

Task description

- Language recognition on very short (1s) test utterances.
- Severe domain mismatch (esp. recording conditions) between training and test utterances.

Motivation & Contribution

- Front-end: Speaker adversarial multi-task learning (AMTL)
- Phonetic bottleneck features (BNFs) outperform spectral features in i-vector training. – Speaker variation is implicitly suppressed by phonetic BNF learning.
- Speaker AMTL aims explicitly at speaker-invariant BNF learning.
- **Back-end:** Unsupervised adaptation of probabilistic linear discriminant analysis (PLDA)
- -Commonly used back-end models e.g. LDA and Gaussian linear classifier suffer from severe performance degradation due to domain mismatch.
- Unsupervised PLDA adaptation is effective in alleviating domain mismatch in speaker recognition [1].

• General framework:



Model Structures

• Speaker-invariant feature learning:



-During training, parameters of M_u , M_s and M_h , denoted as θ_u , θ_s and θ_h , are updated as,

$$\begin{aligned} \theta_{y} &\leftarrow \theta_{y} - \delta \frac{\partial \mathcal{L}_{y}}{\partial \theta_{y}}, \\ \theta_{s} &\leftarrow \theta_{s} - \delta \frac{\partial \mathcal{L}_{s}}{\partial \theta_{s}}, \\ \theta_{h} &\leftarrow \theta_{h} - \delta \Big[\frac{\partial \mathcal{L}_{y}}{\partial \theta_{h}} - \lambda \frac{\partial \mathcal{L}_{s}}{\partial \theta_{h}} \Big], \end{aligned}$$

- where δ is the learning rate, \mathcal{L}_u and \mathcal{L}_s are cross-entropy loss values of senone and speaker classification tasks, λ is the adversarial weight.
- -After training, BNF representation learnt by M_h is speaker-invariant and phoneticallydiscriminative.

Senone labels:

- Generated by an out-of-domain (OOD) phone recognizer.
- -Language-independent senone labels.
- To control the output layer size of M_u .
- GMM-UBM/i-vector training:
- Input features are BNFs extracted from speaker AMTL.

Adversarial Multi-Task Deep Features and Unsupervised Back-End Adaptation for Language Recognition

Zhiyuan Peng^{*}, Siyuan Feng^{*} and Tan Lee DSP-Speech Technology Lab, Dept. of Electronic Engineering, The Chinese University of Hong Kong (CUHK) {jerrypeng1937, fengsym.ee}@gmail.com, tanlee@ee.cuhk.edu.hk (*Equal contribution)

► Scoring

(1)(2)(3)

• Back-end PLDA estimation:

PLDA assumes an i-vector ω_{ij} (*j*-th utterance in *i*-th language) generated as,

 $oldsymbol{\omega}_{oldsymbol{i}oldsymbol{j}} = oldsymbol{\mu} + {
m Fh}_{f i} + oldsymbol{\epsilon}_{oldsymbol{i}oldsymbol{j}},$ $\mathbf{h_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$ $\epsilon_{ij} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),$

where $\boldsymbol{\omega}_{ij} \in \mathbb{R}^D$, $\mathbf{F} \in \mathbb{R}^{D \times P}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$.

- -Columns of F provide the basis for the language-specific subspace, or *eigen-language*.
- -P is the subspace dimension, normally smaller than #classes (#languages in this work).
- -Based on Eqt. (4), an i-vector is assumed drawn from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{F}\mathbf{F}^{\mathsf{T}})$, where $\boldsymbol{\Sigma}$ and $\mathbf{F}\mathbf{F}^{\mathsf{T}}$ are within- and between-class variability. μ is global mean and can be precomputed and removed.
- -PLDA parameters $\{\mathbf{F}, \boldsymbol{\Sigma}\}\$ are estimated by an EM algorithm [2].
- During scoring phase, PLDA computes the similarity score of a *trial* (ω_t , *i*) composed of a test i-vector ω_t and language *i* as,

$$\mathcal{R}(\boldsymbol{\omega_t}, i) = \log \frac{p(\mathbf{a_t})}{p(\mathbf{\omega^i}|\mathbf{F})}$$

where ω^i is the average of training i-vectors that belong to language *i*.

