Human-in-the-loop Speaker Adaptation for DNN-based Multi-speaker TTS

Kenta Udagawa, Yuki Saito, Hiroshi Saruwatari
The University of Tokyo, Japan.

MOONSHOT RESEARCH & DEVELOPMENT PROGRAM

AVATAR SYMBIOTIC SOCIETY

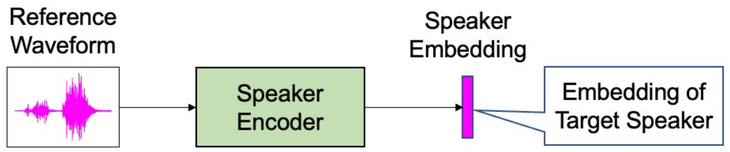
Introduction

- **Speaker adaptation of speech synthesis** is a technique for synthesizing speech of unseen speaker with small amounts of data.
- The conventional speaker adaptation method cannot be used without the reference speech of the unseen speaker.
- We propose a speaker adaptation method which incorporates humans' perceptual feedback and does not feed reference speech to speaker encoder.
- Experimental results indicate that the proposed method can synthesize speech with the same or better quality than the conventional method, depending on the user and test speaker.

Speech samples page:
http://sython.org/demo/udagawa22interspeech/demo_slstts.html

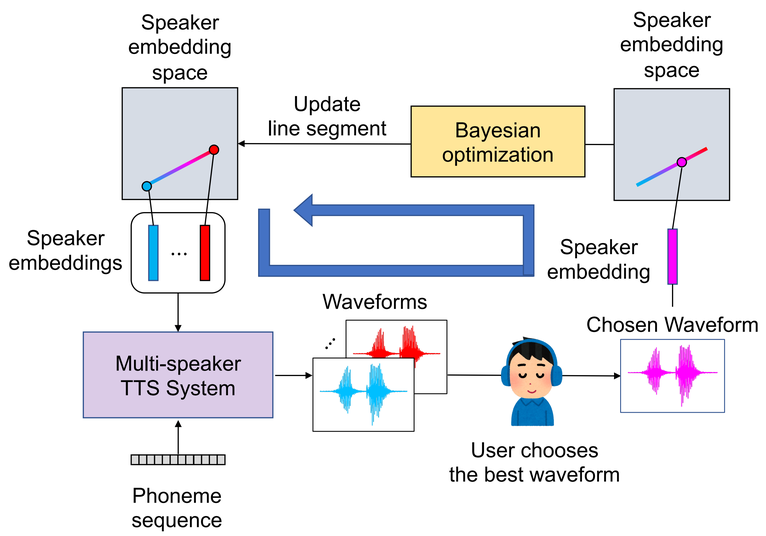
Conventional Method

Transfer-learning-based method [1] extracts speaker embedding **from reference audio with speaker encoder** pretrained in the speaker verification task.
Issue: The conventional method cannot be used without the reference speech of the target speaker.



Proposed Method

- Target speaker embedding is explored with **Sequential Line Search (SLS)** [2] algorithm.
- **Reference speech is not fed to speaker encoder.**
- Two measures are taken to search for a more natural voice.
  1. Initialization with mean male/female speaker
  2. Setting the search space to **a quantile of the training data**



Experimental Evaluations

Experimental Condition

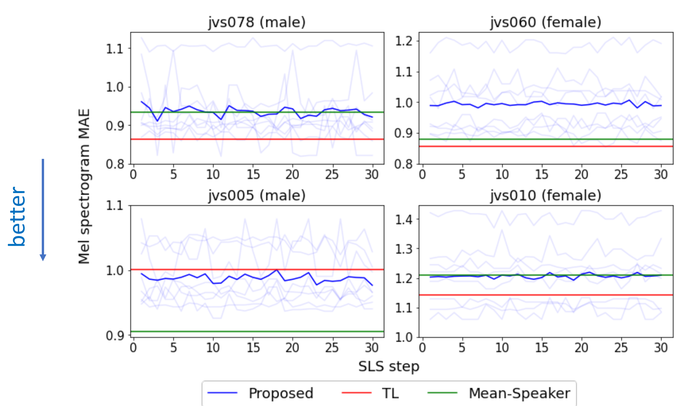
Corpus (speaker encoder)		The Corpus of Spontaneous Japanese (CSJ) [3] (947 men and 470 women, 660 hours)
TTS model		FastSpeech 2 [4]
Corpus (TTS model)		The parallel data of the Japanese Versatile Speech (JVS) corpus [5] (49 men and 51 women, 22 hours 100 sentences per speaker)
Data splitting	train	90 speakers (44 men, 46 women)
	test	4 speakers (2 men, 2 women)
	val	6 speakers (3 men, 3 women)
Vocoder		Pretrained universal model of HiFi-GAN [6]

