30 August – 3 September 2021
INTERSPEECH 2021
BRNO | CZECHIA
Speech everywhere!

Multi-speaker Emotional Text-to-speech Synthesizer

Sungjae Cho[1,†], Soo-Young Lee[2]

[1]Korea Institute of Science and Technology, Republic of Korea

[2]Korea Advanced Institute of Science and Technology, Republic of Korea

sj.cho@snu.ac.kr, sylee@kaist.ac.kr

Show & Tell Session @InterSpeech2021

16:00-18:00 (GMT+2), September 1, 2021

[†]work done at KAIST

Flow of video

Summarize our paper.

Introduce our demo pages.

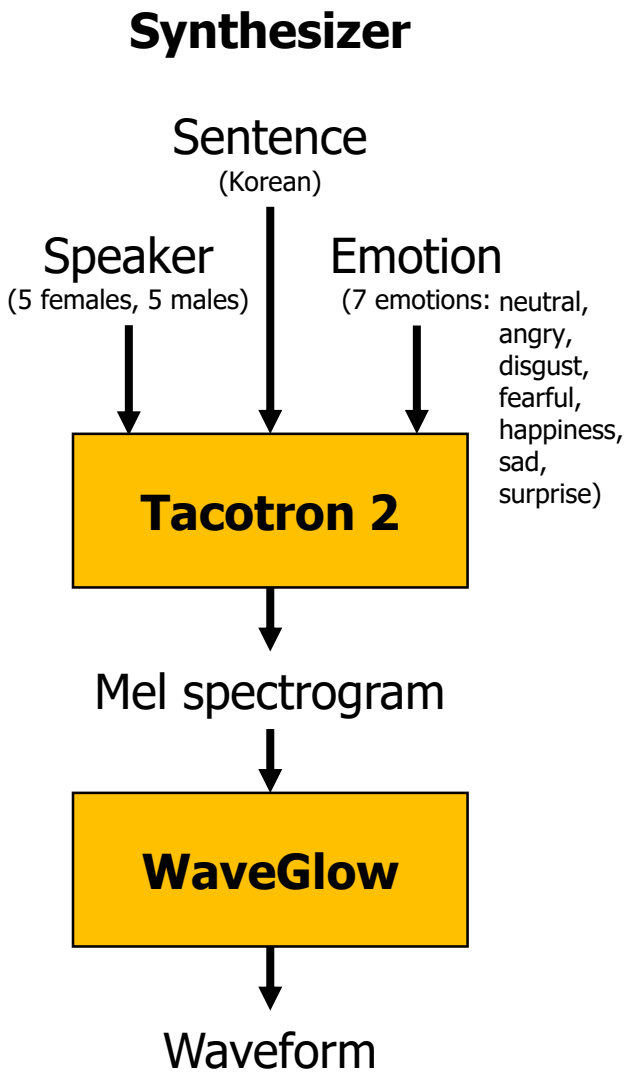
Motivations: Why we did this work

1. We wanted to develop a **methodology** to make multi-speaker emotional text-to-speech (TTS) systems, given imbalanced data distributions for multiple speakers and emotions.
   - Most studies on emotional TTS have trained models on
     - a small number of speakers or
     - balanced class distributions.

Motivations: Why we did this work

1. We wanted to develop a **methodology** to make multi-speaker emotional text-to-speech (TTS) systems, given imbalanced data distributions for multiple speakers and emotions.
   - Most studies on emotional TTS have trained models on
     - a small number of speakers or
     - balanced class distributions.
2. We wanted to **demonstrate** our synthesized audios for all possible speakers and emotions.
   - Most studies just offer audios for a small number of cases.

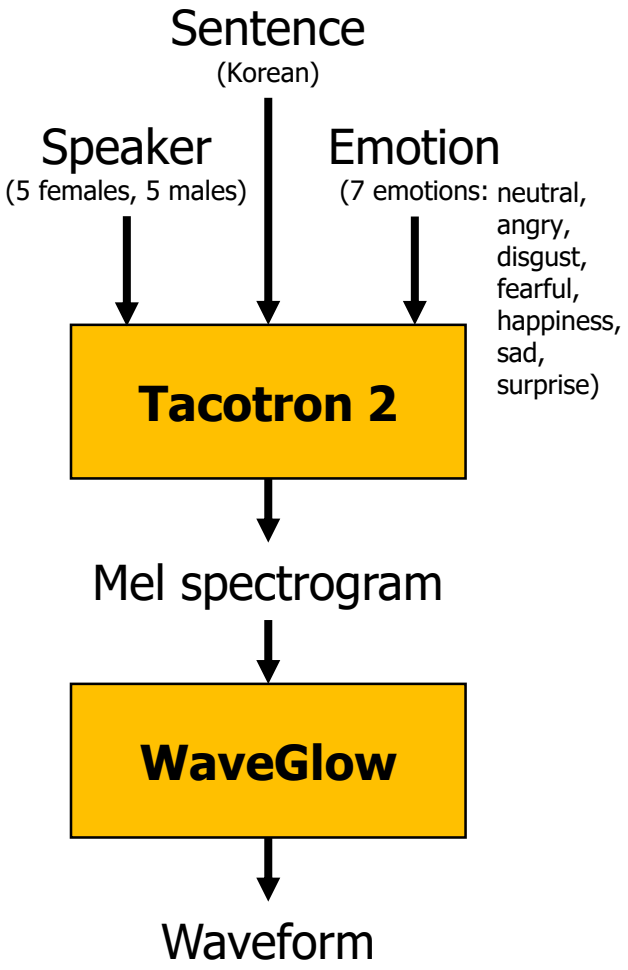
Methodology for multi-speaker emotional synthesis



