ICASSP 2020

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

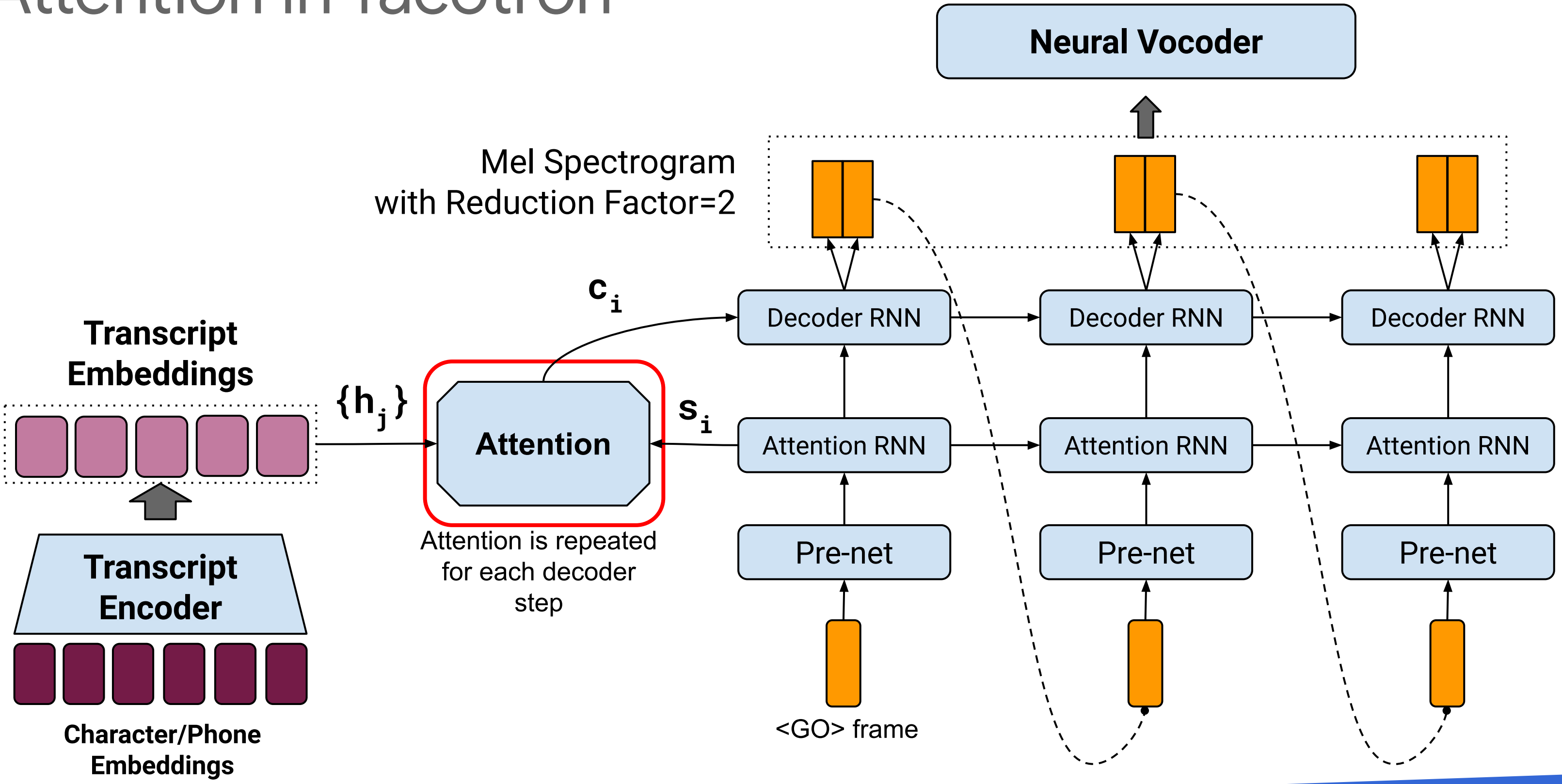
Eric Battenberg, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, Tom Bagby

Google Research

Audio Test

Google

Attention in Tacotron



Attention in Tacotron

- Computing the attention weights and context vector.

