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

- Our baseline is state-of-the-art system based on iVector + PLDA paradigm.
- PLDA parameters are re-discriminatively trained.
- 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
  $\mu = \mathbf{m} + \mathbf{T}$
  is constrained to live in single subspace $\mathbf{T}$ spanning both speaker and channel variability → no need for speaker labels to train $\mathbf{T}$.
- iVector – point estimate of $\mathbf{l}$ 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 distribution of iVectors.
  - We consider only simple variant of PLDA, making LDA-like assumptions.

Evaluation of verification score

- Bayesian model comparison:
  - For trial represented by pair of iVectors $i_1$ and $i_2$, 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:
    \[
    s = \log \frac{p(i_1 | i_2, H_s)}{p(i_1, i_2 | H_d)} = \log \frac{p(i_1 | i_2, H_s)}{p(i_1, i_2 | H_d)}
    \]
  - 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:
    \[
    s = \mathbf{w}^T \mathbf{vec}(\mathbf{A}) \mathbf{vec}(\mathbf{y} \mathbf{x}^T),
    \]
    i.e. linear classifier represented by weights $\mathbf{w}$ applied to nonlinear expansion of iVector pair $\varphi(i_1, i_2)$.
- Linear classifier:
  - Using $\mathbf{x}^T \mathbf{A}^T \mathbf{y} = \mathbf{vec}(\mathbf{A})^T \mathbf{vec}(\mathbf{y} \mathbf{x}^T)$, we can express the score as
    \[
    s = \mathbf{w}^T \varphi(i_1, i_2) = \begin{bmatrix} \mathbf{vec}(\mathbf{A})^T \mathbf{vec}(\mathbf{y} \mathbf{x}^T) \\ \mathbf{vec}(\mathbf{i}_1^T \mathbf{1} + \mathbf{i}_2^T \mathbf{1}) \\ \mathbf{vec}(\mathbf{i}_1^T \mathbf{1} + \mathbf{i}_2^T \mathbf{1}) \\ \mathbf{i}_1^T \mathbf{1} + \mathbf{i}_2^T \mathbf{1} \end{bmatrix}
    \]

Discriminative training

- Training examples are trials - different-and same-speaker iVector pairs.
- Labels $u \in \{-1, 1\}$ correspond to different- and same-speaker trials.
- Score $s$ is log-likelihood ratio $\rightarrow$ log probability of correctly classifying trial
  \[
  \log p(i_1, i_2) = -\log(1 + \exp(-st))
  \]
  (for simplicity, we assume equal priors for both hypothesis $H_s$ and $H_d$).
- Logistic regression maximizes (log) probability of correctly classifying all training examples correctly (i.e. sum of the terms above over all training examples):
  \[
  E(w) = \sum_{n=1}^{N} \alpha_n E_{LR}(t_n, s_n) + \frac{\lambda}{2} \left\| w \right\|^2
  \]
  (proportion of target and non-target trials can be balanced by weight $\alpha_n$).
- Alternatively, SVM objective is obtained by replacing logistic regression loss $E_{LR}$ with hinge loss
  \[
  E_{SV}(t) = \max(0, 1 - ts)
  \]

Efficient gradient evaluation

- Our training set (Switchboard and NIST SRE data) comprises 20k female and 16k male recordings $\rightarrow$ 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.

\[
\nabla E(w) = \begin{bmatrix} \nabla E_{LR} \\ \nabla L \\ \nabla L \\ \nabla L \end{bmatrix} = \begin{bmatrix} 2 \cdot \nabla \Phi G \Phi^T (\mathbf{G} \Phi^T \mathbf{G} + \lambda \mathbf{I}) \end{bmatrix} + \lambda \mathbf{w}
\]

where $\Phi$ is matrix of all training iVectors and $\mathbf{G}_{ij} = \alpha_{ij} \delta_{ij} (\mathbf{1}_N^T \mathbf{1}_N) / \delta_{ij}$ 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.

NIST SRE 2010, tel-tele condition (DET5)