Zero-Shot Multi-Speaker Text-to-Speech with State-of-the-Art Neural Speaker Embeddings

Erica Cooper*, Cheng-I Jeff Lai**,
Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, Junichi Yamagishi

NII, Tokyo, Japan*; MIT CSAIL, Cambridge, MA, USA**

ICASSP 2020 Speech Synthesis and Voice Conversion I

What to Expect in This Talk

- Background on multi-speaker TTS setup, neural speaker embeddings and the Learnable Dictionary Encoding (LDE) method.

What to Expect in This Talk

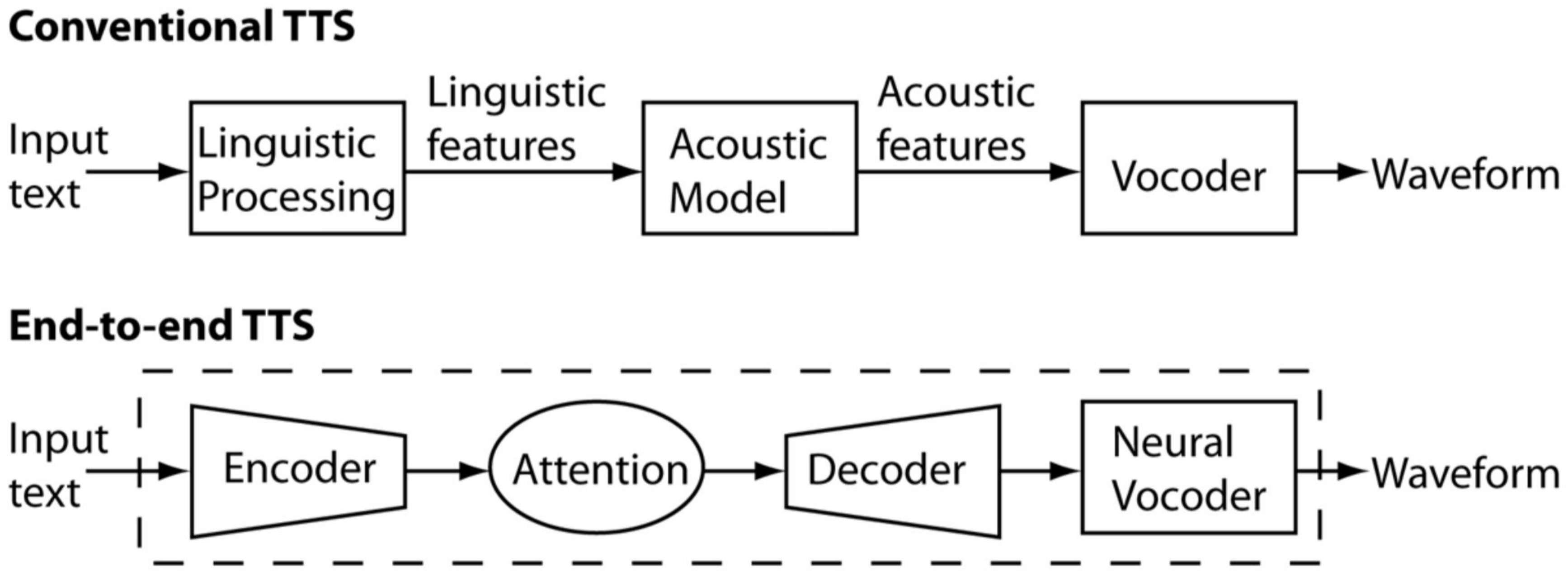
- Background on multi-speaker TTS setup, neural speaker embeddings and the Learnable Dictionary Encoding (LDE) method.
- Incorporating neural speaker embeddings into Tacotron-based TTS systems; experiments on zero-shot speaker similarity.

What to Expect in This Talk

- Background on multi-speaker TTS setup, neural speaker embeddings and the Learnable Dictionary Encoding (LDE) method.
- Incorporating neural speaker embeddings into Tacotron-based TTS systems; experiments on zero-shot speaker similarity.
- A large-scale listening test to demonstrate the effectiveness of these neural speaker embeddings

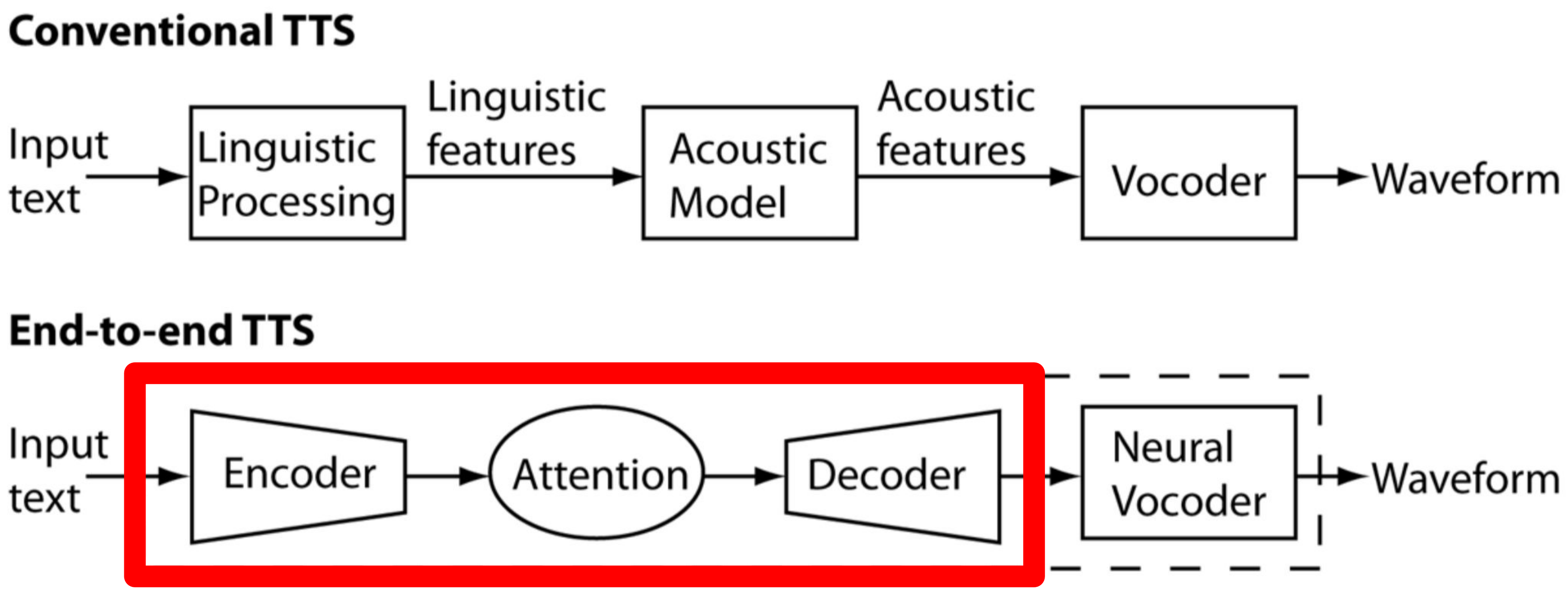
End-to-end TTS: Tacotron + Vocoder

- Tacotron: Learns mappings from char/phonemes to mel spectrogram
- Vocoder (Wavenet): Converts mel spectrogram to waveforms



