

Learning to Transfer Multi-speaker Emotional Prosody to a Neutral Speaker

Sungjae Cho, Sejik Park, Tae-Ho Kim, Soo-Young Lee

KAIST

Show and Tell Session @ICASSP2020

Content

- Data
- Model
- Training settings
- Demo 1: Multi-speaker emotional TTS demo
- Demo 2: Emotional TTS spoken by a neutral speaker

Why is transferring emotional prosody valuable?

Acquiring emotional speech is expensive because it requires professional acting ability.

it would be economical, beneficial to transfer emotional prosody to neutral voices.

Data

LJ-Speech-1.1

Neutral single female speaker

<https://keithito.com/LJ-Speech-Dataset/>

Total Clips	13,100
Total Duration	23:55:17
Mean Clip Duration	6.57 sec

Train clips = 12,500
Validation clips = 100
Test clips = 500

EmoV-DB

Emotional two female speaker

<https://github.com/numediart/EmoV-DB>

Speaker	Emotion	Clips	mean(dur)	Sum(dur)
bea	amused	296	2.97	00:14:37
bea	angry	304	2.76	00:13:59
bea	disgusted	333	3.85	00:21:22
bea	neutral	357	3.93	00:23:22
bea	sleepy	496	4.00	00:33:04
jenie	amused	222	3.31	00:12:14
jenie	angry	496	3.32	00:27:26
jenie	disgusted	189	3.59	00:11:17
jenie	neutral	417	4.53	00:31:30
jenie	sleepy	466	3.38	00:26:13

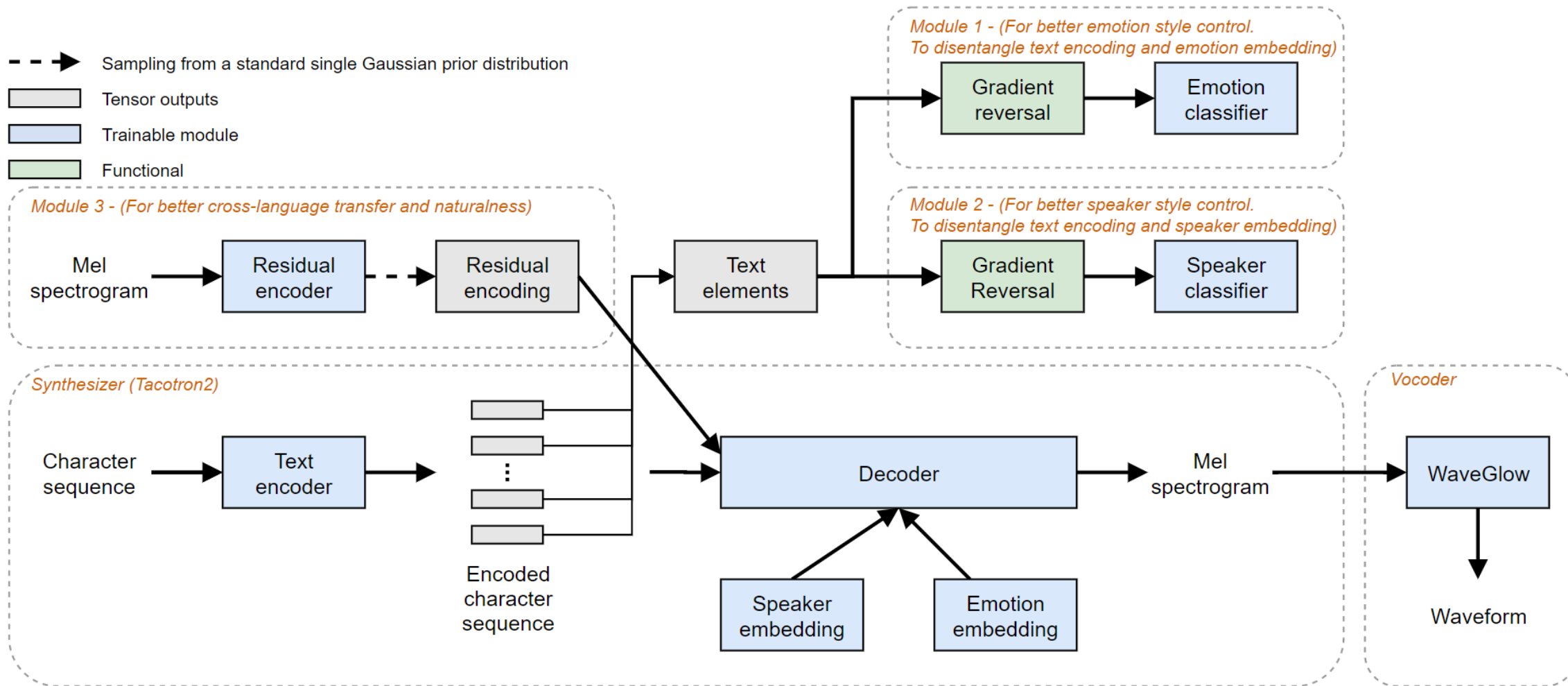
Speaker	Clips	mean(dur)	Sum(dur)
bea	1786	17.51	1:46:24
jenie	1712	17.85	1:44:01

Emotion	Clips	mean(dur)	Sum(dur)
amused	259	3.14	0:13:26
angry	400	3.04	0:20:42
disgusted	261	3.72	0:16:19
neutral	387	4.23	0:27:26
sleepy	481	3.69	0:29:39

Train : Val : Test = 6 : 2 : 2

Model

Our model follows Zhang et al. (2019) except for emotion embedding, Module 1 and vocoder.



Zhang et al. (2019) Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. InterSpeech 2019.

Ver. 4/17/2020













Learning to Transfer Multi-speaker Emotional Prosody
to a Neutral Speaker

Training settings











- Used frozen pre-trained WaveGlow vocoder
 - Provided in <https://github.com/NVIDIA/tacotron2>
- Initialized Tacotron2 weights pre-trained by LJ-Speech-1.1
 - Provided in <https://github.com/NVIDIA/tacotron2>
- Upsampling w.r.t. (speaker, emotion) pairs
 - Samples per batch equally distributed w.r.t. (speaker, emotion) pairs
- Followed training settings of [Zhang et al., 2019]
 - Batch size 256, exponential learning rate decay, following loss
- Loss
$$\text{MSE}(\text{MelSpec}) + 0.02 \cdot \text{CE}(\text{SpkClsf}) + 0.02 \cdot \text{CE}(\text{EmoClsf}) + \beta \cdot \text{KLD}(\text{ResidualEncoding} || N(0,1))$$
- β is set according to cyclical annealing schedule [Fu et al., 2019]

Fu et al., (2019) Cyclical annealing schedule: A simple approach to mitigating KL vanishing. NAACL 2019.













Demo 1: Multi-speaker emotional TTS demo

Speaker	B		J	
Emotion	Neutral	Amused	Neutral	Amused
Script 1	I'm amused. I'm really amused.			
Audio 1				
Script 2	What is that?			
Audio 2				
Script 3	I have first seen this in my life.			
Audio 3				













Demo 1: Multi-speaker emotional TTS demo

Speaker	B		J	
Emotion	Neutral	Angry	Neutral	Angry
Script 1	I'm angry. I'm really angry.			
Audio 1				
Script 2	It makes me so upset.			
Audio 2				
Script 3	What is this stupid thing?			
Audio 3				







Demo 1: Multi-speaker emotional TTS demo

Speaker	B		J	
Emotion	Neutral	Disgusted	Neutral	Disgusted
Script 1	I'm disgusted. I'm really disgusted.			
Audio 1				
Script 2	It smells so disgusting.			
Audio 2				
Script 3	It feels really strange.			
Audio 3				






Demo 1: Multi-speaker emotional TTS demo

Speaker	B		J	
Emotion	Neutral	Sleepy	Neutral	Sleepy
Script 1	I'm sleepy. I'm really sleepy.			
Audio 1				
Script 2	It is too late. It is time to go to bed.			
Audio 2				
Script 3	I didn't remember that scene. It was so bored.			
Audio 3				







Demo 2: Emotional TTS spoken by a neutral speaker

Speaker	L	
Emotion	Neutral	Amused
Script 1	I'm amused. I'm really amused.	
Audio 1		
Script 2	What is that?	
Audio 2		
Script 3	I have first seen this in my life.	
Audio 3		







Demo 2: Emotional TTS spoken by a neutral speaker

Speaker	L	
Emotion	Neutral	Angry
Script 1	I'm angry. I'm really angry.	
Audio 1		
Script 2	It makes me so upset.	
Audio 2		
Script 3	What is this stupid thing?	
Audio 3		

Demo 2: Emotional TTS spoken by a neutral speaker

Speaker	L	
Emotion	Neutral	Disgusted
Script 1	I'm disgusted. I'm really disgusted.	
Audio 1		
Script 2	It smells so disgusting.	
Audio 2		
Script 3	It feels really strange.	
Audio 3		

Demo 2: Emotional TTS spoken by a neutral speaker

Speaker	L	
Emotion	Neutral	Sleepy
Script 1	I'm sleepy. I'm really sleepy.	
Audio 1		
Script 2	It is too late. It is time to go to bed.	
Audio 2		
Script 3	I didn't remember that scene. It was so bored.	
Audio 3		

End