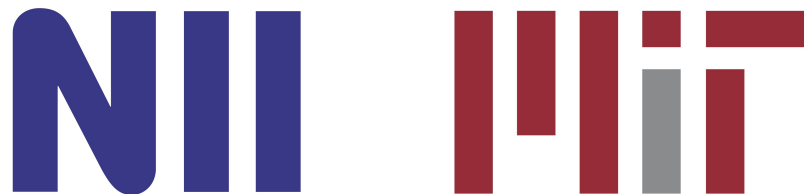


Can Speaker Augmentation Improve Multi-Speaker End-to-End TTS?

Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Junichi Yamagishi
INTERSPEECH 2020

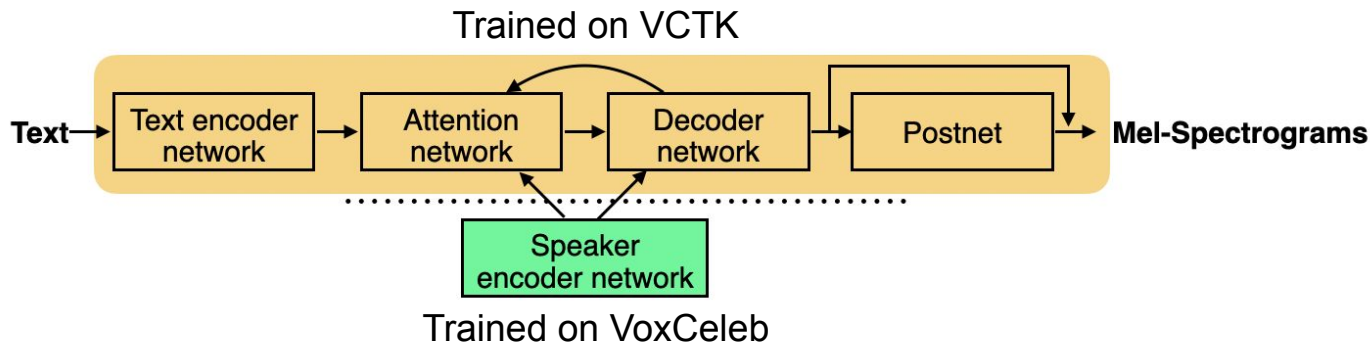


Overview

- **Background:** Zero-shot adaptation of multi-speaker Tacotron
- **Speaker augmentation:** Two approaches
- **Modifications** to Tacotron
- **Experiments**
- **Results**
- **Conclusions**

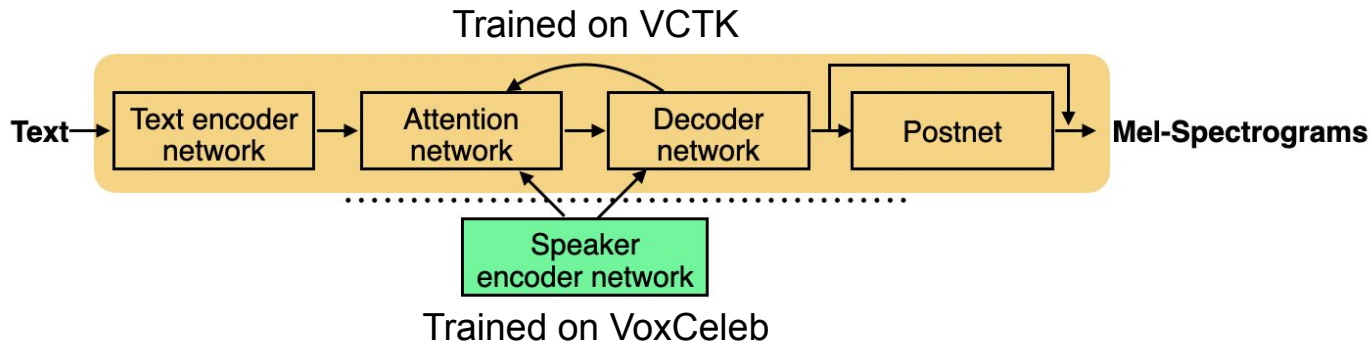
Background: Zero-shot speaker adaptation for Tacotron

- Model speakers using a **speaker embedding** extracted from a separately-trained speaker encoder network



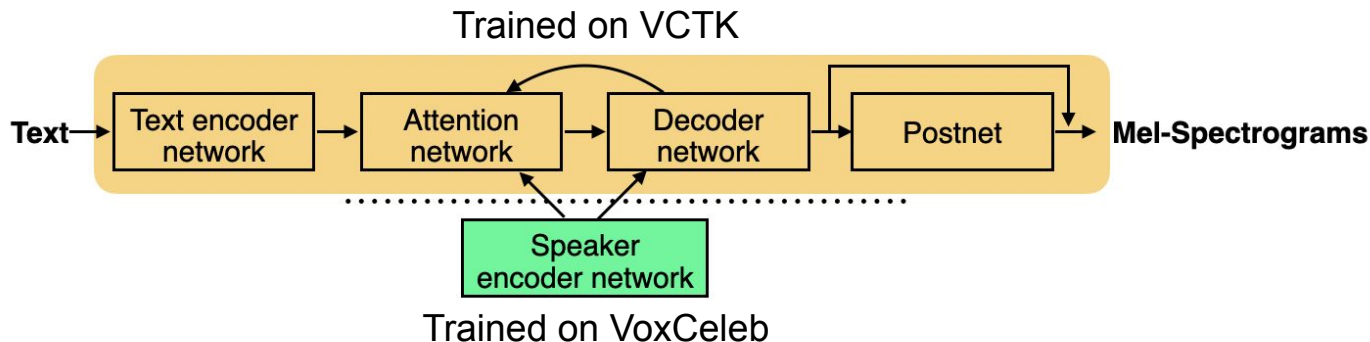
Background: Zero-shot speaker adaptation for Tacotron

- Model speakers using a **speaker embedding** extracted from a separately-trained speaker encoder network
- VCTK: ~100 speakers; Tacotron overfits and does not generalize to **unseen speakers** well.



