IEEE ICASSP 2020

A Generalized Framework for Domain Adaptation of PLDA in Speaker Recognition

Qiongqiong Wang, Koji Okabe, Kong Aik Lee, and Takafumi Koshinaka

Biometrics Research Laboratories,
NEC Corporation, Japan

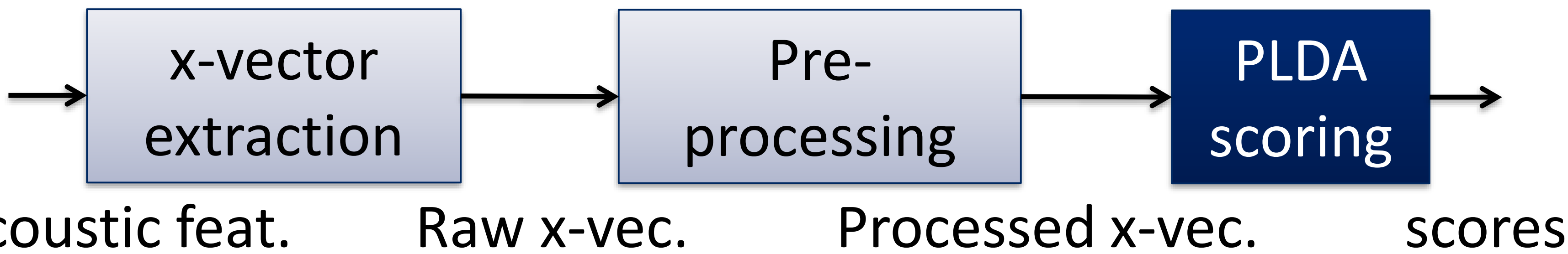
Contents

Background

Speaker recognition: to recognize a person for his/her voices

Promising framework

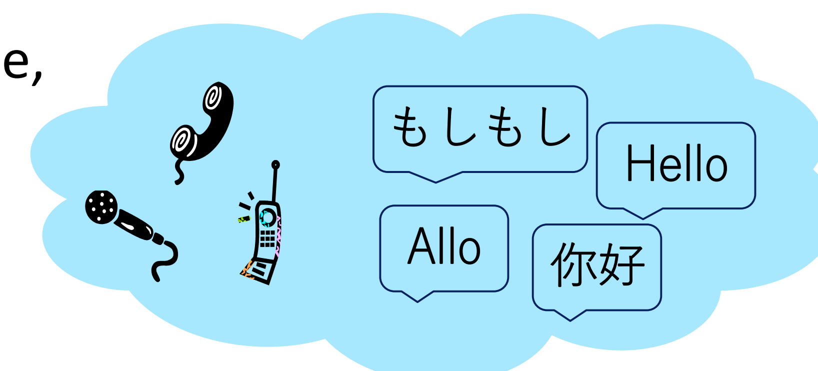
- Deep speaker embedding + PLDA (Probabilistic Linear Discriminant Analysis)



Domain mismatch

- Domain: recording condition, language, emotion, transmission channel…
- Degradation in EER

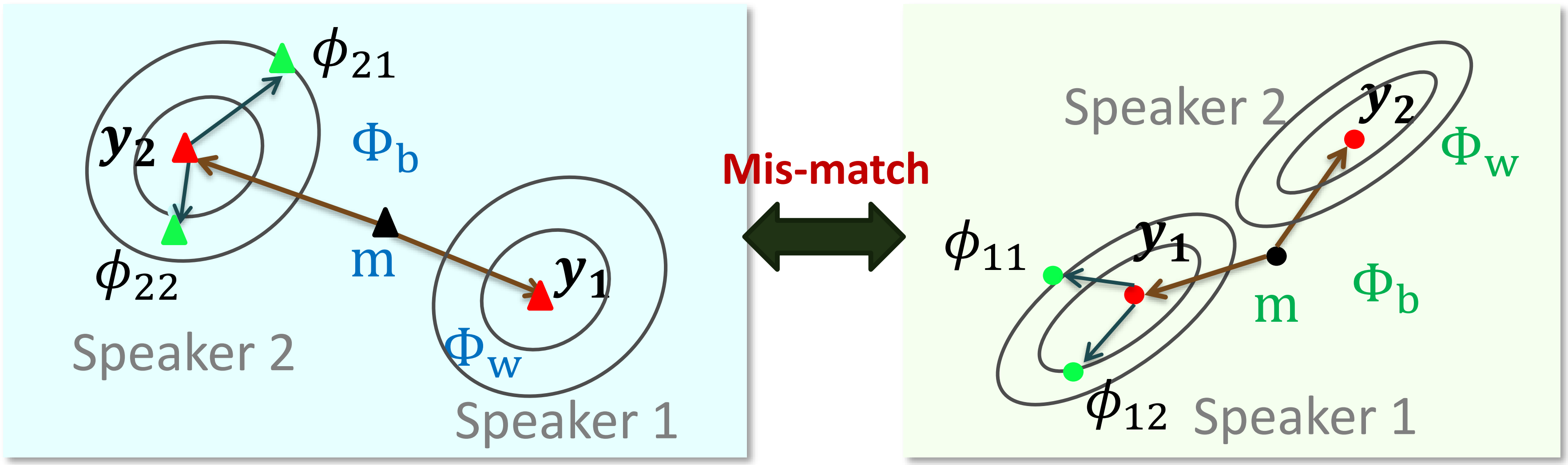
  - 2~3 times [Garcia-Romero+ 2014]