Details are presented in our paper.

Methodology for multi-speaker emotional synthesis

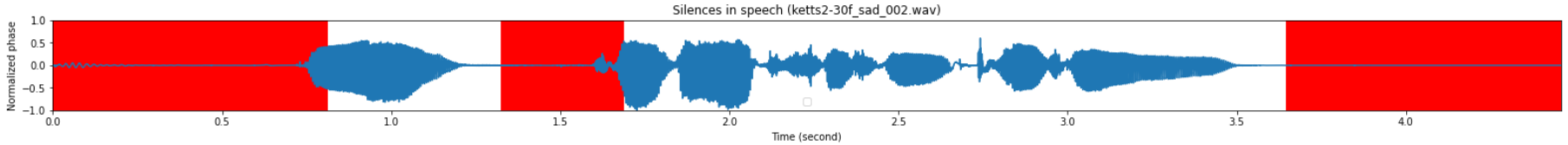
Synthesizer



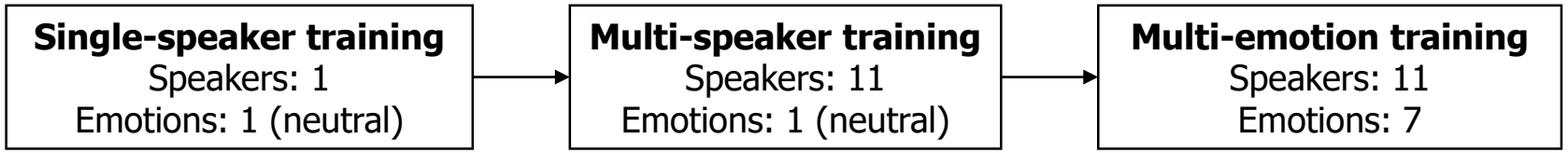
Details are presented in our paper.

Method 1: Silence removal

- Silences (red areas) at the start, end, and middle of training speech are removed.
- This silence removal accelerates learning of Tacotron 2 b/c linear text-speech alignments.

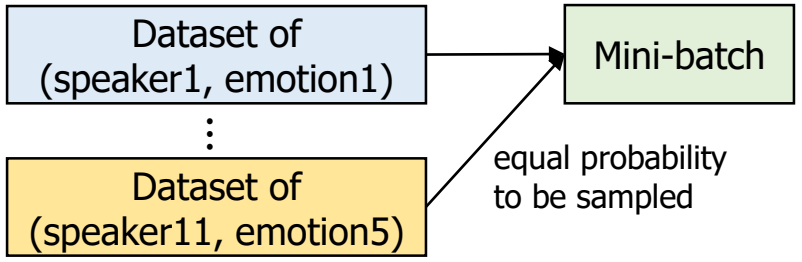


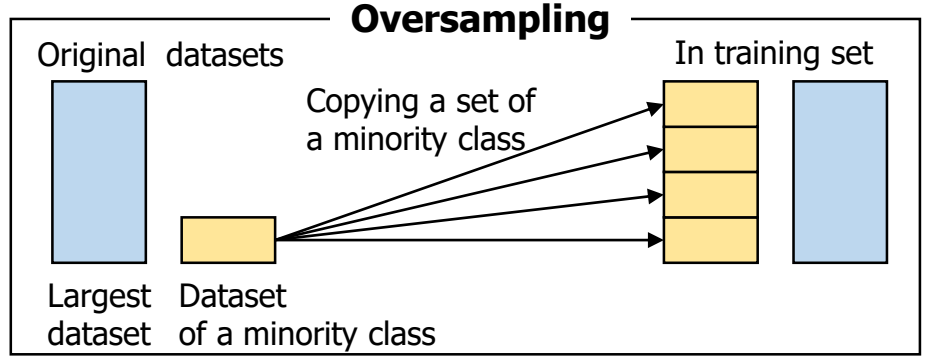
Method 2: Curriculum learning for Tacotron 2



