Perceptual-Similarity-Aware Deep Speaker Representation Learning for Multi-Speaker Generative Modeling SPE-3.5

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

1. Overview

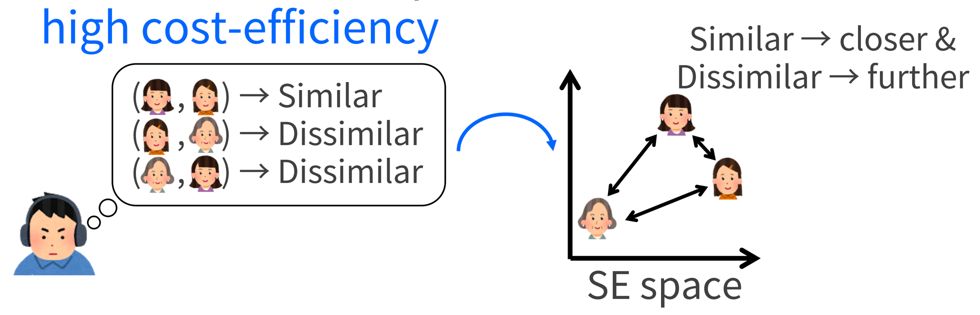
Introduction

- Speaker Embedding (SE)
  - Distributed representation of spkr. ID
  - Primal usage: spkr. *discriminative* tasks
    e.g., spkr. recognition / verification
  - **Our focus: spkr. *generative* tasks**
    e.g., Text-To-Speech (TTS) / Voice Conversion (VC)
- Deep speaker representation learning
  - DNN-based technology for learning SEs
    - d-vector[1] & x-vector[2]
      → Designed for discriminative tasks

Research question: How can we learn SEs suitable for spkr. generative tasks?

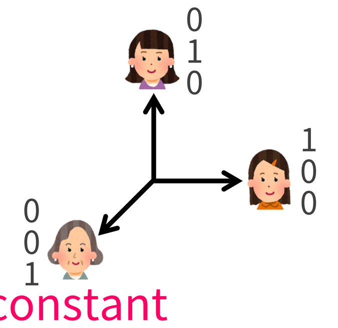
Summary

- Perceptual-similarity-aware SE
  - Trained to represent perceptual similarity among speakers → high interpretability
  - 3 algorithms based on **crowdsourced perceptual similarity score matrix**
  - Human-in-the-loop Active Learning (AL) for high cost-efficiency

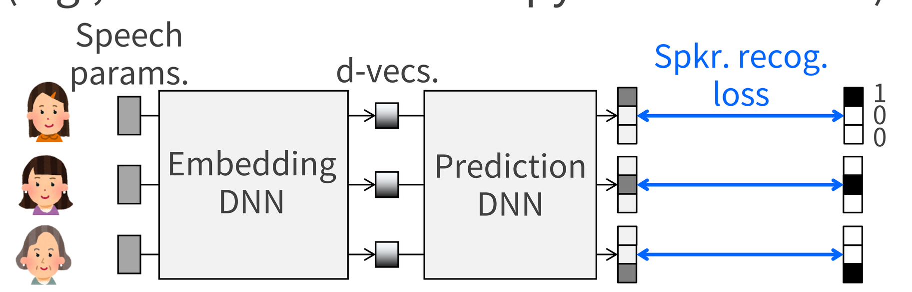


Traditional speaker representations

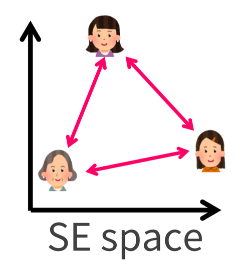
- One-hot speaker code[3]
  - N_s-dim. *discrete* vector (i.e., ID for pre-stored N_s spkrs.)
  - High simplicity when N_s is small
  - Low interpretability & scalability
    - Distance b/w spkrs. = always constant
    - Undefined for unseen spkrs.



