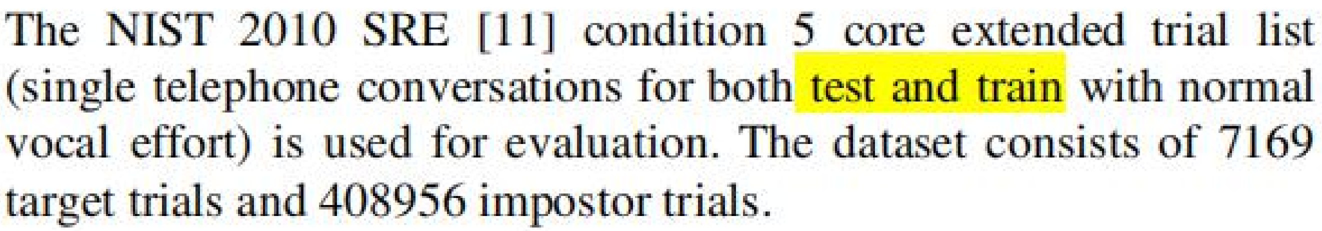
IDVC
(Inter dataset variability compensation)

Data

- The SWB dataset —— mismatched development dataset.
- The MIXER dataset —— matched development dataset.
- The NIST-2010 dataset —— evaluation dataset (train and test).



The NIST 2010 SRE [11] condition 5 core extended trial list (single telephone conversations for both test and train with normal vocal effort) is used for evaluation. The dataset consists of 7169 target trials and 408956 impostor trials.