Background: Zero-shot speaker adaptation for Tacotron

- Model speakers using a **speaker embedding** extracted from a separately-trained speaker encoder network
- VCTK: ~100 speakers; Tacotron overfits and does not generalize to **unseen speakers** well.
- **Speaker augmentation**: can we include **more speakers** during training?



Speaker Augmentation: Two Approaches

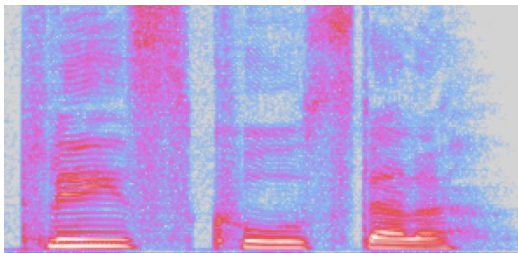
- **Vocal tract length perturbation (VTLP):** speed up and slow down VCTK training data to create additional artificial “speakers” for training

Speaker Augmentation: Two Approaches

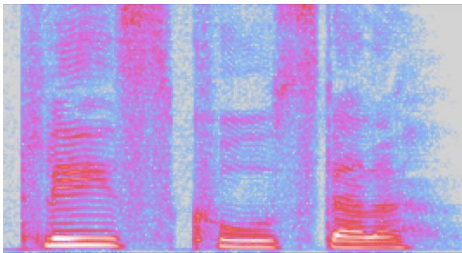
- **Vocal tract length perturbation (VTLP):** speed up and slow down VCTK training data to create additional artificial “speakers” for training
- **Speaker augmentation using low-quality data:** data which was not specifically collected for TTS but contains a large variety of speakers
 - GRID, WSJ1, WSJCAM, TIMIT
 - How should we handle the different channel and recording conditions?
 - Many English dialects; should we model them?

1st Approach: Artificial Speaker Augmentation (VTLP)

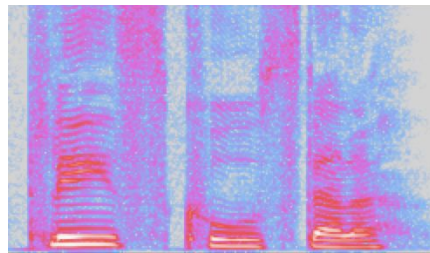
- **SoX:** 'speed' command at rates 0.9 (slower) and 1.1 (faster)
- Re-sampling of waveforms -> different fundamental frequency, speaking rate, formants, and spectra



0.9



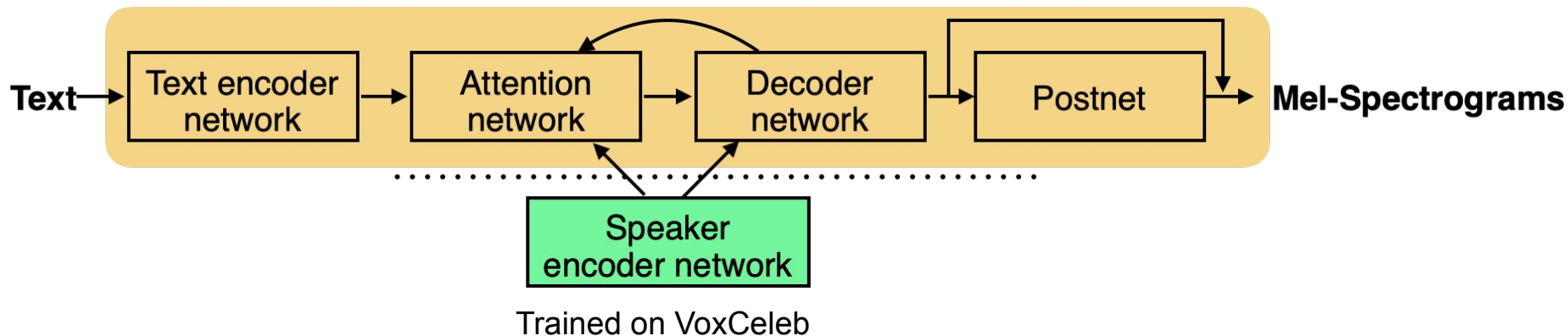
1.0
(original)



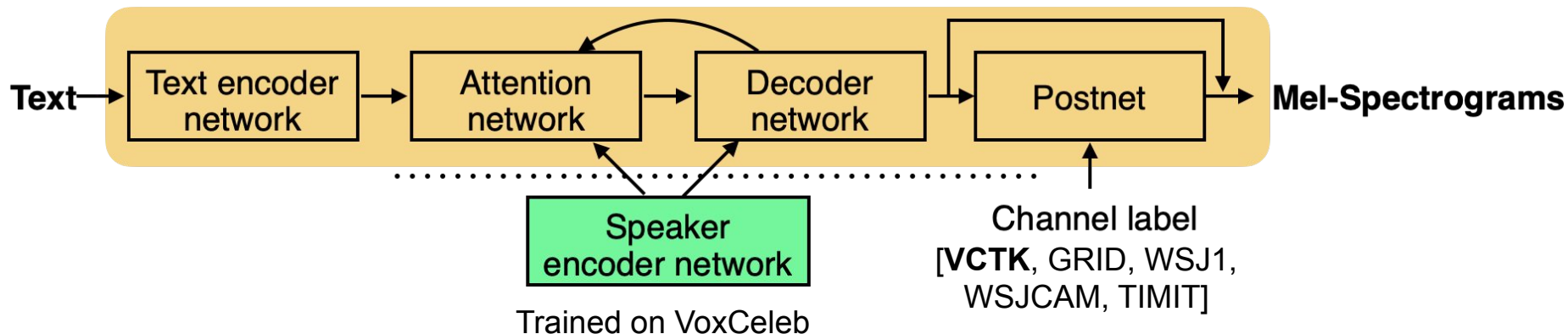
1.1



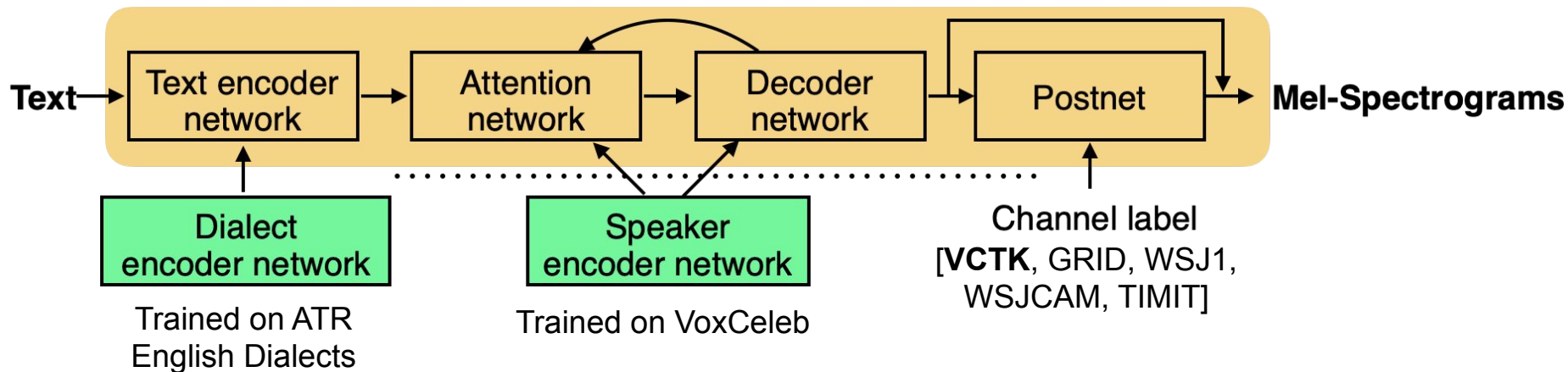
2nd Approach: Speaker augmentation using low-quality data



2nd Approach: Speaker augmentation using low-quality data

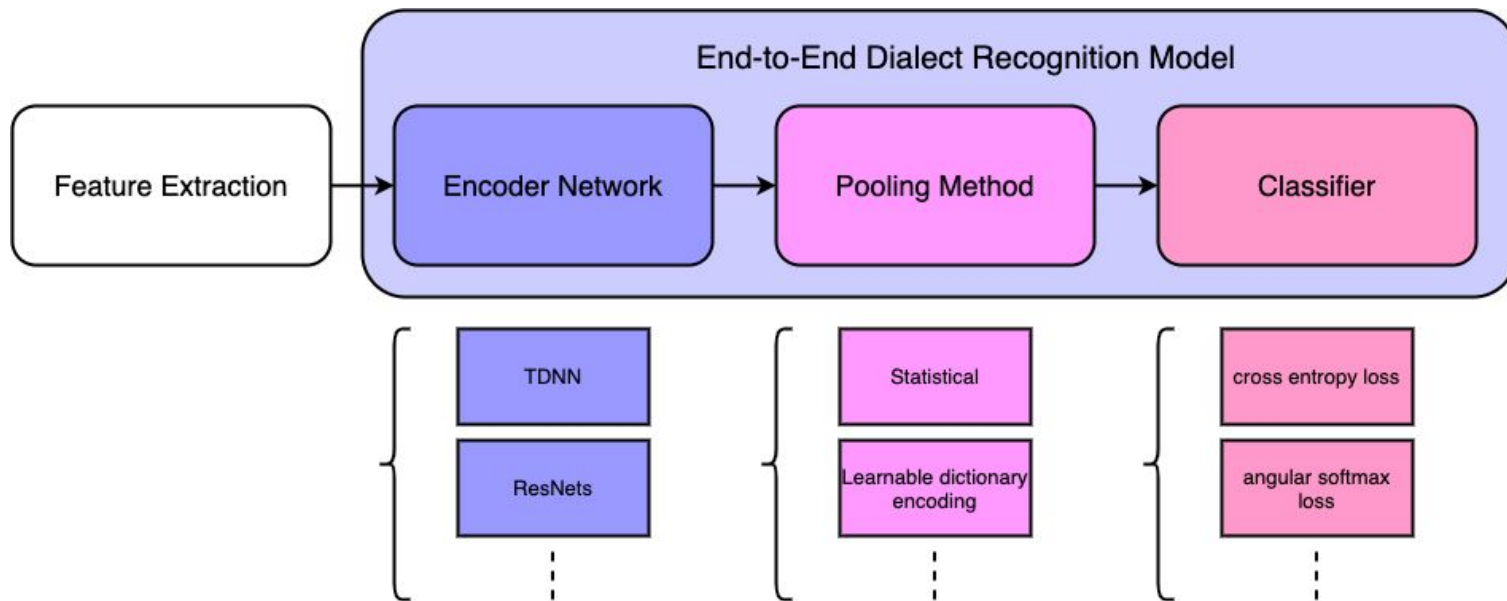


2nd Approach: Speaker augmentation using low-quality data



Extracting Dialect Embeddings from E2E Dialect Recognition Model

Trained on the ATR English Dialect corpus.

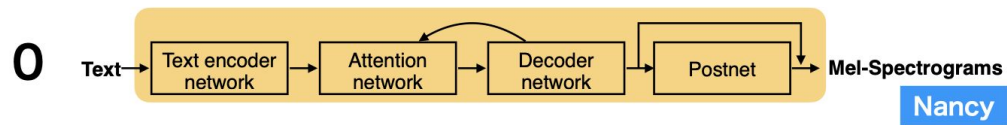


Final Dialect Embeddings (DE) Configurations

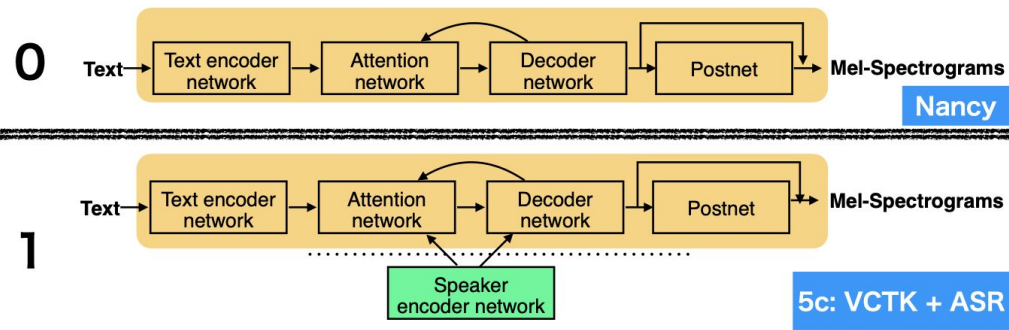
DE Selection Criterion: [cosine-similarity](#) between the embeddings of synthetic speech and target speaker's speech

	Phone			Char		
	dim	pl	dc	dim	pl	dc
DE1	256	m,s	32	128	m,s	32
DE2	256	m	64	256	m	32
DE3	256	m,s	64	32	m,s	64
DE4	32	m,s	64	512	m,s	32
DE5	64	m	64	64	m,s	32

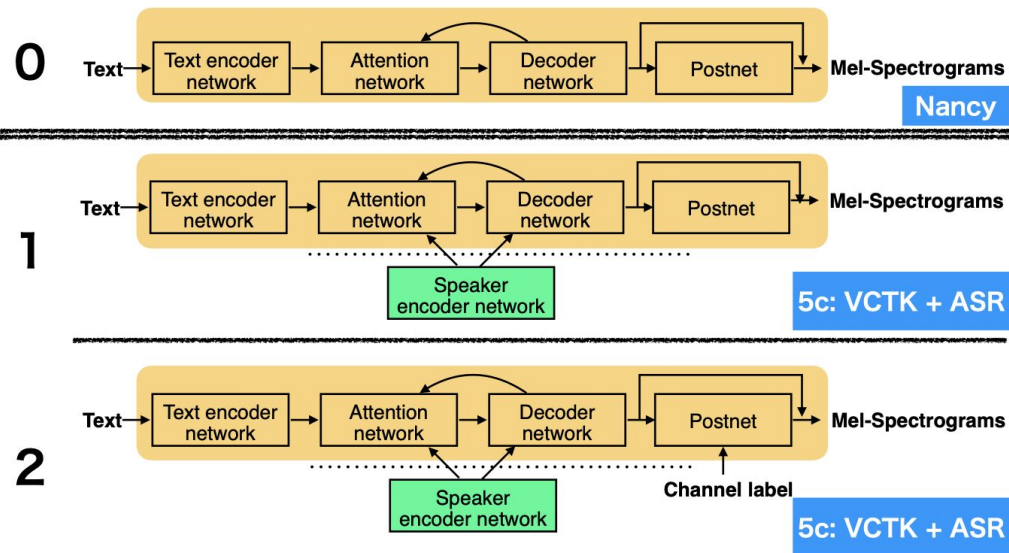
Multi-Step Warm-Start Training



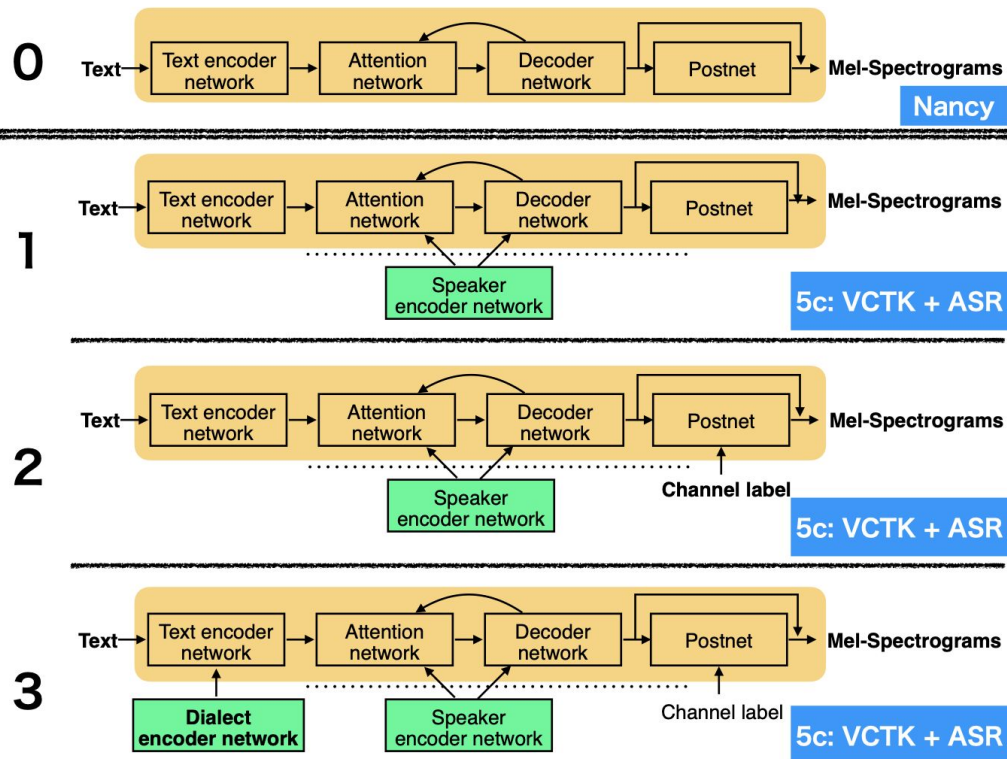
Multi-Step Warm-Start Training



Multi-Step Warm-Start Training



Multi-Step Warm-Start Training



Experiments

- Phone or character input Tacotron models
- Artificial speaker augmentation
- Low-quality data speaker augmentation:
 - GRID: 32 speakers; limited domain; some background noise
 - WSJ1: 50 American dialect speakers; news domain
 - WSJCAM: 85 British dialect speakers; news domain; line noise and reverberation
 - TIMIT: 50 American dialect speakers; small number of phonetically-rich sentences
- Channel labels
 - Training: label which corpus the utterance is from
 - Synthesis: use “VCTK” channel label for best quality
- Dialect encoder
 - 5 best for character input and for phoneme input
- Vocoder: WaveNet trained on VCTK

Evaluation

- **Naturalness:** Mean Opinion Score (1-5)
- **Speaker similarity:** Differential MOS (1-5)



Naturalness

☐ Very Bad ☐ Bad ☐ Fair ☐ Good ☐ Very Good



Same or different speaker?

☐ Definitely different ☐ Possibly different ☐ Not sure ☐ Possibly same ☐ Definitely same



Results: MOS and DMOS

	Naturalness			Speaker Similarity		
system	train	dev	test	train	dev	test
natural	4.5	4.4	4.5	4.6	4.5	4.5
vocoded	4.2	4.2	4.3	4.1	3.9	4.0
phone baseline	3.6	3.7	4.0	3.7	2.0	2.7
phone VTLP	3.7	3.7	4.0	3.6	2.1	2.7
phone 5c	3.8	3.5	3.7	3.4	2.1	2.6
phone 5c+CL	3.7	3.7	3.9	3.5	2.0	2.5
phone 5c+CL+DE1	2.4	2.4	2.5	3.2	2.3	2.4
phone 5c+CL+DE2	2.5	2.4	2.5	3.3	2.2	2.6
phone 5c+CL+DE3	3.9	3.7	3.9	3.6	2.1	2.6
phone 5c+CL+DE4	1.1	1.1	1.1	2.8	2.4	2.5
phone 5c+CL+DE5	3.9	3.7	3.8	3.4	2.0	2.6
char baseline	3.7	3.7	3.9	3.6	1.9	2.8
char VTLP	3.8	3.6	3.9	3.5	2.1	2.5
char 5c	3.7	3.5	3.7	3.3	2.0	2.6
char 5c+CL	3.8	3.8	3.9	3.6	2.1	2.6
char 5c+CL+DE1	2.5	2.5	2.5	3.3	2.1	2.5
char 5c+CL+DE2	4.0	3.7	4.0	3.6	2.0	2.5
char 5c+CL+DE3	3.9	3.4	3.8	3.6	2.1	2.4
char 5c+CL+DE4	4.0	3.7	3.9	3.6	2.1	2.5
char 5c+CL+DE5	3.9	3.8	3.9	3.5	2.0	2.6