Objective Evaluation (During Search)

We calculated mel spectrogram MAE at each step when our method is manipulated by 8 participants up to 30 steps.

Methods:
- **Proposed:** Proposed method
- **TL:** Transfer-learning-based method [1]
- **Mean-Speaker:** Endpoint of initial line segment

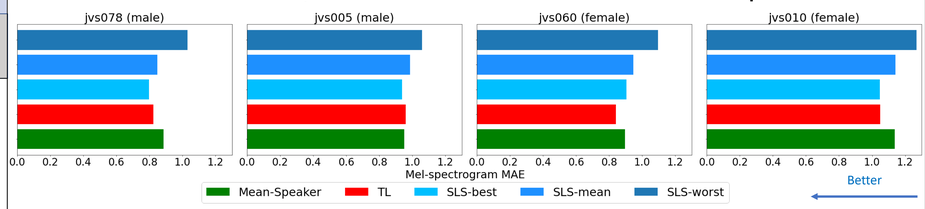
Results:
- Mel spectrogram MAE tended not to improve by human manipulation.
- The proposed method could synthesize speech comparable to TL depending on the participant.



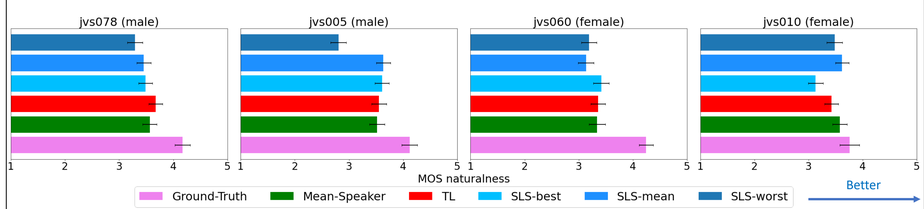
Objective/Subjective Evaluation (Multiple Utterances)

Comparison methods:
- SLS-{best, mean, worst}: The speaker embedding whose mel spectrogram MAE was {minimum, mean, maximum}
- TL: Transfer-learning-based method [1]
- Mean-Speaker: Endpoint of initial line segment

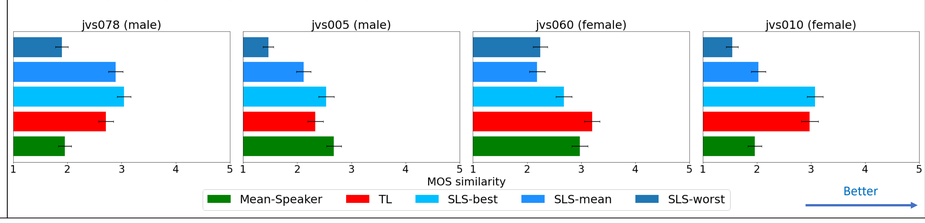
Mel spectrogram MAE results:
- SLS-worst were inferior to TL.
- In several cases, SLS-best and SLS-mean were comparable to TL.



Naturalness MOS results:
- The proposed method could synthesize speech as natural as TL in many cases.



Similarity MOS results:
- In some cases, the proposed method was inferior to Mean-Speaker.
- In some cases, the proposed method could achieve the same quality of speaker adaptation as TL.



[1] Jia et al., *Proc. NeurIPS,* 2018. [2] Koyama et al., *ACM Trans. on Graphics*, 2017. [3] Maekawa, *Proc. SSPR*, 2013. [4] Ren et al., *Proc. ICLR*, 2021. [5] Takamichi et al., *AST*, 2020. [6] Kong et al., *Proc. NeurIPS,* 2018.