End-to-end TTS: Tacotron + Vocoder

- Tacotron: Learns mappings from char/phonemes to mel spectrogram
- Vocoder (Wavenet): Converts mel spectrogram to waveforms



End-to-end Multi-Speaker TTS

- Goal: synthesize 100+ speakers' "voice" with a single model
  - *Without* having to re-train the whole system
  - Small amount of target speaker data
  - Generalize to unseen speakers during training

Two Approaches to Multi-Speaker TTS

- Model Fine-tuning
  - + Works well using a small amount of data (minutes)
  - - Adaptation data must be transcribed and TTS-quality
  - - Requires additional training steps for every new speaker

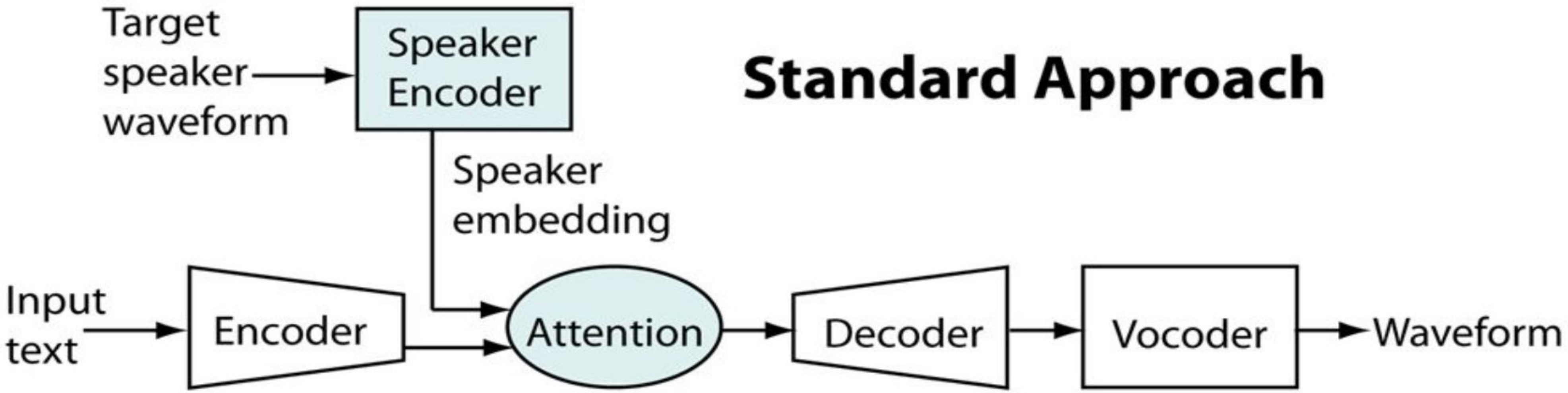
Two Approaches to Multi-Speaker TTS

- Model Fine-tuning
  - + Works well using a small amount of data (minutes)
  - - Adaptation data must be transcribed and TTS-quality
  - - Requires additional training steps for every new speaker

- Transfer Learning from ASV
  - + Requires even less target speaker data (seconds)
  - + Transcripts are not required; adaptation data can be low-quality
  - + ASV systems can be trained on 1000+ of speakers
  - + No additional training steps required for new speakers
  - - Speaker similarity for unseen speakers is not as good

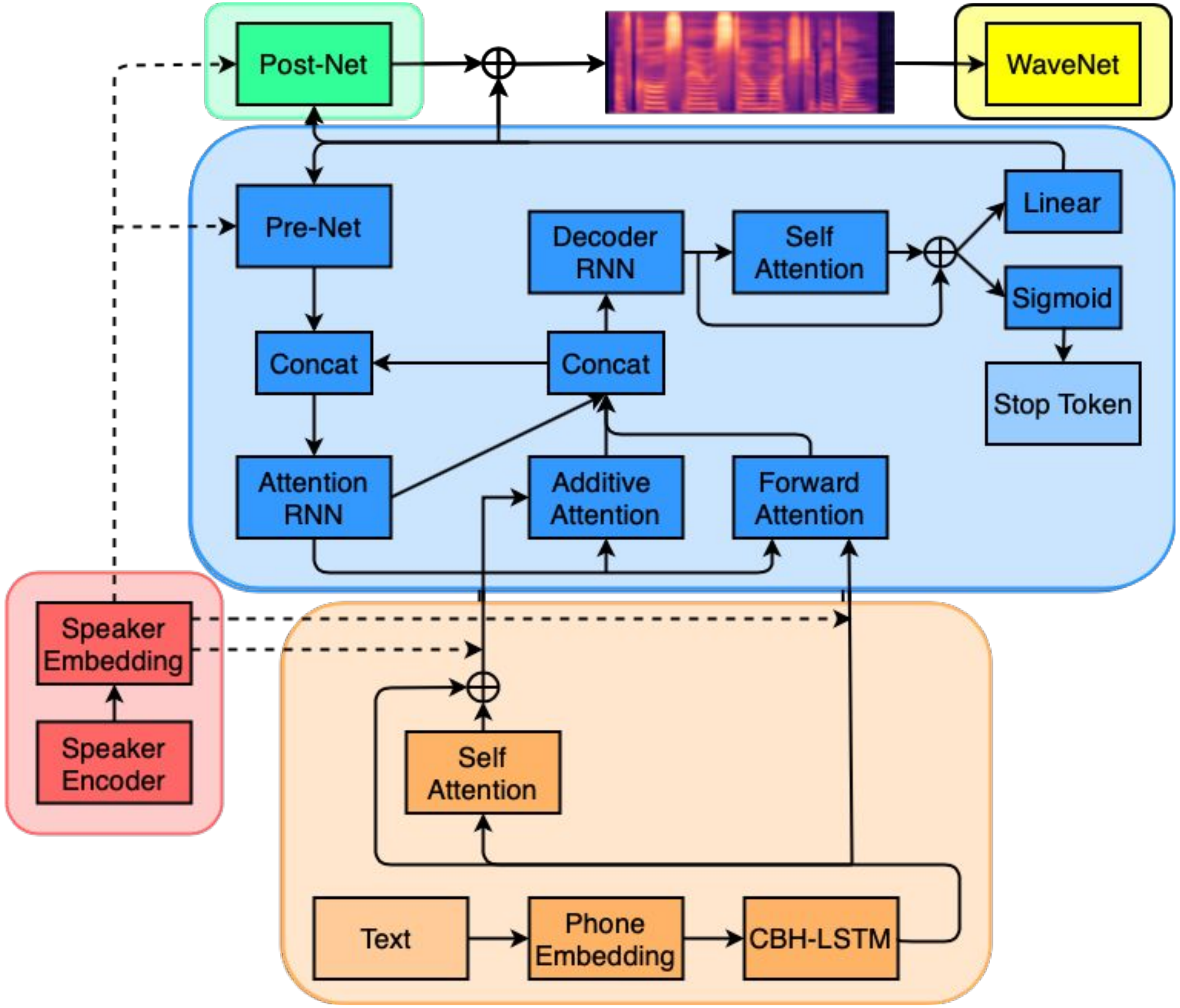
Transfer Learning from ASV to TTS

- Pretrain a speaker recognition model to get speaker embeddings
- Input the speaker embedding to TTS*

