	48.5%	
Multi-speaker	Tanh-128	Same speaker	**6.99**	**27.5%**	$\mathbf{1.307 \pm 0.127}$
Multi-speaker	Baseline	Seen speaker	12.37	64.2%	
Multi-speaker	Tanh-128	Seen speaker	**9.51**	**37.1%**	$\mathbf{0.871 \pm 0.138}$
Multi-speaker	Baseline	Unseen speaker	11.84	60.0%	
Multi-speaker	Tanh-128	Unseen speaker	**10.87**	**41.3%**	$\mathbf{1.146 \pm 0.246}$

Visualizing Prosody Transfer

Single-speaker model with unseen reference speaker

Text: ***Snuffles is a lot happier. And smells a lot better.***



Audio Samples

Please visit our demo page for audio samples:

https://google.github.io/tacotron/publications/end_to_end_prosody_transfer/

Or ask to hear a demo!

0:00 / 0:06

Experiment Setup

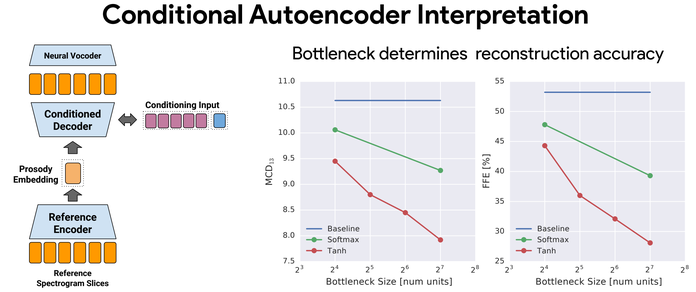
Datasets

- **Single-speaker dataset**: Audio book dataset, 147 hours, 49 books, read in an animated and emotive storytelling style (2013 Blizzard Challenge speaker).
- **Multi-speaker dataset**: Proprietary assistant-style dataset, 296 hours, 44 speakers (5 Australian, 6 British, 1 Indian, 2 Singaporean, and 30 American).

Training

- Train for >200k steps with Adam, batch size 256.
- Learning rate annealed from 1E-3 to 5E-5.
- Converges in 3-4 days.

Conditional Autoencoder Interpretation



Robustness to Text Modifications

- Prosody embeddings work even with modified text input.
- Example modifications:
  - Reference: *For the **first** time in **her** life she had been **danced tired**.*
  - Modified: *For the **last** time in his life he had been **handily embarrassed**.*
- Audio samples show that prosody is transferred to modified text.

Preservation of Target Speaker Identity

Speaker Classifier

- A simple speaker classifier is 99% accurate on ground truth and baseline synthesized audio.
- However, for a prosody model, it **picks the target speaker only 20% of the time**.
- And **picks the reference speaker 61% of the time**.

Disentangling Speaker and Prosody

- Prosody is a part of speaker identity.
- Which aspects of speaker identity should be preserved during prosody transfer?
- **Pitch-relative transfer is a feasible goal.**

Takeaways

- Prosody is a very important aspect of speech.
- Prosody transfer is a natural interface for prosody control.
- End-to-end prosody transfer works well and is robust to text transformations.
- Pitch-relative prosody transfer is a goal for future work.