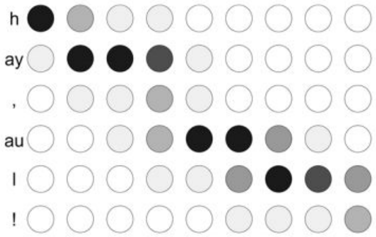
Encoder states (Transcript Embeddings)	$\{\boldsymbol{h}_j\}$
Attention RNN state (Query)	$\boldsymbol{s}_i$
Attention weights	$\boldsymbol{\alpha}_i = \mathrm{Attention}(\boldsymbol{s}_i, \{\boldsymbol{h}_j\}, \boldsymbol{\alpha}_{i-1})$
Context vector (Glimpse)	$\boldsymbol{c}_i = \sum_j \alpha_{i,j} \boldsymbol{h}_j$

Attention Mechanisms for Tacotron

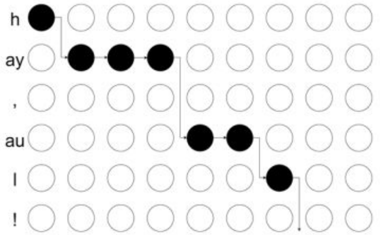
- Common attention mechanisms:
  - Tacotron ➞ Content-based Additive [Bahdanau, 2015]
  - Tacotron 2 ➞ Hybrid Location-Sensitive [Chorowski, 2015]
- However, these **content-based** attention mechanisms sometimes lead to:
  - Missing or repeating words.
  - Incomplete synthesis (stopping early).
  -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

Addressing Attention Problems

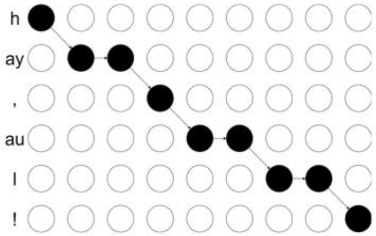
- **Monotonic hard alignment mechanisms**
  - [Raffel, 2017], [Zhang, 2018], [He, 2019].
  - \+ Online, linear-time when using hard alignments.
  - \+ Improved alignment speed/stability, reduction in synthesis errors.
  - \- Recursion required to marginalize across hard alignments.
  - \- Reduced synthesis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



Soft Attention



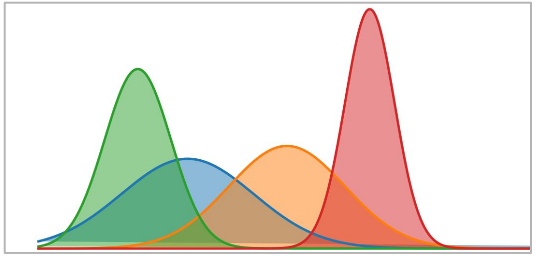
Monotonic Attention



Stepwise Monotonic Attention

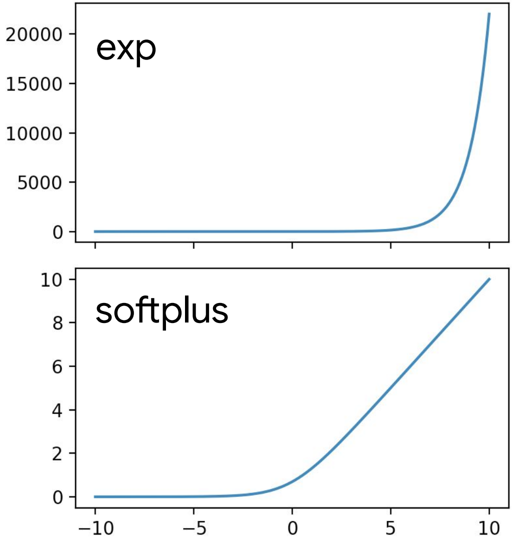
[He, 2019]

GMM-Based Mechanisms



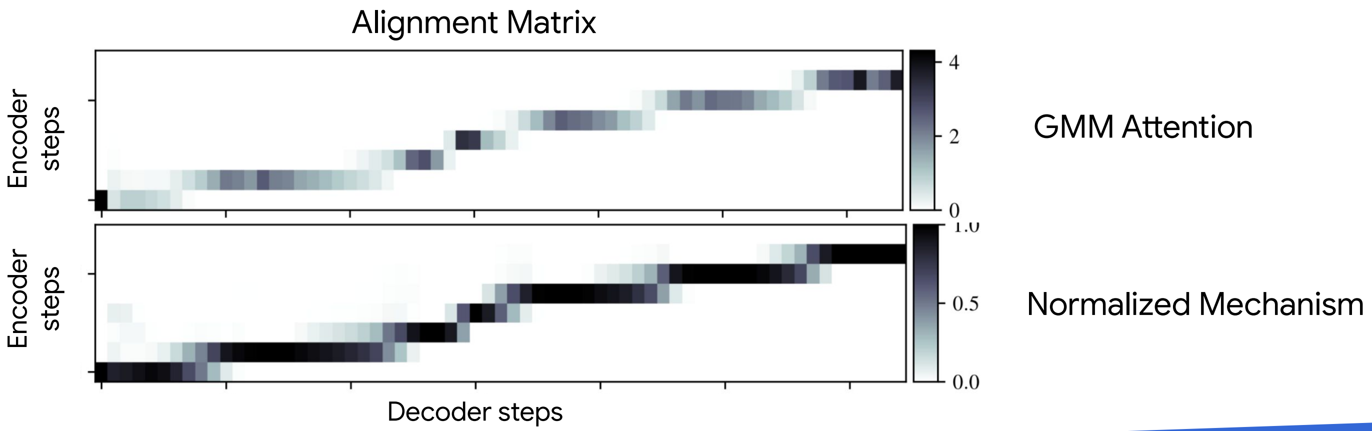
- Attention weights computed using mixture of 1D Gaussians.
- Params computed from $\mathbf{s}_i$ only. (**Location-relative**)
- Monotonic alignment via forward-only movement of means.

- 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. $\boldsymbol{\alpha}_i = \mathrm{softmax}(\boldsymbol{e}_i)$

- Content-Based Additive [Bahdanau, 2015]
  (Tacotron 1)

$$e_{i,j} = \boldsymbol{v}^\intercal \tanh(W\boldsymbol{s}_i + \boxed{V\boldsymbol{h}_j} + \boldsymbol{b})$$

- Hybrid Location-Sensitive [Chorowski, 2015]
  (Tacotron 2)

$$e_{i,j} = \boldsymbol{v}^\intercal \tanh(W\boldsymbol{s}_i + \boxed{V\boldsymbol{h}_j} + U\boldsymbol{f}_{i,j} + \boldsymbol{b})$$

$$\boldsymbol{f}_i = \mathcal{F} * \boldsymbol{\alpha}_{i-1}$$

- Unlike GMM attention, these are both **content-based** (and not location-relative).

Dynamic Convolution Attention (DCA)

$$\boldsymbol{\alpha}_i = \mathrm{softmax}(\boldsymbol{e}_i)$$

- Also in Additive Energy-based Family.

$$e_{i,j} = \boldsymbol{v}^\intercal \tanh(\boxed{U\boldsymbol{f}_{i,j}} + \boxed{T\boldsymbol{g}_{i,j}} + \boldsymbol{b}) + \boxed{p_{i,j}}$$

- Static (but learned) filters.

$$\boldsymbol{f}_i = \mathcal{F} * \boldsymbol{\alpha}_{i-1}$$

- Dynamically-computed filters.

$$\boldsymbol{g}_i = \mathcal{G}(\boldsymbol{s}_i) * \boldsymbol{\alpha}_{i-1}, \quad \mathcal{G}(\boldsymbol{s}_i) = V_{\mathcal{G}} \tanh(W_{\mathcal{G}}\boldsymbol{s}_i + \boldsymbol{b}_{\mathcal{G}})$$

- Fixed prior filter.

