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

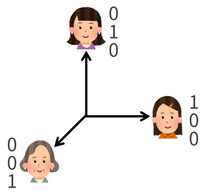
Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari (The University of Tokyo, Japan)

1. Research highlights

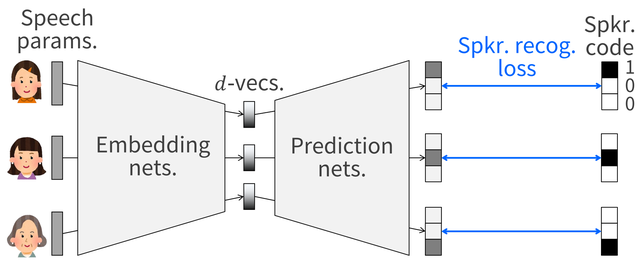
- Purpose: learning speaker representation that is correlated with human speech perception
- Approach: using crowdsourced subjective inter-speaker similarity scores for training speaker embedding model
- Results: obtaining speaker embedding that
  1. is highly correlated with the similarity scores
 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

2. Conventional speaker embedding

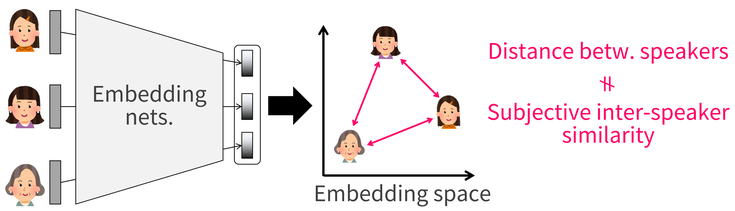
- One-hot speaker code[1]
  - N_s-dim. *discrete* vector (ID for pre-stored N_s speakers)
  - Pros: high simplicity when N_s is small
  - Cons: low interpretability & scalability
    - Distance betw. speakers = constant
    - Sparse representation that cannot deal with unseen speakers



- d-vector[2]
  - N_d-dim. *continuous* vector derived from speaker recognition
    - Applications: speaker verification & voice conversion[3]
  - Training: minimizing speaker recognition loss (cross-entropy)



- Pros: high scalability
  - Low-dim. representation that can deal with unseen speakers
- Cons: still low interpretability
  - Speaker recognition ≠ human speech perception



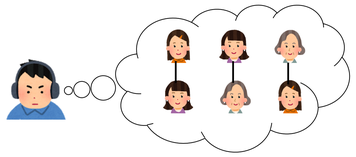
Can we construct an embedding space that preserves subjective inter-speaker similarity?

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

Instruction of the scoring

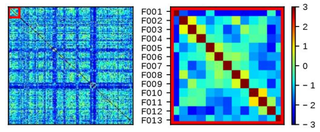
> To what degree do the two speakers' voices sound similar? (−3: dissimilar ∼ +3: similar)



  - Visualization of the obtained scores

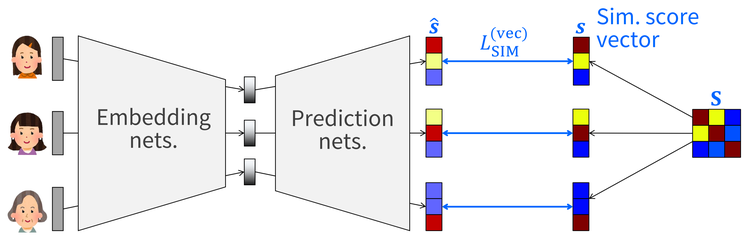
Graph representation (speaker similarity graph)

Matrix representation (speaker similarity score matrix $\mathbf{S}$)



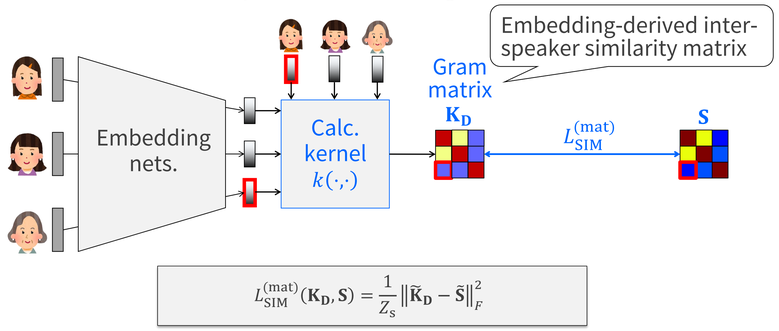
We propose training methods for speaker embedding w/ the matrix $\mathbf{S}$.

