A Principle Solution for Enroll-Test Mismatch in Speaker Recognition

Lantian Li

What is enroll-test mismatch ?

- Two phases of speaker recognition system
  - Enroll (modelling)
  - Test (verification)

- Some typical scenarios of E-T mismatch
  - Enroll on one device, while test on another device.
  - Enroll in one near field, while test in another field.
  - Enroll in one time, while test in a few days later.

Enroll-test mismatch problem

Enroll-Test	Baseline
AN-AN	0.797
AN-Mic	2.146
AN-iOS	1.425
Mic-AN	2.175
Mic-Mic	0.778
Mic-iOS	2.251
iOS-AN	1.599
iOS-Mic	2.216
iOS-iOS	0.920

Cross-channel scenarios

Enroll-Test	Baseline
1m-1m	0.620
1m-3m	3.968
1m-5m	4.866
3m-1m	1.938
3m-3m	0.891
3m-5m	3.244
5m-1m	3.566
5m-3m	2.834
5m-5m	1.135

Near-far scenarios

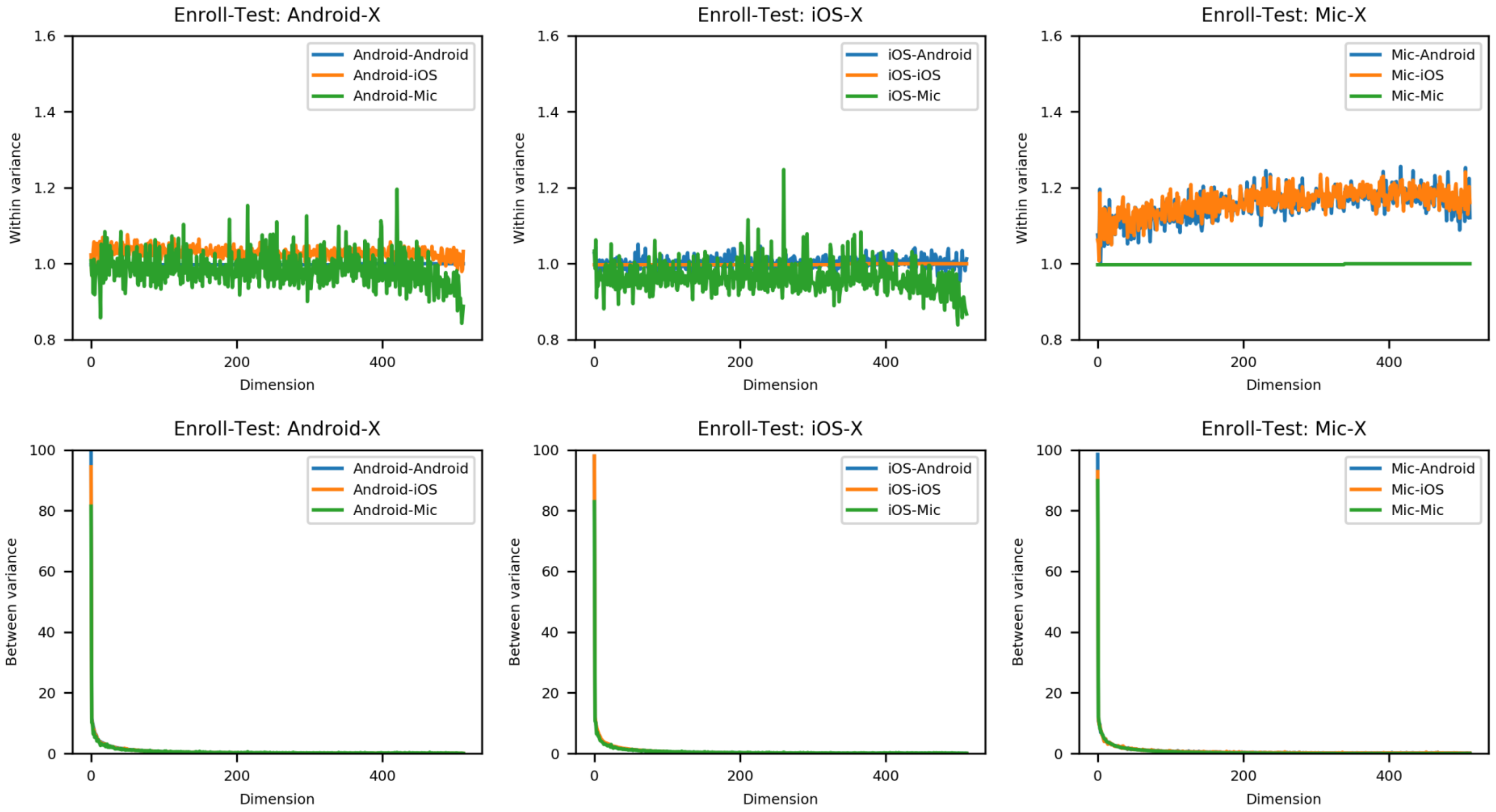
Enroll-Test	Baseline
1st-1st	4.799
1st-2nd	6.400
1st-3rd	6.863
1st-4th	6.884
1st-5th	7.108
1st-6th	7.856
1st-7th	7.906
1st-8th	7.881

Time-varying scenarios

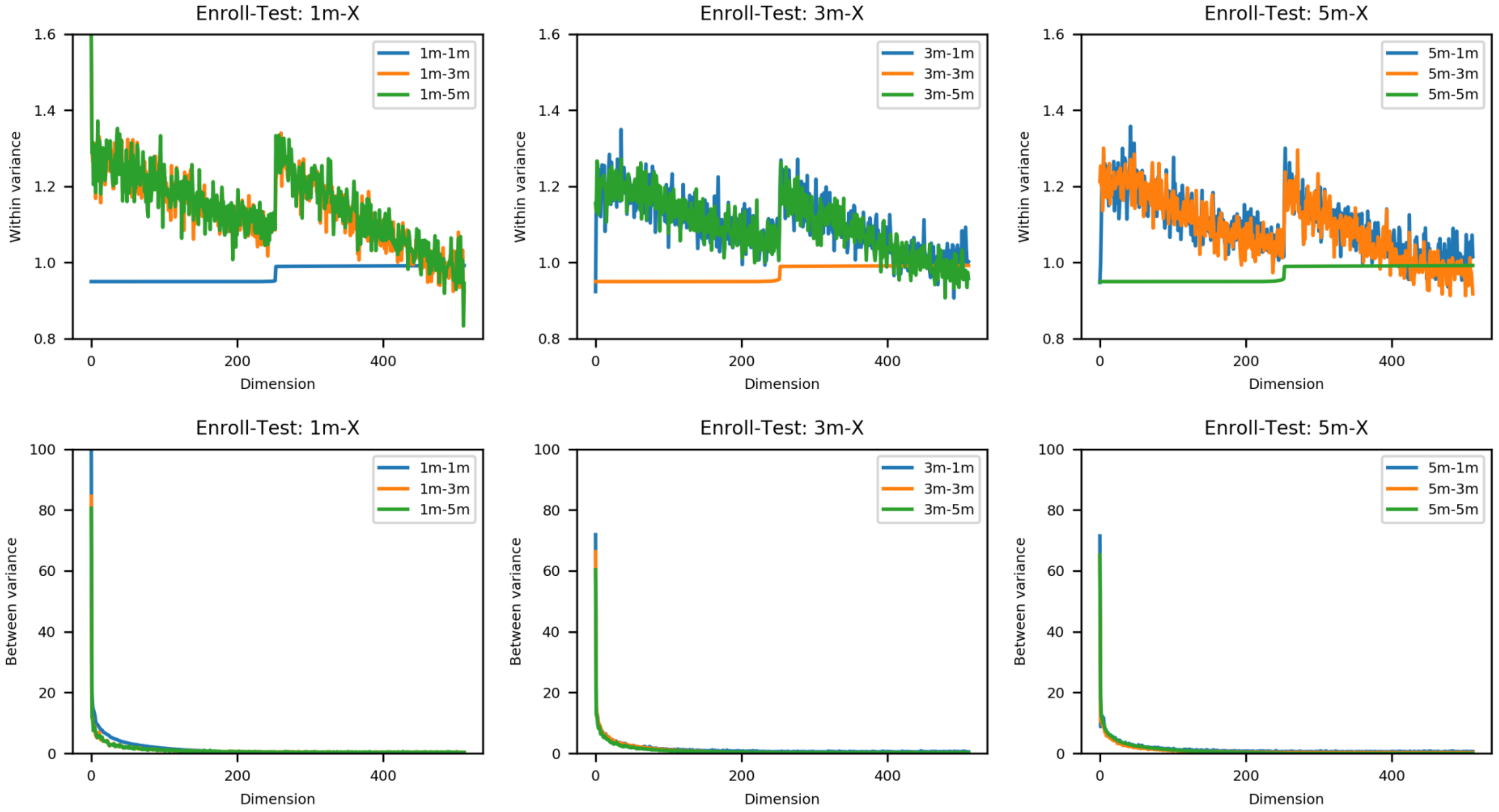
Why performance reduction ?

- Different statistical properties of the enrollment data and test data

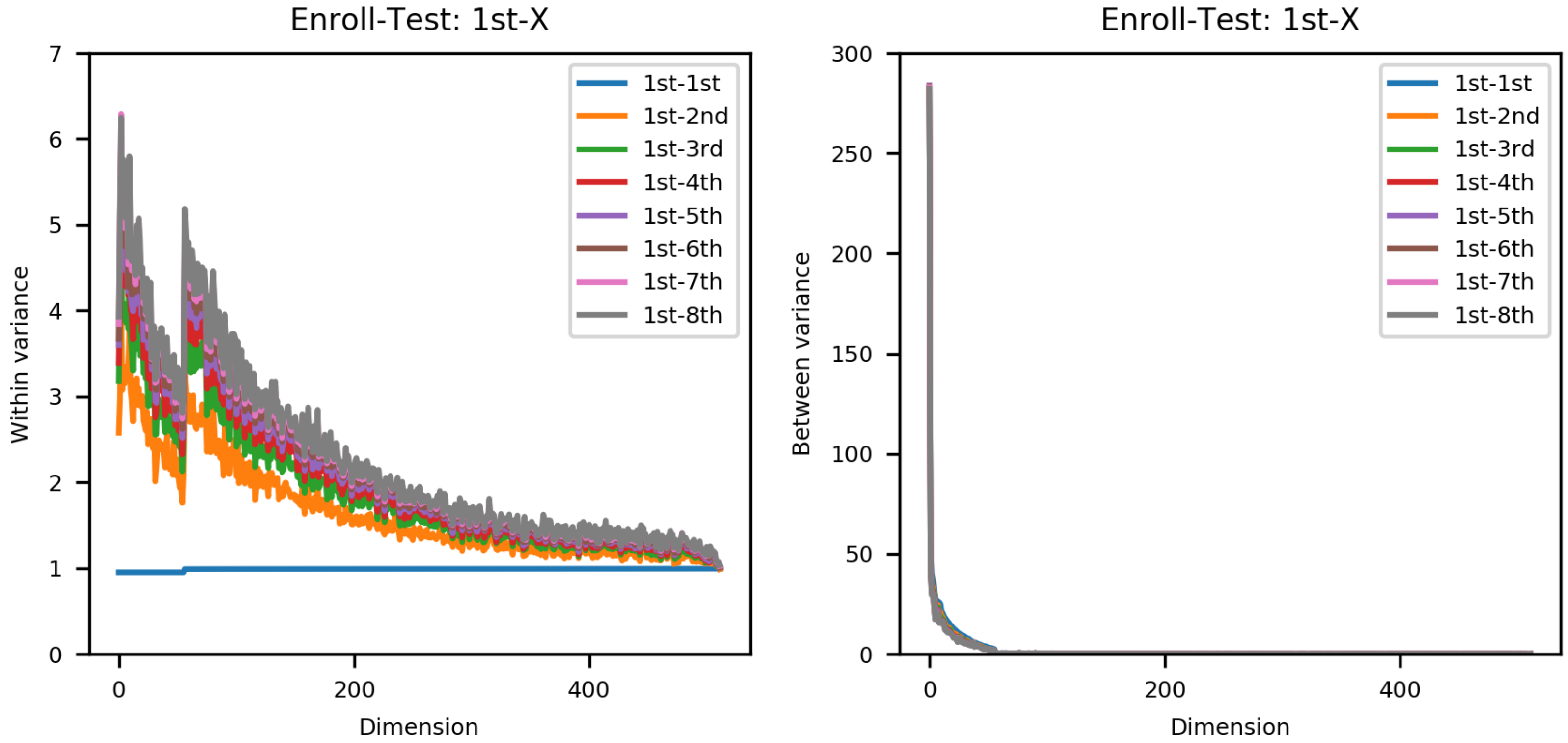
Statistics on cross-channel mismatch



Statistics on near-far mismatch



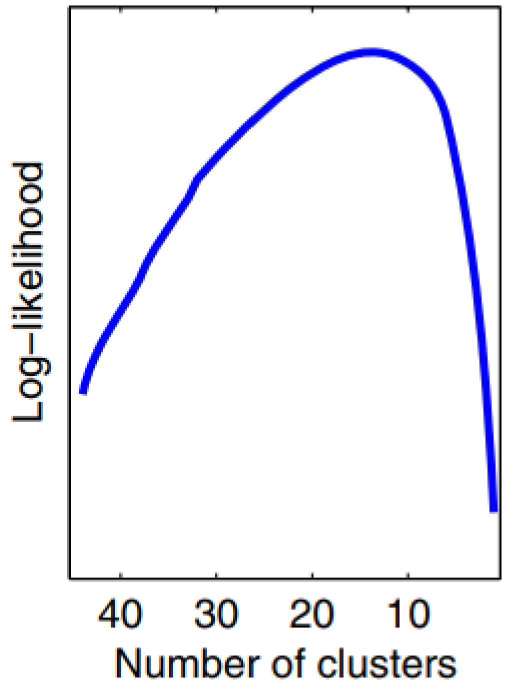
Statistics on time-varying mismatch



Related work

- Unsupervised training
  - Unsupervised PLDA adaption
  - Cluster data
  - Re-train PLDA

- Supervised training
  - Multi-domain training (MDT)
  - Combine multi-domain data
  - Re-train PLDA



Potential problems

- Unsupervised PLDA
  - Low efficiency
  - The statistics are inaccurate in neither enrollment nor test conditions.

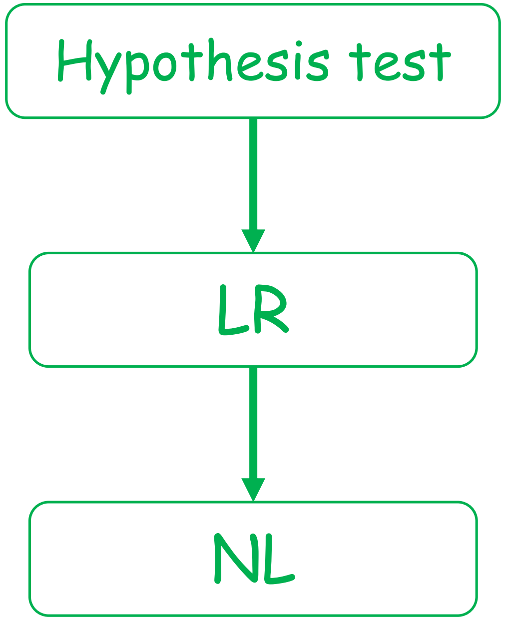
- Supervised MDT
  - A neutralization between enrollment and test
  - The statistics are inaccurate in neither enrollment nor test conditions.

Back to NL scoring

$$p(H0 \mid x) \gtrless p(H1 \mid x)$$

$$LR(x) = \frac{p(x \mid \lambda_{hpy})}{p(x \mid \lambda_{\overline{hpy}})}$$

$$LR(x \mid u_k) = \frac{p_k(x)}{p(x)}$$



$$NL(x \mid u_k) = \frac{p(x \mid u_k)}{p(x)} = \frac{p_k(x)}{p(x)} = \frac{p(x \mid x_k^1, x_k^2, \ldots, x_k^T)}{p(x)}$$

Conditional instead of marginal

Given $X_k^* = \{x_k^1, x_k^2, ..., x_k^T\}$

$$p(u_k \mid x_k^1, x_k^2, ..., x_k^T) = N(u_k; \frac{n_k \varepsilon}{n_k \varepsilon + \sigma} \bar{x}_k, \frac{\varepsilon \sigma}{n_k \varepsilon + \sigma} I)$$

$$\begin{aligned} p_k(x) &= p(x \mid x_k^1, x_k^2, ..., x_k^T) \\ &= \int p(x \mid u_k) p(u_k \mid x_k^1, x_k^2, ..., x_k^T) du_k \\ &= N(x; \frac{n_k \varepsilon}{n_k \varepsilon + \sigma} \bar{x}_k, (\sigma + \frac{\varepsilon \sigma}{n_k \varepsilon + \sigma}) I) \end{aligned}$$

$$\log p_k(x) = -\frac{1}{\sigma + \frac{\varepsilon \sigma}{n_k \varepsilon + \sigma}} \| x - \tilde{u}_k \|^2 + const$$

$$\begin{aligned} \log NL(x \mid u_k) &= \log p_k(x) - \log p(x) \\ &= -\frac{1}{\sigma + \frac{\varepsilon \sigma}{n_k \varepsilon + \sigma}} \| x - \tilde{u}_k \|^2 + \frac{1}{\varepsilon + \sigma} \| x \|^2 + const \end{aligned}$$

Enroll-test mismatch in NL scoring

- Variation of ε and δ in enrollment and test.

$$\log NL(x \mid u_k) = \log p_k(x) - \log p(x)$$

$$= -\frac{1}{\sigma + \frac{\varepsilon\sigma}{n_k\varepsilon + \sigma}} \| x - \tilde{u}_k \|^2 + \frac{1}{\varepsilon + \sigma} \| x \|^2 + const$$

- Estimate accurate ε and δ, and then apply in NL formula.

Methods based on your data prior

- Local label
  - Device label

- Global label
  - Device label
  - Speaker label

Local label on Cross-channel mismatch

- Back to statistics

TABLE I: Statistics on cross-channel enroll-test mismatch.

Enroll	Test	1-Cos(·)	Euc(·)	Avg.σ(512)	Avg.σ(10)	Avg.ϵ(512)	Avg.ϵ(10)
AN(D)	AN(D)	0.000	0.000	0.998	0.997	0.701	16.943
	Mic(D)	0.006	0.109	0.983	0.975	0.733	13.979
	iOS(D)	0.001	0.047	1.029	1.021	0.721	16.013
Mic(D)	AN(D)	0.006	0.109	1.155	1.089	0.829	16.413
	Mic(D)	0.000	0.000	0.998	0.997	0.674	15.804
	iOS(D)	0.005	0.108	1.162	1.088	0.812	15.872
iOS(D)	AN(D)	0.001	0.047	1.002	1.001	0.714	16.282
	Mic(D)	0.005	0.108	0.962	0.978	0.720	14.212
	iOS(D)	0.000	0.000	0.998	0.997	0.692	16.687

Special assumption

- Global shift (GST)

$$\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{b},$$

$$\log NL(\hat{\boldsymbol{x}}|k) \propto -\frac{1}{\sigma + \frac{\epsilon\sigma}{n_k\epsilon+\sigma}}||\hat{\boldsymbol{x}} - \boldsymbol{b} - \tilde{\boldsymbol{\mu}}_k||^2 + \frac{1}{\epsilon + \sigma}||\hat{\boldsymbol{x}} - \boldsymbol{b}||^2,$$