$$\boldsymbol{p}_i = \log(\mathcal{P} * \boldsymbol{\alpha}_{i-1})$$

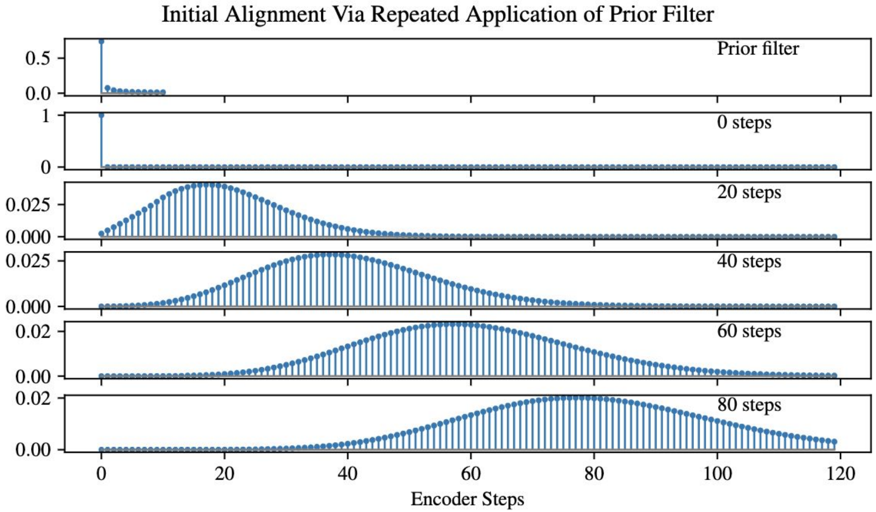
- Attributes
  - Inputs: $\boldsymbol{s}_i$, $\boldsymbol{\alpha}_{i-1}$ (**Location-relative**, not content-based)
  - Normalized weights, unlike GMM-based.

DCA Prior Filter

$$e_{i,j} = \boldsymbol{v}^\intercal \tanh(U\boldsymbol{f}_{i,j} + T\boldsymbol{g}_{i,j} + \boldsymbol{b}) + \boxed{p_{i,j}}$$

$$\boldsymbol{p}_i = \log(\mathcal{P} * \boldsymbol{\alpha}_{i-1})$$

- 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.

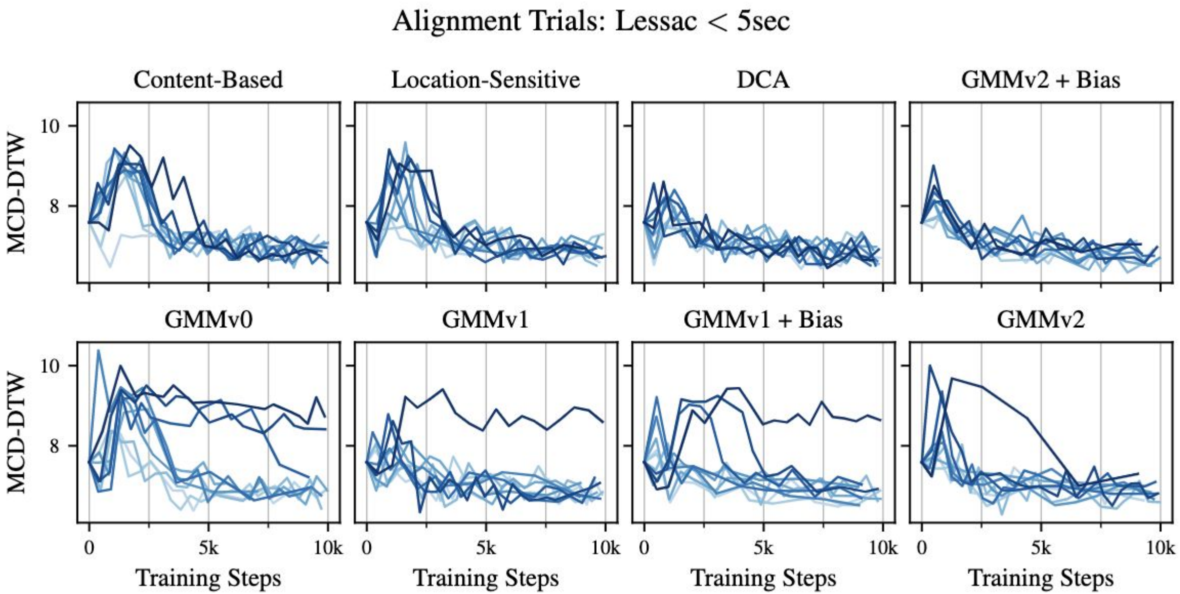


Experiment Setup

- 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.



