Likelihood-Ratio Empirical Kernels for I-Vector Based PLDA-SVM Scoring

Man-Wai MAK and Wei RAO

Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University

Summary

- Likelihood ratio (LR) scoring in PLDA speaker verification systems only uses the information of background speakers **implicitly**. This paper exploits the notion of empirical kernel maps to incorporate background speaker information into the LR scoring process **explicitly**.
- We train a scoring SVM for each target speaker based on empirical kernels.
- We demonstrate that a number of target-speaker i-vectors can be generated by an utterance partitioning and resampling technique, resulting in much better scoring SVMs.
- Results suggest that incorporating background speaker information into PLDA scoring through speaker-dependent SVMs can boost the performance of i-vector based PLDA systems significantly.

Methods

Likelihood Ratio Score for Gaussian PLDA

Given a set of D-dim length-normalized i-vectors:

$$X=\{x_{ij};i=1,\ldots,N,\ j=1,\ldots,H_i\},$$

we aim to estimate the latent variables $Z=\{z_i;i=1,\ldots,N\}$ and parameters $\omega=\{\mu,W,\Sigma\}$ of a factor analyzer:

$$x_{ij}=\mu+Wz_i+\varepsilon_{ij}$$

Given a test i-vector x_t and target-speaker's i-vector x_s, the likelihood ratio (LR) score is:

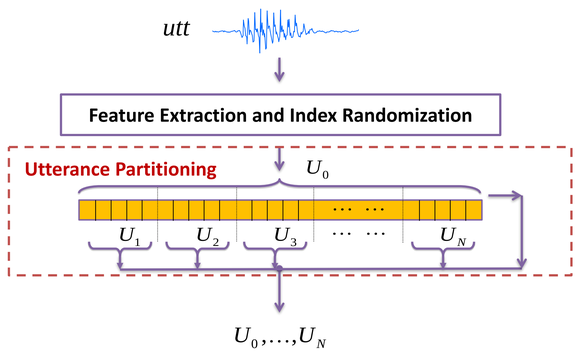
$$S_{\text{LR}}(x_s,x_t)=\text{const}+x_s^{\text{T}}Qx_s+x_t^{\text{T}}Qx_t+2x_s^{\text{T}}Px_t$$

where

$$P=\Lambda\Gamma\left(\Lambda-\Gamma\Lambda^{-1}\Gamma\right)^{-1};\ \Lambda=WW^{\text{T}}+\Sigma;\ Q=\Lambda^{-1}-\left(\Lambda-\Gamma\Lambda^{-1}\Gamma\right)^{-1};\ \Gamma=WW^{\text{T}}$$

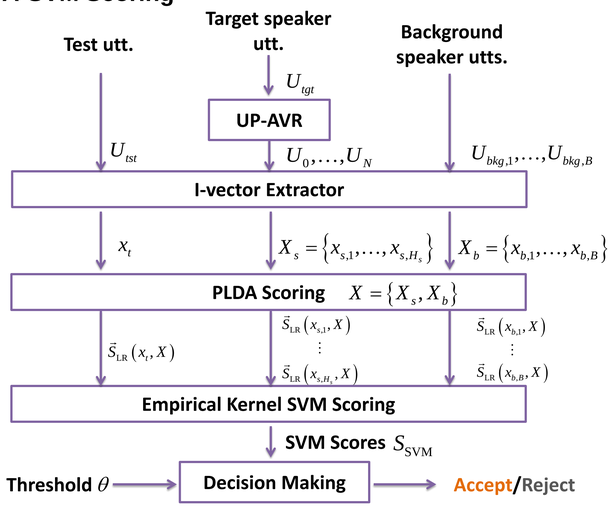
Background information

UP-AVR



Methods

PLDA-SVM Scoring



$$S_{\text{SVM}}(x_t,X_s,X_b)=\sum_{j\in\text{SV}_s}\alpha_{s,j}K(x_t,x_{s,j})-\sum_{j\in\text{SV}_b}\alpha_{s,j}K(x_t,x_{b,j})+d_s$$

