Discriminatively Trained Probabilistic Linear Discriminant Analysis for Speaker Verification

Lukáš Burget[1], Oldřich Plchot[1], Sandro Cumani[2], Ondřej Glembek[1], Pavel Matějka[1], Niko Bümmer[3]
[1]Brno University of Technology, Czech Republic, {burget,iplchot,glembek,matejkap}@fit.vutbr.cz,
[2]Politecnico di Torino, Italy, sandro.cumani@polito.it, [3]AGNITIO, South Africa, niko.brummer@gmail.com

Discriminative training was successfully implemented optimizing the true objective of the speaker verification task: discrimination between same-speaker and different-speaker trials.

- Our baseline is state-of-the-art system based on iVector + PLDA paradigm
- PLDA parameters are re-trained discriminatively.
- Cross-entropy or hinge loss is optimized for binary classifier addressing the true objective of the task: same- vs. different-speaker trial classification.
- This is the first time such "true" discriminative training was successfully applied to speaker verification.

Previous work on discriminative training in SRE

•**SVM based systems** (e.g. GMM-SVM)
- •Discriminatively trained model for each enrollment speaker ➔ very limited number of positive examples (usually only one)
- •Does not address the "true" speaker verification objective

•**Discriminative training of JFA hyper-parameters**
- •Preliminary work done and JHU 08 summer workshop
- •Very limited gains (too many parameters to train, gains canceled by score normalization that is necessary in the case of JFA)

•**Discriminative score fusion**
- •Only score fusion weights are trained discriminatively

iVector + PLDA Baseline

•**IVector extractor** – model similar to JFA, where GMM mean supervector

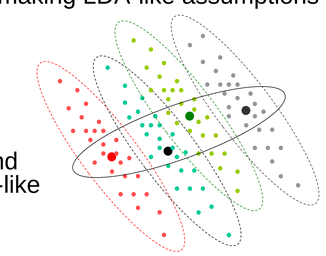
$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\mathbf{i}$$

is constrained to live in single subspace **T** spanning both speaker and channel variability ➔ no need for speaker labels to train **T**

•**iVector** – point estimate of **i** adapting GMM to a segment
- •extracted for every recording as its low-dimensional, fixed-length representation (typically 400 dimensions)
- •contains information about both speaker and channel

•**Probabilistic Linear Discriminant Analysis (PLDA)**
- •Simple generative model is used to model distributoion of iVectors
- •We consider only simple variant of PLDA, making LDA-like assumptions

$$p(\mathbf{r}) = \mathcal{N}(\mathbf{r}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_{ac})$$
$$p(\mathbf{i}|\mathbf{r}) = \mathcal{N}(\mathbf{i}|\mathbf{r}, \boldsymbol{\Sigma}_{wc})$$



- •Note that the original formulation uses subspaces **V** and **U** to describe speaker and channel variability ➔ single Gaussian JFA-like model:

$$\mathbf{i} = \boldsymbol{\mu} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\Sigma}_{ac} = \mathbf{V}\mathbf{V}^T \qquad \boldsymbol{\Sigma}_{wc} = \mathbf{U}\mathbf{U}^T + \mathrm{cov}(\boldsymbol{\epsilon})$$

Evaluation of verification score

•**Bayesian model comparison:**
- •For trial represented by pair of iVectors $\mathbf{i}_1$ and $\mathbf{i}_2$, compare likelihoods for two hypothesis:
  - •H_s – both recordings come from the same speaker
  - •H_d – recordings come from different speakers

•I.e. log-likelihood ratio verification score is:

$$s = \log \frac{p(\mathbf{i}_1, \mathbf{i}_2|\mathcal{H}_s)}{p(\mathbf{i}_1, \mathbf{i}_2|\mathcal{H}_d)} = \log \frac{\int p(\mathbf{i}_1|\mathbf{r})p(\mathbf{i}_2|\mathbf{r})p(\mathbf{r})\mathrm{d}\mathbf{r}}{\int p(\mathbf{i}_1|\mathbf{r})p(\mathbf{r})\mathrm{d}\mathbf{r} \int p(\mathbf{i}_2|\mathbf{r})p(\mathbf{r})\mathrm{d}\mathbf{r}}$$

- •Note the symmetrical role of both recordings, which is in contrast to training speaker model on one recordings and evaluating it on the other one.
- •For PLDA, the log-likelihood ratio formula has simple analytical solution:

$$s = \log \mathcal{N}\left(\begin{bmatrix}\mathbf{i}_1\\ \mathbf{i}_2\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{tot} & \boldsymbol{\Sigma}_{ac}\\ \boldsymbol{\Sigma}_{ac} & \boldsymbol{\Sigma}_{tot}\end{bmatrix}\right) - \log \mathcal{N}\left(\begin{bmatrix}\mathbf{i}_1\\ \mathbf{i}_2\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{tot} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{tot}\end{bmatrix}\right)$$

and after some manipulation we obtain formula allowing for extremely fast evaluation of the score:

$$s = \mathbf{i}_1^T \boldsymbol{\Lambda} \mathbf{i}_2 + \mathbf{i}_1^T \boldsymbol{\Gamma} \mathbf{i}_1 + \mathbf{i}_2^T \boldsymbol{\Gamma} \mathbf{i}_2 + (\mathbf{i}_1 + \mathbf{i}_2)^T \mathbf{c} + k,$$

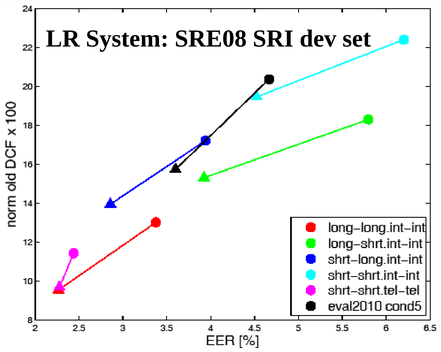
where $\boldsymbol{\Lambda}$, $\boldsymbol{\Gamma}$, $\mathbf{c}$ and k are parameters derived from PLDA parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}_{ac}$ and $\boldsymbol{\Sigma}_{wc}$ (see the paper for more details).

