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)

Task description

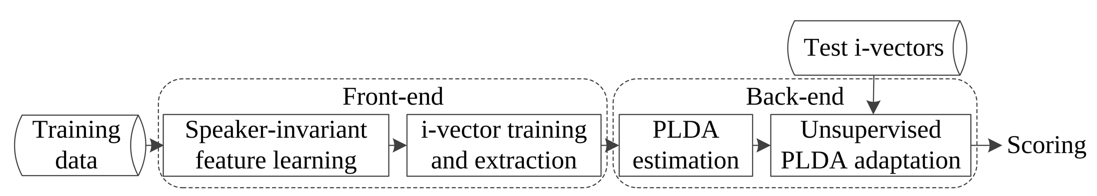
- Language recognition on very short (1s) test utterances.
- Severe domain mismatch (esp. recording conditions) between training and test utterances.

Motivation & Contribution

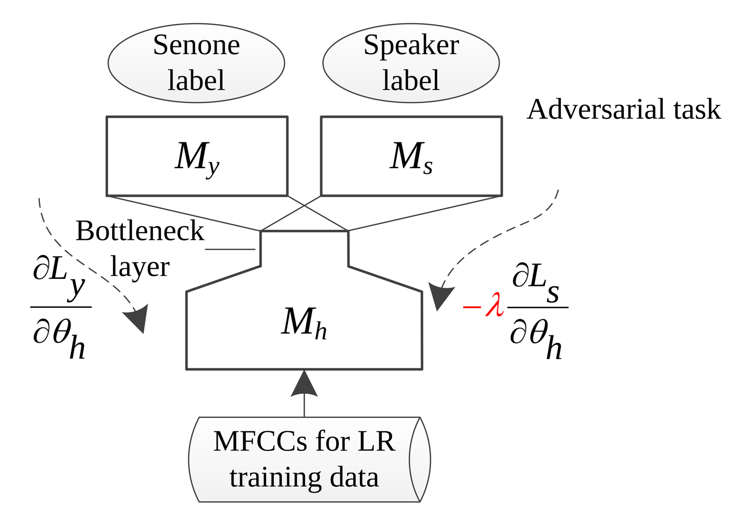
- **Front-end: Speaker adversarial multi-task learning (AMTL)**
  - Phonetic bottleneck features (BNFs) outperform spectral features in i-vector training.
  - Speaker variation is implicitly suppressed by phonetic BNF learning.
  - Speaker AMTL aims explicitly at speaker-invariant BNF learning.
- **Back-end: 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].

Model Structures

- **General framework:**



- **Speaker-invariant feature learning:**



- During training, parameters of M_y, M_s and M_h, denoted as θ_y, θ_s and θ_h, are updated as,

$$\theta_y \leftarrow \theta_y - \delta \frac{\partial \mathcal{L}_y}{\partial \theta_y}, \tag{1}$$

$$\theta_s \leftarrow \theta_s - \delta \frac{\partial \mathcal{L}_s}{\partial \theta_s}, \tag{2}$$

$$\theta_h \leftarrow \theta_h - \delta \Big[\frac{\partial \mathcal{L}_y}{\partial \theta_h} - \lambda \frac{\partial \mathcal{L}_s}{\partial \theta_h}\Big], \tag{3}$$

where δ is the learning rate, $\mathcal{L}_y$ 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 phonetically-discriminative.

Senone labels:

- Generated by an out-of-domain (OOD) phone recognizer.
- Language-independent senone labels.
- To control the output layer size of M_y.

- **GMM-UBM/i-vector training:**
  - Input features are BNFs extracted from speaker AMTL.

- **Back-end PLDA estimation:**

PLDA assumes an i-vector $\boldsymbol{\omega_{ij}}$ (j-th utterance in i-th language) generated as,

$$\begin{aligned} \boldsymbol{\omega_{ij}} &= \boldsymbol{\mu} + \mathbf{F}\mathbf{h_i} + \boldsymbol{\epsilon_{ij}}, \\ \mathbf{h_i} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ \boldsymbol{\epsilon_{ij}} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{4}$$