Domain Adaptation for PLDA

Different distributions in feature (speaker embedding) space

- Training: Out-of-domain (OOD) data   - Eval: In-domain (InD) data



Problem: costly to collect large labelled InD data

Solution

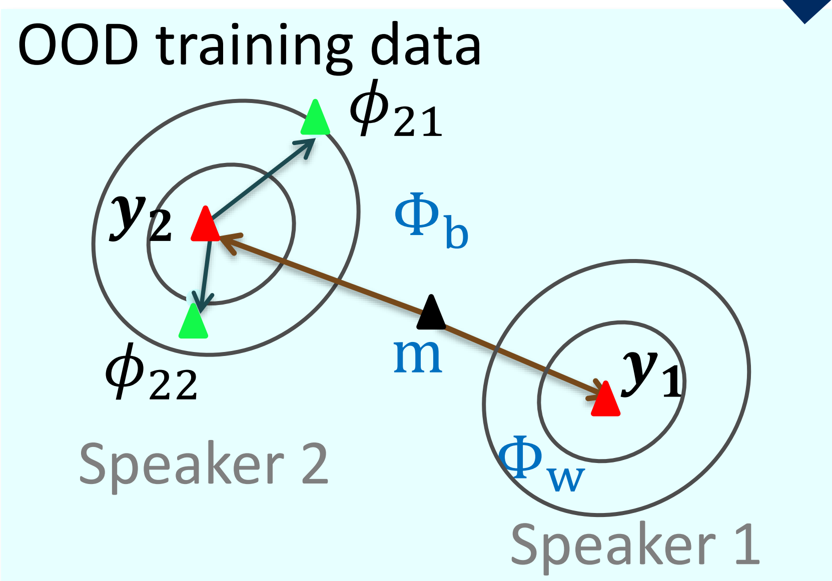
OOD model → **Domain adaptation** → Adapted model
Small InD dataset → **Domain adaptation**

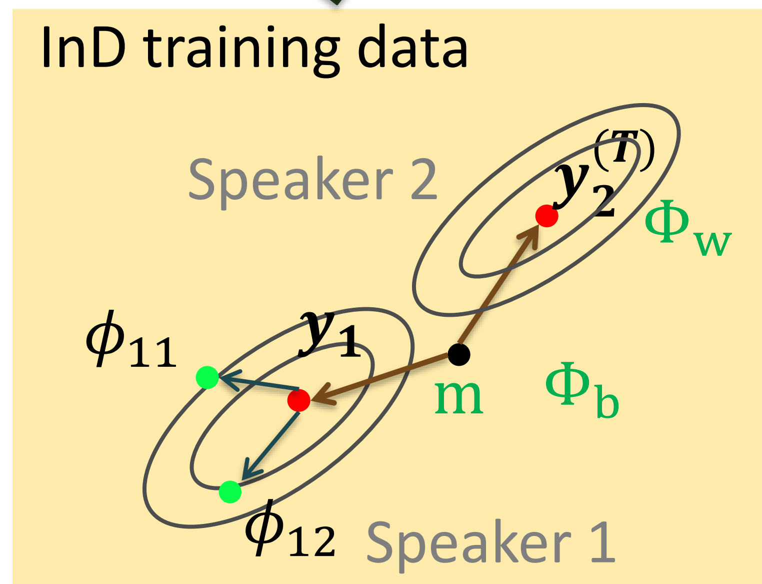
- Adapting backend is preferred => More effective and less costly
- PLDA: mean, between- and within-covariance $\{\boldsymbol{m}, \Phi_b, \Phi_w\}$

