Google AI

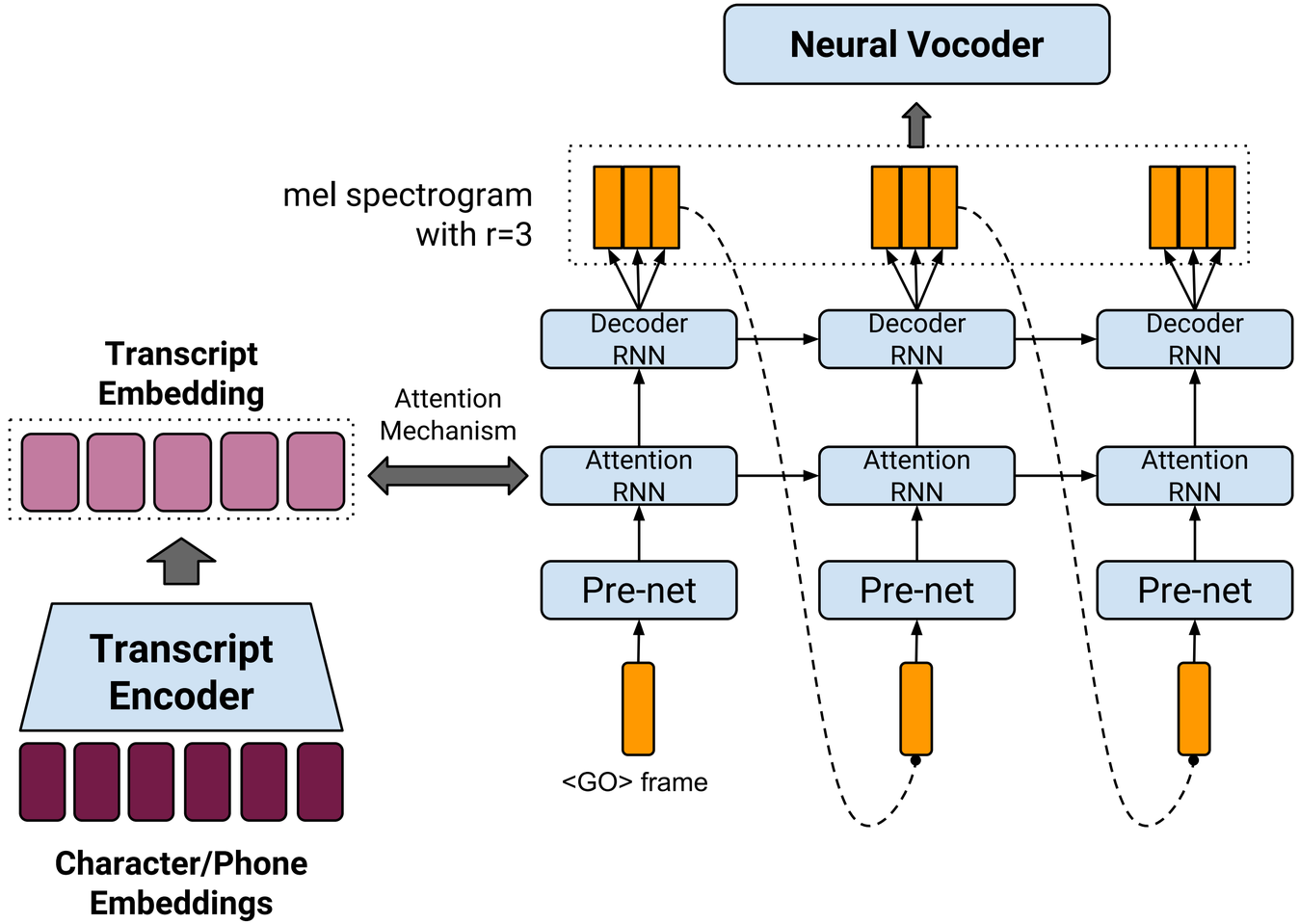
Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron

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

Audio Test

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):

