Location-Relative Attention Mechanisms For Robust Long-Form Speech Synthesis

Eric Battenberg, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, Tom Bagby

Google Research
Attention in Tacotron

Transcript Encoder

Transcript/Phone Embeddings

{h_j}

Attention

Attention is repeated for each decoder step

s_i

c_i

Mel Spectrogram with Reduction Factor=2

Decoder RNN

Attention RNN

Pre-net

<GO> frame

Neural Vocoder
Attention in Tacotron

- Computing the attention weights and context vector.

Encoder states
(Transcript Embeddings)
\{h_j\}

Attention RNN state
(Query)
\(s_i\)

Attention weights
\(\alpha_i = \text{Attention}(s_i, \{h_j\}, \alpha_{i-1})\)

Context vector
(Glimpse)
\(c_i = \sum_j \alpha_{i,j} h_j\)
Attention Mechanisms for Tacotron

- Common attention mechanisms:
  - Tacotron → Content-based Additive [Bahdanau, 2015]
  - Tacotron 2 → Hybrid Location-Sensitive [Chorowski, 2015]
- However, these **content-based** attention mechanisms sometimes lead to:
  - Missing or repeating words.
  - Incomplete synthesis (stopping early).
  - Inability to generalize to longer utterances.
Addressing Attention Problems

- **Monotonic hard alignment mechanisms**
  - [Raffel, 2017], [Zhang, 2018], [He, 2019].
  - + Online, linear-time when using hard alignments.
  - + Improved alignment speed/stability, reduction in synthesis errors.
  - - Recursion required to marginalize across hard alignments.
  - - Reduced synthesis quality in hard alignment mode.
  - Still content-based.

- **GMM-based mechanisms**
  - Based on [Graves, 2013] original sequence-to-sequence work.
  - Attention weights computed using a mixture of Gaussians.
  - Location-relative, not content-based.
GMM-Based Mechanisms

- Attention weights computed using mixture of 1D Gaussians.
- Params computed from $s_i$ only.  (Location-relative)
- Monotonic alignment via forward-only movement of means.
- In the paper, we test 5 GMM-based variants.
- The best performing was **GMMv2b**:
  - Uses softplus (instead of exp) to compute positive parameters.
  - Uses biases to encourage:
    - Forward movement of means.
    - Initial standard deviations of 10.
GMM-Based Mechanisms

- Issues with GMM Attention:
  - Lack of strict monotonicity.
    - A wide Gaussian can look “backward” (or too far forward).
  - Discretization of continuous PDF → Attention weights don't sum to 1.
    - Can lead to "holes and spikes" in attention trajectory if decoder lingers on an encoder step.
Additive Energy-Based Mechanisms

- Transform energies to weights using softmax. \( \alpha_i = \text{softmax}(e_i) \)

- Content-Based Additive [Bahdanau, 2015] (Tacotron 1)
  \( e_{i,j} = v^T \tanh(Ws_i + \boxed{Vh_j + b}) \)

- Hybrid Location-Sensitive [Chorowski, 2015] (Tacotron 2)
  \( e_{i,j} = v^T \tanh(Ws_i + \boxed{Vh_j + Uf_{i,j} + b}) \)
  \( f_i = \mathcal{F} \ast \alpha_{i-1} \)

- Unlike GMM attention, these are both \textbf{content-based} (and not location-relative).
Dynamic Convolution Attention (DCA)

- Also in Additive Energy-based Family.
- Static (but learned) filters.
- Dynamically-computed filters.
- Fixed prior filter.