Results: MOS and DMOS

	Naturalness			Speaker Similarity		
system	train	dev	test	train	dev	test
natural	4.5	4.4	4.5	4.6	4.5	4.5
vocoded	4.2	4.2	4.3	4.1	3.9	4.0
phone baseline	3.6	3.7	4.0	3.7	2.0	2.7
phone VTLP	3.7	3.7	4.0	3.6	2.1	2.7
phone 5c	3.8	3.5	3.7	3.4	2.1	2.6
phone 5c+CL	3.7	3.7	3.9	3.5	2.0	2.5
phone 5c+CL+DE1	2.4	2.4	2.5	3.2	2.3	2.4
phone 5c+CL+DE2	2.5	2.4	2.5	3.3	2.2	2.6
phone 5c+CL+DE3	3.9	3.7	3.9	3.6	2.1	2.6
phone 5c+CL+DE4	1.1	1.1	1.1	2.8	2.4	2.5
phone 5c+CL+DE5	3.9	3.7	3.8	3.4	2.0	2.6
char baseline	3.7	3.7	3.9	3.6	1.9	2.8
char VTLP	3.8	3.6	3.9	3.5	2.1	2.5
char 5c	3.7	3.5	3.7	3.3	2.0	2.6
char 5c+CL	3.8	3.8	3.9	3.6	2.1	2.6
char 5c+CL+DE1	2.5	2.5	2.5	3.3	2.1	2.5
char 5c+CL+DE2	4.0	3.7	4.0	3.6	2.0	2.5
char 5c+CL+DE3	3.9	3.4	3.8	3.6	2.1	2.4
char 5c+CL+DE4	4.0	3.7	3.9	3.6	2.1	2.5
char 5c+CL+DE5	3.9	3.8	3.9	3.5	2.0	2.6

Results: MOS and DMOS

system	Naturalness			Speaker Similarity		
	train	dev	test	train	dev	test
natural	4.5	4.4	4.5	4.6	4.5	4.5
vocoded	4.2	4.2	4.3	4.1	3.9	4.0
phone baseline	3.6	3.7	4.0	3.7	2.0	2.7
phone VTLP	3.7	3.7	4.0	3.6	2.1	2.7
phone 5c	3.8	3.5	3.7	3.4	2.1	2.6
phone 5c+CL	3.7	3.7	3.9	3.5	2.0	2.5
phone 5c+CL+DE1	2.4	2.4	2.5	3.2	2.3	2.4
phone 5c+CL+DE2	2.5	2.4	2.5	3.3	2.2	2.6
phone 5c+CL+DE3	3.9	3.7	3.9	3.6	2.1	2.6
phone 5c+CL+DE4	1.1	1.1	1.1	2.8	2.4	2.5
phone 5c+CL+DE5	3.9	3.7	3.8	3.4	2.0	2.6
char baseline	3.7	3.7	3.9	3.6	1.9	2.8
char VTLP	3.8	3.6	3.9	3.5	2.1	2.5
char 5c	3.7	3.5	3.7	3.3	2.0	2.6
char 5c+CL	3.8	3.8	3.9	3.6	2.1	2.6
char 5c+CL+DE1	2.5	2.5	2.5	3.3	2.1	2.5
char 5c+CL+DE2	4.0	3.7	4.0	3.6	2.0	2.5
char 5c+CL+DE3	3.9	3.4	3.8	3.6	2.1	2.4
char 5c+CL+DE4	4.0	3.7	3.9	3.6	2.1	2.5
char 5c+CL+DE5	3.9	3.8	3.9	3.5	2.0	2.6

MOS and DMOS for VTLP

system	Naturalness			Speaker Similarity		
	train	dev	test	train	dev	test
natural	4.5	4.4	4.5	4.6	4.5	4.5
vocoded	4.2	4.2	4.3	4.1	3.9	4.0
phone baseline	3.6	3.7	4.0	3.7	2.0	2.7
phone VTLP	3.7	3.7	4.0	3.6	2.1	2.7
phone 5c	3.8	3.5	3.7	3.4	2.1	2.6
phone 5c+CL	3.7	3.7	3.9	3.5	2.0	2.5
phone 5c+CL+DE1	2.4	2.4	2.5	3.2	2.3	2.4
phone 5c+CL+DE2	2.5	2.4	2.5	3.3	2.2	2.6
phone 5c+CL+DE3	3.9	3.7	3.9	3.6	2.1	2.6
phone 5c+CL+DE4	1.1	1.1	1.1	2.8	2.4	2.5
phone 5c+CL+DE5	3.9	3.7	3.8	3.4	2.0	2.6
char baseline	3.7	3.7	3.9	3.6	1.9	2.8
char VTLP	3.8	3.6	3.9	3.5	2.1	2.5
char 5c	3.7	3.5	3.7	3.3	2.0	2.6
char 5c+CL	3.8	3.8	3.9	3.6	2.1	2.6
char 5c+CL+DE1	2.5	2.5	2.5	3.3	2.1	2.5
char 5c+CL+DE2	4.0	3.7	4.0	3.6	2.0	2.5
char 5c+CL+DE3	3.9	3.4	3.8	3.6	2.1	2.4
char 5c+CL+DE4	4.0	3.7	3.9	3.6	2.1	2.5
char 5c+CL+DE5	3.9	3.8	3.9	3.5	2.0	2.6

VTLP does **not**
improve speaker
similarity

MOS and DMOS: ASR Data Speaker Augmentation

system	Naturalness			Speaker Similarity		
	train	dev	test	train	dev	test
natural	4.5	4.4	4.5	4.6	4.5	4.5
vocoded	4.2	4.2	4.3	4.1	3.9	4.0
phone baseline	3.6	3.7	4.0	3.7	2.0	2.7
phone VTLP	3.7	3.7	4.0	3.6	2.1	2.7
phone 5c	3.8	3.5	3.7	3.4	2.1	2.6
phone 5c+CL	3.7	3.7	3.9	3.5	2.0	2.5
phone 5c+CL+DE1	2.4	2.4	2.5	3.2	2.3	2.4
phone 5c+CL+DE2	2.5	2.4	2.5	3.3	2.2	2.6
phone 5c+CL+DE3	3.9	3.7	3.9	3.6	2.1	2.6
phone 5c+CL+DE4	1.1	1.1	1.1	2.8	2.4	2.5
phone 5c+CL+DE5	3.9	3.7	3.8	3.4	2.0	2.6
char baseline	3.7	3.7	3.9	3.6	1.9	2.8
char VTLP	3.8	3.6	3.9	3.5	2.1	2.5
char 5c	3.7	3.5	3.7	3.3	2.0	2.6
char 5c+CL	3.8	3.8	3.9	3.6	2.1	2.6
char 5c+CL+DE1	2.5	2.5	2.5	3.3	2.1	2.5
char 5c+CL+DE2	4.0	3.7	4.0	3.6	2.0	2.5
char 5c+CL+DE3	3.9	3.4	3.8	3.6	2.1	2.4
char 5c+CL+DE4	4.0	3.7	3.9	3.6	2.1	2.5
char 5c+CL+DE5	3.9	3.8	3.9	3.5	2.0	2.6

Speaker augmentation
with ASR data improves
naturalness for seen
speakers

MOS and DMOS: ASR Data Speaker Augmentation

system	Naturalness			Speaker Similarity		
	train	dev	test	train	dev	test
natural	4.5	4.4	4.5	4.6	4.5	4.5
vocoded	4.2	4.2	4.3	4.1	3.9	4.0
phone baseline	3.6	3.7	4.0	3.7	2.0	2.7
phone VTLP	3.7	3.7	4.0	3.6	2.1	2.7
phone 5c	3.8	3.5	3.7	3.4	2.1	2.6
phone 5c+CL	3.7	3.7	3.9	3.5	2.0	2.5
phone 5c+CL+DE1	2.4	2.4	2.5	3.2	2.3	2.4
phone 5c+CL+DE2	2.5	2.4	2.5	3.3	2.2	2.6
phone 5c+CL+DE3	3.9	3.7	3.9	3.6	2.1	2.6
phone 5c+CL+DE4	1.1	1.1	1.1	2.8	2.4	2.5
phone 5c+CL+DE5	3.9	3.7	3.8	3.4	2.0	2.6
char baseline	3.7	3.7	3.9	3.6	1.9	2.8
char VTLP	3.8	3.6	3.9	3.5	2.1	2.5
char 5c	3.7	3.5	3.7	3.3	2.0	2.6
char 5c+CL	3.8	3.8	3.9	3.6	2.1	2.6
char 5c+CL+DE1	2.5	2.5	2.5	3.3	2.1	2.5
char 5c+CL+DE2	4.0	3.7	4.0	3.6	2.0	2.5
char 5c+CL+DE3	3.9	3.4	3.8	3.6	2.1	2.4
char 5c+CL+DE4	4.0	3.7	3.9	3.6	2.1	2.5
char 5c+CL+DE5	3.9	3.8	3.9	3.5	2.0	2.6

Dialect embedding and channel label are required to see improvements with ASR data

MOS and DMOS: ASR Data Speaker Augmentation

system	Naturalness			Speaker Similarity		
	train	dev	test	train	dev	test
natural	4.5	4.4	4.5	4.6	4.5	4.5
vocoded	4.2	4.2	4.3	4.1	3.9	4.0
phone baseline	3.6	3.7	4.0	3.7	2.0	2.7
phone VTLP	3.7	3.7	4.0	3.6	2.1	2.7
phone 5c	3.8	3.5	3.7	3.4	2.1	2.6
phone 5c+CL	3.7	3.7	3.9	3.5	2.0	2.5
phone 5c+CL+DE1	2.4	2.4	2.5	3.2	2.3	2.4
phone 5c+CL+DE2	2.5	2.4	2.5	3.3	2.2	2.6
phone 5c+CL+DE3	3.9	3.7	3.9	3.6	2.1	2.6
phone 5c+CL+DE4	1.1	1.1	1.1	2.8	2.4	2.5
phone 5c+CL+DE5	3.9	3.7	3.8	3.4	2.0	2.6
char baseline	3.7	3.7	3.9	3.6	1.9	2.8
char VTLP	3.8	3.6	3.9	3.5	2.1	2.5
char 5c	3.7	3.5	3.7	3.3	2.0	2.6
char 5c+CL	3.8	3.8	3.9	3.6	2.1	2.6
char 5c+CL+DE1	2.5	2.5	2.5	3.3	2.1	2.5
char 5c+CL+DE2	4.0	3.7	4.0	3.6	2.0	2.5
char 5c+CL+DE3	3.9	3.4	3.8	3.6	2.1	2.4
char 5c+CL+DE4	4.0	3.7	3.9	3.6	2.1	2.5
char 5c+CL+DE5	3.9	3.8	3.9	3.5	2.0	2.6