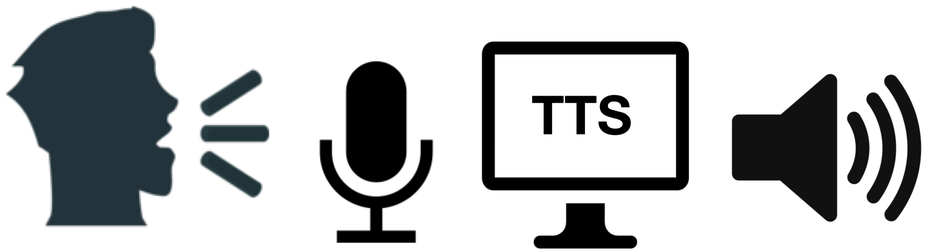
Definition. *Prosody is the variation in speech signals that remains after accounting for variation due to phonetics, speaker identity, and channel effects (i.e. the recording environment).*

Prosody isn't:
- **What**'s being said.
- **Who**'s saying it.
- **Where** it's being said.

Prosody is:
- **How** it's said.

Prosody Transfer

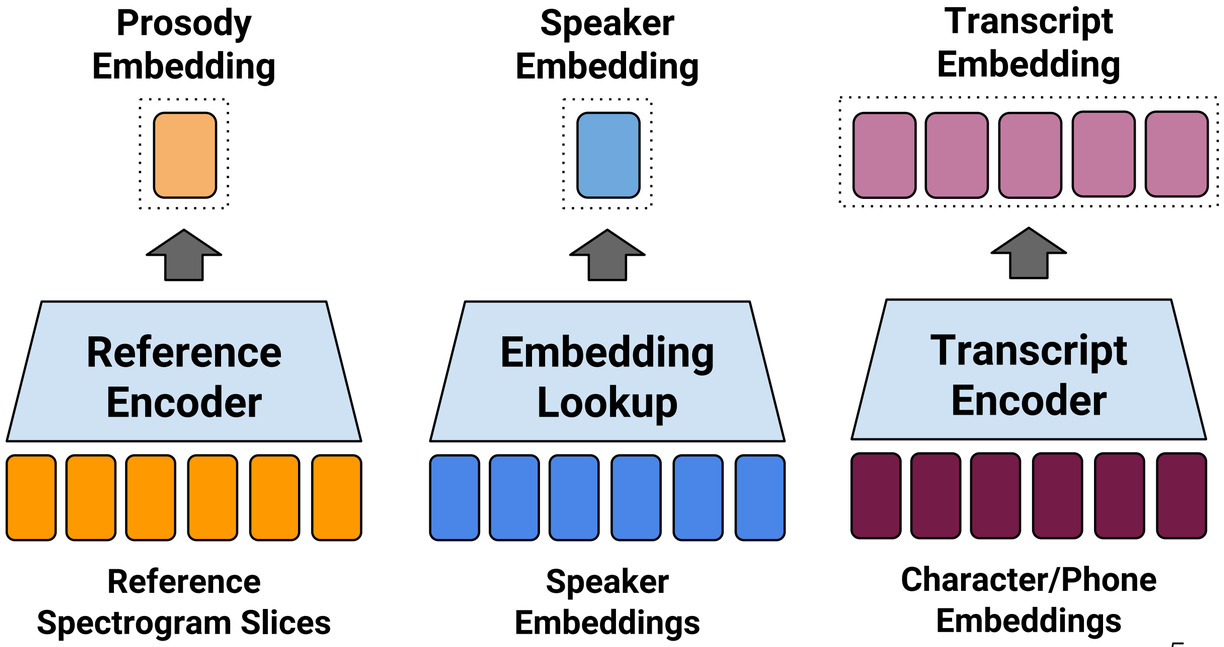
- Various way to control prosody:
  - Prosody annotations (e.g., ToBI)
  - Linguistic features (pitch, energy, duration).
  - **Prosody transfer ("Say it like this")**

