2022 IEEE International Conference on Acoustics, Speech and Signal Processing

Using Multi-talker Neural TTS to Synthesize Speech for Dysarthric Speech Recognition

Mohammad Soleymanpour[1], Michael T. Johnson[1], Rahim Soleymanpour[2], Jeffrey Berry[3]

[1] University of Kentucky, Lexington, KY, USA

[2] University of Connecticut, Storrs, CT, USA

[3] Marquette University, Milwaukee, WI, USA

OUTLINE:

1. Introduction
2. Multi-talker Neural TTS to Synthesize Speech
3. Experiments
4. Results and Discussion
5. Conclusion

INTRODUCTION:

1. Dysarthria and Severity Level
2. Problem: Challenges of Dysarthric ASR
3. Aim of this project
4. Solution: Dysarthric Speech Augmentation/Synthesis

INTRODUCTION:

TORGO dataset:

- TORGO, dysarthric database with aligned acoustic/articulatory.
- 15 speakers
  - 8 Dysarthric: 5 males, 3 females
  - 7 control: 4 males, 3 females
- The dataset contains about 23 hours.

Severity Level	Intelligibility Category	Speaker ID
Normal	Intelligible	FC01
		FC02
		FC03
		MC01
		MC02
		MC03
		MC04
Very low		F03
		F04
		M03
Low	Unintelligible	F01
		M05
Medium		M01
		M02
		M04

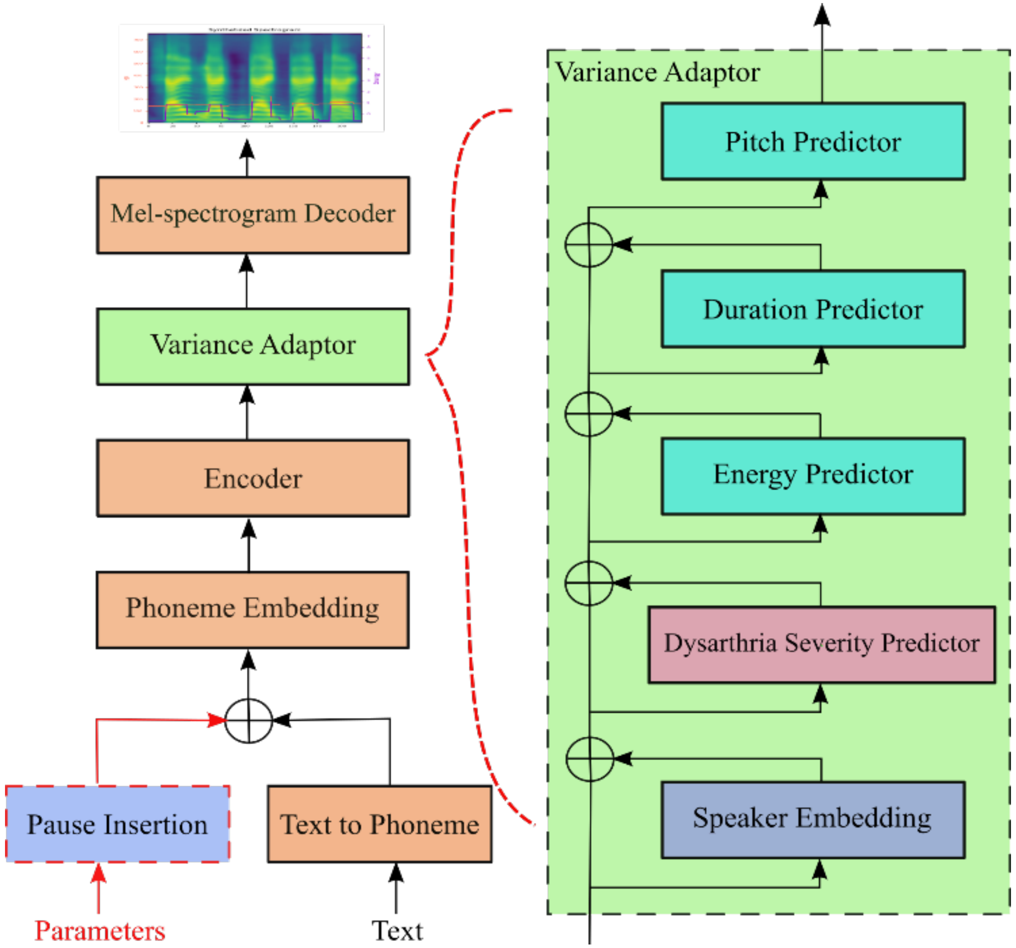
INTRODUCTION:

1. Required Parameters: Pitch, Duration, Energy, Pause, Severity level
2. Aim to improve multi-speaker end-to-end TTS systems to synthesize dysarthric speech
3. Advantages of this work?
4. Evaluation: DNN-HMM based ASR models and audio samples at our demo page

USING MULTI-TALKER NEURAL TTS TO SYNTHESIZE SPEECH:

- Adjusted Model:
  - Contains 4 feed-forward transformer blocks in the encoder and decoder.
  - The decoder generates an 80-dimensional mel-spectrogram from hidden state.
  - The size of phoneme embedding is 256.
  - Trained the adjusted model with a GeForce RTX 2080 Ti.

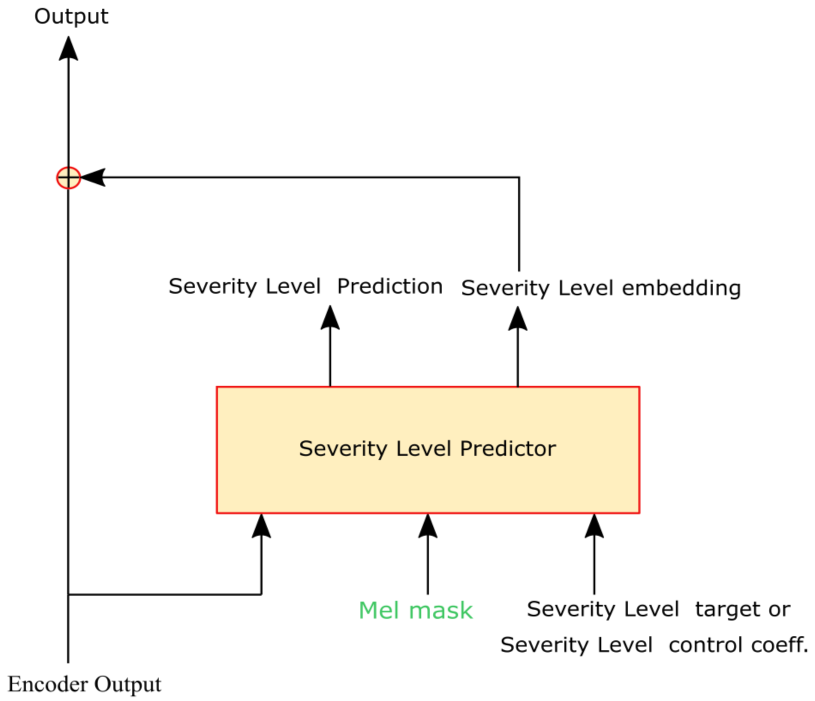
- Predictors:
  - 2-layer 1D-convolutional network
  - Each followed by a normalization and a dropout layer,
  - An extra linear layer to project the hidden states into the output sequence.



An overview of the proposed architecture

USING MULTI-TALKER NEURAL TTS TO SYNTHESIZE SPEECH:

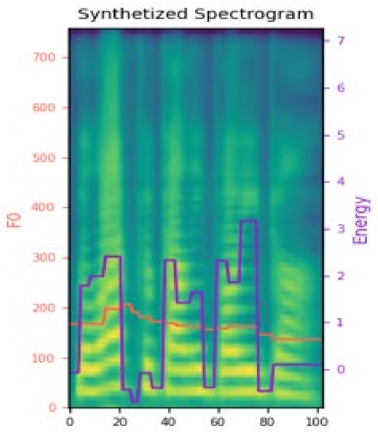
- Dysarthric Severity Predictors:
  - 2-layer 1D-convolutional network
  - Each followed by the layer normalization and a dropout layer,
  - An extra linear layer to project the hidden states into the output sequence.