GMMv2b

In-Domain Naturalness

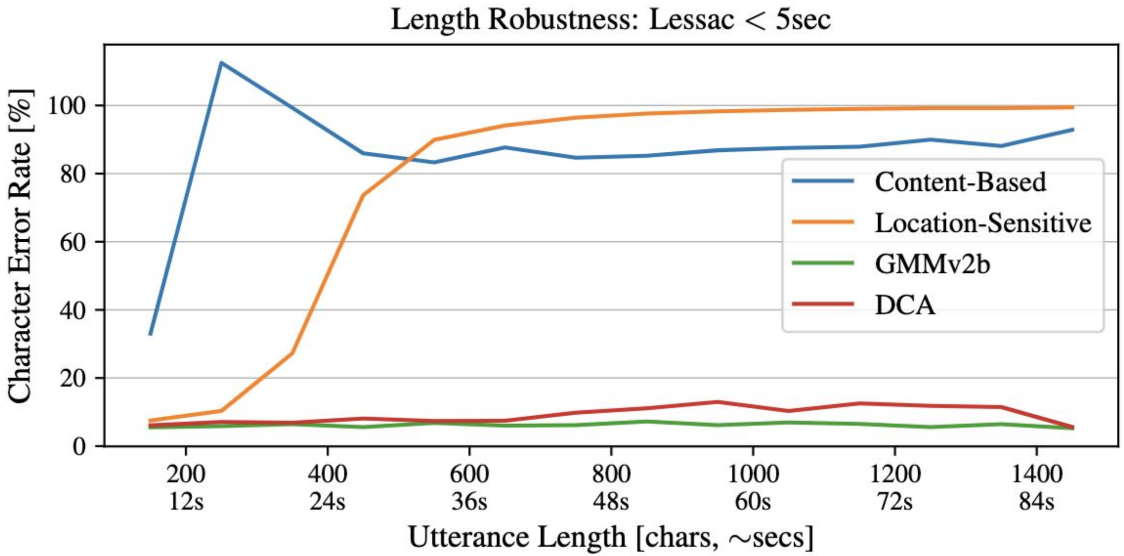
- Crowd-sourced MOS naturalness ratings.
- Test set: Hold-out from same dataset.

	Lessac	LJ
Content-Based	4.07 ± 0.08	4.19 ± 0.06
Location-Sensitive	4.31 ± 0.06	4.34 ± 0.06
GMMv2b	4.32 ± 0.06	4.29 ± 0.06
DCA	4.31 ± 0.06	4.33 ± 0.06
Ground Truth	4.64 ± 0.04	4.55 ± 0.04

- 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

- Harry Potter novels: 1034 utts, (58-1648 chars each).
- 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.

Content-Based

Location-Sensitive

DCA

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

- J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang,Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions," ICASSP, 2018.
- Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio,"Neural Machine Translation by Jointly Learning to Align and Translate," ICLR, 2015.
- Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based Models for Speech Recognition," NIPS, 2015.
- Colin Raffel, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, and Douglas Eck, "Online and Linear-time Attention by Enforcing Monotonic Alignments," ICML, 2017.
- J. Zhang, Z. Ling, and L. Dai, "Forward attention in sequence-to-sequence acoustic modeling for speech synthesis," ICASSP, 2018.
- Mutian He, Yan Deng, and Lei He, "Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS," INTERSPEECH, 2019.
- Alex Graves, "Generating Sequences With Recurrent Neural Networks," arXiv, 2013.