ICASSP2022 @ Singapore [SPE-L3.9]

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


**Yuki Saito**

Shinnosuke Takamichi

Hiroshi Saruwatari

The University of Tokyo, Japan.

# Overview

## Deep Speaker Representation Learning (DSRL)

- DNN-based technology for learning Speaker Embeddings (SEs)
  - Feature extraction for *discriminative* tasks (e.g., [Variani+14])
  - Control of spkr. identity in *generative* tasks (e.g., [Jia+18])


Discriminative task
(e.g., automatic speaker verification: ASV)

**Generative task**
(e.g., text-to-speech: TTS)

## This talk: method to learn SEs suitable for generative tasks

- Purpose: improving quality & controllability of synthetic speech
- Core idea: introducing human listeners for learning SEs that are highly correlated with **perceptual similarity among spkrs.**

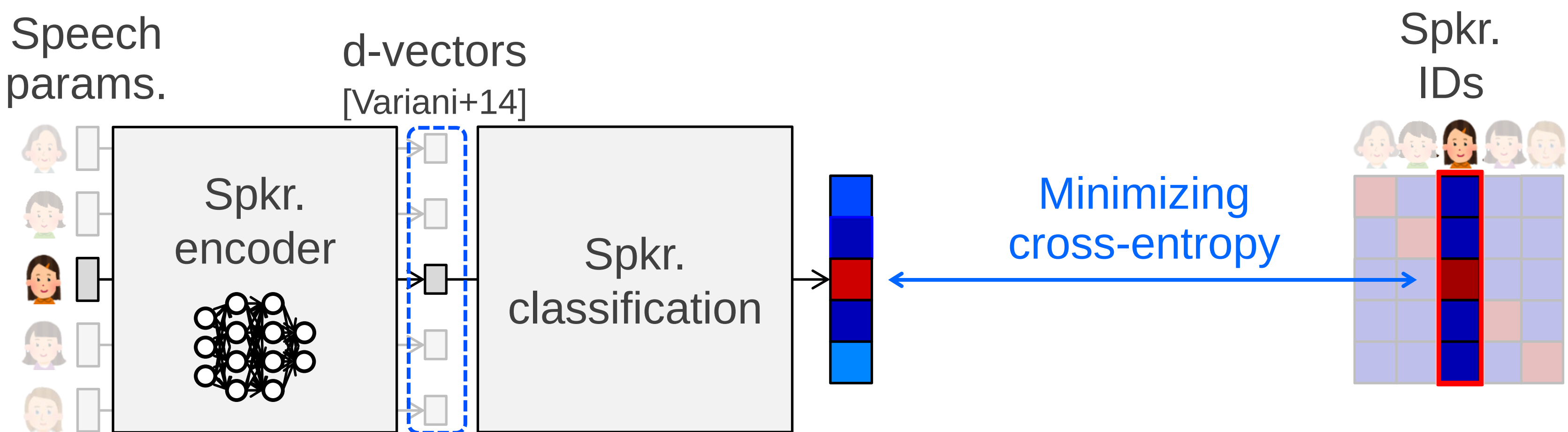
DNN: Deep Neural Network

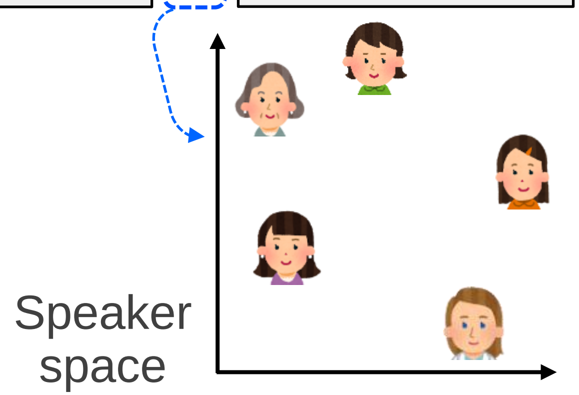
# Conventional Method: Speaker-Classification-Based DSRL

**Learning to predict speaker ID from input speech parameters**

SEs suitable for speaker classification → **also suitable for TTS/VC?**

One reason: low interpretability of SEs



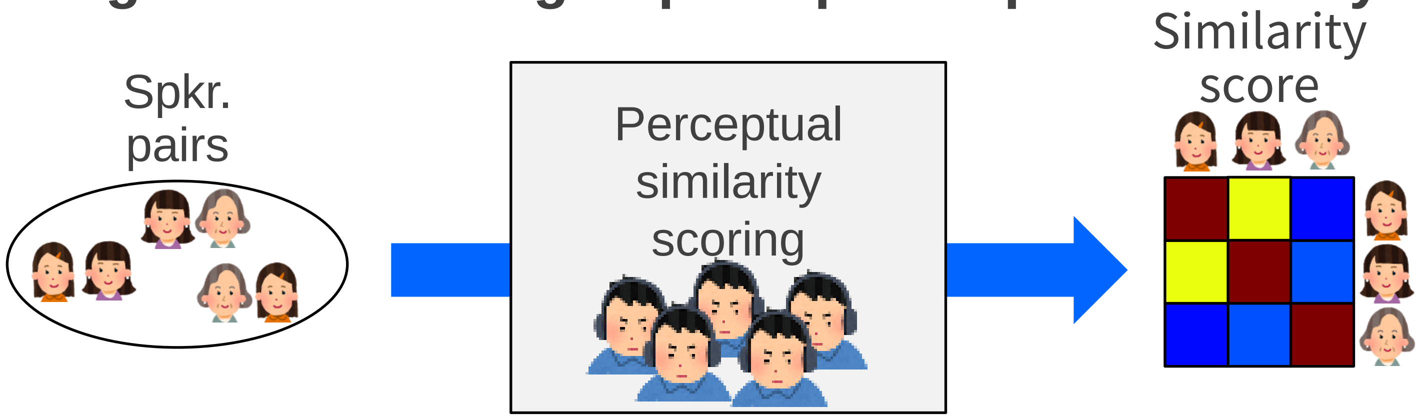

**Distance metric in spkr. space**

**≠**

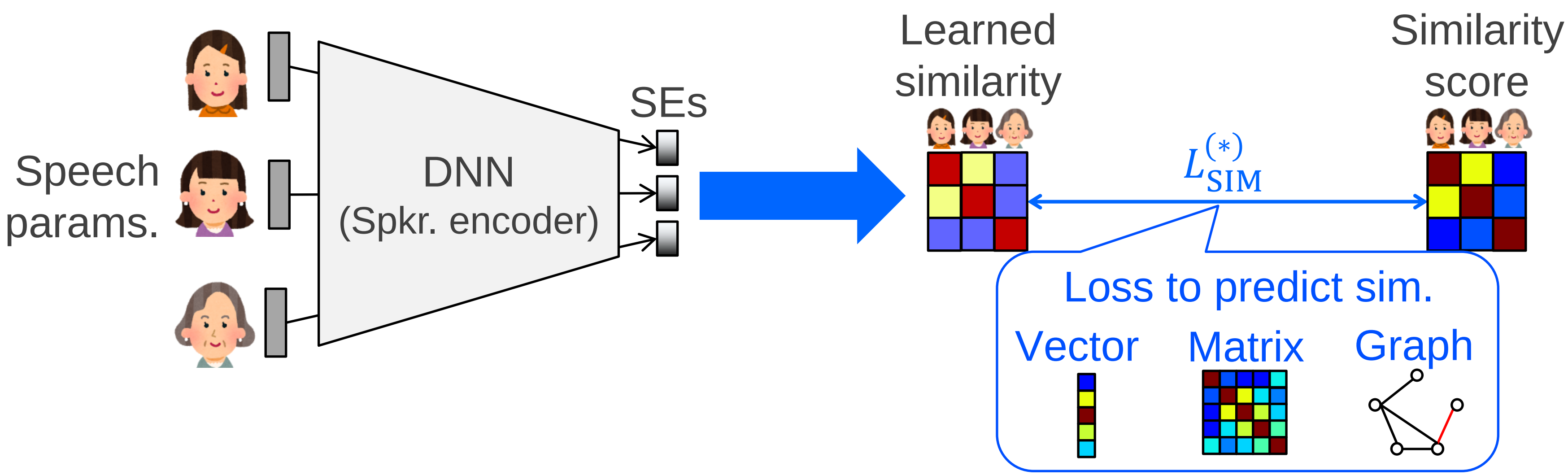
**Perceptual metric (i.e., speaker similarity)**

# Our Method:
# Perceptual-Similarity-Aware DSRL

## 1. Large-scale scoring of perceptual spkr. similarity




## 2. SE learning considering the similarity scores



# Large Scale Scoring of Perceptual Speaker Similarity

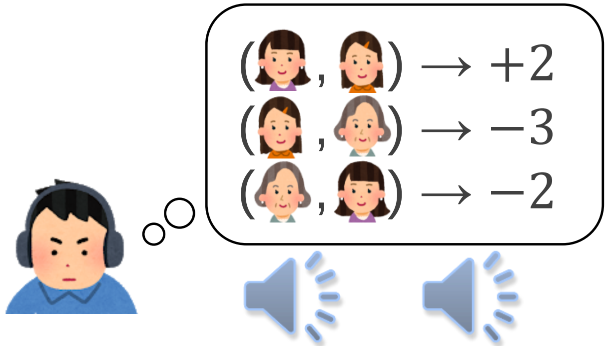
## Crowdsourcing of perceptual speaker similarity scores

Dataset we used: 153 females in JNAS corpus [Itou+99]

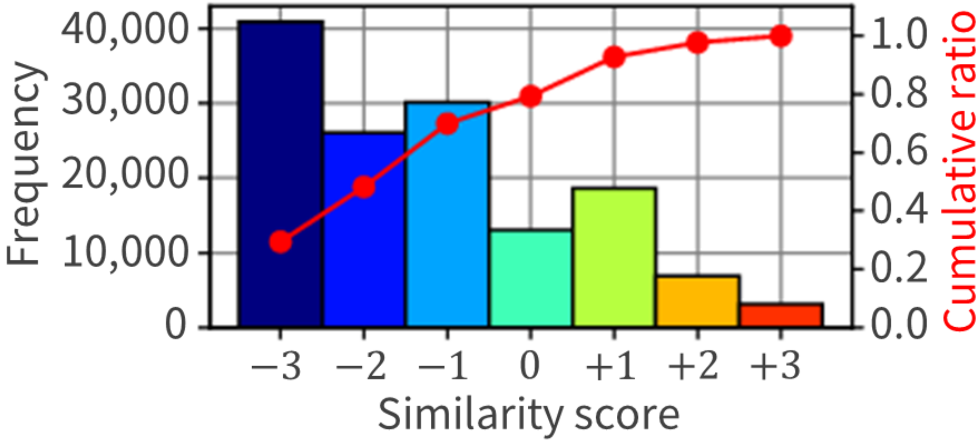
**4,000↑** listeners scored the similarity of two speakers' voices.

Instruction of the scoring

**To what degree do these two speakers' voices sound similar?**
**($-3$: dissimilar 〜 $+3$: similar)**




## Histogram of the collected scores



# Perceptual Speaker Similarity Matrix

**Similarity matrix** $\boldsymbol{S} = \left[\boldsymbol{s}_1, \cdots, \boldsymbol{s}_i, \cdots, \boldsymbol{s}_{N_s}\right]$

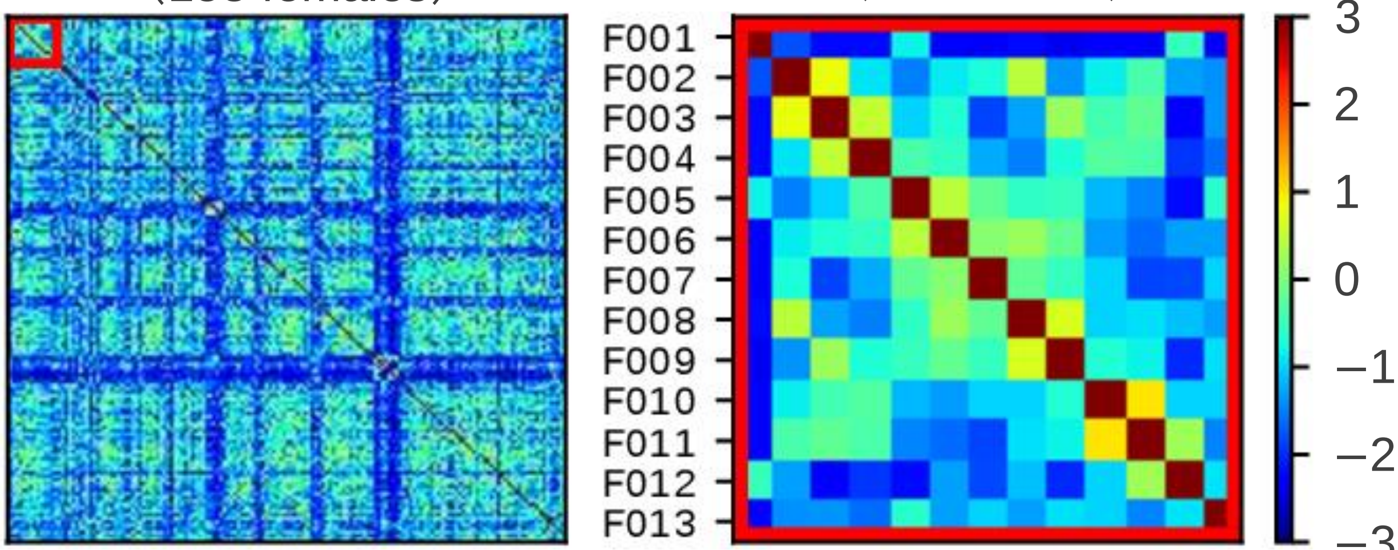
$N_s$: # of pre-stored (i.e., closed) speakers