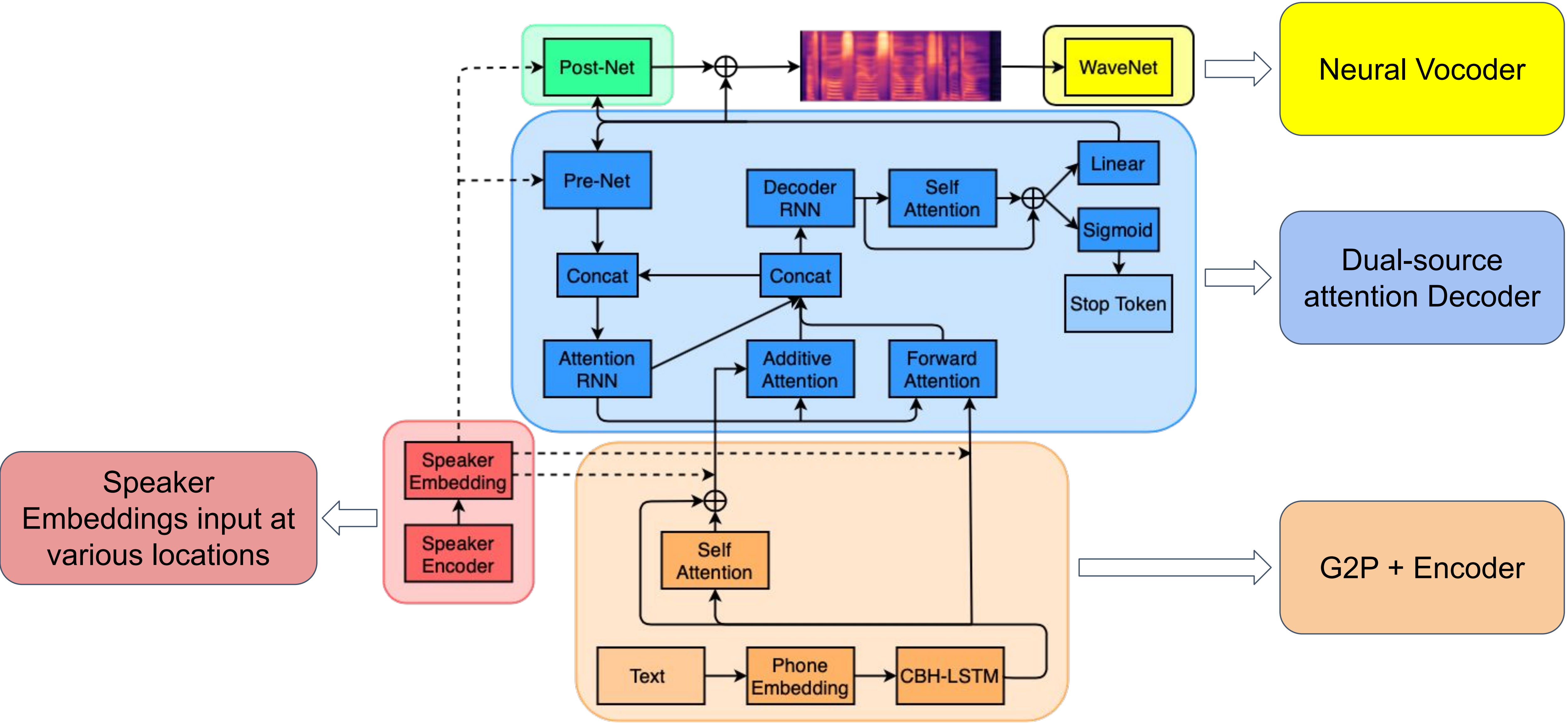


*During inference, target speakers need not be seen during training

Tacotron2 with Dual-Source Attention*



*Yasuda et al. 2019: Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language


Post-Net
WaveNet
Neural Vocoder
Pre-Net
Decoder RNN
Self Attention
Linear
Sigmoid
Concat
Concat
Stop Token
Dual-source attention Decoder
Attention RNN
Additive Attention
Forward Attention
Speaker Embedding
Speaker Encoder
Speaker Embeddings input at various locations
Self Attention
G2P + Encoder
Text
Phone Embedding
CBH-LSTM

Experiments: Embeddings Input Location and Training Strategy

- Input location: Prenet (**pre**), Decoder Attention (**attn**), Both (**pre+attn**), Both+Postnet (**pre+attn+post**)

Experiments: Embeddings Input Location and Training Strategy

- Input location: Prenet (**pre**), Decoder Attention (**attn**), Both (**pre+attn**), Both+Postnet (**pre+attn+post**)
- Training strategy:
  - Train from scratch or pre-train?
  - Gender-independent or gender-dependent?

Experiments: Embeddings Input Location and Training Strategy

- Input location: Prenet (**pre**), Decoder Attention (**attn**), Both (**pre+attn**), Both+Postnet (**pre+attn+post**)
- Training strategy:
  - Train from scratch or pre-train?
  - Gender-independent or gender-dependent?
- Objective evaluation: Speaker similarity between original voices and synthesized voices $\longrightarrow$ cosine similarity
  - Unseen speakers are most important
- Data: VCTK corpus (English; 109 speakers)

Results 1: From Scratch vs. Warm Start

- **Training from scratch:**
  - Train on VCTK data only
  - **~4 days** to get reasonable quality and speaker similarity

Results 1: From Scratch vs. Warm Start

- **Training from scratch:**
  - Train on VCTK data only
  - **~4 days** to get reasonable quality and speaker similarity

- **Warm-start training:**
  - Initialize model parameters from a well-trained single-speaker model (Blizzard 2011 "Nancy,"; 3x larger vocabulary)
  - **~1 day** of additional training with VCTK data to get about equivalent quality and speaker similarity