Conventional Method 1: LIP

Linear interpolation (LIP) [Garcia-Romero+ 2014]

$$\Phi_{\mathrm{b/w}} = (1-\alpha)\Phi_{\mathrm{OOD}} + \alpha\Phi_{\mathrm{InD}}$$





Problem:

- Assumption: OOD is not far from InD
- Variance of OOD can be under-estimated

Conventional Method 2: CORAL+

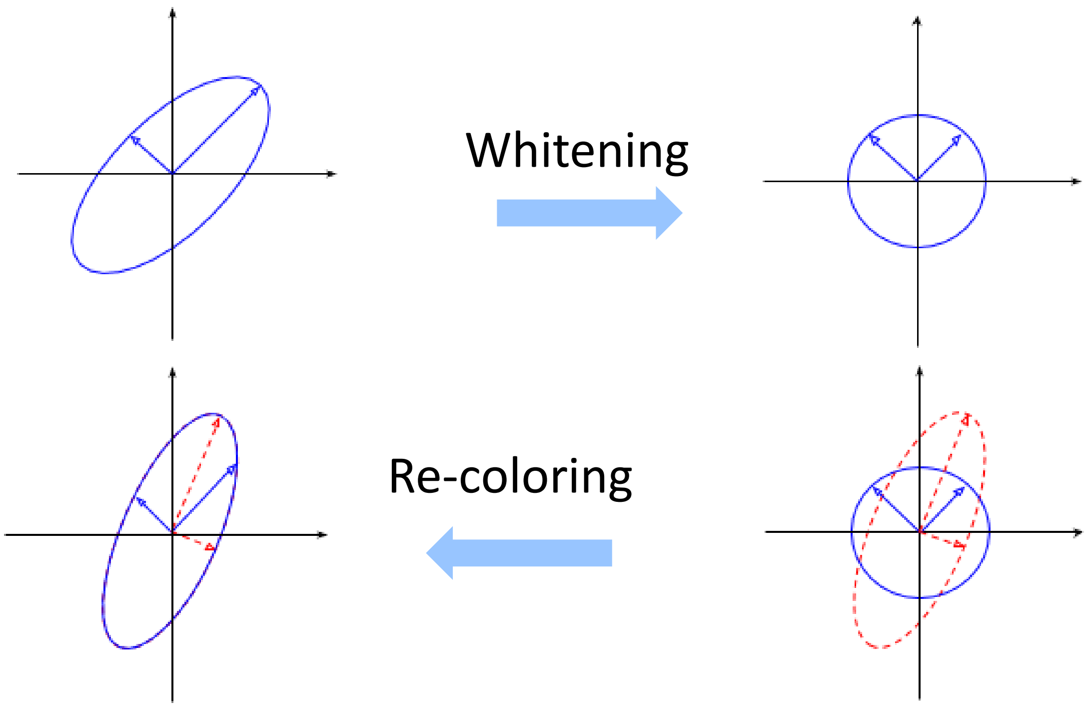
Correlation alignment (CORAL) [Sun et al, 2016] [Alam et al, 2018]

- Align OOD covariance matrices to match the InD feature vectors

$$\Phi_{\mathrm{InD}}^{\mathrm{pseudo}} = C_{\mathrm{InD}}^{1/2}\left(C_{\mathrm{OOD}}^{-1/2} \Phi_{\mathrm{OOD}} C_{\mathrm{OOD}}^{-1/2} \right) C_{\mathrm{InD}}^{1/2}$$