$\boldsymbol{s}_i = \left[s_{i,1}, \cdots, s_{i,j}, \cdots, s_{i,N_s}\right]^\top$: the $i$th similarity score vector

$s_{i,j}$: similarity of the $i$th & $j$th speakers $\left(-v \leq s_{i,j} \leq v\right)$

(a) Full score matrix (153 females)

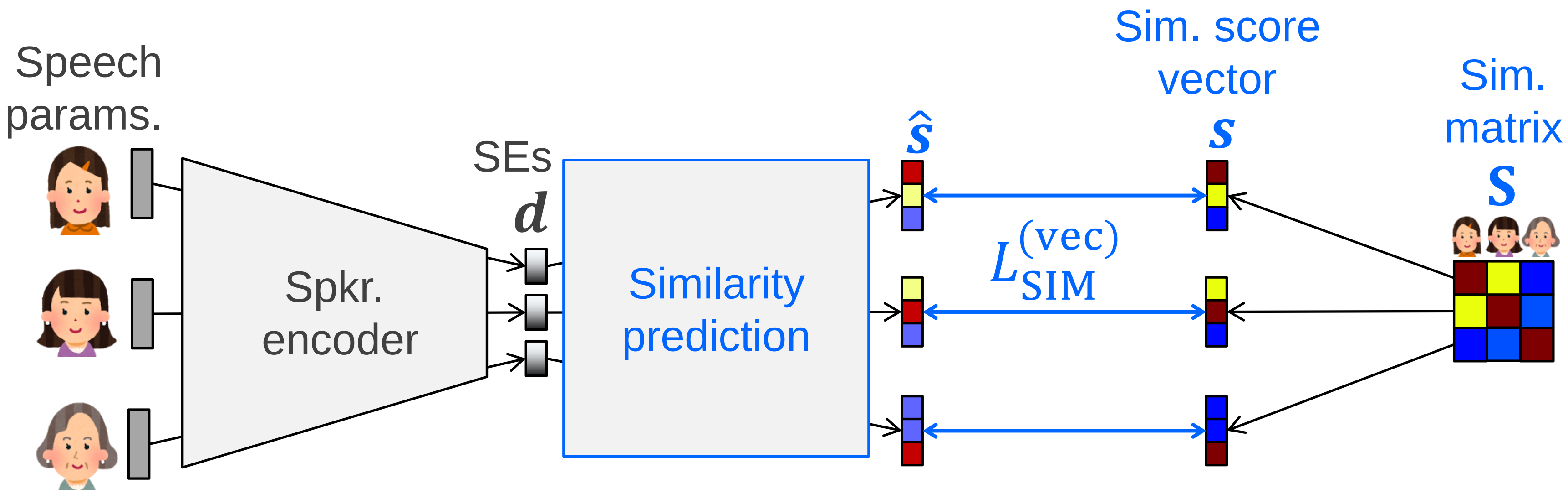
(b) Sub-matrix of (a) (13 females)




**I'll present three algorithms to learn the similarity.**

# Algorithm 1: Similarity *Vector* Embedding

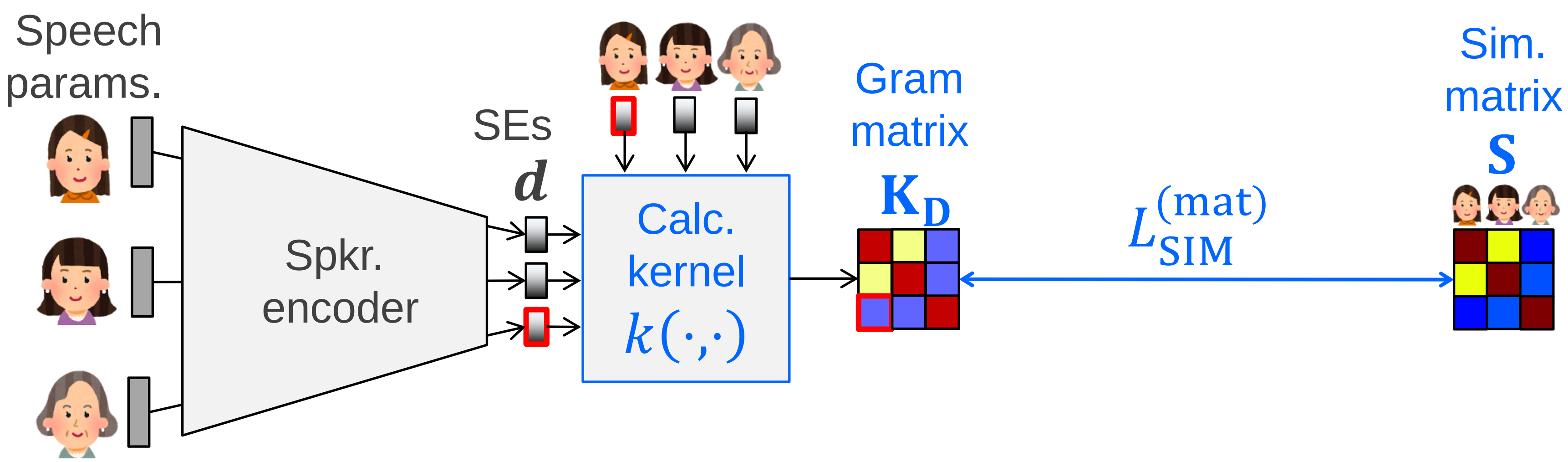
## Predict a vector of the matrix $S$ from speech parameters