Motivation

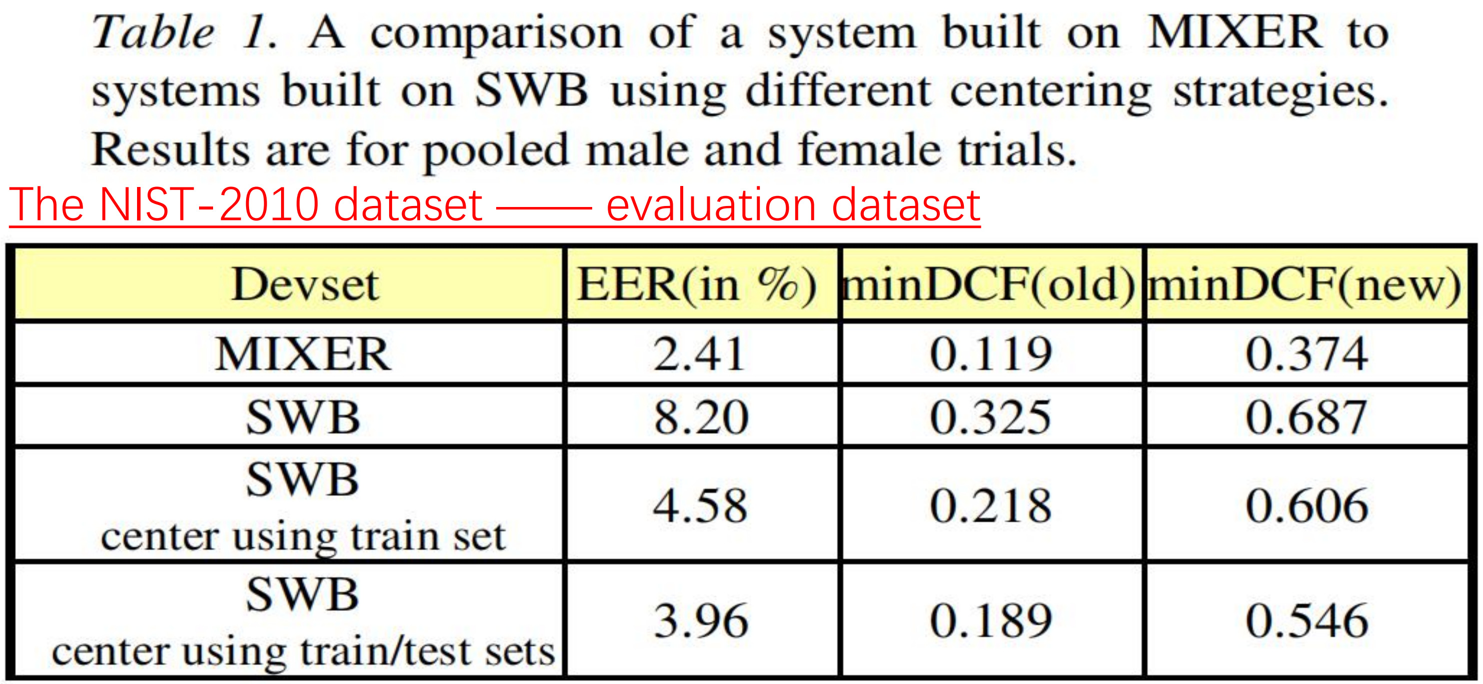


Table 1. A comparison of a system built on MIXER to systems built on SWB using different centering strategies. Results are for pooled male and female trials.

The NIST-2010 dataset —— evaluation dataset

Devset	EER(in %)	minDCF(old)	minDCF(new)
MIXER	2.41	0.119	0.374
SWB	8.20	0.325	0.687
SWB center using train set	4.58	0.218	0.606
SWB center using train/test sets	3.96	0.189	0.546

- Findings :
  - EER is cut by 50% by just doing proper centering motivates IDVC approach.

Why we need IDVC ?

- Background :
  - Many times is a dismatch between the development data and the evaluation data.
  - PLDA framework do not optimally cope with dataset shift.

- Ideal:
  - Modeling dataset variation in the i-vector space.
  - Compensating it as a pre-processing cleanup step.

- So we need IDVC!

How to implement IDVC ?

1. The development data set (such as SWB) is divided into subsets.

2. For each subset all i-vectors are averaged.

3. Use PCA to find a basis for subspace spanned by the centers(12).

4. The subspace is removed from the development and evaluation data as a pre-processing stage.

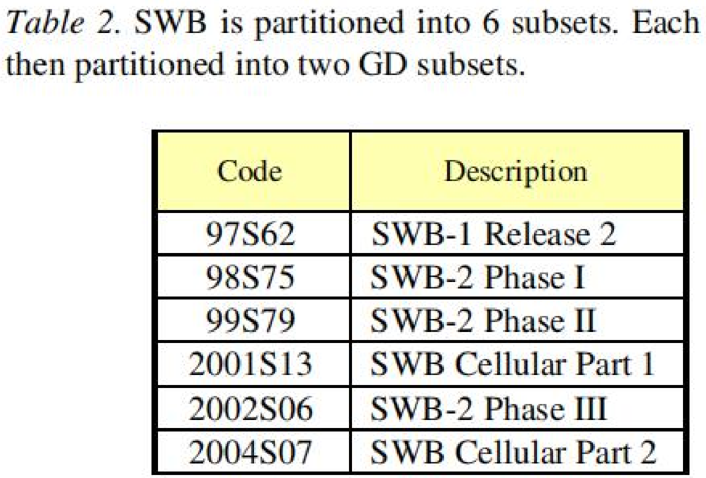


Table 2. SWB is partitioned into 6 subsets. Each subset is then partitioned into two GD subsets.

Code	Description
97S62	SWB-1 Release 2
98S75	SWB-2 Phase I
99S79	SWB-2 Phase II
2001S13	SWB Cellular Part 1
2002S06	SWB-2 Phase III
2004S07	SWB Cellular Part 2