- Unsupervised PLDA adaptation:
- -Leverage test (in-domain) i-vectors for adapting PLDA parameters $\{F_0, \Sigma_0\}$ estimated from training (out-of-domain) i-vectors.
- -Key issue: test i-vectors lack labels.
- -Solution: Agglomerative hierarchical clustering (AHC) towards test i-vectors to obtain labels.
- -Distance between a pair of i-vectors η_1 and η_2 is defined based on $\{F_0, \Sigma_0\}$ as follows,

$$d(\boldsymbol{\eta_1}, \boldsymbol{\eta_2}) = -\log \frac{p(\boldsymbol{\eta_1})}{p(\boldsymbol{\eta_1}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_0}|\mathbf{F_$$

- -AHC with *complete-linkage criterion* is performed until a pre-defined cluster number is reached.
- -In-domain PLDA $\{F_{ad}, \Sigma_{ad}\}$ are estimated by test i-vectors and their cluster labels. -Final scoring based on $\{\mathbf{F}_{ad}, \boldsymbol{\Sigma}_{ad}\}$.

AP17-OLR Task Description

- AP17-OLR challenge dataset [3]: 10 oriental languages, each with 10 hours recorded by mobile phones.
- **Training**: 54, 266 utterances, 79 hours.
- **Dev_1s:** 17, 948 utterances, 5 hours.
- Test_1s: 22,051 utterances, 6 hours.
- Evaluation metric: C_{avg} and Equal Error Rate (EER).

$$C_{avg} = \frac{1}{N} \sum_{L_t} 0.5 \cdot \left[P_{MS}(L_t) + \frac{1}{N-1} \sum_{L_n} P_{FA}(L_t, L_n) \right],$$

where N is the number of languages, L_t and L_n denote the target and non-target languages, P_{MS} and P_{FA} are the missing and false alarm probabilities.

- Measuring the mismatch between training and dev_1s: A demo experiment is conducted to show the domain mismatch between training and developemnt/test data. – Setup:
 - *** Pseudo-dev:** a 12-hour subset randomly selected from **training set**.
 - * Training-part: the remaining 67-hour subset from training set.
 - * **Pseudo-dev_1s** and **training-part_1s**: utterances are trimmed to 1 second.
- * Front-end: 100-dim i-vectors extracted from 60-dim voiced MFCCs+ Δ + $\Delta\Delta$ without CMVN.
- * Back-end: one-layer MLP with 512 neurons, followed by softmax output.
- -Results (C_{avq} /EER%):

(4)

 $(oldsymbol{\omega^i},oldsymbol{\omega_t}|\mathbf{F}\mathbf{F}^\intercal,oldsymbol{\Sigma})$ (5) $\mathbf{F}^{\mathsf{T}}, \mathbf{\Sigma}) p(\boldsymbol{\omega_t} | \mathbf{F} \mathbf{F}^{\mathsf{T}}, \mathbf{\Sigma})$

$(\boldsymbol{\eta_1}, \boldsymbol{\eta_2} \mathbf{F_0} \mathbf{F_0}^T, \boldsymbol{\Sigma_0})$	(6)
$\mathbf{F_0}^{T}, \mathbf{\Sigma_0}) p(\boldsymbol{\eta_2} \mathbf{F_0} \mathbf{F_0}^{T}, \mathbf{\Sigma_0})^{T}.$	(0)