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

# Algorithm 2: Similarity *Matrix* Embedding

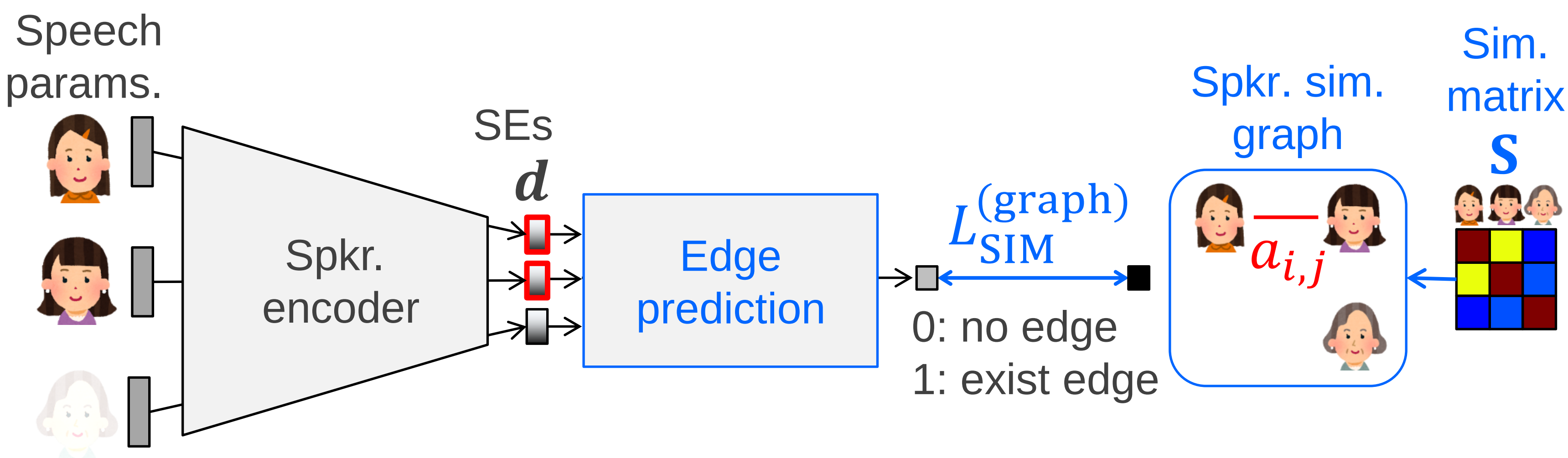
## Associate the Gram matrix of SEs with the matrix $\boldsymbol{S}$




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

# Algorithm 3: Similarity *Graph* Embedding

**Learn the structure of speaker similarity graph from SE pairs**



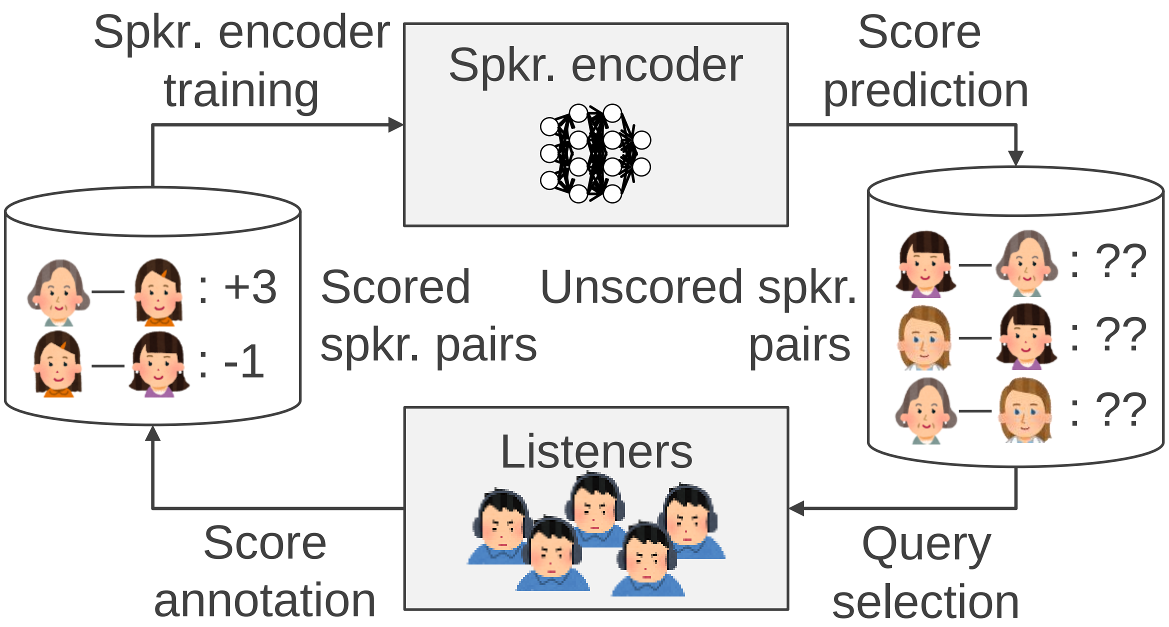

$$L_{\mathrm{SIM}}^{(\mathrm{graph})}(\boldsymbol{d}_i, \boldsymbol{d}_j) = -a_{i,j} \log p_{i,j} - (1 - a_{i,j}) \log(1 - p_{i,j})$$

$p_{i,j} = \exp\left(-\|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2\right)$: edge probability (referring to [Li+18])

# Human-In-The-Loop Active Learning (AL) for Perceptual-Similarity-Aware SEs

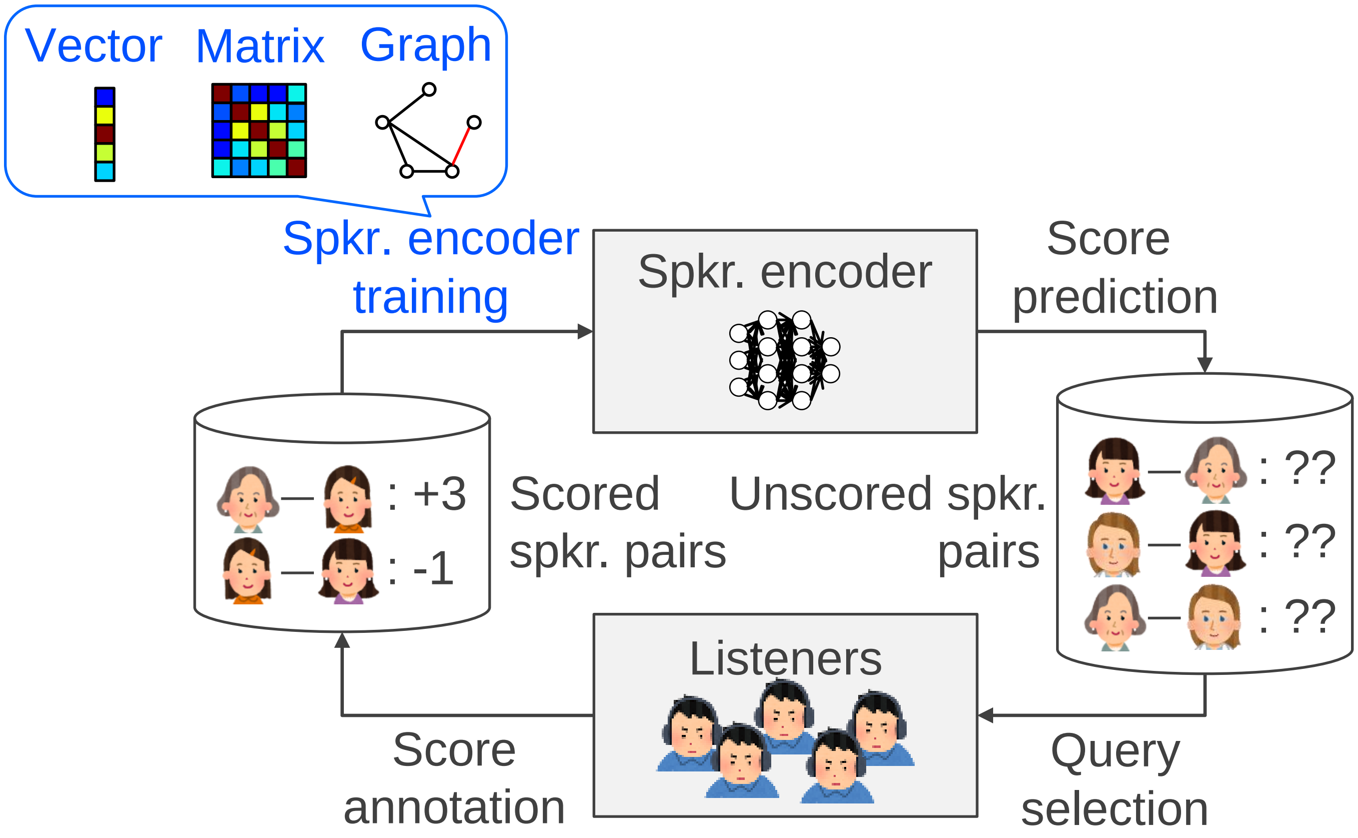
## Overall framework: iterate similarity scoring & SE learning

Obtaining better SEs while reducing costs of scoring & learning



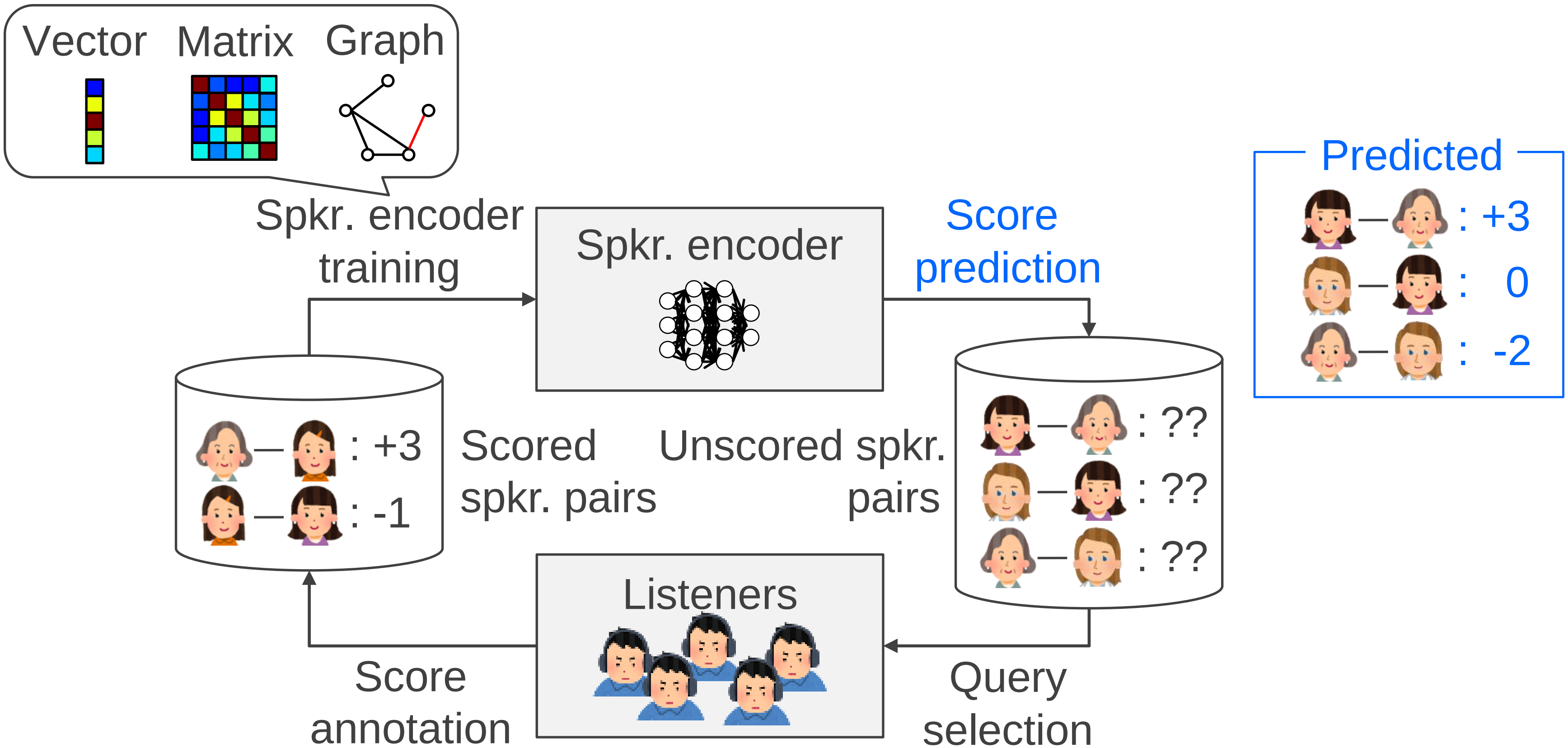
# Human-In-The-Loop Active Learning (AL) for Perceptual-Similarity-Aware SEs

## AL step 1: train spkr. encoder using partially observed scores



# Human-In-The-Loop Active Learning (AL) for Perceptual-Similarity-Aware SEs

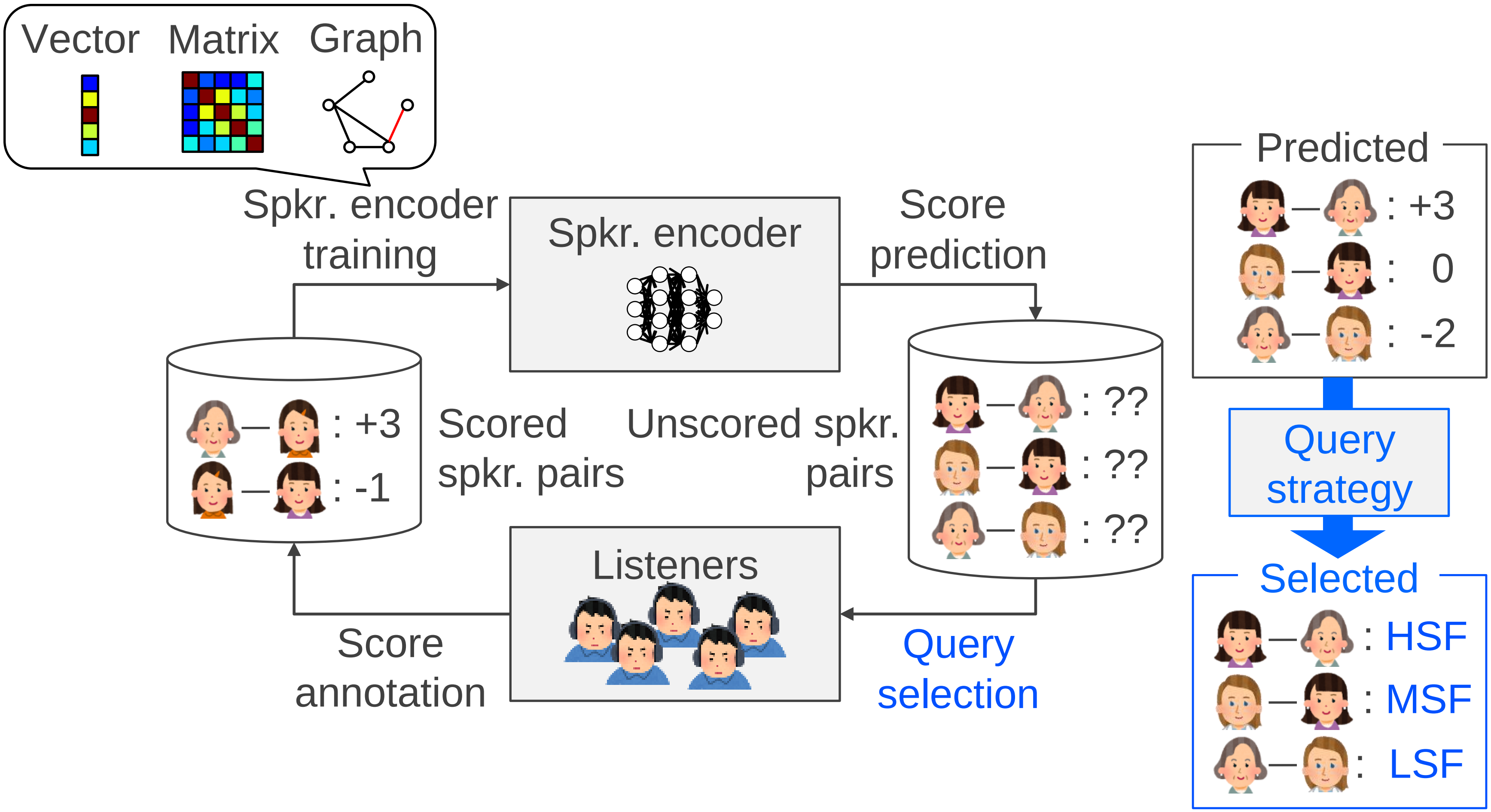
## AL step 2: predict similarity scores for unscored spkr. pairs



# Human-In-The-Loop Active Learning (AL) for Perceptual-Similarity-Aware SEs

## AL step 3: select unscored pairs to be scored next

Query strategy: criterion to determine priority of scoring

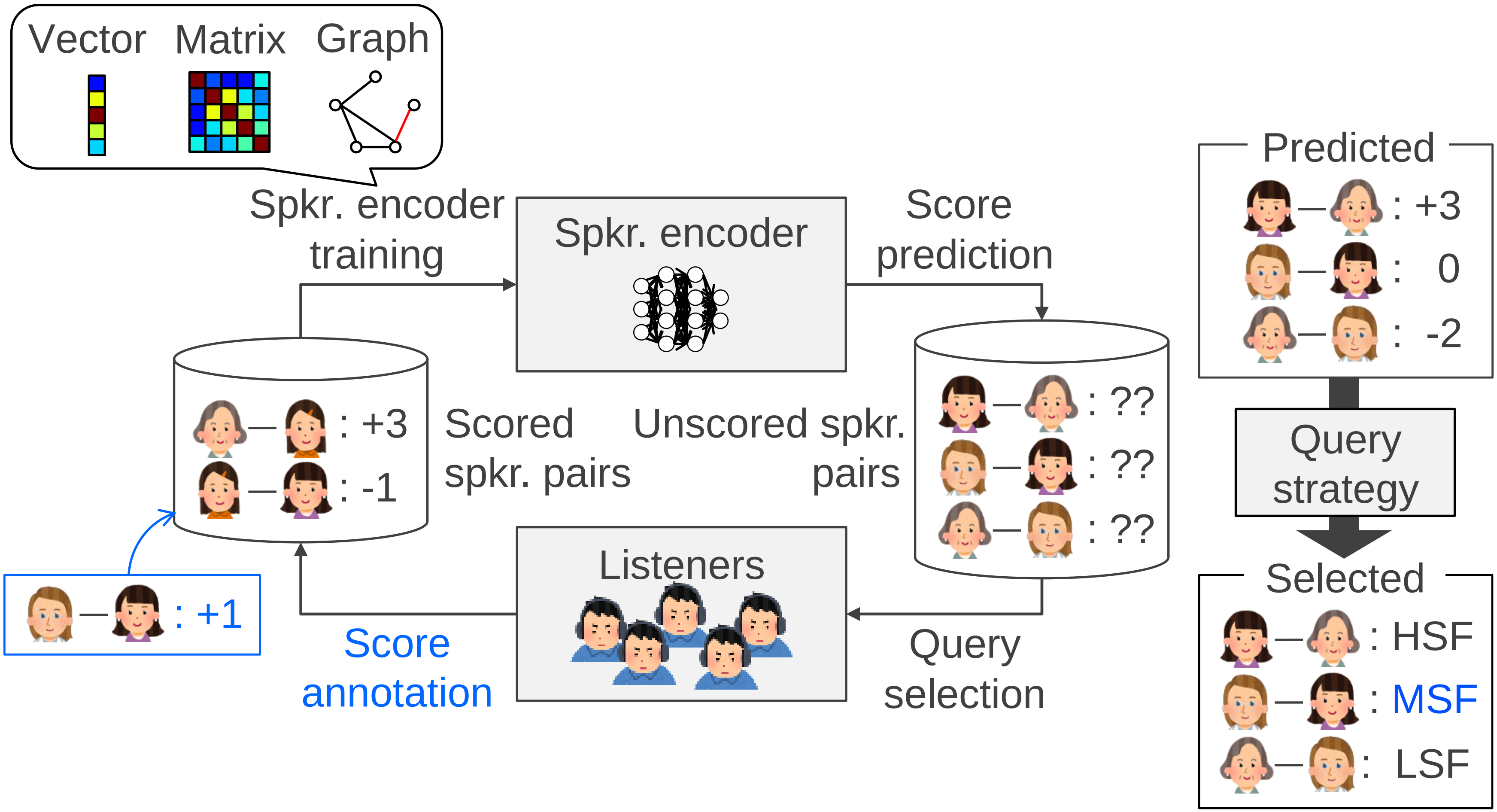



{ Higher, Middle, Lower }-Similarity First

# Human-In-The-Loop Active Learning (AL) for Perceptual-Similarity-Aware SEs

## AL step 4: annotate similarity scores to selected spkr. pairs

→ return to AL step 1



# Experimental Evaluations

# Experimental Conditions

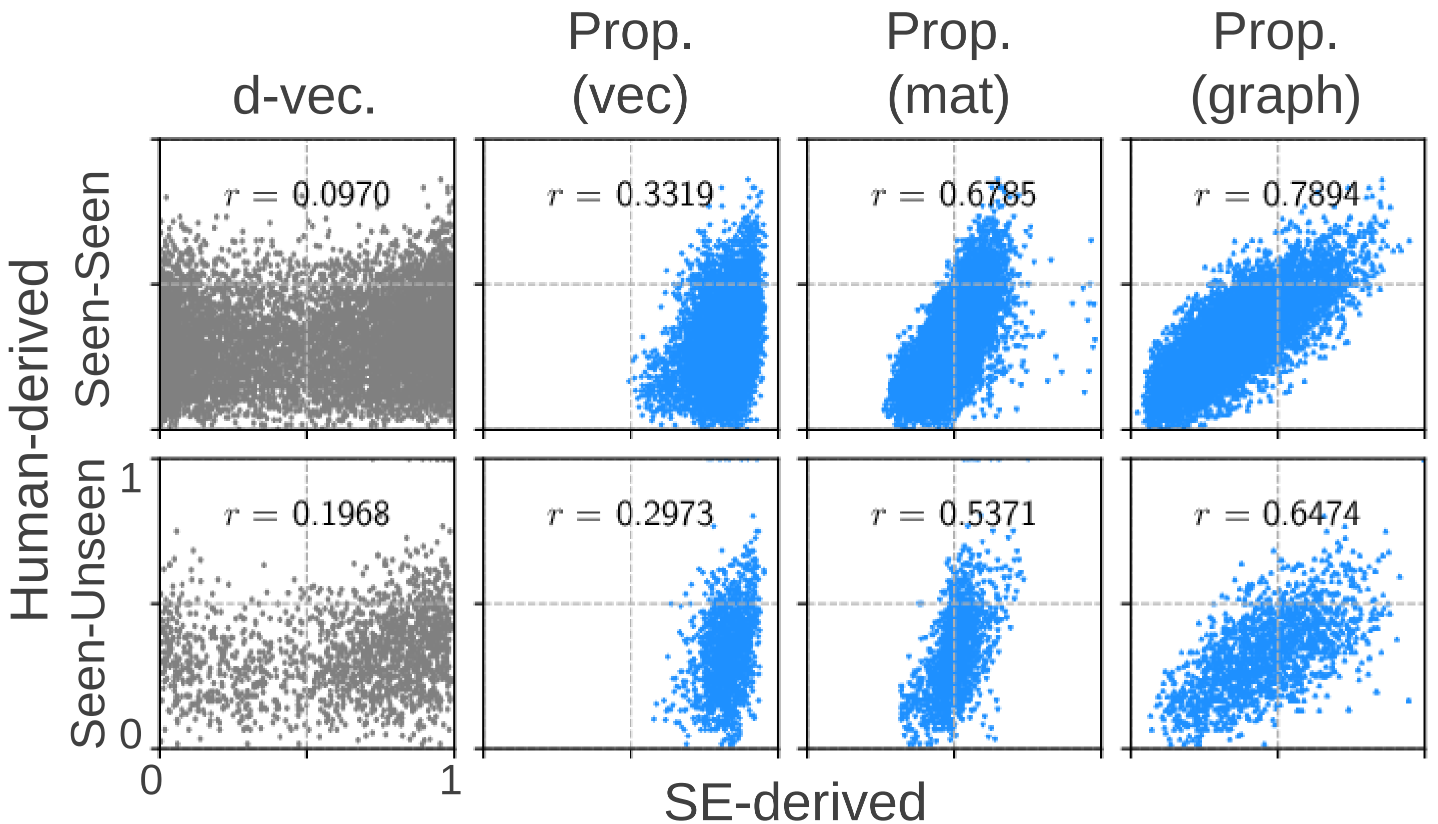
| | |
|---|---|
| Dataset<br>(16 kHz sampling) | JNAS [Itou+99] 153 female speakers<br>5 utterances per speaker for scoring<br>About 130 / 15 utterances for DSRL & evaluation<br>(F001 ~ F013: unseen speakers for evaluation) |
| Similarity score | －3 (dissimilar) ~ ＋3 (similar)<br>(Normalized to [－1, ＋1] or [0, 1] in DSRL) |
| Speech parameters | 40-dimensional mel-cepstra, F0, aperiodicity<br>(extracted by STRAIGHT analysis [Kawahara+99] |
| DNNs | Fully-connected (for details, please see our paper) |
| Dim. of SEs | 8 |
| AL setting | Pool-based simulation<br>(Using binary masking for excluding unobserved scores) |
| **DSRL methods** | **Conventional: d-vectors** [Variani+14]<br>**Ours: Prop. (vec), Prop. (mat), or Prop. (graph)** |

# Evaluation 1: SE Interpretability

**Scatter plots of human-/SE-derived similarity scores**

Prop. (*) highly correlated with the human-derived sim. scores.