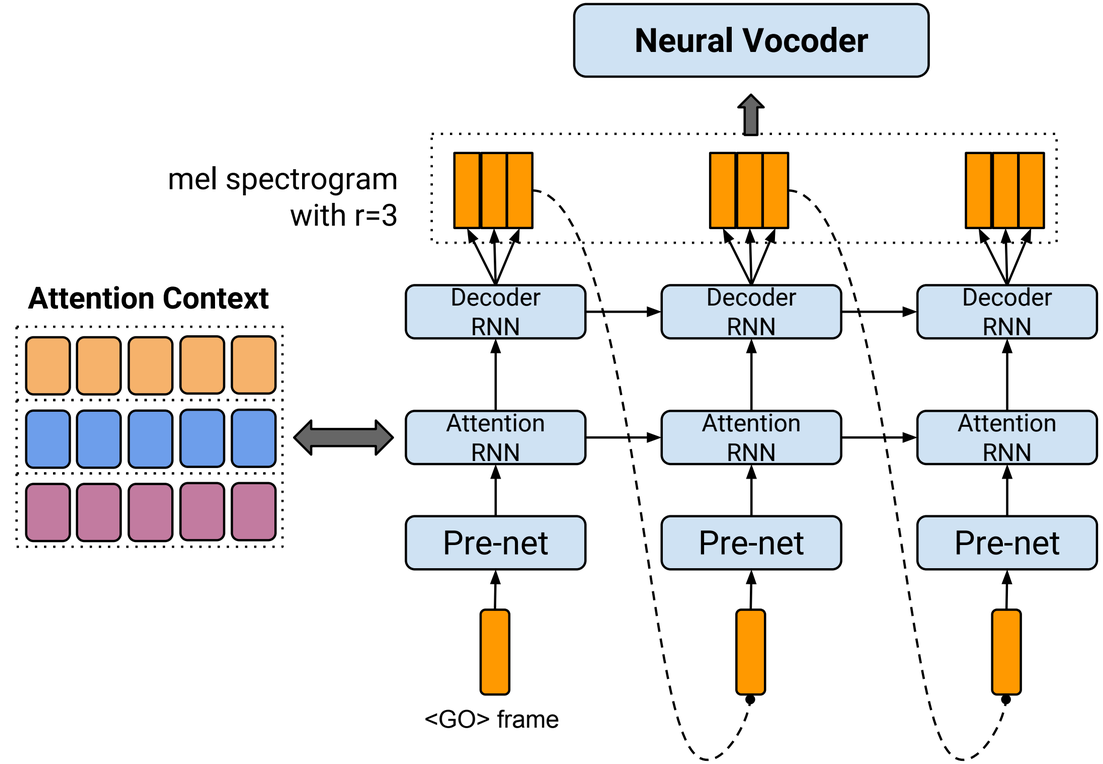


- <u>Prosody transfer desired features</u>:
  - **Pitch relative transfer** (output is within a speaker's natural pitch range).
  - **Robust to text transformations** (one reference for many sentences, makes it scalable).
  - **Meaningful embedding space** (for sampling or control via other systems).

End-to-End Prosody Transfer

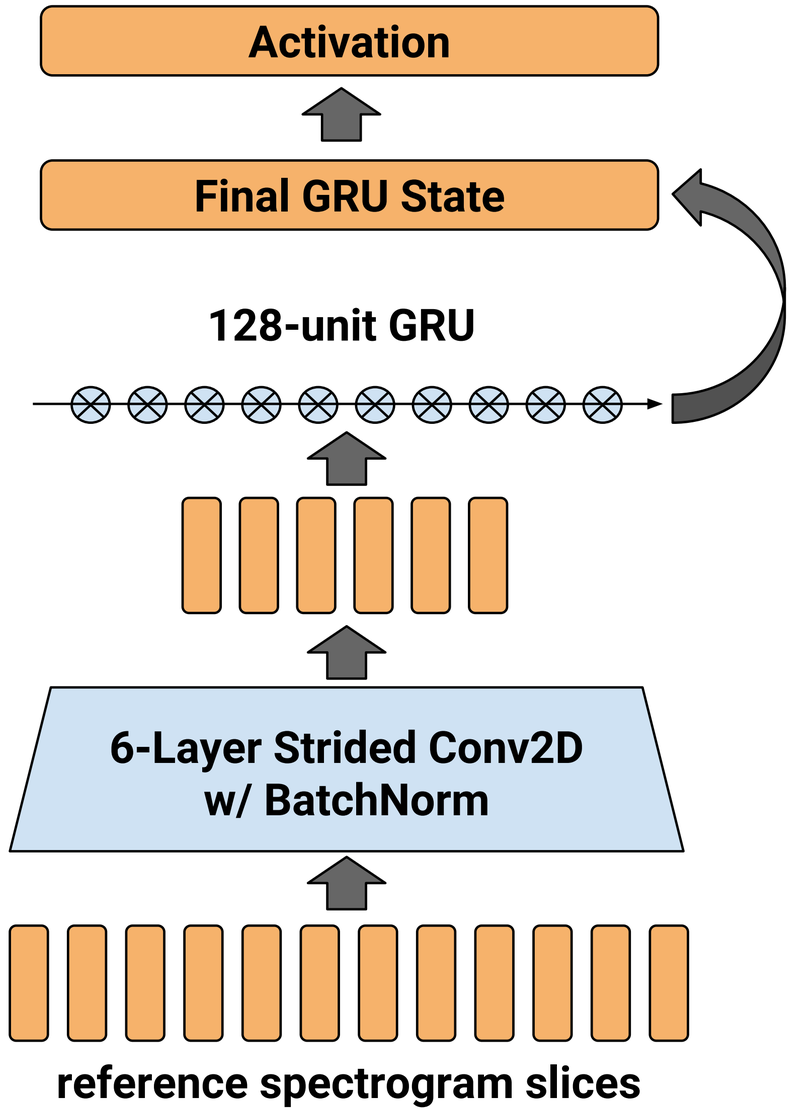
- Prosody Embeddings are computed using a Reference Encoder.
- Speaker embeddings are used for multi-speaker models.
- Both are broadcast-concatenated to the transcript embeddings.
- Reference and target speaker are the same during training.
  (but can be different during inference)