$\{10, 50, 100, 200, 500\}.$ Results and Analysis Comparison of C_{avg} /EER% with different adversarial weights evaluated on dev 1s (b simple cosine scoring) $\int_{10}^{12} \int_{10}^{10} \int_{10}$		Dev_1s	Pseudo-dev_1s	Pseudo-dev	aining data	Traini	
Speaker-invariant BNFs: - Input: 40-dim MECCs without cepstral truncation. - Senone labels: obtained from a Czech phone recognizer [4], 135 senones in total. - Speaker labels: obtained from training data, 641 speakers in total. - DNN configuration: M_h is a 6-Ret.U-layer TDNN, 1024 neurons per layer (64 neur layer), layer-wise context: $\{-2, -1, 0, 1, 2\}, \{0\}, \{-1, 2\}, \{-3, -3\}, \{-7, -2\}, \{0\};$ M_y and M_z have 1 RetU layer followed by a softmax output layer. i-vector extractor: 2048-mixture UBM, 400-dimension i-vector extractor. Unsupervised PLDA adaptation: - Out-of-domain PLDA: estimated on training i-vectors and ground-truth labels. - In-domain PLDA: estimated on dev 1s i-vectors and cluster labels. - AHC: cluster dev 1s i-vectors to a pre-defined number of clusters ra {10, 50, 100, 200, 500}. Results and Analysis Comparison of C_{atcg} /EER% with different adversarial weights evaluated on dev 1s (b simple cosine scoring) Comparison of C_{atcg} /EER% with different adversarial weights evaluated on dev 1s (b simple cosine scoring)		/	/	1	• •		
- Input: 40-dim MFCCs without cepstral truncation. - Senone labels: obtained from a Czech phone recognizer [4], 135 senones in total. - Speaker labels: obtained from training data, 641 speakers in total. - DNN configuration: M_h is a 6-ReLU-layer TDNN, 1024 neurons per layer (64 neur layer), layer-wise context: {-2, -1, 0, 1, 2}, {0}, {-1, 2}, {-3, -3}, {-7, -2}, {0}; M_y and M_s have 1 ReLU layer followed by a softmax output layer. - vector extractor: 2048-mixture UBM, 400-dimension i-vector extractor. Unsupervised PLDA adaptation: - Out-of-domain PLDA: estimated on training i-vectors and ground-truth labels. - In-domain PLDA: estimated on dev 1s i-vectors and cluster labels. - AHC: cluster dev 1s i-vectors to a pre-defined number of clusters ra {10, 50, 100, 200, 500}. Results and Analysis Comparison of $C_{avg}/EER\%$ with different adversarial weights evaluated on dev 1s (b simple cosine scoring) $\frac{12}{\sqrt{1-0}}$ $\frac{12}{\sqrt{-0}}$ $\frac{12}$			tal Setup	Experime]		
- Senone labels: obtained from a Czech phone recognizer [4], 135 senones in total. - Speaker labels: obtained from training data, 641 speakers in total. - DNN configuration: M_b is a 6-ReLU-layer TDNN, 1024 neurons per layer (64 neur layer), layer-wise context: $\{-2, -1, 0, 1, 2\}$, $\{0\}$, $\{-1, 2\}$, $\{-3, -3\}$, $\{-7, -2\}$, $\{0\}$: M_y and M_z have 1 ReLU layer followed by a softmax output layer. i-vector extractor: 2048-mixture UBM, 400-dimension i-vector extractor. Unsupervised PLDA adaptation: - Out-of-domain PLDA: estimated on training i-vectors and ground-truth labels. - In-domain PLDA: estimated on dev 1s i-vectors and cluster labels. - AHC: cluster dev_ls i-vectors to a pre-defined number of clusters re $\{10, 50, 100, 200, 500\}$. Results and Analysis Comparison of C_{avg} /EER% with different adversarial weights evaluated on dev 1s (b simple cosine scoring) $12 \frac{12 \frac{12}{1-5} \frac{1}{1-5} \frac{1}{1-5}$					Fs:	ariant BNFs:	oeaker-inv