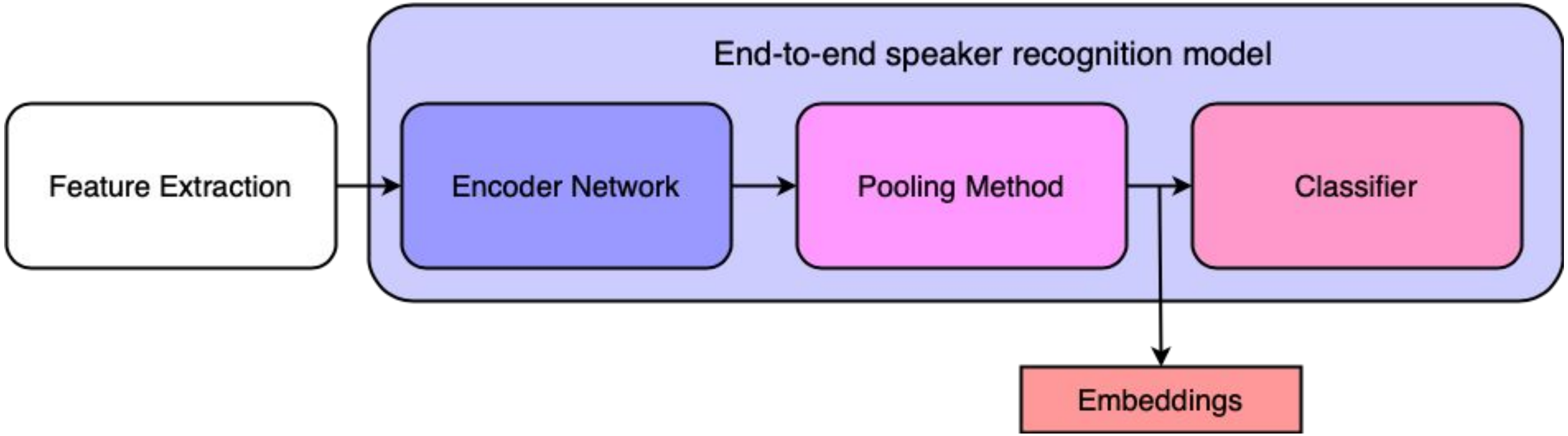
Results 2: Unseen Speaker Similarity [-1, +1]

Input location	Gender-ind		Gender-dep	
	train	dev	train	dev
pre	0.357	0.402	0.438	0.361
attn	0.709	0.490	0.711	0.476
pre+attn	0.676	0.489	0.708	**0.533**
pre+attn+post	0.684	0.480	0.717	0.477

Results 2: Unseen Speaker Similarity [-1, +1]

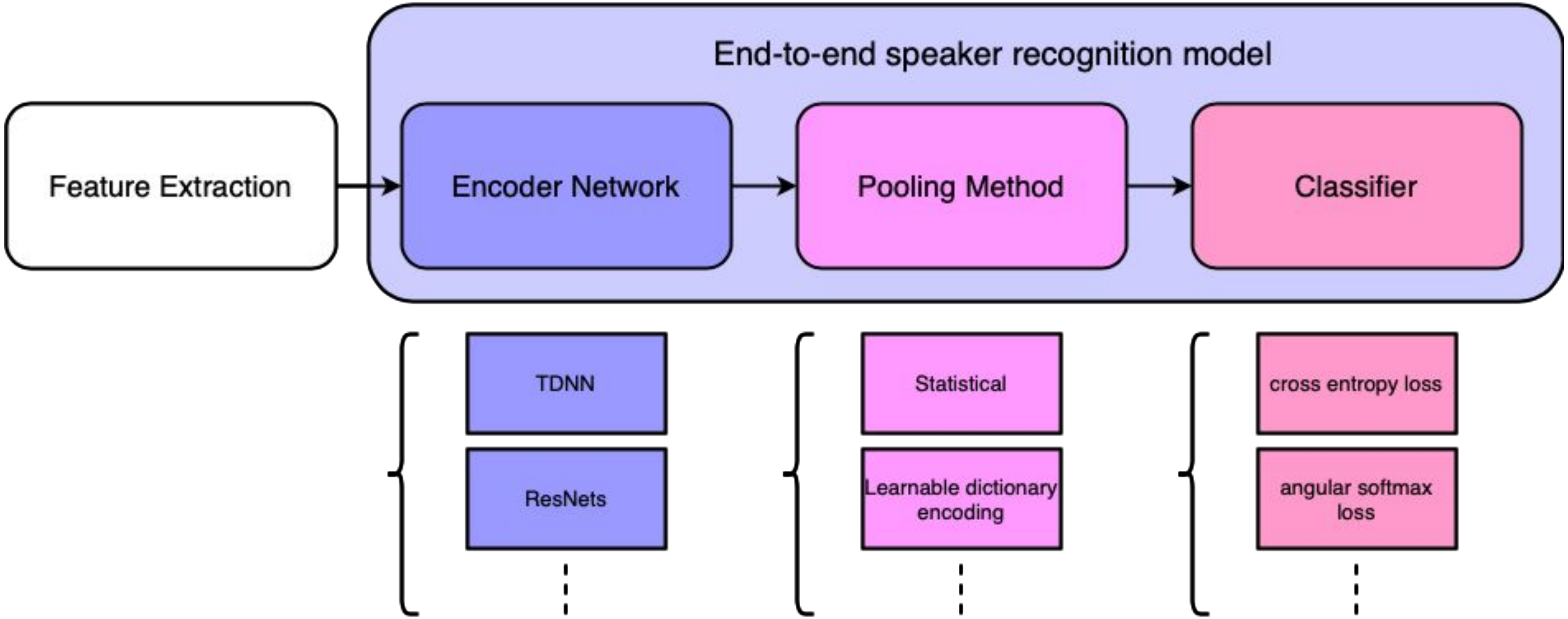
Input location	Gender-ind		Gender-dep	
	train	dev	train	dev
pre	0.357	0.402	0.438	0.361
attn	0.709	0.490	0.711	0.476
pre+attn	0.676	0.489	0.708	**0.533**
pre+attn+post	0.684	0.480	0.717	0.477

Neural Speaker Embeddings: Overview



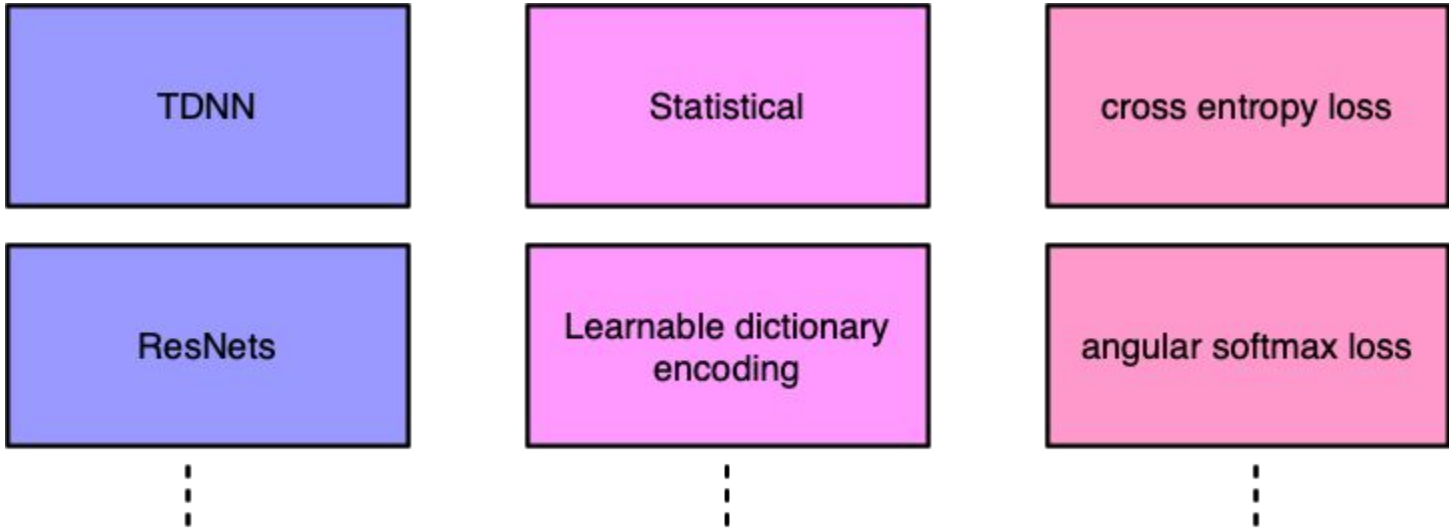
Villalba et al. 2019: State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18