- Within-variance adaptation (WVA)

$$\log NL(\boldsymbol{x}|k) \propto -\frac{1}{\hat{\sigma} + \frac{\epsilon\sigma}{n_k\epsilon+\sigma}}||\boldsymbol{x} - \tilde{\boldsymbol{\mu}}_k||^2 + \frac{1}{\epsilon + \hat{\sigma}}||\boldsymbol{x}||^2,$$

Performance on cross-channel test

Enroll-Test	Baseline	GST	WVA
AN-AN	0.797	0.797	0.797
AN-Mic	2.146	1.764	2.165
AN-iOS	1.425	1.382	1.401
Mic-AN	2.175	1.665	2.033
Mic-Mic	0.778	0.778	0.778
Mic-iOS	2.251	1.892	2.081
iOS-AN	1.599	1.430	1.590
iOS-Mic	2.216	1.759	2.231
iOS-iOS	0.920	0.920	0.920
Mean	1.590	1.376	1.555
Var.	0.361	0.172	0.327

- Performance tendency is consistent with statistical properties.

Global label on Cross-channel mismatch

- From enroll to test (MLE-A)

$$\hat{\boldsymbol{x}} = \mathbf{M}\boldsymbol{x} + \boldsymbol{b}$$

$$p(\boldsymbol{\mu}_k|\boldsymbol{x}_1^k, ...\boldsymbol{x}_{n_k}^k) = N(\boldsymbol{\mu}_k; \frac{n_k\epsilon}{n_k\epsilon+\sigma}\bar{\boldsymbol{x}}_k, \frac{\epsilon\sigma}{n_k\epsilon+\sigma}\mathbf{I}).$$

$$p'(\hat{\boldsymbol{\mu}}_k|\boldsymbol{x}_1^k, ...\boldsymbol{x}_{n_k}^k) = N(\hat{\boldsymbol{\mu}}_k; \frac{n_k\epsilon}{n_k\epsilon+\sigma}\mathbf{M}\bar{\boldsymbol{x}}_k+\boldsymbol{b}, \frac{\epsilon\sigma}{n_k\epsilon+\sigma}\mathbf{M}\mathbf{M}^T).$$

$$\begin{aligned} p'_k(\hat{\boldsymbol{x}}) &= \int p'(\hat{\boldsymbol{x}}|\hat{\boldsymbol{\mu}}_k)p'(\hat{\boldsymbol{\mu}}_k|\boldsymbol{x}_1^k, ...\boldsymbol{x}_{n_k}^k)\mathrm{d}\hat{\boldsymbol{\mu}}_{\mathrm{k}} \\ &= N(\hat{\boldsymbol{x}}; \frac{n_k\epsilon}{n_k\epsilon+\sigma}\mathbf{M}\bar{\boldsymbol{x}}_k + \boldsymbol{b}, \hat{\sigma}\mathbf{I} + \frac{\epsilon\sigma}{n_k\epsilon+\sigma}\mathbf{M}\mathbf{M}^T) \end{aligned}$$

- Optimization
  - MLE (Maximum Likelihood

$$\mathcal{L}(M, b) = \sum_{k=1}^{K}\sum_{i=1}^{N} \log p_k(\hat{\boldsymbol{x_{ik}}}; M, b)$$

- NL scoring

$$\log NL \propto -(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{\mu}}_k)^T\tilde{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{\mu}}_k) + \frac{1}{\hat{\epsilon}+\hat{\sigma}}||\hat{\boldsymbol{x}}||^2$$

Global label on Cross-channel mismatch

- From test to enroll (MLE-B)

$$\boldsymbol{x} = \mathbf{M}\hat{\boldsymbol{x}} + \boldsymbol{b}$$

$$\begin{aligned} p_k(\boldsymbol{x}) &= p(\boldsymbol{x}|\boldsymbol{x}_1^k, ..., \boldsymbol{x}_{n_k}^k) \\ &= \int p(\boldsymbol{x}|\boldsymbol{\mu}_k)p(\boldsymbol{\mu}_k|\boldsymbol{x}_1^k, ...\boldsymbol{x}_{n_k}^k)\mathrm{d}\boldsymbol{\mu}_\mathrm{k} \\ &= N(\boldsymbol{x}; \frac{n_k\epsilon}{n_k\epsilon+\sigma}\bar{\boldsymbol{x}}_k, (\sigma + \frac{\epsilon\sigma}{n_k\epsilon+\sigma})\mathbf{I}) \end{aligned}$$