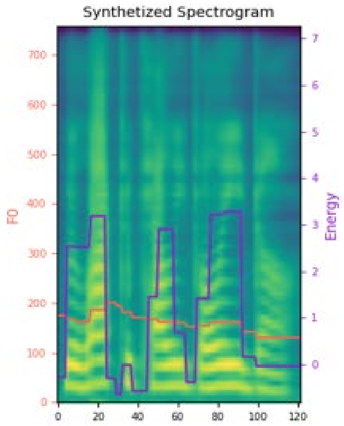
An overview of Dysarthric Severity Predictors

USING MULTI-TALKER NEURAL TTS TO SYNTHESIZE SPEECH:

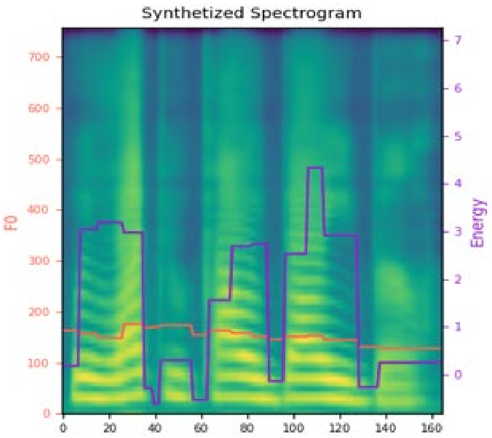
Input text: “We would like to paly volleyball”



Severity level: 0.0 (Normal)



Severity level: 1.0 (Combined Levels of very low and low)



Severity level: 2.0 (Moderate)

Effect of dysarthria severity coefficients in synthesizing dysarthric speech for speaker MC04

Demo page: https://mohammadelc.github.io/SpeechGroupUKY/

EXPERIMENT FOR DYSARTHRIC SPEECH RECOGNITION:

- Exp.1 included augmented speech across 3 severities with pause insertion.
- Exp.2 included augmented speech across severity, pause, pitch, energy, and duration.
- For ASR, Pytorch-kaldi is used to the train model.
- A light Gated Recurrent Unit (liGRU) architecture is implemented with fMLLR transformed features.
- For testing, a leave-one-speaker-out cross-validation procedure was applied.

The prosody coefficients for synthesizing dysarthric speech in the two experiments

Coef.	Baseline	Exp. 1	Exp. 2
Pitch	-	1.0	[0.1, 0.6, 1.2, 1.75]
Energy	-	1.0	[0.1, 1.0, 2.0]
Duration	-	1.0	[1.0, 1.3, 1.6, 1.8]
Severity level	-	[0.0, 1.0, 2.0]	[0.0, 1.0, 2.0]
Pause insertion	-	Yes	Yes
Total utterance	~ 16000	~ ×3	~ ×10

RESULTS AND DISCUSSION:

- In the first experiment that only used severity synthesis and pause insertion, the synthesized speech used for augmenting ASR training improved from 44.5% to 41.6%.

- In the second experiment, average WER performance across all speakers improved from 44.5% to 39.2%.

- On average, the first and second experiments reduced WER by 6.5 %, 12.2% with the respect to the baseline.

WER of each test speaker for the two augmentation experiments

Severity Level	Test Spk	WER (%)				
		Baseline	Exp. 1	Exp. 2	[22]	[23]
Very low	F04	16.8	16.3	14.5	18.3	13.1
	M03	10.9	12.7	10.7	18.2	17.7
Low	F03	46.6	39.3	36.8	44.2	39.1
Moderate	F01	58.3	52.4	50.4	71.5	39.6
	M01	55.4	51.3	50.3	69.3	62.2
	M02	44	43.1	38.4	70.9	42.9
	M04	65.8	64.2	62	79.9	69.0
	M05	58.2	53.6	49.6	77.2	62.6
Overall Average		44.5	41.6	39.2	56.2	43.3

RESULTS AND DISCUSSION:

- This table shows that augmentation using synthetic speech at three dysarthria levels with pause insertion improved the WER of each severity level on average except for the group with the low severity.

- Augmentation using synthetic speech at three severity levels plus pause insertion, further varying energy, pitch, and duration improved WER across all severity levels.

WER of each severity level for the two augmentation experiments.

Severity level	**baseline**	**Exp. 1**	**Exp. 2**	**Improvement**	
				Exp.1	**Exp.2**
Very Low	13.8	14.5	12.6	-4.7%	**9%**
Low	46.6	39.3	36.8	**7.3%**	**21%**
Moderate	56.3	52.9	50.1	**6%**	**11%**
All	44.5	41.6	39.2	**6.5%**	**12.2%**

CONCLUSION:

- **Main contribution**: Adding a dysarthria severity level coefficient and a pause insertion model to synthesize dysarthric speech for varying severity levels.
- **Result**: A DNN-HMM ASR model trained on additional synthetic dysarthric speech achieves WER improvement of 12.2% compared to the baseline.
- **For the future**: we intend to combine the all dysarthric dataset to have more data with aligning their severity categories and also try Zero-shot learning to enrich speaker in different severity groups.

THANK YOU FOR YOUR ATTENTION