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
Tacotron: End-to-End TTS

• Tacotron [Wang 2017]:
  • Convert spectrogram to samples using Griffin-Lim algorithm.
  • End-to-end TTS sounds pretty good.

• Tacotron 2 [Shen 2017]:
  • Convert spectrogram to samples using WaveNet
  • End-to-end TTS can sound really good.

• Is TTS Solved?
Prosody in Speech

- What’s prosody?
- Intonation, rhythm, pitch, stress, loudness.
- Conveys emotion, emphasis, and additional meaning.
- Examples:
  - The cat sat on the mat.
  - End-to-end TTS sounds pretty good.
- Our working definition (subtractive):

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 isn’t:
- What’s being said.
- Who’s saying it.
- Where it’s being said.

Prosody is:
- How it’s said.
Prosody Transfer

• Various way to control prosody:
  • Prosody annotations (e.g., ToBI)
  • Linguistic features (pitch, energy, duration).
  • Prosody transfer (“Say it like this”)

• Prosody transfer desired features:
  • Pitch relative transfer (output is within a speaker’s natural pitch range).
  • Robust to text transformations (one reference for many sentences, makes it scalable).
  • Meaningful embedding space (for sampling or control via other systems).
End-to-End Prosody Transfer

- Prosody Embeddings are computed using a Reference Encoder.
- Speaker embeddings are used for multi-speaker models.
- Both are broadcast-concatenated to the transcript embeddings.
- Reference and target speaker are the same during training. (but can be different during inference)
Prosody Encoder

- Input: mel spectrogram
- Strided 2D convolutions
  - (Make sure they’re padding invariant)
- RNN aggregation (GRU)
  - Summarize conv features into a single vector.
- Fully connected + activation (tanh)
  - Project vector to desired dimensionality.

Activation

Final GRU State

128-unit GRU

6-Layer Strided Conv2D w/ BatchNorm

reference spectrogram slices
Experiment Setup

• Datasets:
  
  • **Single-speaker** audiobook, 147 hours, emotive speech (Blizzard Challenge)
  
  • **Multi-speaker** voice assistant, 296 hours, 44 English speakers (Proprietary)

• (Some) Training details:
  
  • Train for at least 200k steps with batch size 256 and Adam optimizer (3-4 days).
Evaluation Metrics

• How well does the prosody embedding capture prosodic variation?

• Compare synthesized audio with reference audio.

• Quantitative metrics:

  • Mel Cepstral Distortion (MCD$_{13}$): Sum squared differences over first 13 MFCCs.

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

• Subjective evaluation:

  • Anchored side-by-side prosody similarity comparisons on a scale of [-3 to 3]
The **tanh-128** model uses a 128-dimensional prosody embedding.

<table>
<thead>
<tr>
<th>VOICE</th>
<th>MODEL</th>
<th>REFERENCE</th>
<th>MCD\textsubscript{13}</th>
<th>FFE</th>
<th>SUBJECTIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINGLE-SPEAKER</td>
<td>BASELINE</td>
<td>SAME SPEAKER</td>
<td>10.63</td>
<td>53.2%</td>
<td>1.611 ± 0.164</td>
</tr>
<tr>
<td>SINGLE-SPEAKER</td>
<td>TANH-128</td>
<td>SAME SPEAKER</td>
<td>7.92</td>
<td>28.1%</td>
<td></td>
</tr>
<tr>
<td>SINGLE-SPEAKER</td>
<td>BASELINE</td>
<td>UNSEEN SPEAKER</td>
<td>11.22</td>
<td>59.6%</td>
<td>1.465 ± 0.132</td>
</tr>
<tr>
<td>SINGLE-SPEAKER</td>
<td>TANH-128</td>
<td>UNSEEN SPEAKER</td>
<td>8.89</td>
<td>38.0%</td>
<td></td>
</tr>
<tr>
<td>MULTI-SPEAKER</td>
<td>BASELINE</td>
<td>SAME SPEAKER</td>
<td>9.93</td>
<td>48.5%</td>
<td>1.307 ± 0.127</td>
</tr>
<tr>
<td>MULTI-SPEAKER</td>
<td>TANH-128</td>
<td>SAME SPEAKER</td>
<td>6.99</td>
<td>27.5%</td>
<td></td>
</tr>
<tr>
<td>MULTI-SPEAKER</td>
<td>BASELINE</td>
<td>SEEN SPEAKER</td>
<td>12.37</td>
<td>64.2%</td>
<td>0.871 ± 0.138</td>
</tr>
<tr>
<td>MULTI-SPEAKER</td>
<td>TANH-128</td>
<td>SEEN SPEAKER</td>
<td>9.51</td>
<td>37.1%</td>
<td></td>
</tr>
<tr>
<td>MULTI-SPEAKER</td>
<td>BASELINE</td>
<td>UNSEEN SPEAKER</td>
<td>11.84</td>
<td>60.0%</td>
<td>1.146 ± 0.246</td>
</tr>
<tr>
<td>MULTI-SPEAKER</td>
<td>TANH-128</td>
<td>UNSEEN SPEAKER</td>
<td>10.87</td>
<td>41.3%</td>
<td></td>
</tr>
</tbody>
</table>
## Audio Examples