Results

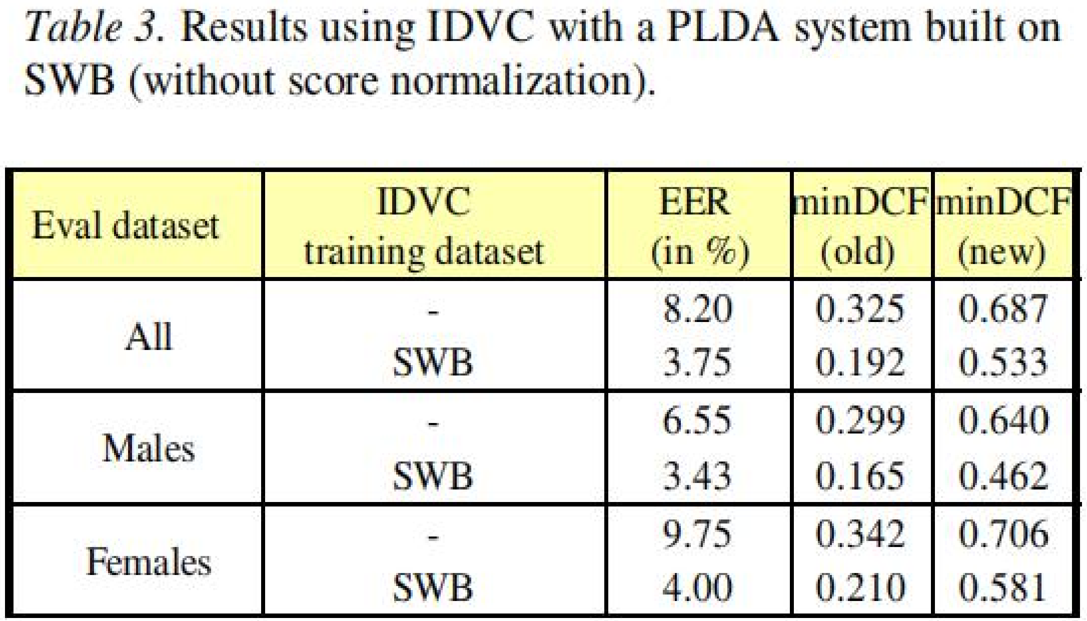


Table 3. Results using IDVC with a PLDA system built on SWB (without score normalization).

Eval dataset	IDVC training dataset	EER (in %)	minDCF (old)	minDCF (new)
All	-	8.20	0.325	0.687
	SWB	3.75	0.192	0.533
Males	-	6.55	0.299	0.640
	SWB	3.43	0.165	0.462
Females	-	9.75	0.342	0.706
	SWB	4.00	0.210	0.581

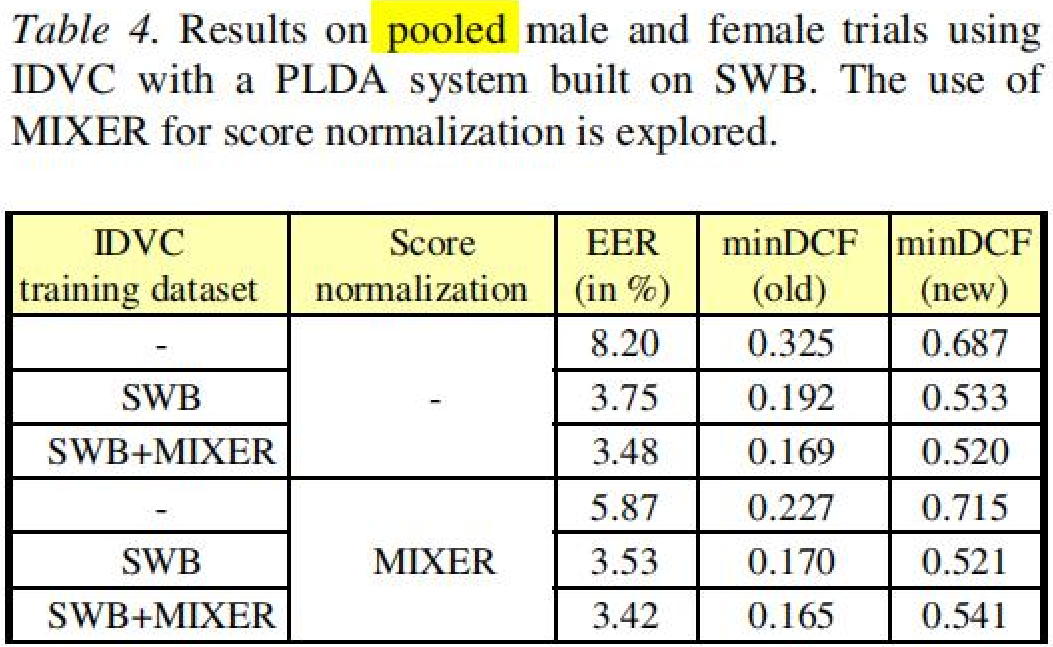


Table 4. Results on pooled male and female trials using IDVC with a PLDA system built on SWB. The use of MIXER for score normalization is explored.