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.



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]

Evaluation Results

The **tanh-128** model uses a 128-dimensional prosody embedding.

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%**	$\mathbf{1.146 \pm 0.246}$

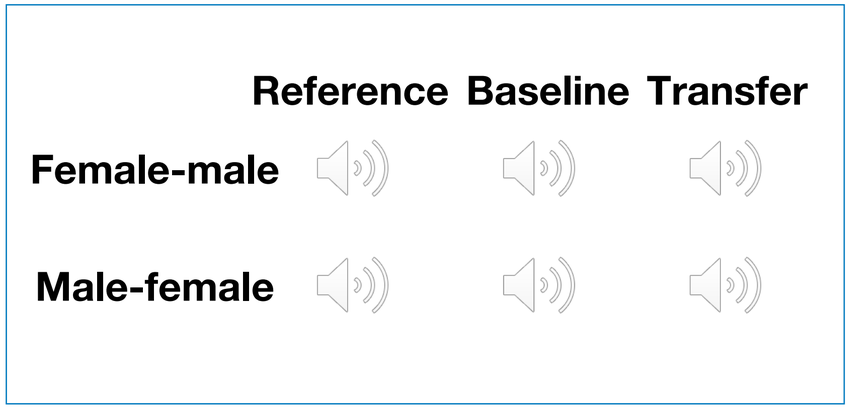
Audio Examples

Text	Reference	Baseline	Prosody Embedding
Single-speaker model: Reference from unseen speaker			
The past, the present, and the future walk into a bar. It was tense.	Aus F 	Les 	Les 
Multi-speaker model: Reference from seen speaker			
***Is** that Utah travel agency?*	Aus F 	US F  GB F  Ind F 	US F  GB F  Ind F 
Only one was deployed, while they need a hundred teams.	Ind F 	US F  Aus F  GB F 	US F  Aus F  GB F 
Multi-speaker model: Reference from unseen speaker			
*It will be good for **both** of you.*	Les 	Aus F  GB F  US M 	Aus F  GB F  US M 
I've swallowed a pollywog.	Les 	Aus F  GB F  US M 	Aus F  GB F  US M 

More audio examples available at: 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.



Robustness to Text Transformations

Text	Reference	Baseline	Prosody Embedding
Reference: *“I can now,” said the **Leopard**.* Perturbed: *“I can now,” said the **Porcupine**.*			
Reference: *For the **first** time in **her** life **she** had been **danced tired**.* Perturbed: *For the **last** time in **his** life **he** had been **handily embarrassed**.*			
Reference: ***Second**--Her family was very **ancient** and **noble**.* Perturbed: ***First**--Her family was very **sarcastic** and **horrible**.*			
Reference: *Never again shall **Eleanor Lavish** be a **friend** of mine.* Perturbed: *Never again shall **Bartholomew Bigglesby** be a **son** of mine.*			
Reference: ***Alice** was not much surprised at this, **she** was getting so used to **queer things happening**.* Perturbed: ***Eric** was not much surprised at this, **he** was getting so used to **TensorFlow breaking**.*			

More Audio Examples!!?

- 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

- [1] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards End-to-End Speech Synthesis,” *arXiv.org*, vol. cs.CL. 29-Mar-2017.