- Speaker labels: obtained from training data, 641 speakers in total. - DNN configuration: M_h is a 6-ReLU-layer TDNN, 1024 neurons per layer (64 neur layer), layer-wise context: {-2, -1, 0, 1, 2}, {0}, {-1, 2}, {-3, -3}, {-7, -2}, {0}; M_g and M_a have 1 ReLU layer followed by a softmax output layer. i-vector extractor: 2048-mixture UBM, 400-dimension i-vector extractor. Unsupervised PLDA adaptation: - Out-of-domain PLDA: estimated on training i-vectors and ground-truth labels. - In-domain PLDA: estimated on dev_1s i-vectors and cluster labels. - AHC: cluster dev_1s i-vectors to a pre-defined number of clusters rates {10, 50, 100, 200, 500}. Results and Analysis Comparison of C_{axg} /EER% with different adversarial weights evaluated on dev 1s (b simple cosine scoring) $\frac{12}{15, \frac{1}{9, 0}, \frac{1}{9, -0}, \frac{1}{25, 0}, $					Ĩ		•
- DNN configuration: M_h is a 6-ReLU-layer TDNN, 1024 neurons per layer (64 neur layer), layer-wise context: {-2, -1, 0, 1, 2}, {0}, {-1, 2}, {-3, -3}, {-7, -2}, {0}; M_y and M_y have 1 ReLU layer followed by a softmax output layer. i-vector extractor: 2048-mixture UBM, 400-dimension i-vector extractor. Unsupervised PLDA adaptation: - Out-of-domain PLDA: estimated on training i-vectors and ground-truth labels. - In-domain PLDA: estimated on dev 1s i-vectors and cluster labels. - AHC: cluster dev 1s i-vectors to a pre-defined number of clusters re {10, 50, 100, 200, 500}. Comparison of $C_{wwy}/EER\%$ with different adversarial weights evaluated on dev_1s (b simple cosine scoring) Comparison of $C_{wwy}/EER\%$ with/without unsupervised PLDA adaptation evaluate dev 1s and test 1s sets (same front-end configuration, $\lambda = 0.250$) No Adaptation			e •	A			
layer), layer-wise context: $\{-2, -1, 0, 1, 2\}$, $\{0\}$, $\{-1, 2\}$, $\{-3, -3\}$, $\{-7, -2\}$, $\{0\}$; M_y and M_a have 1 ReLU layer followed by a softmax output layer. i-vector extractor: 2048-mixture UBM, 400-dimension i-vector extractor. Unsupervised PLDA adaptation: = Out-of-domain PLDA: estimated on training i-vectors and ground-truth labels. = In-domain PLDA: estimated on dev_1s i-vectors and cluster labels. = AHC: cluster dev 1s i-vectors to a pre-defined number of clusters ra $\{10, 50, 100, 200, 500\}$. Results and Analysis Comparison of $C_{acg}/EER\%$ with different adversarial weights evaluated on dev 1s (b simple cosine scoring) $\frac{12}{0.5} \frac{10}{0.5} \frac{100}{0.5} \frac{100}{0.$	irons		•	C			-
i-vector extractor: 2048-mixture UBM, 400-dimension i-vector extractor. Unsupervised PLDA adaptation: = Out-of-domain PLDA: estimated on training i-vectors and ground-truth labels. = In-domain PLDA: estimated on dev_1s i-vectors and cluster labels. = AHC: cluster dev 1s i-vectors to a pre-defined number of clusters rates and Analysis Comparison of C_{aug} /EER% with different adversarial weights evaluated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine scoring) = Out-of-domain PLDA: estimate dev_1s and test 1s ests (same front-end configuration, $\lambda = 0.250$) = Out-of-domain PLDA: estimated on dev_1s (b simple cosine score				•		e	
Unsupervised PLDA adaptation: = Out-of-domain PLDA: estimated on training i-vectors and ground-truth labels. = In-domain PLDA: estimated on dev_Is i-vectors and cluster labels. = AHC: cluster dev Is i-vectors to a pre-defined number of clusters rates and the cluster of the cluster of clusters rates and the cluster of the cluster of clusters rates and the cluster of the cluster of clusters rates and the cluster of the cluster of clusters rates and the cluster of the cluster of clusters and the cluster of the cluster of clusters and the cluster of the cluster of the cluster of clusters and the cluster of		er.	tmax output lay	owed by a sc	eLU layer folle	, have 1 ReLU	M_y and M_y
- Out-of-domain PLDA: estimated on training i-vectors and ground-truth labels. - In-domain PLDA: estimated on dev_1s i-vectors and cluster labels. - AHC: cluster dev 1s i-vectors to a pre-defined number of clusters ration (10, 50, 100, 200, 500). Results and Analysis Comparison of C_{avg} /EER% with different adversarial weights evaluated on dev 1s (b) simple cosine scoring) $12 \int_{0}^{12} \int_{0}^{12} \int_{0}^{10} \int_{0}^$		ctractor.	nsion i-vector e	M , 400-dim	48-mixture UE	actor: 2048-n	ector exti
- In-domain PLDA: estimated on dev 1s i-vectors and cluster labels. - AHC: cluster dev 1s i-vectors to a pre-defined number of clusters ra {10, 50, 100, 200, 500}. Results and Analysis Comparison of C_{avy} /EER% with different adversarial weights evaluated on dev 1s (b simple cosine scoring) $12 - \frac{12}{11.5} - $					adaptation:	ed PLDA ada	nsupervise
- AHC: cluster dev_1s i-vectors to a pre-defined number of clusters ra {10, 50, 100, 200, 500}. Results and Analysis Comparison of C_{avg} /EER% with different adversarial weights evaluated on dev_1s (b simple cosine scoring) $\frac{12}{1.5} - \frac{1}{1.5} -$			e	e			
{10, 50, 100, 200, 500}. Results and Analysis Comparison of C_{ang} /EER% with different adversarial weights evaluated on dev_1s (b simple cosine scoring) $\frac{12}{11.5} - \frac{12}{10.5} - \frac{12}{10.5} - \frac{10}{10.5} - \frac{10}{10$							
Results and AnalysisComparison of C_{avg} /EER% with different adversarial weights evaluated on dev 1s (bsimple cosine scoring)1212141415Comparison of C_{avg} /EER% with/without unsupervised PLDA adaptation evaluatedev_1s and test_Is sets (same front-end configuration; cluster number in AHC 10SOIDev_1s8.25/7.566.68/6.846.61/6.656.47/6.497.36/7.53ConclusionsSpeaker AMTL suppresses speaker variation, which is beneficial to the LR task.	rangii	per of clusters rai	e-defined num	to a pi		_	_
Comparison of C_{avg} /EER% with different adversarial weights evaluated on dev 1s (b) simple cosine scoring) $12 \\ 11.5 \\ 10 \\ 10.5 \\ 10 \\ 9.5 \\ 9.0 \\ 2.0.125 \\ 1.0.250 \\ 1$					•	, 200, 000 J.	(10,00,100)