•**Linear classifier:**
- •Using $\mathbf{x}^T\mathbf{A}\mathbf{y} = \mathrm{vec}(\mathbf{A})^T \mathrm{vec}(\mathbf{y}\mathbf{x}^T)$, we can express the score as

$$s = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{i}_1, \mathbf{i}_2) = \begin{bmatrix}\mathrm{vec}(\boldsymbol{\Lambda})\\ \mathrm{vec}(\boldsymbol{\Gamma})\\ \mathbf{c}\\ k\end{bmatrix}^T \begin{bmatrix}\mathrm{vec}(\mathbf{i}_1\mathbf{i}_2^T + \mathbf{i}_2\mathbf{i}_1^T)\\ \mathrm{vec}(\mathbf{i}_1\mathbf{i}_1^T + \mathbf{i}_2\mathbf{i}_2^T)\\ \mathbf{i}_1 + \mathbf{i}_2\\ 1\end{bmatrix}$$

i.e. linear classifier represented by weights $\mathbf{w}$ applied to nonlinear expansion of iVector pair $\boldsymbol{\varphi}(\mathbf{i}_1,\mathbf{i}_2)$

•We will train weights $\mathbf{w}$ discriminatively as logistic regression or SVM .



Discriminative training

•Training examples are trials - different-and same-speaker iVector pairs
•Labels $t \in \{-1, 1\}$ correspond to different-, and same-speaker trials.
•Score s is log-likelihood ratio ➔ log probability of correctly classifying trial

$$\log p(t|\mathbf{i}_1, \mathbf{i}_2) = -\log(1 + \exp(-st))$$

(for simplicity, we assume equal priors for both hypothesis H_s and H_d)

•**Logistic regression** maximizes (log) probability of classifying all training examples correctly (i.e. sum of the terms above over all training examples):

$$E(\mathbf{w}) = \sum_{n=1}^{N} \alpha_n E_{LR}(t_n s_n) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$
$$E_{LR}(ts) = \log(1 + \exp(-ts))$$

(proportion of target and non-target trials can be balanced by weight α_n)

•Alternatively, **SVM** objective is obtained by replacing logistic regression loss E_{LR} with hinge loss $E_{SV}(ts) = \max(0, 1 - ts)$

Efficient gradient evaluation

•Our training set (Switchboard and NIST SRE data) comprises 20k female and 16k male recordings ➔ we create almost a billion training examples (trials) from all possible pairs of training recordings

•Fortunately, the gradient (and similarly Hessian) necessary for the optimization can be evaluated very efficiently

$$\nabla E(\mathbf{w}) = \begin{bmatrix}\nabla_\Lambda L\\ \nabla_\Gamma L\\ \nabla_c L\\ \nabla_k L\end{bmatrix} = \begin{bmatrix}2 \cdot \mathrm{vec}\left(\boldsymbol{\Phi}\mathbf{G}\boldsymbol{\Phi}^T\right)\\ 2 \cdot \mathrm{vec}\left(\boldsymbol{\Phi}[\boldsymbol{\Phi}^T \circ (\mathbf{G}\mathbf{1}\mathbf{1}^T)]\right)\\ 2 \cdot \mathbf{1}^T[\boldsymbol{\Phi}^T \circ (\mathbf{G}\mathbf{1}\mathbf{1}^T)]\\ \mathbf{1}^T\mathbf{G}\mathbf{1}\end{bmatrix} + \lambda\mathbf{w}$$

where $\boldsymbol{\Phi}$ is matrix of all training iVectors and $\mathbf{G}_{ij} = \alpha_{ij}\frac{\partial E(t_{ij}s_{ij})}{\partial s_{ij}}$
i.e $\mathbf{G}_{ij} = \alpha_{ij}t_{ij}\sigma(-t_{ij}s_{ij})$ for E_{LR}.

Results and Conclusions

•Gains across conditions obtained with both logistic regression and SVM
•Gains from discriminative training are comparable to Kenny's Heavy Tailed PLDA, which is much slower to evaluate
•Recently, however, similar improvements were obtained with "ad-hoc" modifications to standard iVector+PLDA approach (e.g. iVector length norm.)
•Currently, we focus on discriminative training of earlier stages such as iVector extraction.

	Female Set			Male Set			Pooled		
System	minDCF	oldDCF	EER	minDCF	oldDCF	EER	minDCF	oldDCF	EER
PLDA	0.40	0.15	3.57	0.42	0.13	2.86	0.41	0.14	3.23
LR	0.40	0.12	2.94	0.39	0.10	2.22	0.40	0.11	2.62
SVM	0.39	0.11	2.35	0.31	0.08	1.55	0.37	0.10	1.94
HT-PLDA	0.34	0.11	2.22	0.33	0.08	1.47	0.34	0.10	1.88

NIST SRE 2010, tel-tel condition (DET5)