Neural Speaker Embeddings: Overview



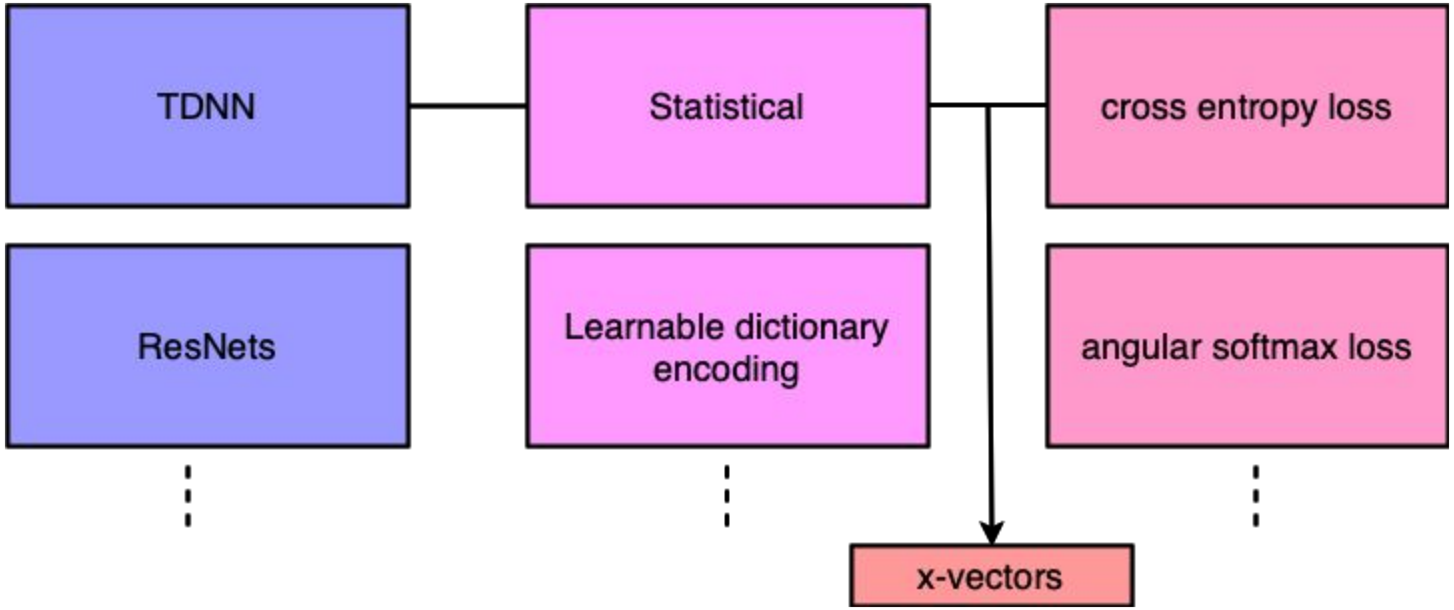
Villalba et al. 2019: State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18

Neural Speaker Embeddings: Overview



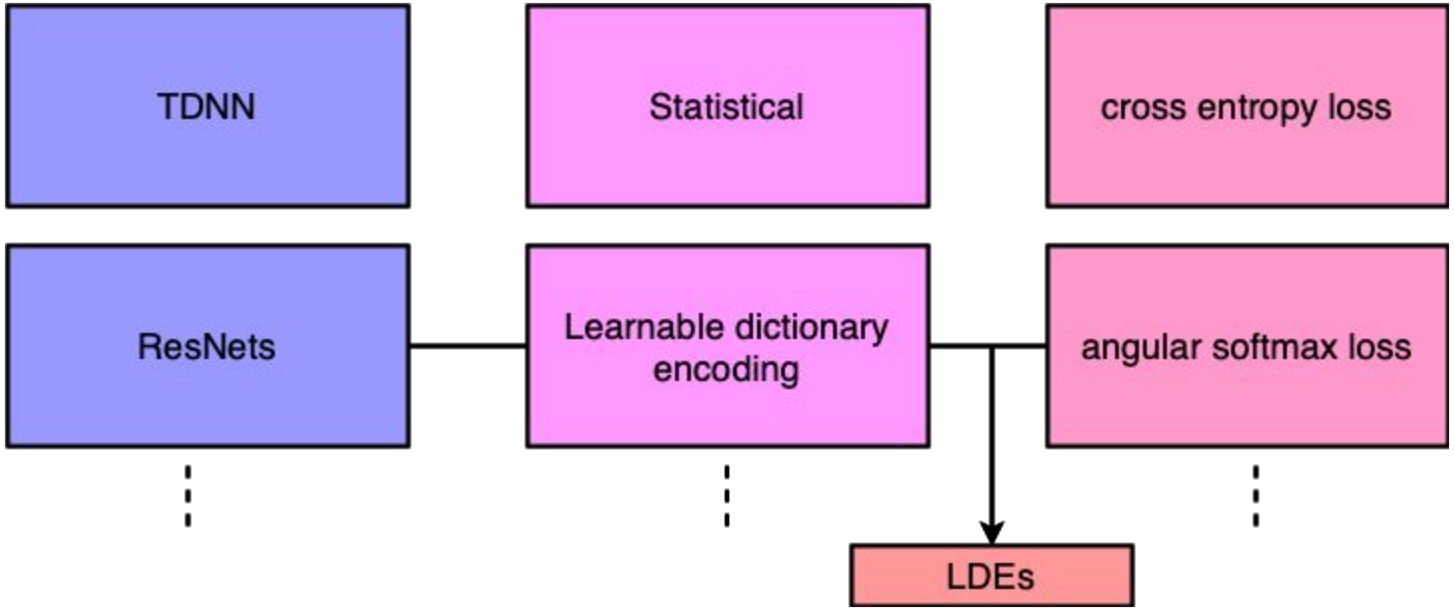
Villalba et al. 2019: State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18