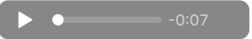
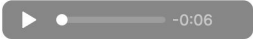
Architecture of
dialect encoder
matters



Evaluation: Dialect

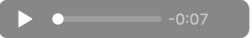
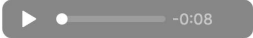
- Dialect: Multiple choice from set of VCTK dialects
 - References provided for listeners
 - [Frobenius distance](#) between synthesized speech and natural speech of confusion matrices of dialects

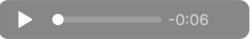
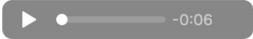


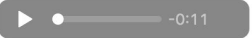
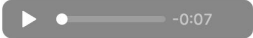
☐ American ☐ Canadian ☐ English ☐ Irish ☐ Northern Irish ☐ Scottish

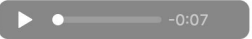
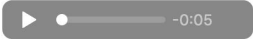
American


Canadian


English


Irish


Northern Irish


Scottish


Results: Dialects

system	Dialect confusion		
	train	dev	test
vocoded	0.06	0.32	0.32
phone baseline	0.20	1.06	1.12
phone VTLP	0.31	0.86	1.20
phone 5c	0.19	0.93	0.93
phone 5c+CL	0.19	0.84	0.99
phone 5c+CL+DE1	0.42	0.84	0.88
phone 5c+CL+DE2	0.34	0.95	0.81
phone 5c+CL+DE3	0.13	0.93	0.95
phone 5c+CL+DE4	0.44	0.88	0.90
phone 5c+CL+DE5	0.20	0.92	0.79
char baseline	0.25	0.96	0.86
char VTLP	0.12	1.00	1.17
char 5c	0.25	0.96	0.86
char 5c+CL	0.25	0.86	0.79
char 5c+CL+DE1	0.41	0.91	0.83
char 5c+CL+DE2	0.17	0.92	1.33
char 5c+CL+DE3	0.18	0.92	1.02
char 5c+CL+DE4	0.21	0.91	1.15
char 5c+CL+DE5	0.22	1.02	1.08

Results: Dialects

system	Dialect confusion		
	train	dev	test
vocoded	0.06	0.32	0.32
phone baseline	0.20	1.06	1.12
phone VTLP	0.31	0.86	1.20
phone 5c	0.19	0.93	0.93
phone 5c+CL	0.19	0.84	0.99
phone 5c+CL+DE1	0.42	0.84	0.88
phone 5c+CL+DE2	0.34	0.95	0.81
phone 5c+CL+DE3	0.13	0.93	0.95
phone 5c+CL+DE4	0.44	0.88	0.90
phone 5c+CL+DE5	0.20	0.92	0.79
char baseline	0.25	0.96	0.86
char VTLP	0.12	1.00	1.17
char 5c	0.25	0.96	0.86
char 5c+CL	0.25	0.86	0.79
char 5c+CL+DE1	0.41	0.91	0.83
char 5c+CL+DE2	0.17	0.92	1.33
char 5c+CL+DE3	0.18	0.92	1.02
char 5c+CL+DE4	0.21	0.91	1.15
char 5c+CL+DE5	0.22	1.02	1.08

Results: Dialects

system	Dialect confusion		
	train	dev	test
vocoded	0.06	0.32	0.32
phone baseline	0.20	1.06	1.12
phone VTLP	0.31	0.86	1.20
phone 5c	0.19	0.93	0.93
phone 5c+CL	0.19	0.84	0.99
phone 5c+CL+DE1	0.42	0.84	0.88
phone 5c+CL+DE2	0.34	0.95	0.81
phone 5c+CL+DE3	0.13	0.93	0.95
phone 5c+CL+DE4	0.44	0.88	0.90
phone 5c+CL+DE5	0.20	0.92	0.79
char baseline	0.25	0.96	0.86
char VTLP	0.12	1.00	1.17
char 5c	0.25	0.96	0.86
char 5c+CL	0.25	0.86	0.79
char 5c+CL+DE1	0.41	0.91	0.83
char 5c+CL+DE2	0.17	0.92	1.33
char 5c+CL+DE3	0.18	0.92	1.02
char 5c+CL+DE4	0.21	0.91	1.15
char 5c+CL+DE5	0.22	1.02	1.08

Audio Samples

Natural speech



Phone baseline



phone 5c+CL+DE3



Conclusions

- We investigated **artificial speaker augmentation (VTLP)** and **speaker augmentation using real data from lower-quality corpora**
- We revised the **postnet** and **encoder** of Tacotron to support **channel** and **dialect** variations from the low-quality data
- Use of low-quality data with a **variety of speakers and dialects** is an effective augmentation strategy
- Contrary to our initial expectations, **naturalness of seen speakers has been improved**
- Listeners' ratings of perceived dialects are better matched to natural speech for unseen speakers

Thank You!

Audio samples:

<https://nii-yamagishilab.github.io/samples-multi-speaker-tacotron/augment.html>

GitHub:

<https://github.com/nii-yamagishilab/multi-speaker-tacotron>