- d-vector[2]
  - N_d-dim. *continuous* vector learned by spkr. recog.
  - Training objective: spkr. recog. loss (e.g., softmax cross-entropy or GE2E loss[4])



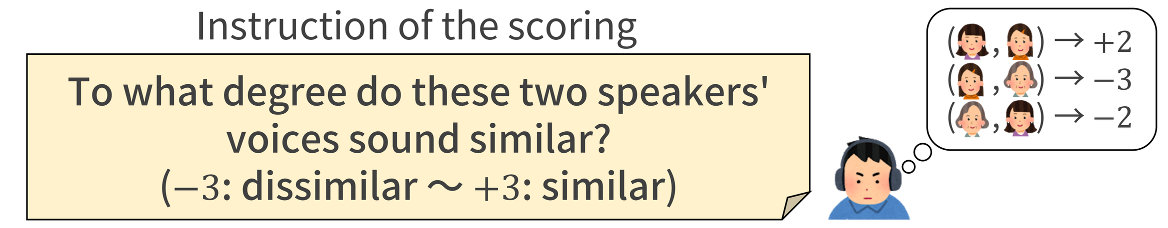
  - High scalability
    - Low-dim. SE that can deal with unseen spkrs.
  - Still low interpretability
    - Spkr. recog. ≠ human speech perception (i.e., similar speaker-pairs → further)
  - Uninterpretable SE leads to...
    - Less controllability & adaptability in speaker generative tasks



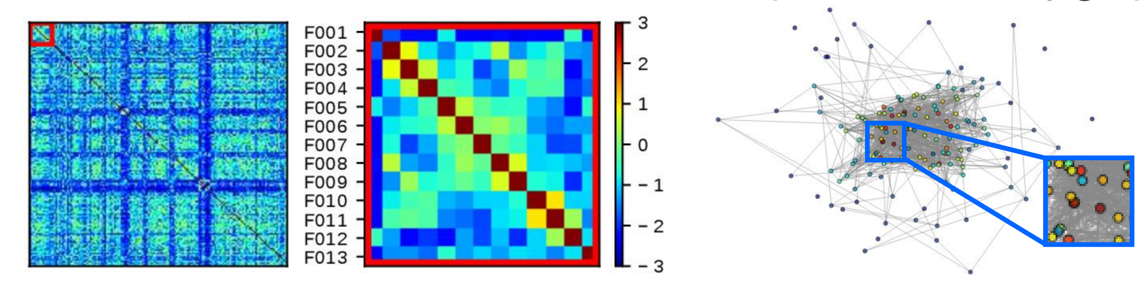
2. Methods

Proposed method: Perceptual-similarity-aware SE learning

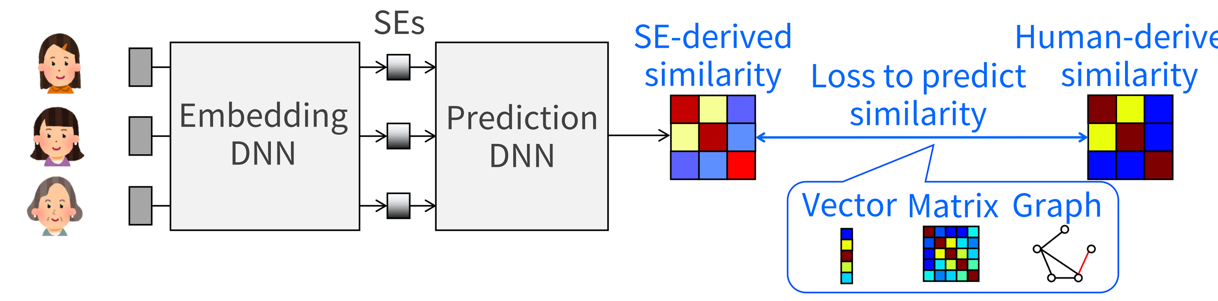
- Large scale scoring of perceptual speaker similarity
  - Crowdsourcing similarity scores **involving 4,000+ listeners**
    - 1 speaker pair → scored by more than 10 listeners