$$\begin{aligned} p_k(\hat{\boldsymbol{x}}; M, b) &= p(M\hat{\boldsymbol{x}} + b|\boldsymbol{x}_1^k, ..., \boldsymbol{x}_{n_k}^k) \\ &= \int p(M\hat{\boldsymbol{x}} + b|\boldsymbol{\mu}_k)p(\boldsymbol{\mu}_k|\boldsymbol{x}_1^k, ...\boldsymbol{x}_{n_k}^k)\mathrm{d}\boldsymbol{\mu}_\mathrm{k} \\ &= N(M\hat{\boldsymbol{x}} + b; \frac{n_k\epsilon}{n_k\epsilon+\sigma}\bar{\boldsymbol{x}}_k, (\sigma + \frac{\epsilon\sigma}{n_k\epsilon+\sigma})\mathbf{I}) \end{aligned}$$

- Optimization
  - MLE (Maximum Likelihood

$$\mathcal{L}(M, b) = \sum_{k=1}^{K}\sum_{i=1}^{N} \log p_k(\hat{\boldsymbol{x_{ik}}}; M, b)$$

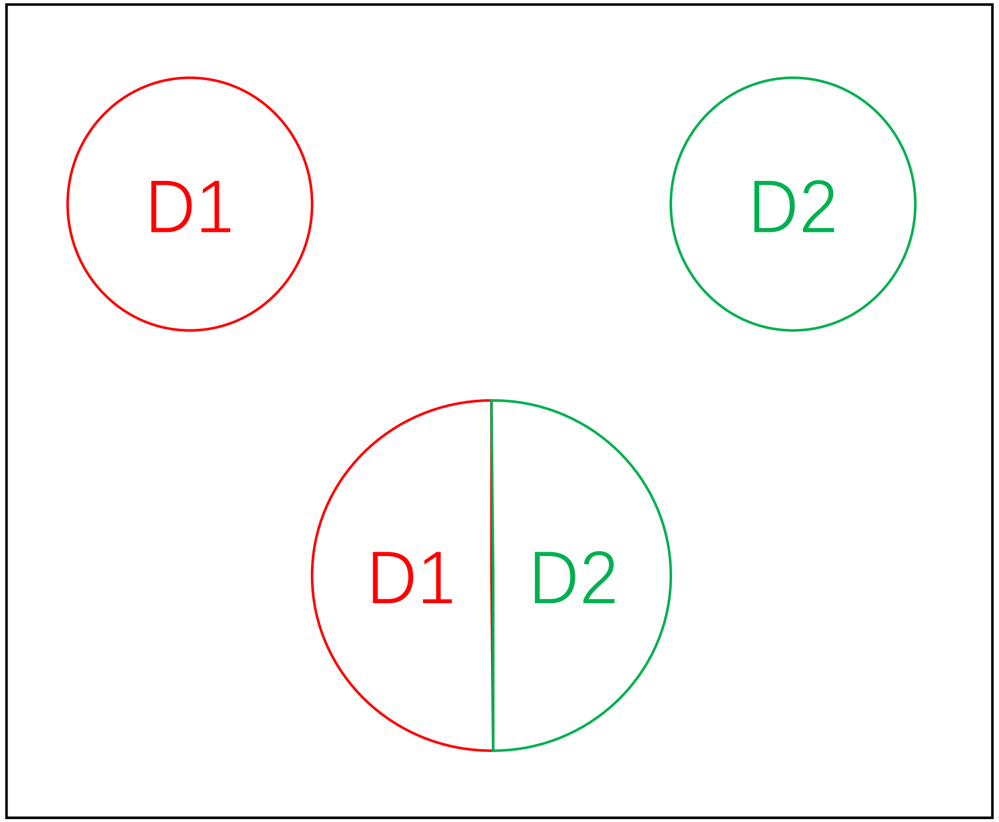
- NL scoring

$$\log NL \propto -(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{\mu}}_k)^T \tilde{\mathbf{\Sigma}}^{-1}(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{\mu}}_k) + \frac{1}{\hat{\epsilon}+\hat{\sigma}}||\hat{\boldsymbol{x}}||^2$$