<table>
<thead>
<tr>
<th>Text</th>
<th>Reference</th>
<th>Baseline</th>
<th>Prosody Embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single-speaker model: Reference from unseen speaker</strong>&lt;br&gt;The past, the present, and the future walk into a bar. It was tense.</td>
<td>Aus F</td>
<td>Les</td>
<td>Les</td>
</tr>
<tr>
<td><strong>Multi-speaker model: Reference from seen speaker</strong>&lt;br&gt;<em>Is that Utah travel agency?</em>&lt;br&gt;Only one was deployed, while they need a hundred teams.</td>
<td>Aus F</td>
<td>US F</td>
<td>GB F</td>
</tr>
<tr>
<td><strong>Multi-speaker model: Reference from unseen speaker</strong>&lt;br&gt;<em>It will be good for both of you.</em>&lt;br&gt;I've swallowed a pollywog.</td>
<td>Les</td>
<td>Aus F</td>
<td>GB F</td>
</tr>
</tbody>
</table>

More audio examples available at: [https://google.github.io/tacotron/publications/end_to_end_prosody_transfer/](https://google.github.io/tacotron/publications/end_to_end_prosody_transfer/)
Is Speaker Identity Preserved?

• Simple speaker classifier is 99% accurate on ground truth and baseline output.
• But for the prosody model, it only chooses the target speaker 20% of the time.
  • (Chooses the reference speaker 61% of the time.)
• Speaker identity is entangled with prosody in a complicated way.
• Preserving a target speaker’s pitch range is a more concrete goal.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Baseline</th>
<th>Transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female-male</td>
<td><img src="Sound_Female-Male.png" alt="Sound" /></td>
<td><img src="Sound_Baseline.png" alt="Sound" /></td>
</tr>
<tr>
<td>Male-female</td>
<td><img src="Sound_Male-Female.png" alt="Sound" /></td>
<td><img src="Sound_Baseline.png" alt="Sound" /></td>
</tr>
</tbody>
</table>
## Robustness to Text Transformations

<table>
<thead>
<tr>
<th>Text</th>
<th>Reference</th>
<th>Baseline</th>
<th>Prosody Embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference: “I can now,” said the <strong>Leopard</strong>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perturbed: “I can now,” said the <strong>Porcupine</strong>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference: <em>For the first time in her life</em> she had been <strong>danced</strong> <em>tired</em>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perturbed: <em>For the last time in his life</em> he had been <strong>handily</strong> <strong>embarrassed</strong>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference: <strong>Second</strong>--Her family was very <strong>ancient</strong> and <strong>noble</strong>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perturbed: <strong>First</strong>--Her family was very <strong>sarcastic</strong> and <strong>horrible</strong>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference: Never again shall <strong>Eleanor Lavish</strong> be a <strong>friend</strong> of mine.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perturbed: Never again shall <strong>Bartholomew Bigglesby</strong> be a <strong>son</strong> of mine.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference: <strong>Alice</strong> was not much surprised at this, <strong>she</strong> was getting so used to <strong>queer things happening</strong>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perturbed: <strong>Eric</strong> was not much surprised at this, <strong>he</strong> was getting so used to <strong>TensorFlow breaking</strong>.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Come check out our poster (#43) for more.

• A final fun example!

• There are no examples of singing in the single-speaker training data.

• What if the reference contains singing?

Text:
Sweet dreams are made of these.  
Friendly Assistants who work hard to please.

Reference  Baseline  Prosody Embedding

More audio examples available at: https://google.github.io/tacotron/publications/end_to_end_prosody_transfer/
Summary

- 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.
- **Stick around for the Style Tokens talk next!**
References


Extra Slides
Tacotron Configuration

• Transcript Encoder:
  • Phoneme inputs
  • CBHG [Wang 2017]

• Attention Mechanism:
  • GMM [Graves 2013]

• Sample Generation:
  • Griffin-Lim or WaveNet
Visual Comparisons

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