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

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

**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.

**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

**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.

---