Empirical LR Kernel I:

$$K(x_t,x_{s,j})=\mathbb{K}\left(\vec{S}_{\text{LR}}(x_t,X_s),\vec{S}_{\text{LR}}(x_{s,j},X_s)\right)$$

where

$$X_s=\{x_{s,1},\ldots,x_{s,H_s}\};\ \vec{S}_{\text{LR}}(x_t,X_s)=\begin{bmatrix}S_{\text{LR}}(x_t,x_{s,1})\\S_{\text{LR}}(x_t,x_{s,2})\\\vdots\\S_{\text{LR}}(x_t,x_{s,H_s})\end{bmatrix}$$

Empirical LR Kernel II:

$$K(x_t,x_{s,j})=\mathbb{K}\left(\vec{S}_{\text{LR}}(x_t,X),\vec{S}_{\text{LR}}(x_{s,j},X)\right)$$

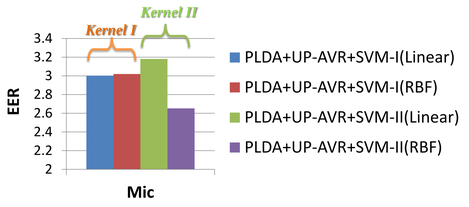
where

$$X=\{X_s,X_{b'}\};\qquad \vec{S}_{\text{LR}}(x_t,X)=\begin{bmatrix}S_{\text{LR}}(x_t,x_{s,1})\\\vdots\\S_{\text{LR}}(x_t,x_{s,H_s})\\S_{\text{LR}}(x_t,x_{b,1})\\\vdots\\S_{\text{LR}}(x_t,x_{b,B'})\end{bmatrix}$$

$$X_{b'}=\{x_{b,1},\ldots,x_{b,B'}\};$$

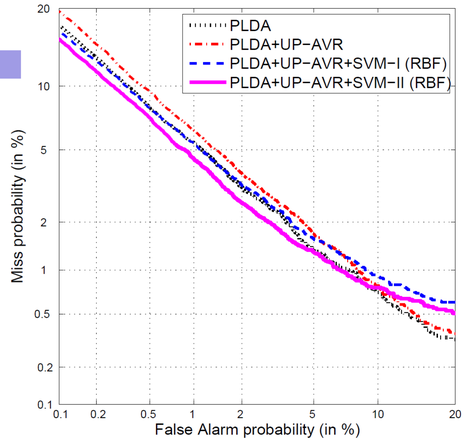
$$B'\le B;$$

Results

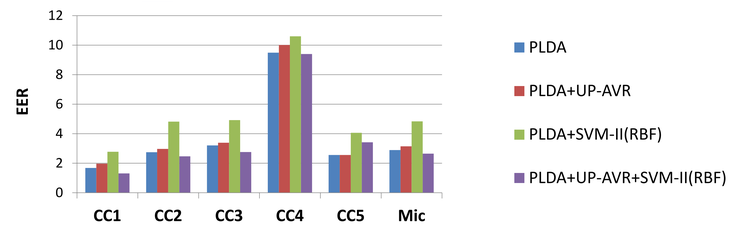


- *Linear kernel performs better in the **empirical LR kernel I**.*
- *Non-linear kernel (RBF) performs better in the **empirical LR kernel II**.*

PLDA+UP-AVR+SVM-II(RBF) significantly outperforms PLDA LR scoring.



- *UP-AVR not only helps to alleviate the data-imbalance problem in SVM training, but also enriches the information content of the scoring vectors by increasing the number of LR scores derived from the target speaker.*
- *UP-AVR is not beneficial to LR scoring.*



References

[1] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. of Odyssey: Speaker and Language RecognitionWorkshop*, *Brno, Czech Republic*, June 2010.

[2] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech'2011*, pp. 249–252, 2011.

[3] W. Rao and M. W. Mak, "Boosting the performance of i-vector based speaker verification via utterance partitioning," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 1012–1022, 2013.