Neural Speaker Embeddings: x-vectors



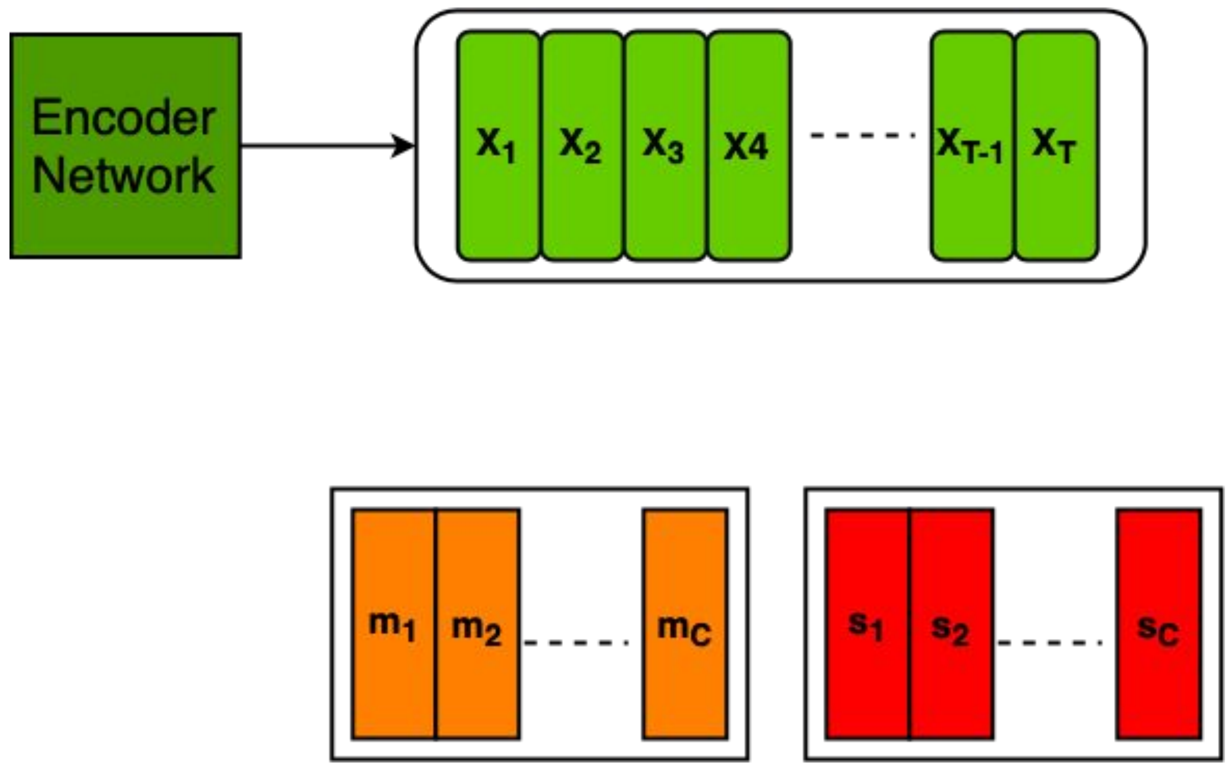
Villalba et al. 2019: State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18

Neural Speaker Embeddings: LDEs



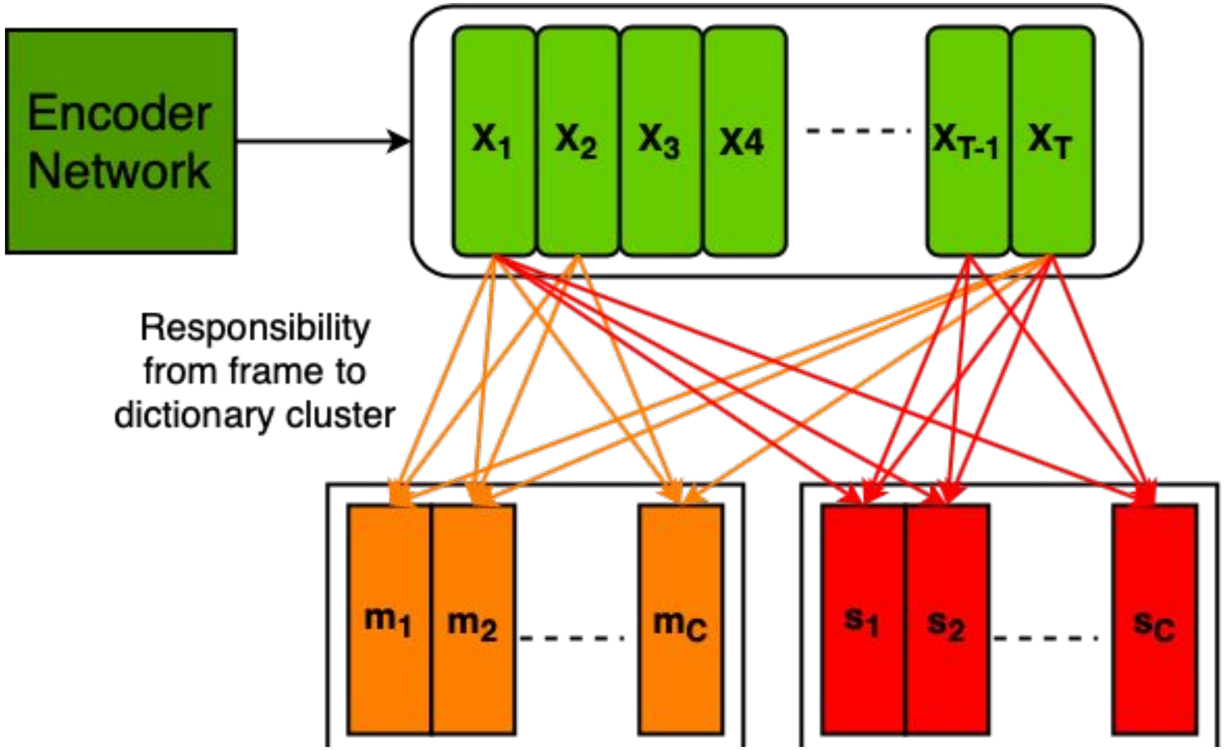
Villalba et al. 2019: State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18