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 $\mathbf{F}$ 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 $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{F}\mathbf{F}^\intercal)$, where $\boldsymbol{\Sigma}$ and $\mathbf{F}\mathbf{F}^\intercal$ are within- and between-class variability. $\boldsymbol{\mu}$ 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* $(\boldsymbol{\omega_t}, i)$ composed of a test i-vector $\boldsymbol{\omega_t}$ and language i as,

$$\mathcal{R}(\boldsymbol{\omega_t}, i) = \log \frac{p(\overline{\boldsymbol{\omega^i}}, \boldsymbol{\omega_t} | \mathbf{F}\mathbf{F}^\intercal, \boldsymbol{\Sigma})}{p(\overline{\boldsymbol{\omega^i}} | \mathbf{F}\mathbf{F}^\intercal, \boldsymbol{\Sigma}) p(\boldsymbol{\omega_t} | \mathbf{F}\mathbf{F}^\intercal, \boldsymbol{\Sigma})}, \tag{5}$$

where $\overline{\boldsymbol{\omega^i}}$ 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 $\{\mathbf{F_0}, \boldsymbol{\Sigma_0}\}$ 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 $\boldsymbol{\eta_1}$ and $\boldsymbol{\eta_2}$ is defined based on $\{\mathbf{F_0}, \boldsymbol{\Sigma_0}\}$ as follows,

$$d(\boldsymbol{\eta_1}, \boldsymbol{\eta_2}) = -\log \frac{p(\boldsymbol{\eta_1}, \boldsymbol{\eta_2} | \mathbf{F_0}\mathbf{F_0}^\intercal, \boldsymbol{\Sigma_0})}{p(\boldsymbol{\eta_1} | \mathbf{F_0}\mathbf{F_0}^\intercal, \boldsymbol{\Sigma_0}) p(\boldsymbol{\eta_2} | \mathbf{F_0}\mathbf{F_0}^\intercal, \boldsymbol{\Sigma_0})}. \tag{6}$$

  - AHC with *complete-linkage criterion* is performed until a pre-defined cluster number is reached.
  - In-domain PLDA $\{\mathbf{F_{ad}}, \boldsymbol{\Sigma_{ad}}\}$ 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).

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

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+$\Delta + \Delta\Delta$ without CMVN.
    - ∗ Back-end: one-layer MLP with 512 neurons, followed by softmax output.
  - Results (C_{avg}/EER%):

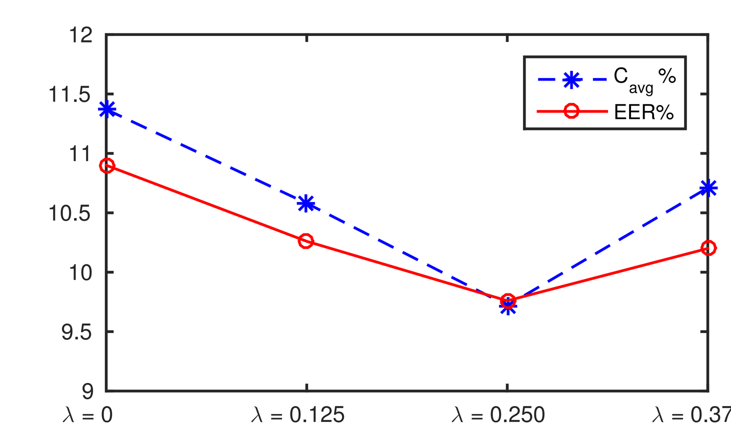
Training data	Pseudo-dev	Pseudo-dev_1s	Dev_1s
Training-part	3.50/3.97	7.78/9.56	13.42/13.18
Training-part_1s	—	7.61/8.94	14.01/13.88

Experimental Setup

- **Speaker-invariant BNFs:**
  - **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 neurons in BN 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.
- **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 ranging in $\{10, 50, 100, 200, 500\}$.

Results and Analysis

- Comparison of C_{avg}/EER% with **different adversarial weights** evaluated on **dev_1s** (back-end is simple cosine scoring)



- Comparison of C_{avg}/EER% with/without **unsupervised PLDA adaptation** evaluated on both **dev_1s** and **test_1s** sets (same front-end configuration, $\lambda = 0.250$)

	No Adaptation	With adaptation; cluster number in AHC					SOTA [5]
		10	50	100	200	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/A
Test_1s	9.46/8.78	—	—	**7.36/7.53**	—	—	7.65/7.91

Conclusions

- Speaker AMTL suppresses speaker variation, which is beneficial to the LR task.
- Unsupervised PLDA adaptation alleviates train-test domain mismatch and contributes significantly to performance improvement on short-duration LR task.
- Effectiveness of PLDA adaptation is insensitive to the number of clusters.

References

[1] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Proc. Odyssey*, 2014, pp. 260–264.

[2] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*, 2007, pp. 1–8.

[3] Z. Tang, D. Wang, Y. Chen, and Q. Chen, "AP17-OLR challenge: Data, plan, and baseline," in *Proc. APSIPA*, 2017, pp. 749–753.

[4] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Brno University of Technology, Brno, Czech Republic, 2009.

[5] "AP17-OLR challenge results," accessed: 2018-10-22. [Online]. Available: http://cslt.riit.tsinghua.edu.cn/mediawiki/index.php/OLRChallenge2017