Step 1: Φ_{OOD}

Step 2: $\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}$

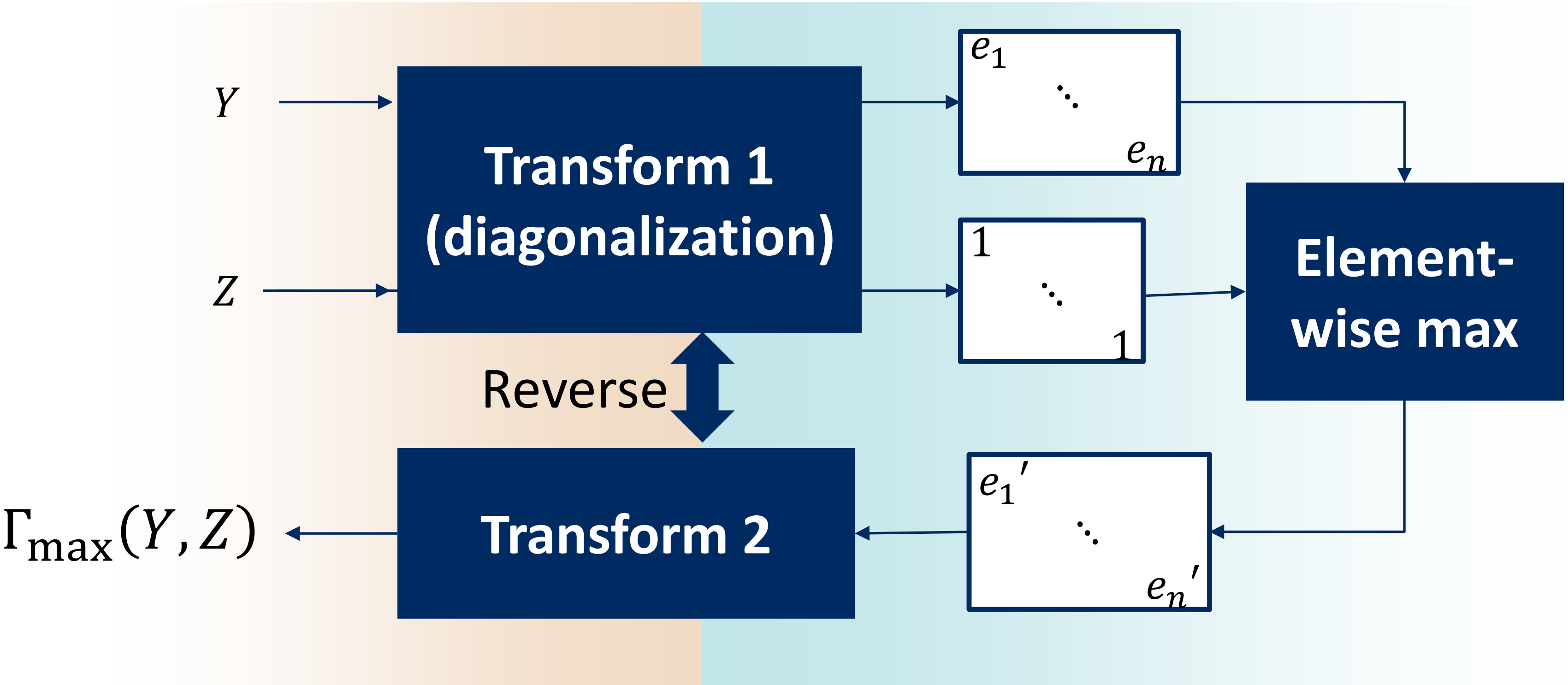


Conventional Method 2: CORAL+

CORAL+ [Lee+ 2019]: Unsupervised

$$\Phi_{\mathrm{InD}}^{+} = \alpha\Phi_{\mathrm{OOD}} + (1-\alpha)\Gamma_{\max}(\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}, \Phi_{\mathrm{OOD}})$$

- Linear interpolation w/ OOD model
- $\Gamma_{\max}(\mathrm{Y},\mathrm{Z})$: guarantee the variances and uncertainty to increase



Proposed Techniques

Robust domain adaptation with supervised and unsupervised manner

1. <u>Correlation-alignment-based interpolation (CIP)</u>
   - Pseudo InD PLDA is closer to a true InD PLDA than OOD PLDA
2. <u>Covariance regularization (reg)</u>
   - Guarantee to propagate the uncertainty seen in OOD data to PLDA

LIP

$$\Phi^{+} = \alpha\Phi_{\mathrm{InD}} + (1-\alpha)\Phi_{\mathrm{OOD}}$$

CIP

$$\Phi^{+} = \alpha\Phi_{\mathrm{InD}} + (1-\alpha)\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}$$

LIP-reg

$$\Phi^{+} = \alpha\Phi_{\mathrm{InD}} + (1-\alpha)\,\Gamma_{\max}(\Phi_{\mathrm{OOD}}, \Phi_{\mathrm{InD}})$$