IDVC training dataset	Score normalization	EER (in %)	minDCF (old)	minDCF (new)
-	-	8.20	0.325	0.687
SWB		3.75	0.192	0.533
SWB+MIXER		3.48	0.169	0.520
-	MIXER	5.87	0.227	0.715
SWB		3.53	0.170	0.521
SWB+MIXER		3.42	0.165	0.541

Use extensions

- Background :
  - To compensate inter-dataset variability attributed to additional PLDA hyper-parameters.

- Hypothesis :
  - Some directions in the i-vector space are more sensitive to dataset mismatch than other directions.

- Aim:
  - Finding and removing a low-dimensional subspace which is spanned by directions in i-vector space which are relatively sensitive to dataset mismatch.

How to implement IDVC？

- Inter-dataset variability subspace estimation ：

1. Partition the development dataset into *n* subsets.
2. Estimate PLDA hyper-parameters {μi , Bi , Wi} for each subset *i*.
3. Estimate i-vector subspace Sμ corresponding to the set {μi}
4. Estimate i-vector subspace Sw corresponding to the set {Wi}
5. Estimate i-vector subspace SB corresponding to the set {Bi}
6. Join subspaces to form a single subspace: $S = S\mu \cup Sw \cup SB$

How to implement IDVC？

- Estimating subspace Sw：

1. For a set of n covariance matrices {Wi} we denote the mean of the set by W.
2. Whiten the i-vector space.
3. Compute $\Omega = \frac{1}{n}\sum W_i^2$
4. Find the *k* largest eigenvalues of Ω. The corresponding eigenvectors span subspace Sw.

How to implement IDVC ?

- PLDA training :
    - Remove subspace S from the i-vectors of the development set.
    - Train PLDA using the standard scheme.

- PLDA scoring :
    - Remove subspace S from the i-vectors of the evaluation set.

Results

- Baseline:

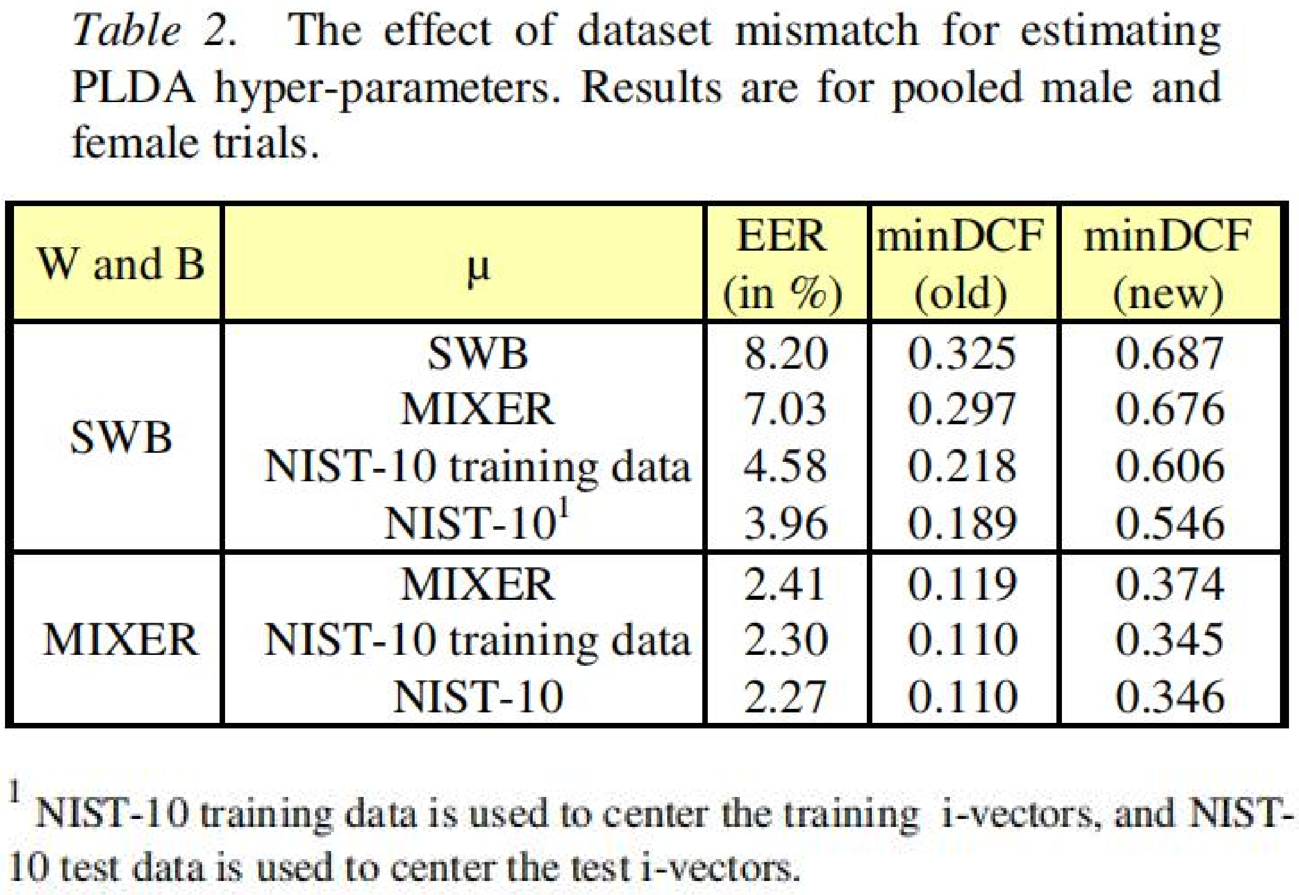
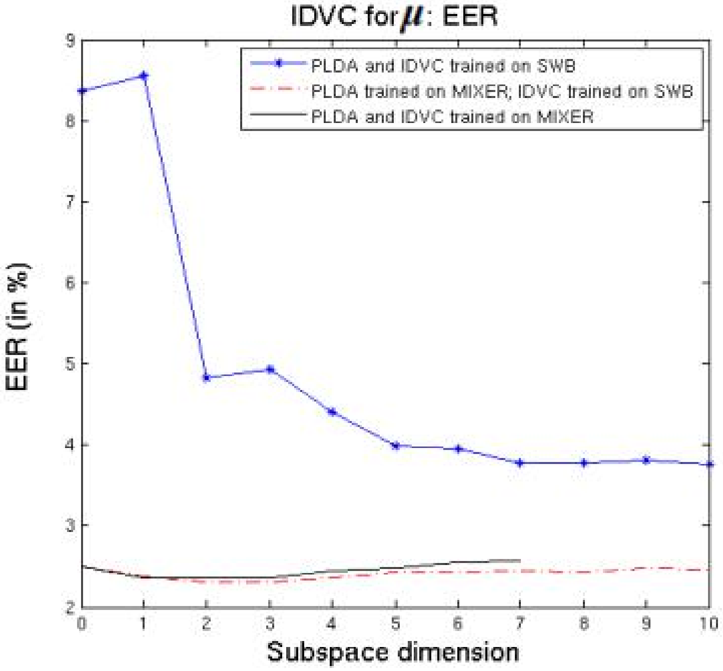