- Similarity *vector* embedding
  - Training DNN to predict a similarity score vector $\boldsymbol{s} \in \mathbf{S}$



$$L_{\mathrm{SIM}}^{(\mathrm{vec})}(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \frac{1}{N_s}(\hat{\boldsymbol{s}} - \boldsymbol{s})^{\top}(\hat{\boldsymbol{s}} - \boldsymbol{s})$$

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



$$L_{\mathrm{SIM}}^{(\mathrm{mat})}(\mathbf{K_D}, \mathbf{S}) = \frac{1}{Z_s}\left\|\tilde{\mathbf{K}}_{\mathbf{D}} - \tilde{\mathbf{S}}\right\|_F^2$$

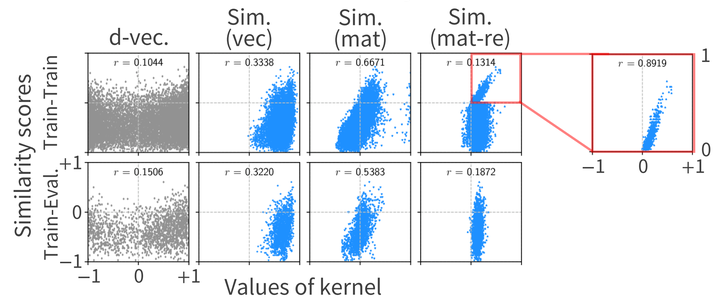
Z_s: normalization term
$\tilde{\mathbf{K}}_{\mathbf{D}}, \tilde{\mathbf{S}}$: matrices the same as $\mathbf{K_D}, \mathbf{S}$ w/o their diagonal components

4. Experimental evaluation

- Experimental conditions

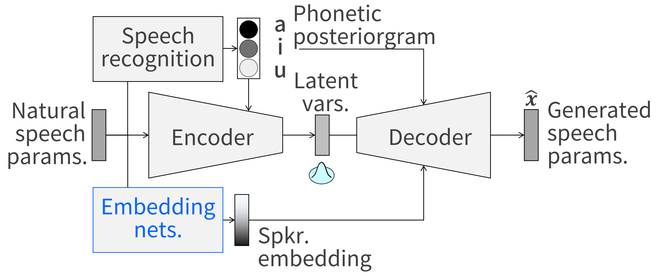
Dataset (16 kHz sampling)	JNAS[4] 153 Japanese females Training (seen) data: 140 females except for F001—F013 Evaluation (unseen) data: 13 females (F001—F013)
Vocoder	STRAIGHT[5]
DNN input	1—39 dim. mel-cepstra (including Δ)
DNN architecture	Feed-Forward (see our paper for details)
Methods	(1) d-vec. : trained by speaker recognition (2) Sim. (vec): trained by similarity vector embedding (3) Sim. (mat): trained by similarity matrix embedding (4) Sim. (mat-re): trained by (3) w/o dissimilar spkr. pairs
Spkr. embedding	8-dimensional vector
Kernel in Sim. (mat)	Sigmoid kernel: $k(\boldsymbol{x}, \boldsymbol{y}) = \tanh(\boldsymbol{x}^{\top}\boldsymbol{y})$

- Correlation coef. betw. similarity scores & values of kernel



Our algorithms learn speaker embedding that is highly correlated with the similarity scores!

- Evaluation in DNN-based multi-speaker speech synthesis
  - Acoustic model: variational-autoencoder[7] incorporating pre-trained speech recognition & speaker embedding[3]



  - Results of preference AB/XAB tests on synthetic speech quality
    - A > B: A is significantly higher than B ($p < 0.05$)
    - A ≒ B: there is no significant difference betw. A & B

A vs. B	Naturalness (preference AB) A > B	B > A	A ≒ B	Spkr. similarity (preference XAB) A > B	B > A	A ≒ B
d-vec. vs. Sim. (vec)	0	12	1	0	10	3
d-vec. vs. Sim. (mat)	0	7	6	4	4	5
d-vec. vs. Sim. (mat-re)	0	8	5	4	2	7

(1) Sim. (vec) improves speech quality in almost all spkrs.
(2) Sim. (mat) degrades speech quality in some spkrs.

[References] [1] N. Hojo et al., IEICE Trans. on Information and Systems, 2018. [2] E. Variani et al., Proc. ICASSP, 2014. [3] Y. Saito et al., Proc. ICASSP, 2018. [4] K. Itou et al., Journal of the ASJ (E), 1999. [5] H. Kawahara et al., Speech Communication, 1999. [6] J. Duchi et al., Journal of Machine Learning Research, 2011. [7] D. P. Kingma et al., arXiv, abs/1312.6114, 2013. [8] L. Sun et al., Proc. ICME, 2016.

[Acknowledgements] This research and development was supported by the SECOM Science and Technology Foundation, JSPS KAKENHI Grant Number 18J22090 and 17H06101, the MIC/SCOPE #182103104, and the GAP foundation program of the UTokyo.