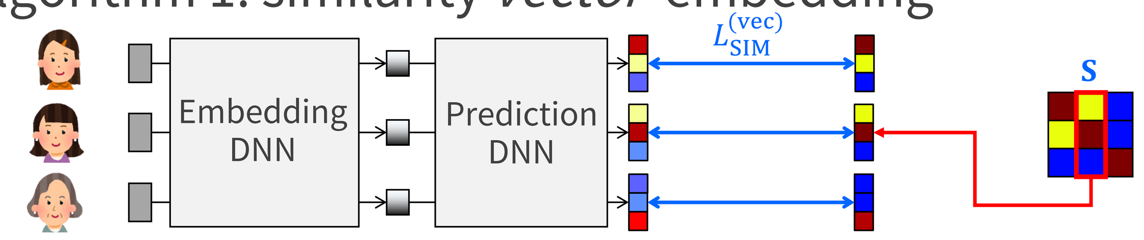
  - Visualization of obtained similarity scores among 153 jp spkrs.



- Learning SEs based on similarity scores
  - Training embedding DNN to predict perceptual similarity from SEs (i.e., human-oriented machine learning)
  - Algorithm: similarity { vector, matrix, graph } embedding

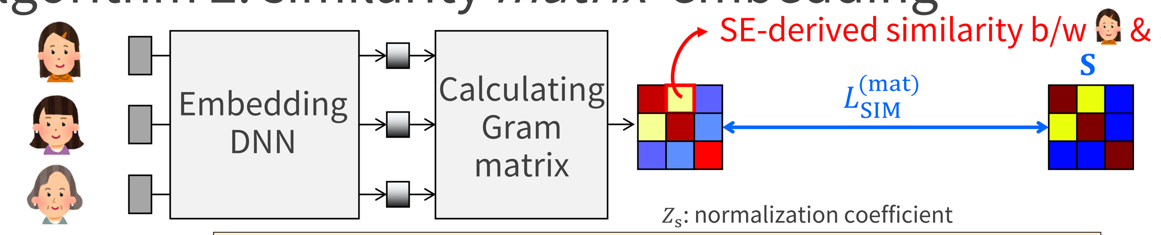


- Algorithm 1: similarity *vector* embedding



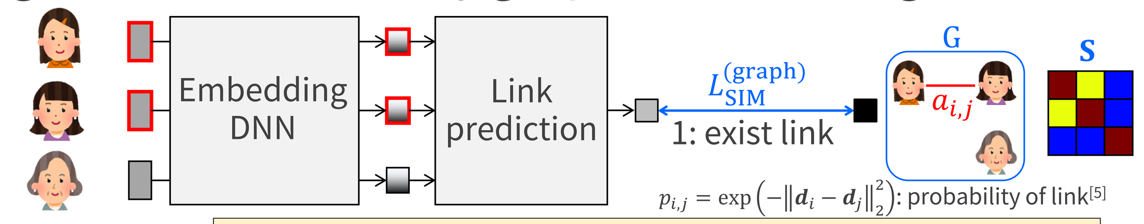
$$L_{\mathrm{SIM}}^{(\mathrm{vec})}(s, \hat{s}) = \frac{1}{N_s}(\hat{s} - s)^\top(\hat{s} - s)$$

- Algorithm 2: similarity *matrix* embedding



$$L_{\mathrm{SIM}}^{(\mathrm{mat})}(\mathbf{S}, \mathbf{K_D}) = \frac{1}{Z_s}\|\mathbf{K_D} - \mathbf{S}\|_F^2$$

- Algorithm 3: similarity *graph* embedding



$$L_{\mathrm{SIM}}^{(\mathrm{graph})}(d_i, d_j) = -a_{i,j}\log p_{i,j} - (1 - a_{i,j})\log(1 - p_{i,j})$$

- Human-in-the-loop AL
  - Iterating SE learning & perceptual similarity scoring
    → learn better SEs while reducing costs
  - Query selection (priority annotation for unscored data)
    - Query strategies: { Higher, Middle, Lower }-Similarity First (SF)

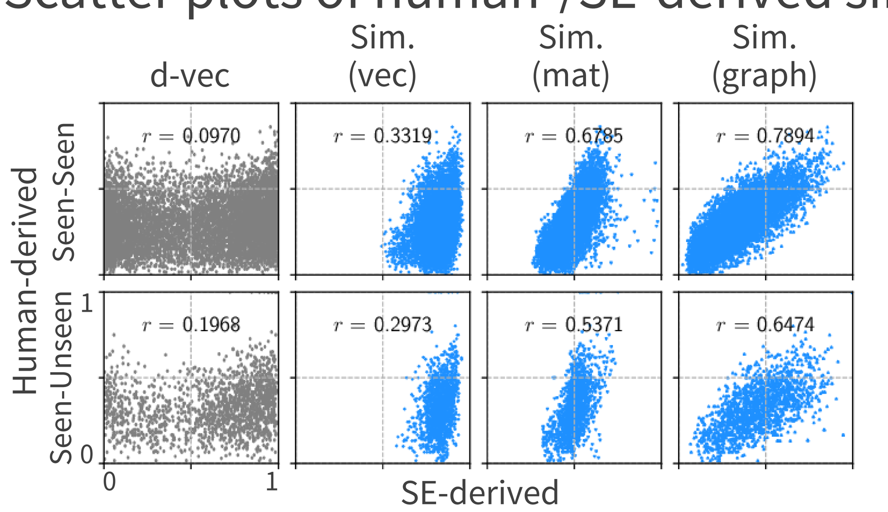
3. Experimental evaluations & results

Experimental conditions

Dataset (16 kHz sampling)	JNAS[6] including 153 Japanese female spkrs. Training (seen) data: 140 females except for F001—F013 Evaluation (unseen) data: 13 females (F001—F013)
Vocoder	STRAIGHT[7]
DNN input	1—39 dim. mel-cepstral coefficients (including Δ)
DNN architecture	Feed-forward (see our paper for details)
Compared SEs (8-dim.)	d-vectors[1] vs. Proposed SEs (i.e., Sim. (vec / mat / graph) embedding)
Kernel for Sim. (mat)	Sigmoid kernel: $k(\boldsymbol{x}, \boldsymbol{y}) = \tanh(\boldsymbol{x}^\top \boldsymbol{y})$

Evaluation1: interpretability of SEs

- Scatter plots of human-/SE-derived similarity (r: PCC)



- d-vec never considers perceptual similarity.
- Sim. (*) achieves higher PCC than d-vec!
  - **Sim. (graph)** works the best among the 3 algorithms.

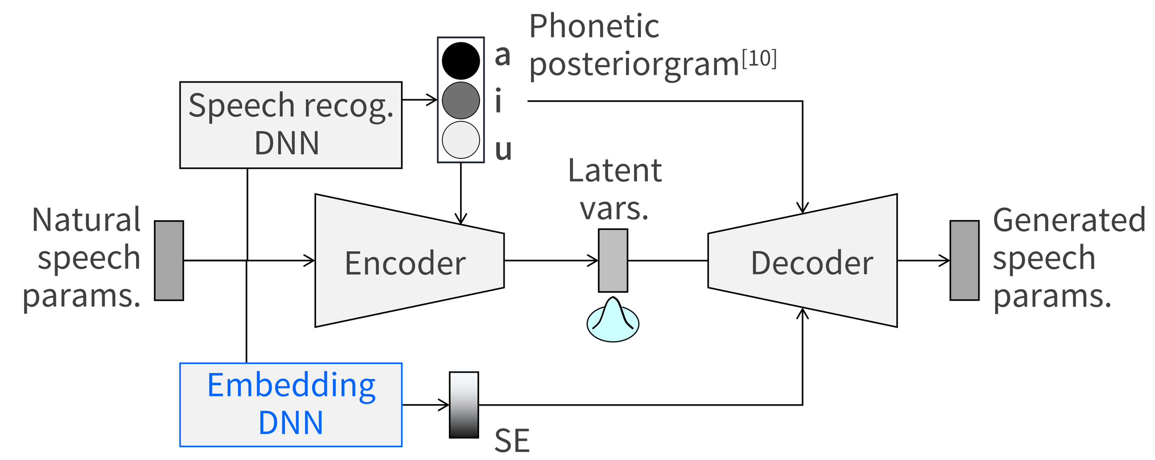
- AUC of similar speaker-pair detection using similarity b/w SEs

	d-vec	Sim. (vec)	Sim. (mat)	Sim. (graph)
Seen-Seen	0.57	0.69	0.89	**0.92**
Seen-Unseen	0.64	0.69	0.77	**0.82**

- Sim. (graph) achieves the highest AUC!

Evaluation 2: quality of multi-speaker generative modeling

- Acoustic model: VAE[8] using pre-trained speech recog. & SE[9]

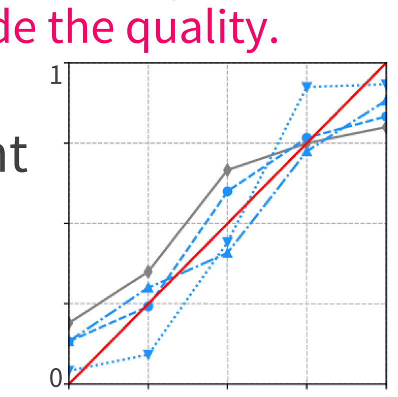


- Quality of speaker adaptation (AB/XAB tests w/ 25 listeners)
  - A > B: A is significantly better than B ($p < 0.05$).
  - A ≒ B: there is no significant difference b/w A & B.

A vs. B	Naturalness A > B	Naturalness B > A	Naturalness A ≒ B	Spkr. similarity A > B	Spkr. similarity B > A	Spkr. similarity 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. (graph)	0	**13**	0	0	**12**	1

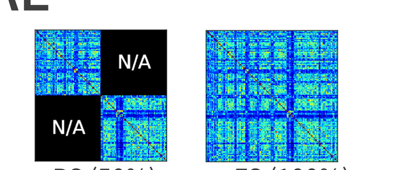
  - Sim. (vec / graph) significantly improve speech quality for almost all spkrs., but Sim. (graph) can degrade the quality.

- Controllability of speaker interpolation
  - Horizontal axis: interpolation coefficient
  - Vertical axis: spkr. similarity score
  - Proposed SEs (blue lines) also improve the controllability of d-vec (grey line)!