CIP-reg

$$\Phi^{+} = \alpha\Phi_{\mathrm{InD}} + (1-\alpha)\Gamma_{\max}(\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}, \Phi_{\mathrm{InD}})$$

Proposed: A Generalized Framework

Three main factors

1) Interpolations of covariance matrices
2) Correlation alignment
3) Covariance regularization

Summarized in a general form

$$\Phi^{+} = \alpha\Phi_0 + (1-\alpha)\Gamma_{\max}(\Phi_1, \Phi_2)$$

- Φ_0: base; Φ_1: developer; Φ_2: reference

Special Cases

Generalized Framework : $\Phi^{+} = \alpha\Phi_0 + (1-\alpha)\Gamma_{\max}(\Phi_1, \Phi_2)$

Special case	Φ_0	Φ_1	Φ_2
CORAL [Alam+ 2018]	$\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}$	$\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}$	$\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}$
CORAL+ [Lee+ 2019]	Φ_{OOD}	$\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}$	Φ_{OOD}
Kaldi *	Φ_{OOD}	$\mathrm{C_i}$	$\Phi_{\mathrm{OOD}}^{\mathrm{b}} + \Phi_{\mathrm{OOD}}^{\mathrm{w}}$
LIP [Garcia-Romero+ 2014]	Φ_{InD}	Φ_{OOD}	Φ_{OOD}
LIP + regularization	Φ_{InD}	Φ_{OOD}	Φ_{InD}
CIP	Φ_{InD}	$\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}$	$\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}$
CIP + regularization	Φ_{InD}	$\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}$	Φ_{InD}
Case 8	Φ_{InD}	$\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}$	Φ_{OOD}
Case 9	Φ_{InD}	$\Gamma_{\max}(\Phi_{\mathrm{InD}}^{\mathrm{pseudo}}, \Phi_{\mathrm{OOD}})$	Φ_{InD}

* Available: https://github.com/kaldiasr/kaldi/tree/master/egs/sre16/v2

Experimental Setting

- Datasets

Dataset				#Speech
Train	X-vector	OOD	SWB, VoxCeleb 1 and 2, MIXER, augmentation	-
	PLDA	OOD	MIXER, augmentation	262,427
		InD	SRE 18 eval	13,451
Score norm			SRE 18 unlabeled	2,332
Eval			SRE 18 dev	1,741

- 40-d acoustic features: energy + 39-d MFCC
- 512-d x-vector extractor: 43-layers TDNN
  - Residual connections and a 2-head attentive statistics pooling
- LDA: 150-d
  - OOD LDA for both InD and OOD PLDAs in interpolations
- Gaussian PLDA: 150-d

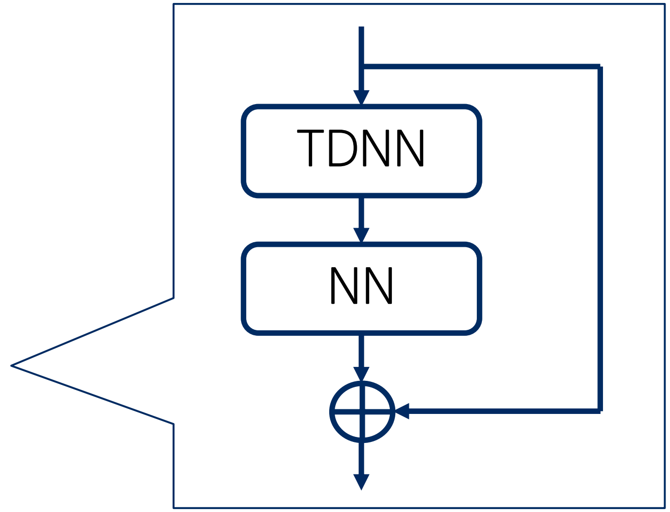
X-vector Extraction

- Very deep 43-layer TDNN
- Residual blocks [He+'16] for avoiding gradient vanishing problem
- Additive margin (AM) softmax [Wang+'18] shows high discriminability

43 layers

	kernel, dilation, output
frame 1	5, 1, 512
frame 2	1, 1, 512
frame 3	$\begin{bmatrix} 3, 2, 512 \\ 1, 1, 512 \end{bmatrix} \times 5$
frame 4	$\begin{bmatrix} 3, 3, 512 \\ 1, 1, 512 \end{bmatrix} \times 5$
frame 5	$\begin{bmatrix} 3, 4, 512 \\ 1, 1, 512 \end{bmatrix} \times 5$
frame 6	$\begin{bmatrix} 3, 5, 512 \\ 1, 1, 512 \end{bmatrix} \times 5$
frame 7	1, 1, 1500
pool	2-head attentive statistics
seg 1	512
seg 2	512
output	additive margin softmax