- [2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” *arXiv.org*, vol. cs.CL. 15-Dec-2017.

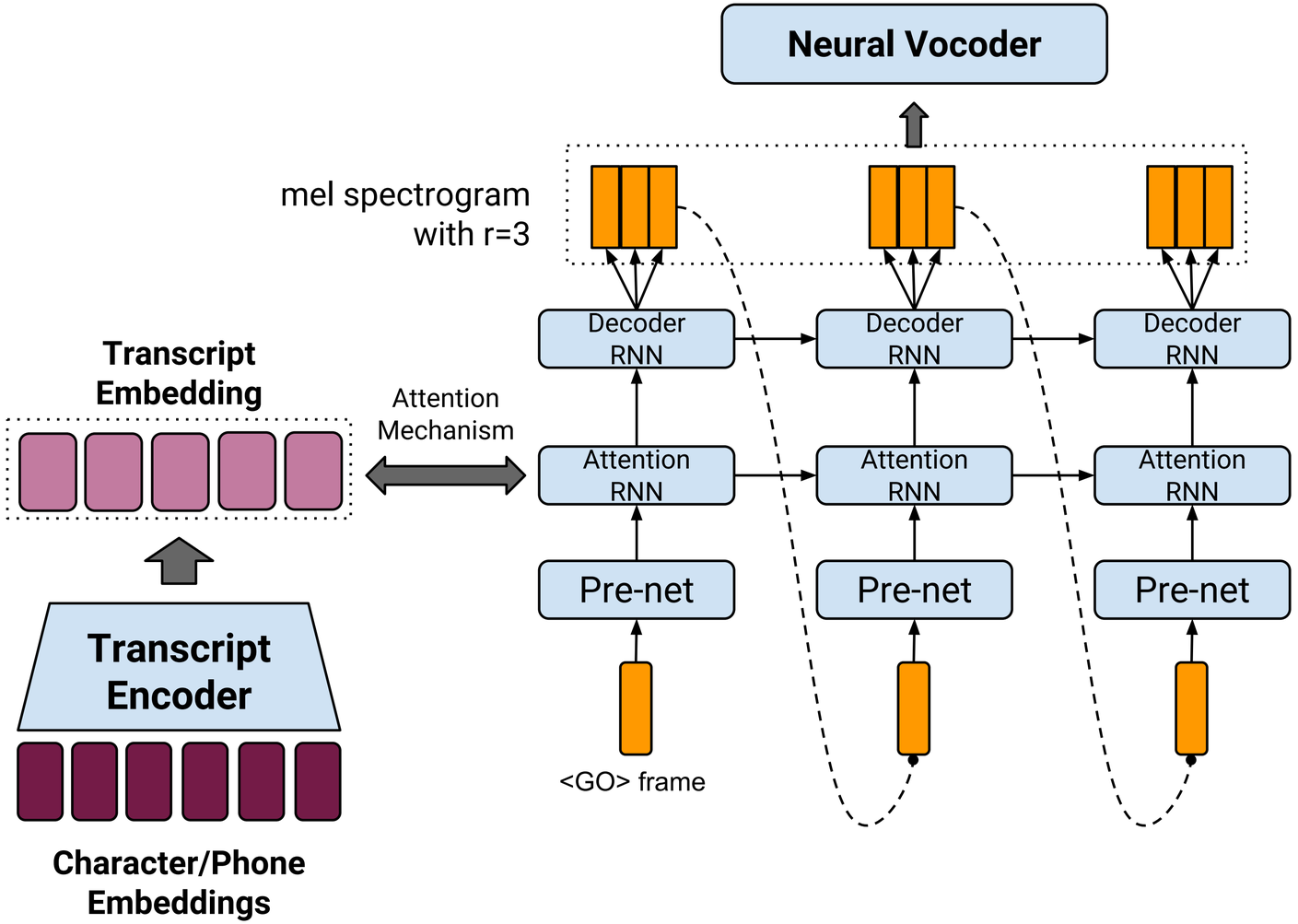
- [3] A. Graves, “Generating Sequences With Recurrent Neural Networks,” arXiv.org. 04-Aug-2013.

- [4] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, “Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis,” *arXiv.org*, vol. cs.CL. 23-Mar-2018.

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.

Reference

Prosody Embedding

Baseline