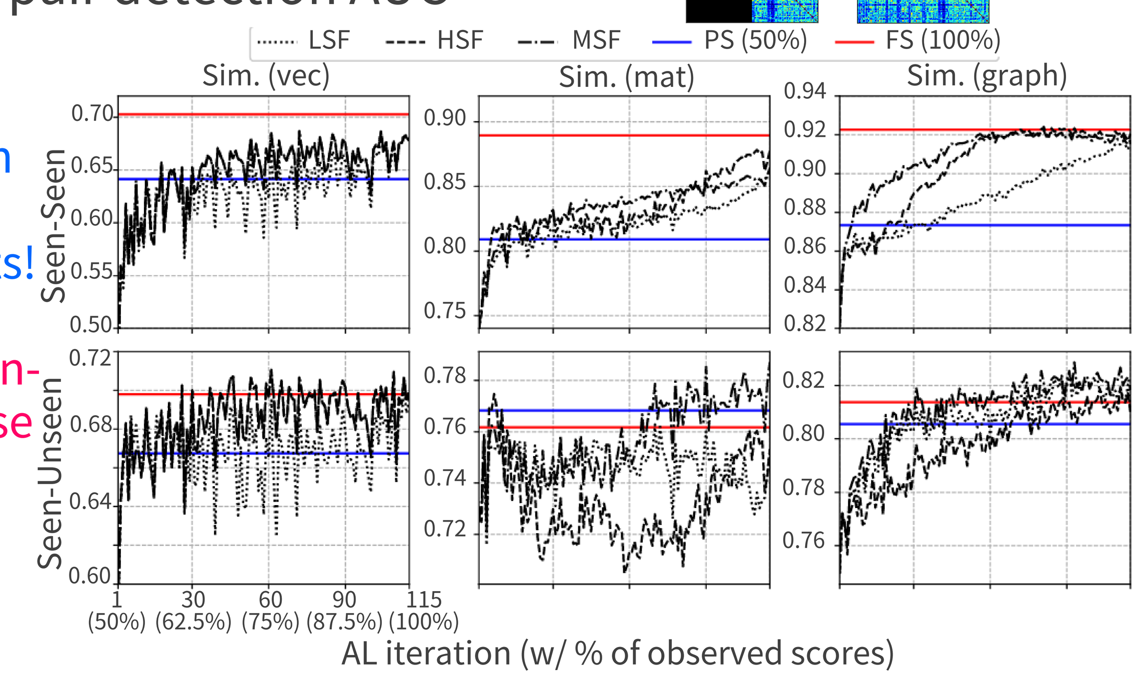


Evaluation 3: cost-efficiency of human-in-the-loop AL

- AL setting: Fully Scored (FS), Partially Scored (PS)



- Curves of similar speaker-pair detection AUC
  - Best query strategy: **MSF**
  - AL in Sim. (vec / graph):
    - AUC comparable with FS, while reducing learning/scoring costs!
  - AL in Sim. (mat):
    - Degraded AUC in Seen-Unseen spkr.-pair case
      → due to high data-dependency?



- Quality of synthetic speech (DMOS test on spkr. similarity w/ 50 listeners)
  - **BOLD** scores ≒ FS ($p < 0.05$)
  - AL in Sim. (vec / graph):
    - DMOS comparable with FS by fewer AL iterations!
  - AL in Sim. (mat):
    - Lowest DMOS among 3 SEs & no improvement

		Sim. (vec)	Sim. (mat)	Sim. (graph)
PS	(50%)	2.85±0.14	**2.90±0.13**	2.86±0.13
MSF	(62.5%)	**2.95±0.14**	**2.93±0.13**	**3.03±0.13**
	(75%)	**3.04±0.14**	**3.00±0.13**	**3.02±0.13**
	(87.5%)	**3.05±0.14**	**3.03±0.13**	**3.06±0.13**
FS	(100%)	3.14±0.14	2.98±0.13	3.08±0.14

- Our project page: http://sython.org/demo/JSPS-DC1/index.html
  - Spkr. similarity matrix for VCTK female spkrs. is also available!

[References] [1] Variani et al., Proc. ICASSP, 2014. [2] Snyder et al., Proc. ICASSP, 2018. [3] Hojo et al., IEICE Trans. on Info. and Systems, 2018. [4] Wan et al., Proc. ICASSP, 2018. [5] Li et al., Proc. IJCAI, 2018.
[6] Itou et al., Journal of the ASJ (E), 1999. [7] Kawahara et al., Speech Communication, 1999. [8] Kingma et al., arXiv, abs/1312.6114, 2013. [9] Saito et al., Proc. ICASSP, 2018. [10] Sun et al., Proc. ICME, 2016.

[Acknowledgements] This research and development was supported by JSPS KAKENHI Grant Number 19H0116, 18J22090 and 19K18100, the MIC/SCOPE #182103104.