Learnable Dictionary Encoding Method



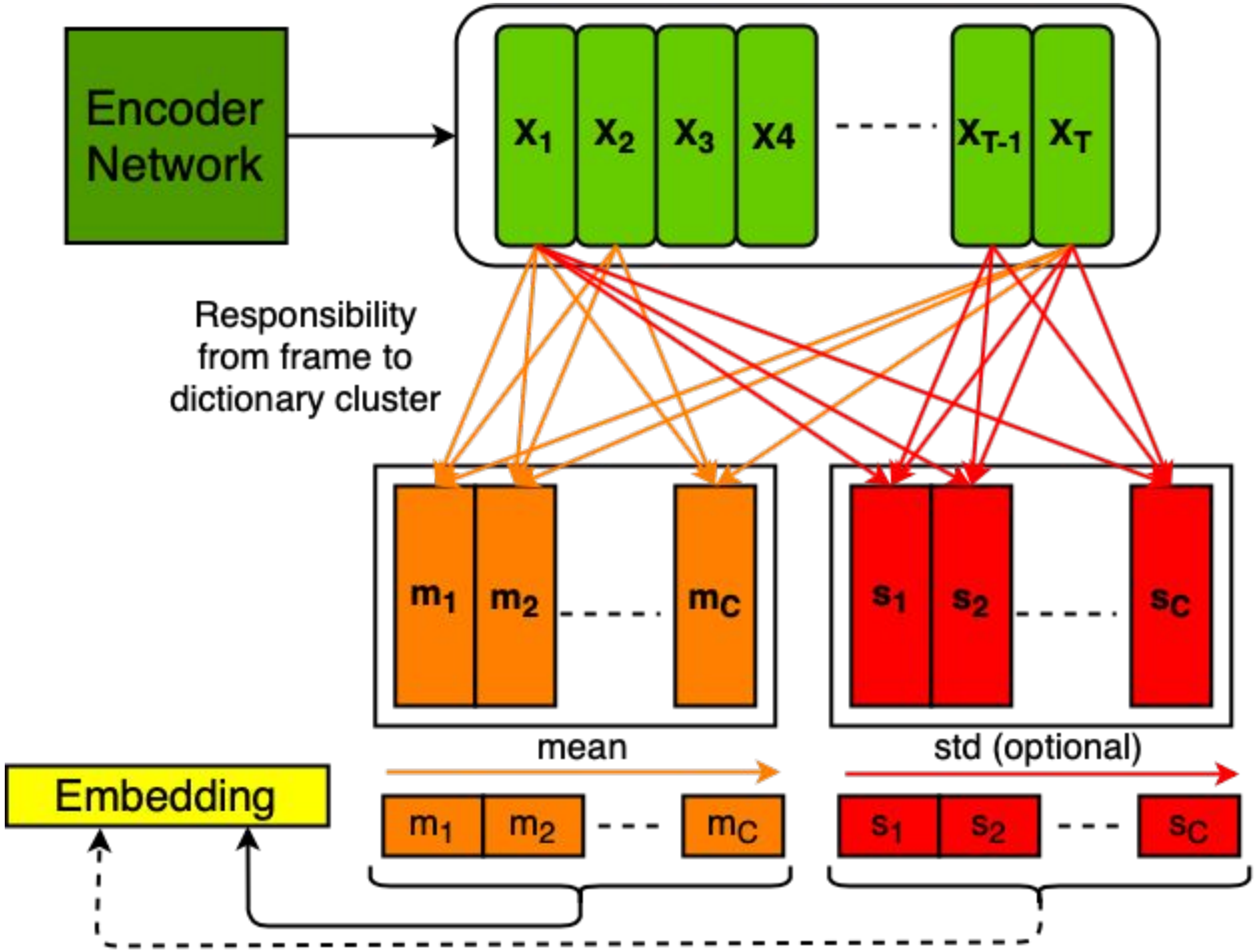
Cai et al. 2019: Exploring the encoding layer and loss function in end-to-end speaker and language recognition system

Learnable Dictionary Encoding Method



Cai et al. 2019: Exploring the encoding layer and loss function in end-to-end speaker and language recognition system

Learnable Dictionary Encoding Method



Cai et al. 2019: Exploring the encoding layer and loss function in end-to-end speaker and language recognition system

Experiments: Speaker Verification EER

- Data: VoxCeleb I+II (7000+ speakers)
- Baselines: i-vectors, x-vectors

Experiments: Speaker Verification EER

- Data: VoxCeleb I+II (7000+ speakers)
- Baselines: i-vectors, x-vectors
- LDEs:
  - Dimension: {512, 256, 200}
  - Loss: {softmax, angular softmax (m=2, 3, 4)}
  - Pooling: {mean, mean+std.dev}
  - Post-processing: {N/A, centering and LDA dim-reduction to 200dim}
- 17 total systems

Results 3: Speaker Verification EER

embed.	dim.	pl.	obj.	norm	EER	$DCF_{0.01}^{min}$
i-Vec^{N}	400	m	EM	✓	5.329	0.493
x-Vec	512	m, s	S		3.298	0.343
x-Vec^{N}	512	m, s	S	✓	3.213	0.342
LDE-1	512	m	S		3.415	0.366
LDE-1^{N}	512	m	S	✓	3.446	0.365
LDE-2	512	m	AS(2)		3.674	0.364
LDE-2^{N}	512	m	AS(2)	✓	3.664	0.386
LDE-3	512	m	AS(3)		**3.033**	**0.314**
LDE-3^{N}	512	m	AS(3)	✓	3.171	0.327
LDE-4	512	m	AS(4)		3.112	0.315
LDE-4^{N}	512	m	AS(4)	✓	3.271	0.327
LDE-5	256	m	AS(2)		3.287	0.343
LDE-5^{N}	256	m	AS(2)	✓	3.367	0.351
LDE-6	200	m	AS(2)		3.266	0.396
LDE-6^{N}	200	m	AS(2)	✓	3.266	0.396
LDE-7	512	m, s	AS(2)		**3.091**	**0.303**
LDE-7^{N}	512	m, s	AS(2)	✓	3.171	0.328

Results 3: Speaker Verification EER

embed.	dim.	pl.	obj.	norm	EER	$DCF_{0.01}^{min}$
i-Vec^{N}	400	m	EM	✓	5.329	0.493
x-Vec	512	m, s	S		3.298	0.343
x-Vec^{N}	512	m, s	S	✓	3.213	0.342
LDE-1	512	m	S		3.415	0.366
LDE-1^{N}	512	m	S	✓	3.446	0.365
LDE-2	512	m	AS(2)		3.674	0.364
LDE-2^{N}	512	m	AS(2)	✓	3.664	0.386
LDE-3	512	m	AS(3)		**3.033**	**0.314**
LDE-3^{N}	512	m	AS(3)	✓	3.171	0.327
LDE-4	512	m	AS(4)		3.112	0.315
LDE-4^{N}	512	m	AS(4)	✓	3.271	0.327
LDE-5	256	m	AS(2)		3.287	0.343
LDE-5^{N}	256	m	AS(2)	✓	3.367	0.351
LDE-6	200	m	AS(2)		3.266	0.396
LDE-6^{N}	200	m	AS(2)	✓	3.266	0.396
LDE-7	512	m, s	AS(2)		**3.091**	**0.303**
LDE-7^{N}	512	m, s	AS(2)	✓	3.171	0.328

Yi et al. 2019: Large Margin Softmax Loss for Speaker Verification.

Experiments: Naturalness and Speaker Similarity

- Ground truth + Wavenets
- Tacotron2 + x-vectors + Wavenets
- Tacotron2 + LDEs + Wavenets

- Naturalness MOS (1-5); Speaker Similarity DMOS (1-4)