Residual block



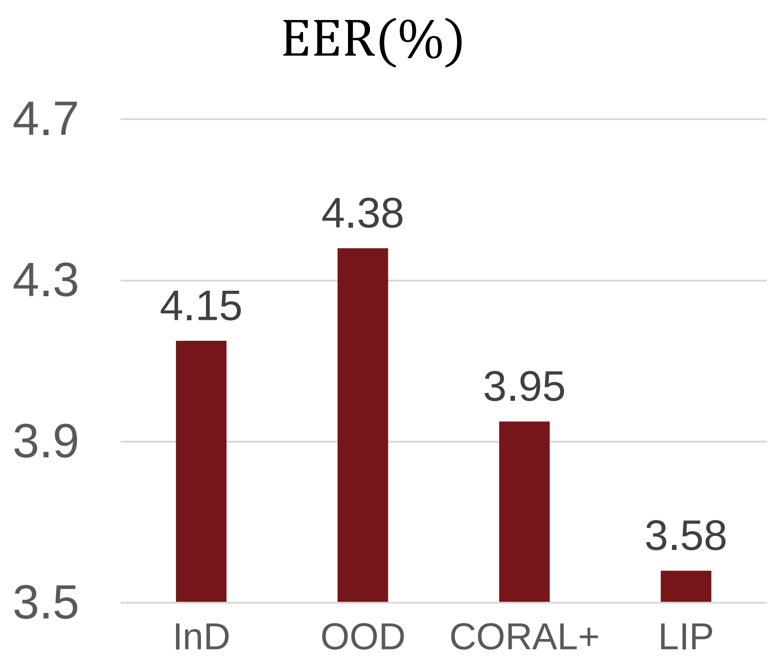
AM-softmax

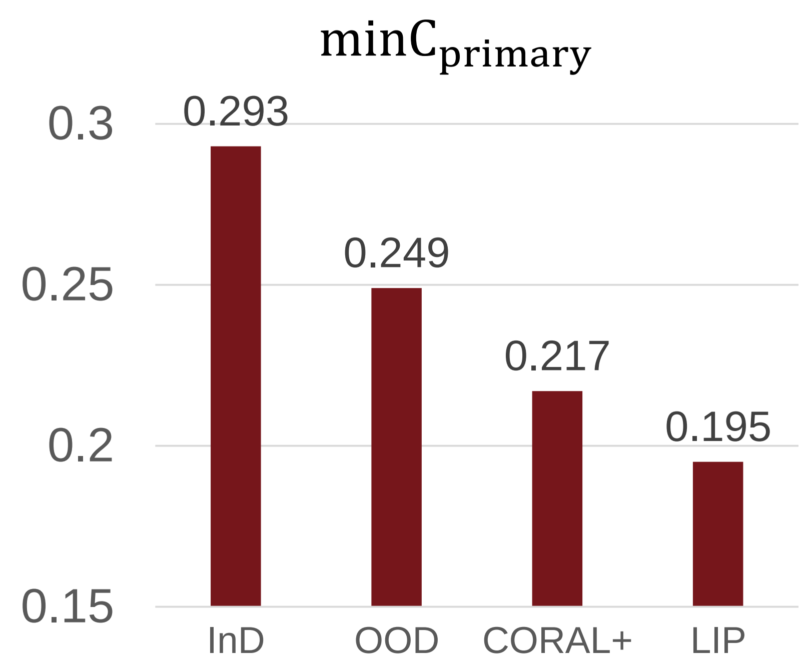
$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\cdot\cos\theta_{y_i} - m}}{e^{s\cdot\cos\theta_{y_i} - m} + \sum_{j\neq y_i} e^{s\cdot\cos\theta_{y_i}}}$$

Experimental Result (1)

❑ PLDA with conventional domain adaptation methods

- InD PLDA did not outperform OOD PLDA due to limited InD training data
- Both domain adaptation methods outperformed any single OOD or InD system
- Supervised linear interpolation would outperform unsupervised CORAL+

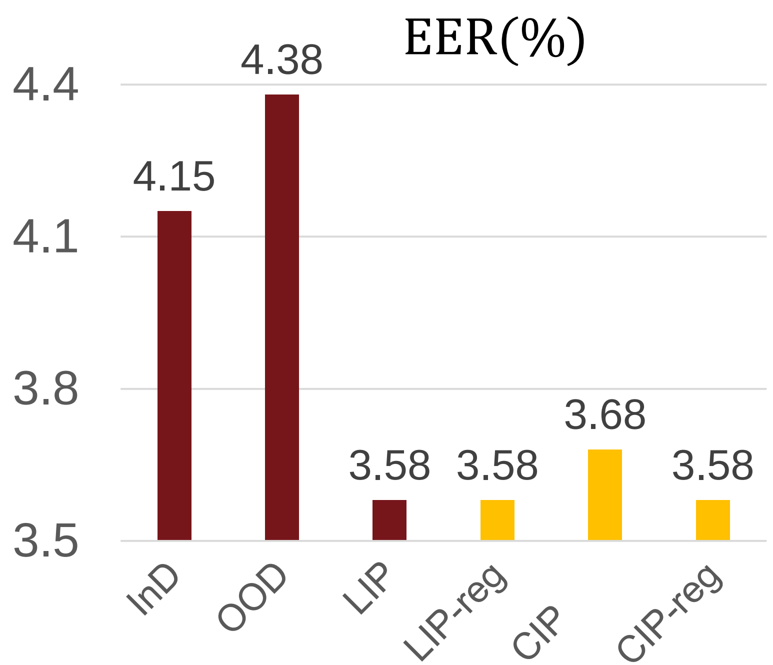


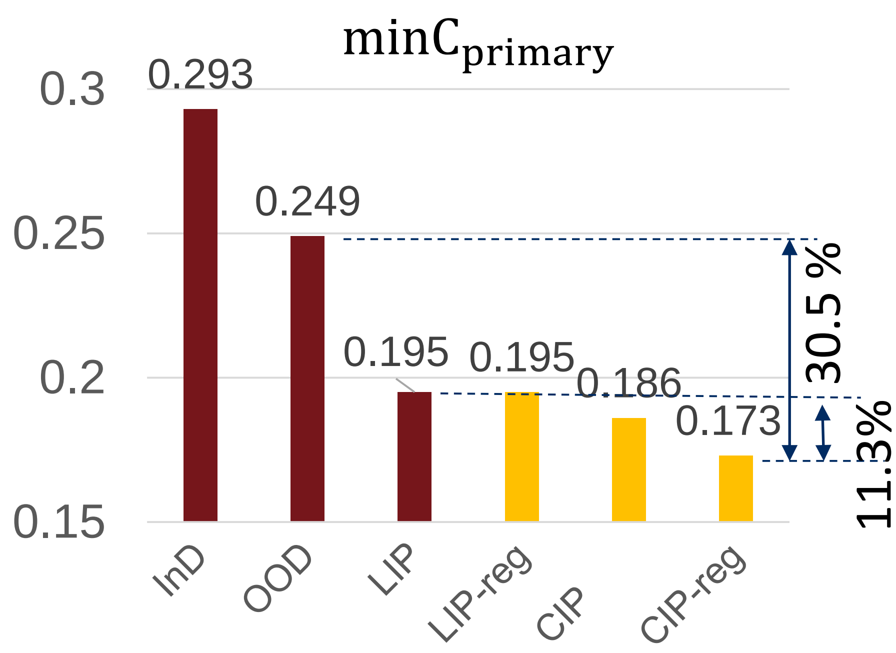


Experimental Result (2)

PLDA with proposed correlation-alignment-based interpolation (CIP) and covariance regularization (reg)

- All of the proposed methods performed better than LIP in $\mathrm{minC_{primary}}$
- CIP-reg reduced $\mathrm{minC_{primary}}$ by 30.5% as compared with the single systems
- CIP-reg lowered $\mathrm{minC_{primary}}$ by 11.3% than that of LIP



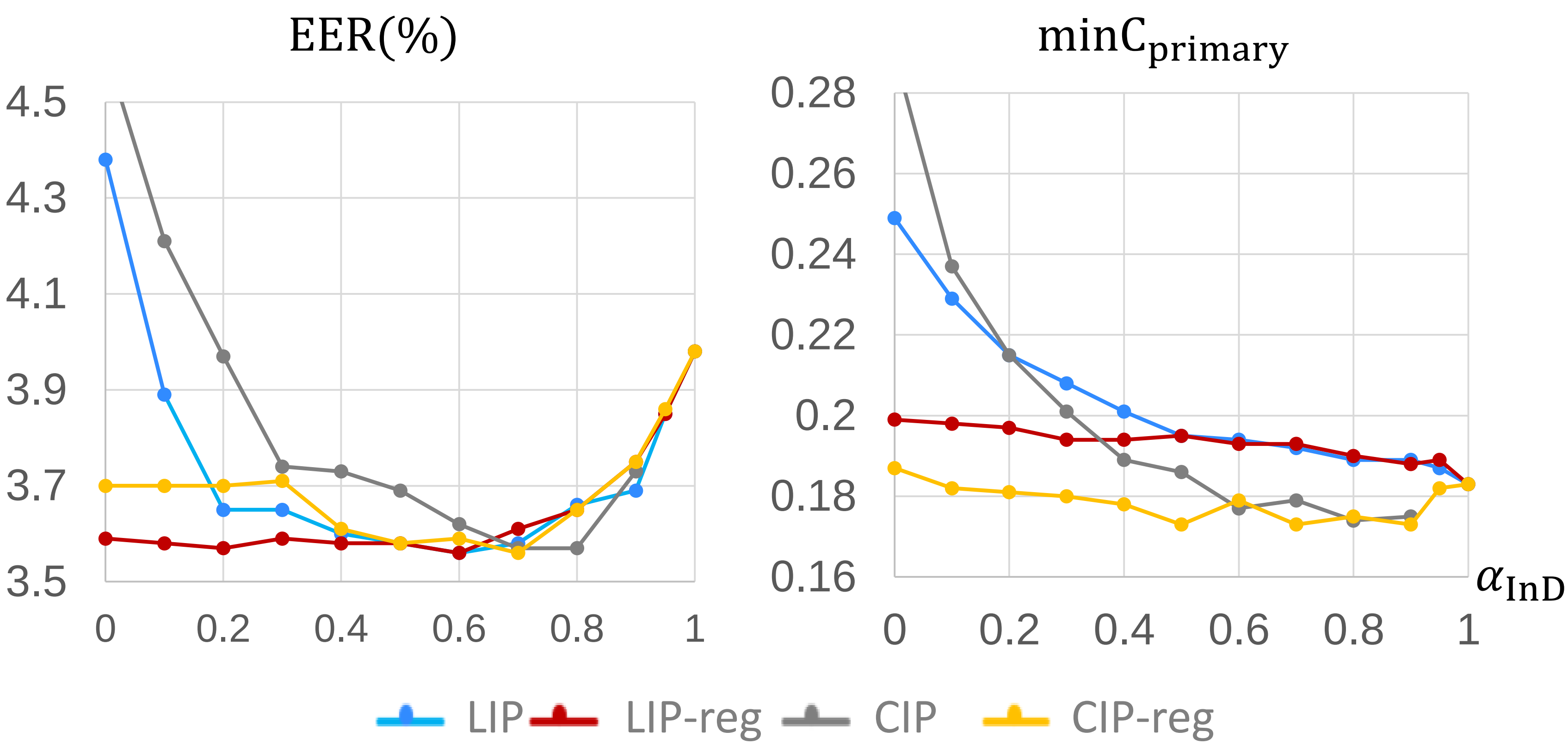


* Interpolation weight is 0.5

Experimental Result (3)

Proposed methods with varying interpolation weights

- Regularization provided more robust performance for both LIP and CIP
- CIP and CIP-reg were better than LIP and LIP-reg in $\mathrm{minC_{primary}}$ with all weights
- Best $\mathrm{minC_{primary}}$ of the CIP reg system was 5.5% lower than LIP's best



Summary

Proposed two techniques for robust domain adaptation of PLDA

1) Correlation-alignment-based interpolation (CIP)
   - Decreases $\mathrm{minC_{primary}}$ up to 30.5% as compared to OOD PLDA
   - 5.5% lower $\mathrm{minC_{primary}}$ than the conventional linear interpolation
2) Covariance regularization
   - Ensures robustness for interpolations w.r.t. varying interpolation weights

Proposed a generalized framework for domain adaptation of PLDA in speaker recognition

- Works with both unsupervised and supervised methods
- Enable to combine the two proposed techniques and several existing methods into a single formulation

\Orchestrating a brighter world

NEC