→ **Our DSRL can learn interpretable SEs better than d-vec!**



# Evaluation 2: Speaker Interpolation Controllability

## Task: generate new speaker identity by mixing two SEs

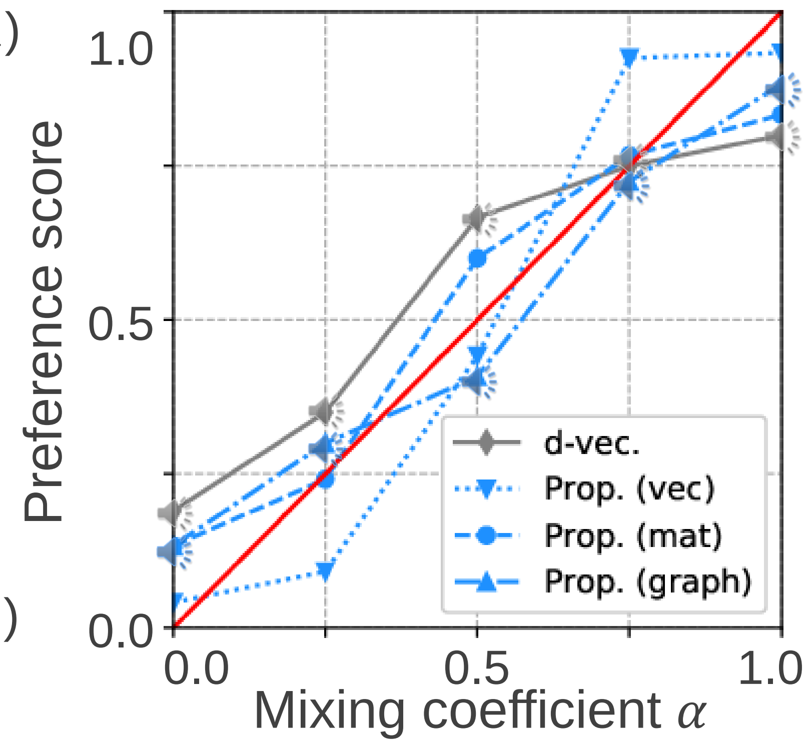
We evaluated spkr. sim. between interpolated speech with $\alpha \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ and original speaker's ($\alpha = 0$ or $1$).

The score curves of Prop. (*) were closer to the red line.

→ **Our SEs achieve higher controllability than d-vec.!**

B (mixed w/ $\alpha = 1$)

A (mixed w/ $\alpha = 0$)

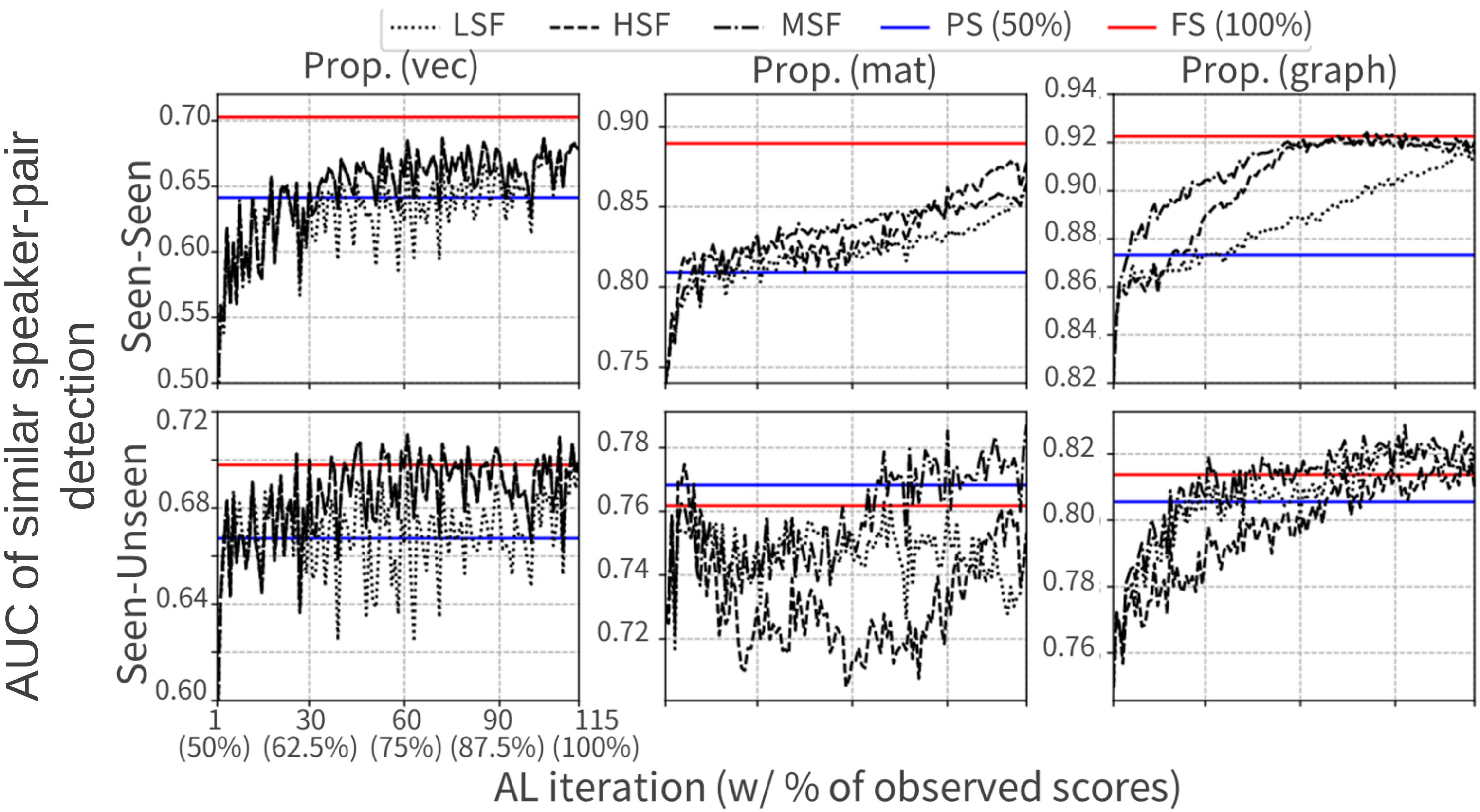


# Evaluation 3: AL Cost Efficacy

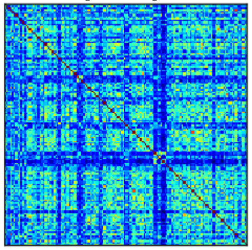
**AL setting: starting DSRL from PS to reach FS situation**

**MSF was the best** query strategy for all proposed methods.

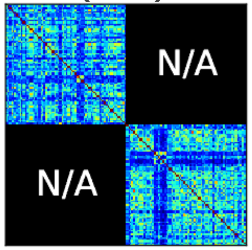
**Prop. (vec / graph) reduced the cost**, but **Prop. (mat) didn't work**




Fully Scored (FS)



Partially Scored (PS)



# Summary

## Purpose

Learning SEs highly correlated with perceptual speaker similarity

## Proposed methods

**1) Perceptual-similarity-aware learning of SEs**

**2) Human-in-the-loop AL for DSRL**

## Results of our methods

**1) learned SEs having high correlation with human perception**

**2) achieved better controllability in speaker interpolation**

**3) reduced costs of scoring/training by introducing AL**

## For detailed discussion...

Please read our TASLP paper (open access)!

**Thank you for your attention!**