- Attributes
  - Inputs: $s_i, \alpha_{i-1}$ (Location-relative, not content-based)
  - Normalized weights, unlike GMM-based.

$$\alpha_i = \text{softmax}(e_i)$$

$$e_{i,j} = v^\top \tanh(U f_{i,j} + T g_{i,j} + b) + p_{i,j}$$

$$f_i = \mathcal{F} \ast \alpha_{i-1}$$

$$g_i = \mathcal{G}(s_i) \ast \alpha_{i-1}, \quad \mathcal{G}(s_i) = V_g \tanh(W_g s_i + b_g)$$

$$p_i = \log(\mathcal{P} \ast \alpha_{i-1})$$
DCA Prior Filter

- Prior filter is a single fixed causal FIR filter.
- We set the taps using the PMF of beta-binomial distribution.
  - Length-11 filter with a mean of 1.
- Prior filter disallows backward movement and excessive forward movement.
- Repeated application quantifies uncertainty in initial alignment.

\[ e_{i,j} = v^T \tanh(Uf_{i,j} + Tg_{i,j} + b) + p_{i,j} \]

\[ p_i = \log(P \ast \alpha_{i-1}) \]
Experiment Setup

- Compare GMM-based and Additive Energy-based families.
- Fixed Tacotron model, but we vary the Attention function.
  - Separately-trained WaveRNN as neural vocoder.
- Datasets
  - Lessac (single-speaker audiobook, 2013 Blizzard Challenge).
    - Train = 37 hours (<5 sec utts), Test = 935 utts.
  - LJ Speech (single-speaker audiobook)
    - Train = 23 hours (<10 sec utts), Test = 130 utts.
- Experiments
  - Alignment speed and consistency during training.
  - In-domain naturalness.
  - Generalization to long utterances.
Alignment Speed/Consistency

- For each mechanism, we run 10 identical trials of 10k training steps.
- Measure MCD-DTW between ground-truth test set and predicted outputs.
- When MCD-DTW drops, model has aligned with text.
In-Domain Naturalness

- Crowd-sourced MOS naturalness ratings.
- Test set: Hold-out from same dataset.

<table>
<thead>
<tr>
<th></th>
<th>Lessac</th>
<th>LJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-Based</td>
<td>4.07 ± 0.08</td>
<td>4.19 ± 0.06</td>
</tr>
<tr>
<td>Location-Sensitive</td>
<td>4.31 ± 0.06</td>
<td>4.34 ± 0.06</td>
</tr>
<tr>
<td>GMMv2b</td>
<td>4.32 ± 0.06</td>
<td>4.29 ± 0.06</td>
</tr>
<tr>
<td>DCA</td>
<td>4.31 ± 0.06</td>
<td>4.33 ± 0.06</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>4.64 ± 0.04</td>
<td>4.55 ± 0.04</td>
</tr>
</tbody>
</table>

- Content-Based slightly worse.
  - Occasional catastrophic attention failures on longer utts.
- Others produced equivalent scores.
  - → No degradation from location-relative mechanisms.
Generalization to **Long** Utterances

- Google Cloud Speech-To-Text\(^1\) used to produce output transcripts.
- Character Error Rate reported (ASR-based eval).

\(^1\) https://cloud.google.com/speech-to-text
Generalization to Long Utterances

- Audio examples

Off camera, he frequently quipped to friends and acquaintances that SCOOP was an acronym for Sensationalism Can Ordinarily Outgun Professionalism. There were reports of a crazy cult leader somewhere out in the California desert who was claiming to be Jesus Christ and had managed to dupe a few prominent personalities, one of whom was Otis Chandler, into assisting Him to promote His scam.

Many more audio examples at:

https://google.github.io/tacotron/publications/location_relative_attention
Discussion

- GMMv2b and DCA able to generalize to very long utterances.
  - While preserving naturalness on shorter utterances.
  - Enables synthesis of entire paragraphs or long sentences.
- Simple to implement, with no dynamic programming to marginalize over alignments.
- Align very quickly during training.
- Compared to GMMv2b, DCA:
  - Can more easily bound its receptive field (due to the prior filter).
  - Has normalized attention weights.
- For monotonic alignment tasks (e.g., TTS, ASR), location-relative attention mechanisms work quite well and should be strongly considered.
Thank You!

Location-Relative Attention Mechanisms For Robust Long-Form Speech Synthesis

Eric Battenberg, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, Tom Bagby

Google Research

Be sure to check out the audio examples at:

https://google.github.io/tacotron/publications/location_relative_attention
References