simple cosine scoring) $ \frac{12}{11.5} + 1$			l Analysis	Results an]		
simple cosine scoring) $ \frac{12}{11.5} + 1$	(back-	valuated on dev 1s (ba	arial weights e	erent adver	ER% with diff	of C_{ava} /EER $^{\circ}$	omparison
$\frac{11.5}{10}$ $\frac{11.5}{10}$ $\frac{11.5}{10}$ $\frac{11.5}{10}$ $\frac{11.5}{10}$ $\frac{10.5}{10}$ $\frac{10.5}{10}$ $\frac{10}{10}$ $$			8				_
$\frac{1}{10,5} + \frac{1}{10,5} + \frac{1}{10,50} + \frac{1}{10,5$					12		
$\frac{1}{10^{5}} \int_{\frac{1}{2}}^{\frac{1}{9}} \int_{\lambda=0.125}^{\frac{1}{2}} \int_{\lambda=0.250}^{\frac{1}{9}} \int_{\lambda=0.375}^{\frac{1}{9}} \int_{\lambda=0.250}^{\frac{1}{9}} \int_{\lambda=0.250}^{\frac{1}{9}} \int_{\lambda=0.375}^{\frac{1}{9}} \int_{\lambda=0.250}^{\frac{1}{9}} \int_{\lambda=0.375}^{\frac{1}{9}} \int_{\lambda=0.375}^{\frac{1}{9}} \int_{\lambda=0.250}^{\frac{1}{9}} \int_{\lambda=0.$			$- * - C_{avg}$		11.5		
$\frac{10}{9.5} \frac{1}{9.0} \frac{1}{\lambda = 0.125} \frac{1}{\lambda = 0.250} \frac{1}{\lambda = 0.375}$ Comparison of C_{avg} /EER% with/without unsupervised PLDA adaptation evaluate dev_1s and test_1s sets (same front-end configuration, $\lambda = 0.250$) $\frac{10}{10} \frac{10}{50} \frac{100}{100} \frac{200}{500} \frac{500}{100}$ $\frac{10}{10} \frac{10}{50} \frac{100}{100} \frac{100}{50} \frac{100}{500}$ $\frac{10}{10} \frac{10}{50} \frac{100}{100} \frac{100}{50} \frac{100}{500}$ $\frac{10}{10} \frac{10}{50} \frac{100}{100} \frac{100}{50} \frac{100}{500}$ $\frac{10}{10} \frac{10}{50} \frac{100}{50} \frac{100}{500}$ $\frac{10}{50} \frac{100}{50} \frac{100}{500}$ $\frac{10}{50} \frac{100}{50} \frac{100}{50}$ $\frac{10}{50} \frac{100}{50} \frac{100}{50}$ $\frac{10}{50} \frac{100}{50} \frac{100}{50}$					11		
$\begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \end{array} \\ \begin{array}{c} \end{array} \\ \end{array} $		- *		*	10.5		
$\int_{9}^{9} \frac{1}{\lambda = 0.125} \frac{1}{\lambda = 0.250} \frac{1}{\lambda = 0.375}$ Comparison of C_{avg} /EER% with/without unsupervised PLDA adaptation evaluate dev_1s and test_1s sets (same front-end configuration, $\lambda = 0.250$) $\boxed{\begin{array}{c c c c c c } No \ Adaptation} & With adaptation; cluster number in AHC & SOT \\ 10 & 50 & 100 & 200 & 500 \\ \hline Dev_1s & 8.25/7.56 & 6.68/6.84 & 6.61/6.65 & 6.47/6.49 & 7.07/6.99 & 7.45/7.26 & N \\ \hline Test_1s & 9.46/8.78 & - & - & 7.36/7.53 & - & - & 7.65 \\ \hline \end{array}}$ Speaker AMTL suppresses speaker variation, which is beneficial to the LR task.				8	10 -		
$\lambda = 0 \qquad \lambda = 0.125 \qquad \lambda = 0.250 \qquad \lambda = 0.375$ Comparison of C_{avg} /EER% with/without unsupervised PLDA adaptation evaluate dev_1s and test_1s sets (same front-end configuration, $\lambda = 0.250$) $\boxed{\begin{array}{c c c c c c } & \text{No Adaptation} & \text{With adaptation; cluster number in AHC} & \text{SOT} \\ \hline 10 & 50 & 100 & 200 & 500 \\ \hline 10 & 50 & 100 & 500 \\ \hline 10 & 50 & 100 & 500 \\ \hline 10 & 50 & 100 & 500 \\ \hline 10 & 50 & 100 & 500 \\ \hline 10 & 50 & 100 & 500 \\ \hline 10 & 50 & 100 & 500 \\ \hline 10 & 50 & 100 & 500 \\ \hline 10 & 50 & 100 & 500 \\ \hline 10 & 50 & 100 & 500 \\ \hline 10 & 50 & 100 & 500 \\ \hline 10 & 50 & 100 & 500 \\ \hline 10 & 50 & 100 \\ \hline 10 & 50 & 100 \\ \hline 10 & 50 & 1$			*		9.5		