Results 4: Naturalness and Speaker Similarity

	Naturalness			Similarity		
system	train	dev	test	train	dev	test
vocoded	3.51	3.41	3.55	3.02	2.79	2.82
x-Vec^{N}	3.20	3.19	3.19	2.93	1.86	2.37
LDE-1	3.15	3.16	3.21	2.87	**2.05**	2.34
LDE-1^{N}	3.04	3.13	**3.46**	2.87	1.97	**2.45**
LDE-2	3.11	**3.28**	3.35	2.84	2.00	2.37
LDE-2^{N}	3.13	3.19	3.33	2.90	2.00	2.35
LDE-3	3.09	3.24	**3.48**	2.89	1.88	**2.46**
LDE-3^{N}	3.14	3.16	3.33	2.91	2.00	2.37
LDE-4	3.08	3.10	3.29	2.94	2.00	2.31
LDE-4^{N}	3.12	3.20	3.29	2.90	1.98	2.39
LDE-5	3.07	**3.26**	**3.40**	2.89	1.99	**2.45**
LDE-5^{N}	3.11	3.07	3.37	2.88	**2.02**	2.41
LDE-6	3.12	3.25	3.33	2.92	1.95	2.43
LDE-6^{N}	3.13	**3.29**	3.23	2.88	1.94	2.39
LDE-7	3.15	3.03	3.18	2.91	1.86	2.28
LDE-7^{N}	3.07	3.02	3.24	2.83	**2.02**	2.42

train, dev and test has separate speaker sets

Results 4: Naturalness and Speaker Similarity

	Naturalness			Similarity		
system	train	dev	test	train	dev	test
vocoded	3.51	3.41	3.55	3.02	2.79	2.82
x-Vec^{N}	3.20	3.19	3.19	2.93	1.86	2.37
LDE-1	3.15	3.16	3.21	2.87	**2.05**	2.34
LDE-1^{N}	3.04	3.13	**3.46**	2.87	1.97	**2.45**
LDE-2	3.11	**3.28**	3.35	2.84	2.00	2.37
LDE-2^{N}	3.13	3.19	3.33	2.90	2.00	2.35
LDE-3	3.09	3.24	**3.48**	2.89	1.88	**2.46**
LDE-3^{N}	3.14	3.16	3.33	2.91	2.00	2.37
LDE-4	3.08	3.10	3.29	2.94	2.00	2.31
LDE-4^{N}	3.12	3.20	3.29	2.90	1.98	2.39
LDE-5	3.07	**3.26**	**3.40**	2.89	1.99	**2.45**
LDE-5^{N}	3.11	3.07	3.37	2.88	**2.02**	2.41
LDE-6	3.12	3.25	3.33	2.92	1.95	2.43
LDE-6^{N}	3.13	**3.29**	3.23	2.88	1.94	2.39
LDE-7	3.15	3.03	3.18	2.91	1.86	2.28
LDE-7^{N}	3.07	3.02	3.24	2.83	**2.02**	2.42

Results 4: Naturalness and Speaker Similarity

	Naturalness			Similarity		
system	train	dev	test	train	dev	test
vocoded	3.51	3.41	3.55	3.02	2.79	2.82
x-Vec^{N}	3.20	3.19	3.19	2.93	1.86	2.37
LDE-1	3.15	3.16	3.21	2.87	**2.05**	2.34
LDE-1^{N}	3.04	3.13	**3.46**	2.87	1.97	**2.45**
LDE-2	3.11	**3.28**	3.35	2.84	2.00	2.37
LDE-2^{N}	3.13	3.19	3.33	2.90	2.00	2.35
LDE-3	3.09	3.24	**3.48**	2.89	1.88	**2.46**
LDE-3^{N}	3.14	3.16	3.33	2.91	2.00	2.37
LDE-4	3.08	3.10	3.29	2.94	2.00	2.31
LDE-4^{N}	3.12	3.20	3.29	2.90	1.98	2.39
LDE-5	3.07	**3.26**	**3.40**	2.89	1.99	**2.45**
LDE-5^{N}	3.11	3.07	3.37	2.88	**2.02**	2.41
LDE-6	3.12	3.25	3.33	2.92	1.95	2.43
LDE-6^{N}	3.13	**3.29**	3.23	2.88	1.94	2.39
LDE-7	3.15	3.03	3.18	2.91	1.86	2.28
LDE-7^{N}	3.07	3.02	3.24	2.83	**2.02**	2.42

Results 4: Naturalness and Speaker Similarity

No drop!

	Naturalness			Similarity		
system	train	dev	test	train	dev	test
vocoded	3.51	3.41	3.55	3.02	2.79	2.82
x-Vec^{N}	3.20	3.19	3.19	2.93	1.86	2.37
LDE-1	3.15	3.16	3.21	2.87	**2.05**	2.34
LDE-1^{N}	3.04	3.13	**3.46**	2.87	1.97	**2.45**
LDE-2	3.11	**3.28**	3.35	2.84	2.00	2.37
LDE-2^{N}	3.13	3.19	3.33	2.90	2.00	2.35
LDE-3	3.09	3.24	**3.48**	2.89	1.88	**2.46**
LDE-3^{N}	3.14	3.16	3.33	2.91	2.00	2.37
LDE-4	3.08	3.10	3.29	2.94	2.00	2.31
LDE-4^{N}	3.12	3.20	3.29	2.90	1.98	2.39
LDE-5	3.07	**3.26**	**3.40**	2.89	1.99	**2.45**
LDE-5^{N}	3.11	3.07	3.37	2.88	**2.02**	2.41
LDE-6	3.12	3.25	3.33	2.92	1.95	2.43
LDE-6^{N}	3.13	**3.29**	3.23	2.88	1.94	2.39
LDE-7	3.15	3.03	3.18	2.91	1.86	2.28
LDE-7^{N}	3.07	3.02	3.24	2.83	**2.02**	2.42

Results 4: Naturalness and Speaker Similarity

	Naturalness			Similarity		
system	train	dev	test	train	dev	test
vocoded	3.51	3.41	3.55	3.02	2.79	2.82
x-Vec^{N}	3.20	3.19	3.19	2.93	1.86	2.37
LDE-1	3.15	3.16	3.21	2.87	**2.05**	2.34
LDE-1^{N}	3.04	3.13	**3.46**	2.87	1.97	**2.45**
LDE-2	3.11	**3.28**	3.35	2.84	2.00	2.37
LDE-2^{N}	3.13	3.19	3.33	2.90	2.00	2.35
LDE-3	3.09	3.24	**3.48**	2.89	1.88	**2.46**
LDE-3^{N}	3.14	3.16	3.33	2.91	2.00	2.37
LDE-4	3.08	3.10	3.29	2.94	2.00	2.31
LDE-4^{N}	3.12	3.20	3.29	2.90	1.98	2.39
LDE-5	3.07	**3.26**	**3.40**	2.89	1.99	**2.45**
LDE-5^{N}	3.11	3.07	3.37	2.88	**2.02**	2.41
LDE-6	3.12	3.25	3.33	2.92	1.95	2.43
LDE-6^{N}	3.13	**3.29**	3.23	2.88	1.94	2.39
LDE-7	3.15	3.03	3.18	2.91	1.86	2.28
LDE-7^{N}	3.07	3.02	3.24	2.83	**2.02**	2.42

Drop!

Results 4: Samples!

https://nii-yamagishilab.github.io/samples-multi-speaker-tacotron/

Conclusions

- Warm-start training works well
- Gender-dependent model training gives better speaker similarity
- Inputting speaker embedding at prenet+attention gives best speaker similarity
- Improved LDE embeddings can improve speaker similarity

Ongoing and Future Work

- Speaker space augmentation
  - SoX speedup and slowdown
  - Additional sources of multi-speaker data
  - Dialect modeling

- Multilingual / cross-lingual
  - Are LDE embeddings trained on English VoxCeleb data model language-independent?

Thanks for listening! Questions?

ecooper@nii.ac.jp (Erica)
clai24@mit.edu (Jeff)