Unsupervised Domain Adaptation for Gender-Aware PLDA Mixture Models

Longxin Li and Man-Wai MAK
Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University

Introduction
- PLDA is still problematic when (1) the model is deployed to a new environment (in-domain) that is very different from the training one (out-of-domain) and (2) there are insufficient labeled data from the new environment.
- This paper proposes using out-of-domain training data to pre-train a PLDA mixture model and applying the mixture model on the in-domain training data to compute a pairwise score matrix for spectral clustering. The hypothesized speaker labels produced by spectral clustering are then used for re-training the mixture model to fit the new environment.
- Experiments on NIST 2016 SRE demonstrate the effectiveness of the proposed framework compared with agglomerative hierarchical clustering (AHC).

Background
- DNN-driven mixture of PLDA (DNN-mPLDA):
  \[ p(x_i) = \sum_{k=1}^{K} \alpha_{ik} N(x_i | m_k, V_k) + \Sigma_k \]
- Target speaker’s utterance
- UBM(V)+Extractor
- DNN
- PLDA Mixture Model
- AHC

DNN-mPLDA.

Spectral Clustering of I-Vectors
- **Step 1** Compute a pairwise PLDA score matrix \( S \) from \( n \) training i-vectors:
  \[ s_{ij} = S_{mPLDA}(x_i, x_j), \quad L = \{1, \ldots, n\}. \]
- **Step 2** Convert \( S \) to an adjacency matrix \( A \) with elements:
  \[ a_{ij} = \begin{cases} \frac{1}{2} \frac{1}{\sigma_{max}} \frac{1}{\sigma_{ij}} & i \neq j \\ 1 & \text{otherwise} \end{cases} \]
  where \( \sigma_{max} \) is the absolute maximum in \( S \).
- **Step 3** Compute a Laplacian matrix:
  \[ L = I - D^{-\frac{1}{2}} AD^{-\frac{1}{2}} \]
  where \( D \) is a diagonal matrix with elements \( d_{ii} = \sum_j a_{ij} \).
- **Step 4** Pack \( K \) eigenvectors of \( L \) with the smallest eigenvalues to form \( V = [v_1 \ldots v_K] \in \mathbb{R}^{n \times K} \).
- **Step 5** Normalize the row of \( V \):
  \[ v_{ij} \leftarrow \frac{v_{ij}}{\sqrt{\sum_j v_{ij}}} \]
- **Step 6** Apply K-means to the \( n \) rows of \( V \).

Cluster Quality
- Silhouette values is used to quantify the quality of clusters. Each sample has a Silhouette value:
  \[ s(i) = \frac{h(i) - \bar{a}(i)}{\max\{a(i), b(i)\}} \]
  where \( a(i) \) is the average dissimilarity of sample \( i \) with respect to other samples in the same cluster and \( b(i) \) is the lowest average dissimilarity of sample \( i \) with respect to any other cluster not containing \( i \).
- **Results**
  - Highest average Silhouette score
  - Less negative Silhouette scores
  - So, Iterative-SC produces clusters with better quality

Results
- Performance of the iterative retaining method for different numbers of iterations on SRE16-dev and SRE16-eval
- Performance of PLDA mixture models on SRE16 using different speaker clustering methods and with and without covariance matrix interpolation (Cov. Interp.)

References: