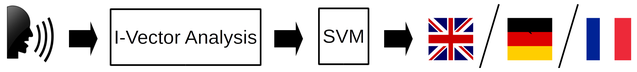
Modified prior i-Vector Estimation for Language Identification of Short Duration Utterances

Ruchir Travadi, Maarten Van Segbroeck, and Shrikanth Narayanan
Speech Analysis and Interpretation Laboratory (SAIL), Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089
travadi@usc.edu, maarten,shri@sipi.usc.edu

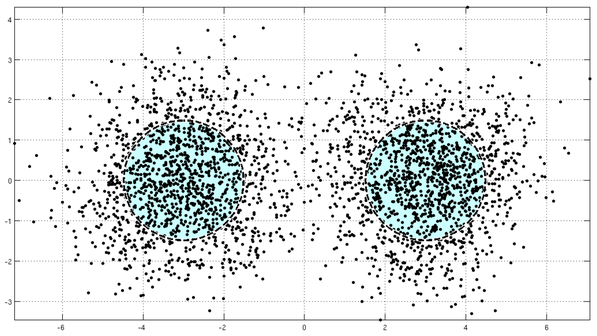
Introduction

- **Goal** : Identify spoken language from utterances
- **Challenge** : Short and noisy utterances
- **Framework** : Total Variability i-Vector Modeling + SVM

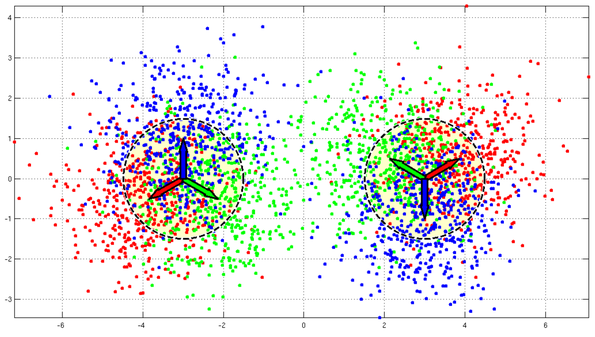


Total Variability i-Vector Modeling

- **Complete data distribution** :
  Gaussian Mixture Model, denoted as **UBM**



- **Utterance-specific data distribution** :
  GMM, with UBM component means shifted *slightly*



- **Model Assumption** :
  Shift in UBM mean supervector is *low-dimensional*

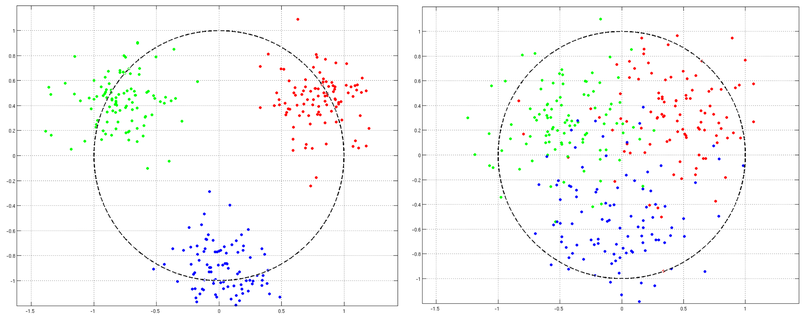
$$\mathbf{M_U} = \mathbf{M_0} + \mathbf{Tx}, \quad P(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- **i-Vector** is the **MAP estimate** of $\mathbf{x}$ given statistics $\mathbf{F}, n$:

$$E\left[\mathbf{x}|\boldsymbol{F}\right] = \left(\mathbf{I} + \mathbf{T}^t\boldsymbol{\Sigma}^{-1}n\mathbf{T}\right)^{-1}\mathbf{T}^t\boldsymbol{\Sigma}^{-1}n\boldsymbol{F}$$

Effect of reducing duration

- **Variance** of the i-Vector estimate increases
- Decisions become error-prone
- Estimate is driven by the **nature of the prior**



Prior Modification

- **Motivation and hypothesis** :
  - **Observation** :
    i-Vectors estimated from ample data form clusters
  - **Standard normal prior** :
    Penalizes probability of *large magnitude* i-Vector estimates
  - **Hypothesis** :
    Better to penalize deviation from cluster centers
  - **Gaussian Mixture** prior is better suited to this purpose

- **Proposed Prior** : GMM, with one component per class

$$P(\mathbf{x}) = \sum_{i=1}^{M} P_C(i)\mathcal{N}\left(\mu_i, \mathbf{C}_i\right)$$

- **i-Vector estimate with GMM Prior** :

$$E\left[\mathbf{x}|\mathbf{F}\right] = \sum_{i=1}^{M} P_C(i)\, \boldsymbol{l}_i^{-1}\mathbf{b}_i$$

$$\mathbf{b}_i = \mathbf{T}^t\boldsymbol{\Sigma}^{-1}n\mathbf{F} + \mathbf{C}_i^{-1}\mu_i, \;\; \boldsymbol{l_i} = \mathbf{C}_i^{-1} + \mathbf{T}^t\boldsymbol{\Sigma}^{-1}n\mathbf{T}$$

- **Prior re-weighting** :
  Provide a parameter(λ) to tune weight of prior relative to data :

$$\mathbf{b}_i = \lambda\mathbf{T}^t\boldsymbol{\Sigma}^{-1}n\mathbf{F} + \mathbf{C}_i^{-1}\mu_i, \;\; \boldsymbol{l_i} = \mathbf{C}_i^{-1} + \lambda\mathbf{T}^t\boldsymbol{\Sigma}^{-1}n\mathbf{T}$$

Proposed Variants

- **GMM i-Vector** : Use i-Vector estimates from long duration utterances to set parameters $P_C(i)$ and $\mu_\mathbf{i}, \mathbf{C}_i$ for the prior.
- **GMM Re-estimation** : Use i-Vector estimates from long duration utterances only to set $\mu_\mathbf{i}, \mathbf{C}_i$ for the prior. Set $P_C(i)$ by first-pass SVM classification.
- **Score Re-estimation** :
  - i-Vector estimates under GMM prior are linear combinations of i-Vector estimates under individual component Gaussian priors
  - Try combining scores after classification as opposed to combining i-Vector estimates before classification

Database and System Description

- **Database : DARPA RATS**
  - Noisy audio recordings from **six** classes :
    - Five **target** languages
    - A class corresponding to 10 **non-target** languages
  - Audio utterances of length : 120s, 30s, 10s, 3s
- **System Description**
  - **UBM Size** : 2048 Components
  - **i-Vector dimension** : 400
  - **Inter-session variability compensation** : WCCN
  - **SVM** : Fifth order polynomical kernel
  - **Utterance duration** :
    - **i-Vector Model $(\mathbf{T}, \boldsymbol{\Sigma})$ Estimation** : 30 s
    - **SVM training and Test** : 3 s

Results

System	*EER*	*DCF*	P_{miss}^{10}	*Accuracy*
Baseline	15.40	15.21	22.19	69.74
GMM i-Vector	15.27	**14.98**	**20.76**	69.44
GMM Re-estimation	16.32	15.82	22.32	**70.11**
Score Re-estimation	**15.14**	15.07	21.28	69.96

Acknowledgment

This research was supported by the Defense Advanced Research Projects Agency (DARPA) and the National Science Foundation (NSF).