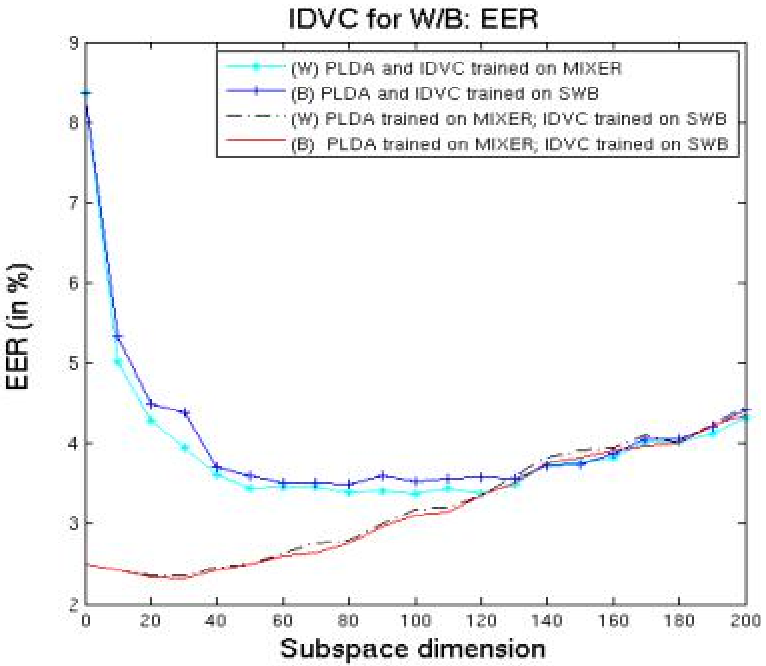
Table 2. The effect of dataset mismatch for estimating PLDA hyper-parameters. Results are for pooled male and female trials.

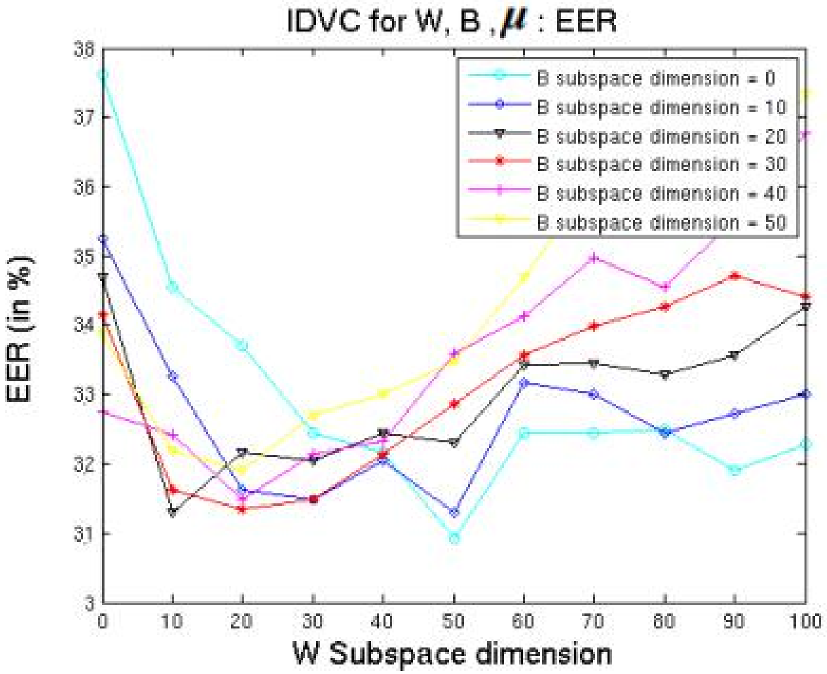
W and B	μ	EER (in %)	minDCF (old)	minDCF (new)
SWB	SWB	8.20	0.325	0.687
	MIXER	7.03	0.297	0.676
	NIST-10 training data	4.58	0.218	0.606
	NIST-10[1]	3.96	0.189	0.546
MIXER	MIXER	2.41	0.119	0.374
	NIST-10 training data	2.30	0.110	0.345
	NIST-10	2.27	0.110	0.346

[1] NIST-10 training data is used to center the training i-vectors, and NIST-10 test data is used to center the test i-vectors.

Results







The dimension for the μ subspace is 10.

Conclusion

- Efficient !
  - Dataset shift in the i-vector domain.
  - The variability in the PLDA hyper-parameters .