Method 3: Oversampling

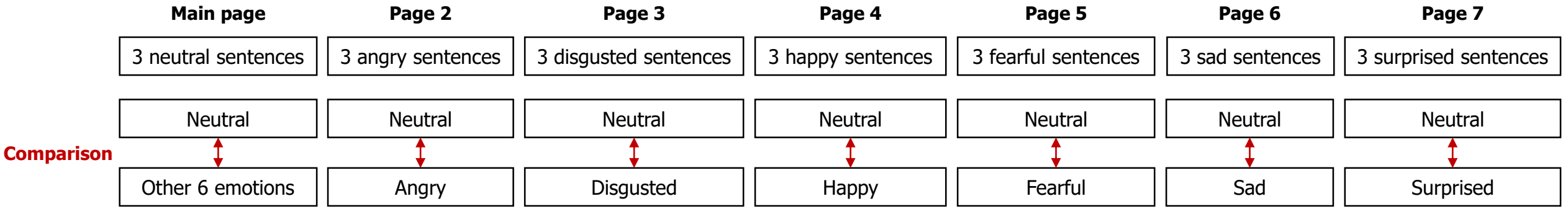
- Training samples of each speaker-emotion pair have equal probability to appear in mini-batches, by oversampling.



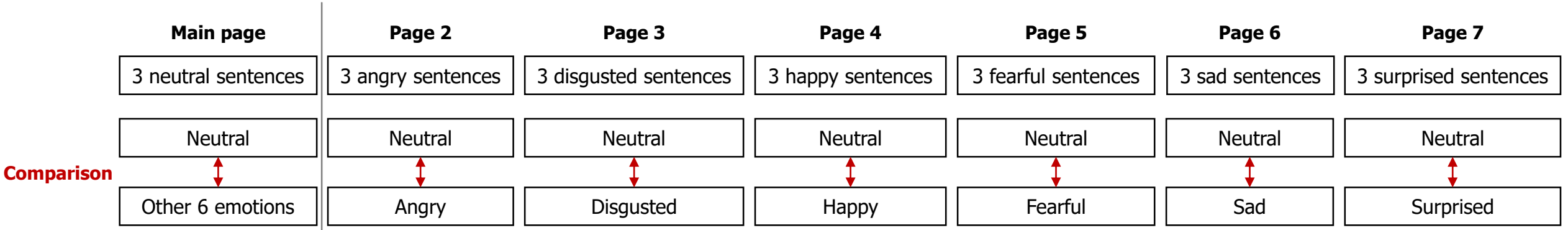


Results

- Results
  - Our system can generate speech for all available 10 speakers and 7 emotions.
  - Disgusted and surprised speech can be synthesized for speakers whose training datasets do not exist for disgust and surprise.
- Demo pages
  - Home: https://github.com/sungjae-cho/InterSpeech2021_STDemo
  - Our demo pages provide the followings for all speakers.



Results



Neutral sentence 1

Sentence	이 음성합성기는 열 명의 화자와 일곱 개의 감정을 합성할 수 있습니다.						
Pronouncing	i eumseonghabseong-gineun yeol myeong-ui hwajawa ilgob gaeui gamjeong-eul habseonghal su issseubnida.						
Meaning	This speech synthesizer can synthesize for ten speakers and seven emotions.						
Speaker	Emotion						
	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
ketts-30f	Play	Play	Play	Play	Play	Play	Play
ketts-30m	Play	Play	Play	Play	Play	Play	Play
ketts2-20m	Play	Play	Play	Play	Play	Play	Play
ketts2-30f	Play	Play	Play	Play	Play	Play	Play
ketts2-40m	Play	Play	Play	Play	Play	Play	Play
ketts2-50f	Play	Play	Play	Play	Play	Play	Play
ketts2-50m	Play	Play	Play	Play	Play	Play	Play
ketts2-60f	Play	Play	Play	Play	Play	Play	Play
ketts3-f	Play	Play	Play	Play	Play	Play	Play
ketts3-m	Play	Play	Play	Play	Play	Play	Play

Angry sentence 1

Sentence	나는 그 행동에 정말 화가 나.	
Pronouncing	naneun geu haengdong-e jeongmal hwaga na.	
Meaning	I'm really angry about that behavior.	
Speaker	Emotion	
	Neutral	Anger
ketts-30f	Play	Play
ketts-30m	Play	Play
ketts2-20m	Play	Play
ketts2-30f	Play	Play
ketts2-40m	Play	Play
ketts2-50f	Play	Play
ketts2-50m	Play	Play
ketts2-60f	Play	Play
ketts3-f	Play	Play
ketts3-m	Play	Play

Disgusted sentence 1

Sentence	나는 정말 이 상황이 너무 싫어.	
Pronouncing	naneun jeongmal i sanghwang-i neomu silh-eo.	
Meaning	I really hate this situation.	
Speaker	Emotion	
	Neutral	Disgust
ketts-30f	Play	Play
ketts-30m	Play	Play
ketts2-20m	Play	Play
ketts2-30f	Play	Play
ketts2-40m	Play	Play
ketts2-50f	Play	Play
ketts2-50m	Play	Play
ketts2-60f	Play	Play
ketts3-f	Play	Play
ketts3-m	Play	Play