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)
{jerroypeng1937, 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.
• Front-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:

  Training data → Speaker-invariant feature learning  → Front-end → Back-end → Test i-vectors  → Scoring

  Front-end: one-layer MLP with 512 neurons, followed by softmax output.
  Back-end: Language adaptive PLDA with 12 neurons in the bottleneck layer.

• Speaker-invariant feature learning:

  BNGS (BNF)  → Bottleneck layer  → Speaker labels

  AMTL: Back-end language-adaptation is learned as a cross-entropy loss.

• Back-end PLDA estimation:

  PLDA assumes an i-vector \( h_i \) (i-th utterance in i-th language) generated as,
  \[
  \omega_i \sim \mu_i + \mathbf{F}_i \mathbf{h}_i + \mathbf{e}_i,
  \]
  \[
  \mathbf{e}_i \sim N(0; \Sigma),
  \]
  \[
  \mathbf{F}_i \in \mathbb{R}^{D \times L}, \mu_* \in \mathbb{R}^D, \Sigma \in \mathbb{R}^{D \times D},
  \]
  where \( \mu_i \in \mathbb{R}^L, \mathbf{F}_i \in \mathbb{R}^{D \times L}, \Sigma \in \mathbb{R}^{D \times D} \).

  – Columns of \( \mathbf{F} \) provide the basis for the language-specific subspace, or eigen-language.
  – \( D \) is the subspace dimension, normally smaller than \( 16384 \) (flangues in this work).

  – Based on Eqn. (4), an i-vector is assumed drawn from \( N(\mu_i, \Sigma \mathbf{F}^T \mathbf{F} \Sigma) \), wherein and between-class variability \( \mu_i \) is global mean and can be precomputed and removed.

  – PLDA parameters \( (\mathbf{F}, \Sigma) \) are estimated by an EM algorithm [2].

  – During scoring phase, PLDA computes the similarity score of a trial \( (\omega, i) \) composed of a test i-vector \( \omega \) and language i as,
  \[
  
  \mathcal{N}(\omega, i) = \log \left( \frac{\mu_i^T \omega \mathbf{F}^T \Sigma \mathbf{F} \mu_i}{\sigma_i^2} \right)
  \]
  where \( \sigma_i^2 \) 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 \( (\mathbf{F}_0, \Sigma_0) \) 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 \( q_0 \) and \( q_2 \) is defined based on \( (\mathbf{F}_0, \Sigma_0) \) as follows,
  \[
  \delta(q_0, q_2) = -\log \left( \frac{\mu_0^T q_0 \mathbf{F}_0 \Sigma_0 \mathbf{F}_0 \Sigma_0 \mu_0}{\sigma_0^2} \right)
  \]
  – AHC with complete-linkage criterion is performed until a pre-defined cluster number is reached.

  – In-domain PLDA \( (\mathbf{F}_d, \Sigma_d) \) are estimated by test-i-vectors and their cluster labels.

  – Final scoring based on \( (\mathbf{F}_d, \Sigma_d) \).

APIT-OLR Task Description

• APIT-OLR challenge dataset [3]: 10 oriental languages, each with 10 hours recorded by mobile phones.

  – Training: 17, 549 utterances, 77 hours.
  – Test-1C: 17, 549 utterances, 7 hours.
  – Test-1S: 22, 051 utterances, 6 hours.

  – Evaluation metric: \( C_{\text{avg}} \) and Equal Error Rate (EER).

  \[
  C_{\text{avg}} = \frac{1}{L} \sum_{l=1}^{L} \left[ P_{\text{IS}}(l) - \frac{1}{N-1} \sum_{l \neq l} P_{\text{IS}}(l, l) \right],
  \]
  where \( N \) is the number of languages, \( L \) is the target and non-target languages, \( P_{\text{IS}} \) and \( P_{\text{SS}} \) are the missing and false alarm rates.

  – Measuring the mismatch between training and dev-\text{Is}:

    • Measuring the mismatch between training and dev-\text{Is}: A demo experiment is conducted to show the domain mismatch between training and development/test data.

      – Setup:
        • \text{Pseudo-dev:} a 12-hour subset randomly selected from training set.
        • \text{Training-part:} the remaining 57-hour subset from training set.
        • \text{Pseudo-dev and training-part:} utterances are trimmed to 1 second.
        • \text{Front-end:} 108-dim i-vectors extracted from 60-dim voiced MFCCs + \( \Delta + \Delta \Delta \) without CMVN.
        • \text{Back-end:} one-layer MLP with 512 neurons, followed by softmax output.
        • Results (3.6% / 11.8%)

Results and Analysis

• Comparison of \( C_{\text{avg}} \) / EER, with different adversarial weights evaluated on dev-\text{Is} (back-end is simple cosine scoring).

• Comparison of \( C_{\text{avg}} \), with and without unsupervised PLDA adaptation evaluated on both dev-\text{Is} and test-\text{Is} sets (same front-end configuration, \( \lambda = 0.250 \)).

Experimental Setup

• Speaker-invariant BNFs:

  – Input: 40-dim MFCCs w/o cepstral truncation.
  – Speaker labels: obtained from a Czech phone recognizer [4], 135 speakers in total.
  – Speaker labels: obtained from training data, 64 (1 speaker in total).

• DNN configuration: \( M_d \) is a 6-RELU layer TDNN, 2048 neurons per layer (64 neurons in BN layer), layer-wise context: \(-2,-1,0,1,2,0\), \(-1,2,-3,2,-1,0\).

• Unsupervised PLDA adaptation:

  – Out-of-domain PLDA: estimated on training i-vectors and ground-truth labels.
  – In-domain PLDA: estimated on dev-\text{Is} i-vectors and cluster labels.

• AHC: cluster dev-\text{Is} i-vectors to a pre-defined number of clusters ranging in \{10, 50, 100, 200, 500\}.

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

ology, Brno, Czech Republic, 2006.