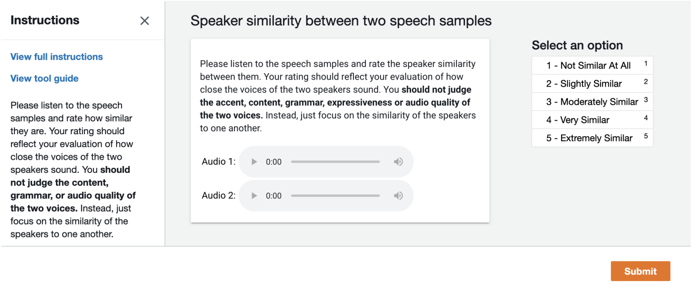
Expressive Neural Voice Cloning - Supplementary

1. User Study Templates

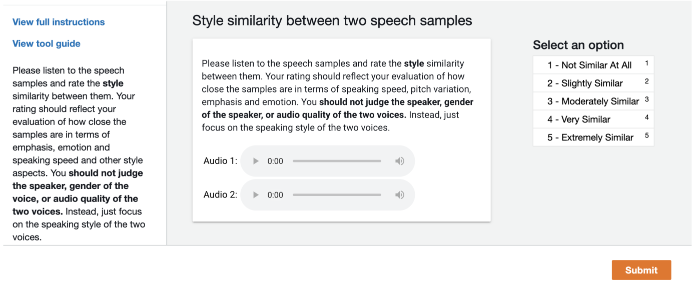


Speaker similarity between two speech samples

Please listen to the speech samples and rate the speaker similarity between them. Your rating should reflect your evaluation of how close the voices of the two speakers sound. You **should not judge the accent, content, grammar, expressiveness or audio quality of the two voices.** Instead, just focus on the similarity of the speakers to one another.

Please listen to the speech samples and rate how similar they are. Your rating should reflect your evaluation of how close the voices of the two speakers sound. You **should not judge the content, grammar, or audio quality of the two voices.** Instead, just focus on the similarity of the speakers to one another.

Figure 1: Interface for MOS evaluations on speaker similarity

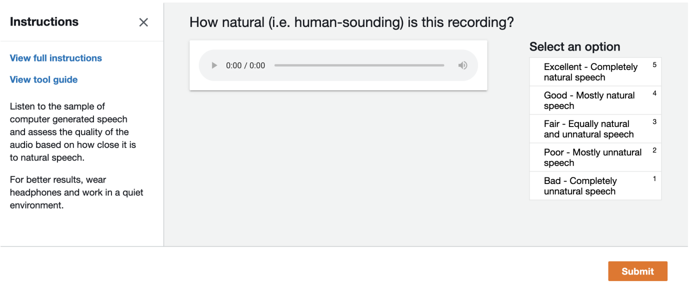


Style similarity between two speech samples

Please listen to the speech samples and rate the **style** similarity between them. Your rating should reflect your evaluation of how close the samples are in terms of speaking speed, pitch variation, emphasis and emotion. You **should not judge the speaker, gender of the speaker, or audio quality of the two voices.** Instead, just focus on the speaking style of the two voices.

Please listen to the speech samples and rate the **style** similarity between them. Your rating should reflect your evaluation of how close the samples are in terms of emphasis, emotion and speaking speed and other style aspects. You **should not judge the speaker, gender of the voice, or audio quality of the two voices.** Instead, just focus on the speaking style of the two voices.

Figure 2: Interface for MOS evaluations on style similarity



How natural (i.e. human-sounding) is this recording?

Listen to the sample of computer generated speech and assess the quality of the audio based on how close it is to natural speech.

For better results, wear headphones and work in a quiet environment.

Instructions
View full instructions
View tool guide
0:00 / 0:00
Select an option
Excellent - Completely natural speech 5
Good - Mostly natural speech 4
Fair - Equally natural and unnatural speech 3
Poor - Mostly unnatural speech 2
Bad - Completely unnatural speech 1
Submit

Figure 3: Interface for MOS evaluations on speech naturalness

2. Speaker Similarity MOS

We conduct a crowd-sourced user study on Amazon Mechanical Turk in which we ask users to rate the speaker similarity between two speech samples on a 5 point scale. This is a cross dataset evaluation (similar to all other evaluations conducted in our paper) where the mel synthesizer model is trained on LibriTTS dataset and evaluated on VCTK dataset. We use 10 target speaker samples for each cloning technique and synthesize 25 samples for the Text task for 20 randomly selected speakers in the VCTK dataset. Table 2 presents our speaker similarity evaluations for different approaches. While our methods outperform the different speaker baseline, they fall short in comparison to the same speaker MOS. This is likely due to the difference in accents of the synthesized speech and the ground truth audio. Our synthesizer is trained on the Libri-TTS dataset containing primarily North American English accented speakers but the VCTK dataset predominantly contains British speakers. The scores obtained for out techniques are similar to those reported in Jia et al. (2018) for their independently conducted study on cross dataset speaker similarity evaluation.

Approach	Speaker similarity MOS
Real data same speaker	3.78 ± 0.11
Real data different speaker	1.61 ± 0.12
Tacotron2 + GST - Zero Shot	2.05 ± 0.10
Proposed Model - Zero Shot	2.21 ± 0.11
Proposed Model - Adaptation Whole	2.98 ± 0.11
Proposed Model - Adaptation Decoder	2.78 ± 0.11

Table 1: Speaker similarity MOS with 95% confidence interval for the Text task.

3. Model Adaptation hyper-parameter details

Technique	Parameters	Samples	Iterations	Wall clock time
Adaptation Whole	32 million	1	100	$\sim$ 3 mins
Adaptation Whole	32 million	5	100	$\sim$ 3 mins
Adaptation Whole	32 million	10	100	$\sim$ 3 mins
Adaptation Whole	32 million	20	200	$\sim$ 6 mins
Adaptation Decoder	21 million	1	100	$\sim$ 3 mins
Adaptation Decoder	21 million	5	200	$\sim$ 5 mins
Adaptation Decoder	21 million	10	200	$\sim$ 5 mins
Adaptation Decoder	21 million	20	200	$\sim$ 6 mins

Table 2: Hyper-parameter details and wall clock time for model adaptation techniques. The column *Samples* refers to the number of *target speaker samples* used for model fine-tuning. We use Adam optimizer with learning rate of 1e-4 across all model adaptation experiments.

References

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *NeurIPS*. 2018.