Comparison of C_{avg} /EER% with/without unsupervised PLDA adaptation evaluate dev_1s and test 1s sets (same front-end configuration, $\lambda = 0.250$)No AdaptationWith adaptation; cluster number in AHC 10 50 100 200 500SOT 200 500Dev_1s8.25/7.56 9.46/8.786.68/6.84 6.61/6.65 7.36/7.536.47/6.49 - 7.36/7.537.07/6.99 7.45/7.26N 7.65ConclusionsSpeaker AMTL suppresses speaker variation, which is beneficial to the LR task.			$\lambda = 0.250$) = 0.125			
dev_1s and test_1s sets (same front-end configuration, $\lambda = 0.250$)No AdaptationWith adaptation; cluster number in AHC 10 50 100 200 500SOT 200 500Dev_1s $8.25/7.56$ $9.46/8.78$ $6.68/6.84$ $ 6.61/6.65$ $ 6.47/6.49$ $7.07/6.99$ $7.45/7.26$ N 7.65 ConclusionsSpeaker AMTL suppresses speaker variation, which is beneficial to the LR task.	_						
No Adaptation With adaptation; cluster number in AHC SOT 10 50 100 200 500 500 Dev_1s $8.25/7.56$ $6.68/6.84$ $6.61/6.65$ $6.47/6.49$ $7.07/6.99$ $7.45/7.26$ N Test 1s $9.46/8.78$ $ 7.36/7.53$ $ 7.65$ Conclusions Speaker AMTL suppresses speaker variation, which is beneficial to the LR task.	ted or	adaptation evaluated		-		J	-
$\begin{array}{c c c c c c c c c c c c c c c c c c c $				•			
Dev_1s $8.25/7.56$ $6.68/6.84$ $6.61/6.65$ $6.47/6.49$ $7.07/6.99$ $7.45/7.26$ N Test_1s $9.46/8.78$ $ 7.36/7.53$ $ 7.65$ Conclusions Speaker AMTL suppresses speaker variation, which is beneficial to the LR task.	ЛА [.			•		No Adaptatio	
Test_1s 9.46/8.78 — — 7.36/7.53 — — 7.65 Conclusions Speaker AMTL suppresses speaker variation, which is beneficial to the LR task.	N/A					8 25 / 7 56	$\overline{\mathbf{D}}$ D \mathbf{u}
Conclusions Speaker AMTL suppresses speaker variation, which is beneficial to the LR task.	55/7.9	· · · · ·	•	4 0.01/0.05	/	1	
Speaker AMTL suppresses speaker variation, which is beneficial to the LR task.						0.10/0.10	
Speaker AMTL suppresses speaker variation, which is beneficial to the LR task.			isions	Concl			
				••••-			
		a the I R task	h is heneficial t	ariation whi	sses sneaker w	TI clinnrecceo	eaker AM
to performance improvement on short-duration LR task.	zionif				-		

References

- speaker recognition," in Proc. Odyssey, 2014, pp. 260–264.
- *ICCV*, 2007, pp. 1–8.
- pp. 749–753.
- ogy, Brno, Czech Republic, 2009.
- [5] "AP17-OLR challenge results," accessed:

Digital Signal Processing & Speech Technology @ EE.CUHK



[1] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised domain adaptation for i-vector [2] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc.*

[3] Z. Tang, D. Wang, Y. Chen, and Q. Chen, "AP17-OLR challenge: Data, plan, and baseline," in Proc. APSIPA, 2017,

[4] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Brno University of Technol-

Available: 2018-10-22. [Online]. http://cslt.riit.tsinghua.edu.cn/mediawiki/index.php/OLRChallenge2017