DNN-based Speaker Embedding Using Subjective Inter-speaker Similarity for Multi-speaker Speech Synthesis

Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari (The University of Tokyo, Japan)

1. Research highlights
- Purpose: learning speaker representation that is correlated with human speech perception
- Approach: using crowdsourced subjective inter-speaker similarity scores for training speaker embedding model
- Results: obtaining speaker embedding that
  1. is highly correlated with the similarity scores
  2. improves speech quality in multi-speaker speech synthesis

2. Conventional speaker embedding
- One-hot speaker code
  - $N_s$-dim. discrete vector (ID for pre-stored $N_s$ speakers)
  - Pros: high simplicity when $N_s$ is small
  - Cons: low interpretability & scalability
- $d$-vector
  - $N_d$-dim. continuous vector derived from speaker recognition
  - Applications: speaker verification & voice conversion
  - Training: minimizing speaker recognition loss (cross-entropy)

3. Proposed speaker embedding
- Large scale scoring of subjective inter-speaker similarity
  - Crowdsourcing the similarity scores involving 4,000+ listeners
  - # of listeners per one speaker pair = at least 10

Evaluation:
- Similarity vector embedding
  - Training DNN to predict a similarity score vector $s \in \mathbb{S}$

- Similarity matrix embedding
  - Training DNN using the similarity score matrix $\mathbb{S}$ as a constraint on coordinates of speaker embedding

4. Experimental evaluation
- Dataset: JNAS - 153 Japanese females
  - Training data: 140 females (F01\text{--}F03)
  - Evaluation data: 13 females (F01\text{--}F03)
- Vocoder: STRAIGHT
- Methods:
  1. d-vec.: trained by speaker recognition
  2. Sim. (vec): trained by similarity vector embedding
  3. Sim. (mat): trained by similarity matrix embedding
- Results: obtaining speaker embedding that is correlated with the similarity scores!
  - Correlation Approach: using crowdsourced subjective inter-speaker scores!

References: