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.
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])

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.
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

Speech params. → d-vectors [Variani+14] → Spkr. encoder → Spkr. classification → Spkr. IDs

Distance metric in spkr. space
Perceptual metric (i.e., speaker similarity)

Minimizing cross-entropy

Speaker space
Our Method: Perceptual-Similarity-Aware DSRL

1. Large-scale scoring of perceptual spkr. similarity

Spkr. pairs

Perceptual similarity scoring

Similarity score

2. SE learning considering the similarity scores

Speech params.

DNN (Spkr. encoder)

SEs

Learned similarity

Similarity score

Loss to predict sim.

Vector

Matrix

Graph

$L_{SIM}^{(*)}$
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** $S = [s_1, \ldots, s_i, \ldots, s_{N_s}]$

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

$s_i = [s_{i,1}, \ldots, s_{i,j}, \ldots, s_{i,N_s}]^T$: the $i$th similarity score vector

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

(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

$$L_{SIM}^{(vec)}(s, \hat{s}) = \frac{1}{N_s} (\hat{s} - s)^T (\hat{s} - s)$$
Algorithm 2: Similarity Matrix Embedding

Associate the Gram matrix of SEs with the matrix $S$

$$L_{SIM}^{(mat)}(D, S) = \frac{1}{Z_S} \|\tilde{K}_D - \tilde{S}\|_F^2$$

$Z_S$: Normalization coefficient ( $\tilde{S}$ represents off-diagonal matrix of $S$)
Algorithm 3: Similarity Graph Embedding

Learn the structure of speaker similarity graph from SE pairs

Speech params.

Speech encoder

SEs

$d$

Edge prediction

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

$p_{i,j} = \exp \left(-\|d_i - d_j\|_2^2\right)$: 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

- Spkr. encoder training
- Score prediction
- Scored spkr. pairs
- Unscored spkr. pairs
- Listeners
- Query selection
- Score annotation

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

Vector

Matrix

Graph

Spkr. encoder

training

Scored spkr. pairs

Scored

Unscored spkr. pairs

Listener

Score

annotation

Query

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

→ return to AL step 1

Vector Matrix Graph

Spkr. encoder training

Scored spkr. pairs

Unscored spkr. pairs

Listeners

Scored spkr. pairs

Query selection

Score prediction

Query strategy

Predicted

Selected

MSF

HSF

LSF

Vector Matrix Graph

Spkr. encoder training

Scored spkr. pairs

Unscored spkr. pairs

Listeners

Scored spkr. pairs

Query selection

Score prediction

Query strategy

Predicted

Selected

MSF

HSF

LSF
Experimental Evaluations
## Experimental Conditions

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

![Diagram showing correlation between human and SE-derived similarity scores](image)
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.!

(20 answers/listener, total $30 \times 2$ listeners, method-wise preference XAB test)
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

In each AL iteration, sim. scores of 43 speaker-pairs were newly annotated.
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!