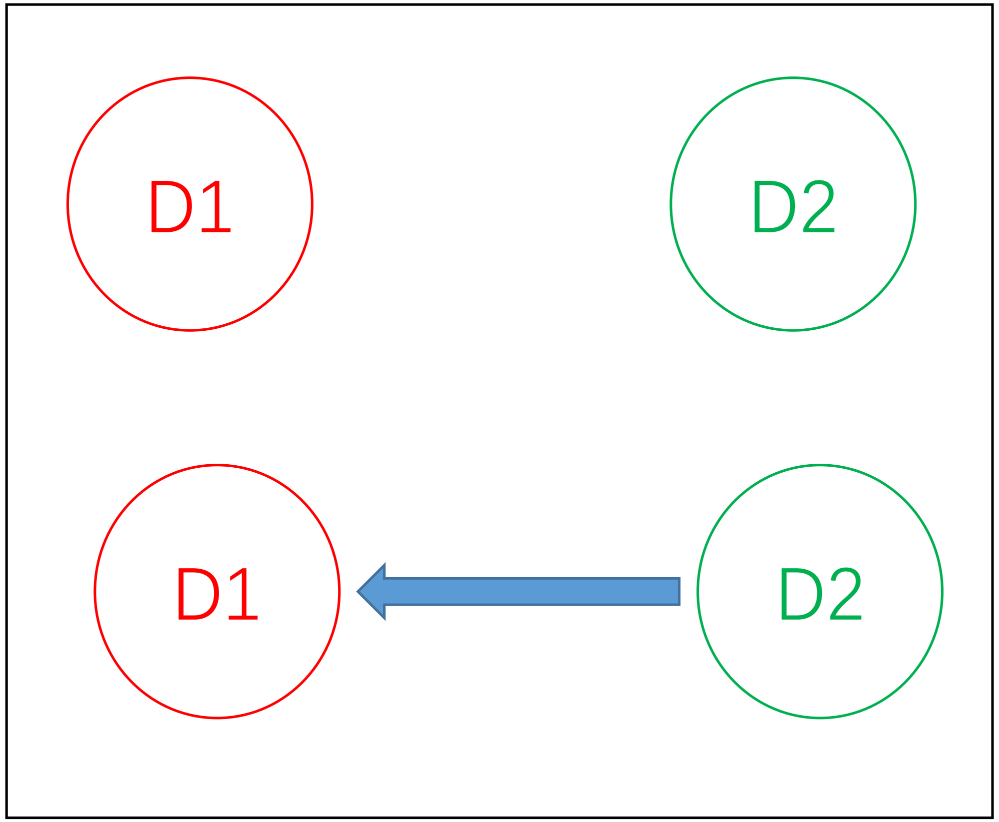
Performance on cross-channel test

Enroll-Test	Baseline	GST	WVA	MCT	MLE-A	MLE-B
AN-AN	0.797	0.797	0.797	0.797	0.797	0.373
AN-Mic	2.146	1.764	2.165	1.151	1.339	0.981
AN-iOS	1.425	1.382	1.401	1.161	0.967	0.628
Mic-AN	2.175	1.665	2.033	1.161	-	0.712
Mic-Mic	0.778	0.778	0.778	0.778	-	0.523
Mic-iOS	2.251	1.892	2.081	1.293	-	0.812
iOS-AN	1.599	1.430	1.590	1.156	0.797	0.755
iOS-Mic	2.216	1.759	2.231	1.137	1.443	1.056
iOS-iOS	0.920	0.920	0.920	0.920	0.920	0.425

MCT vs. MLE



MCT



MLE

MCT is not optimal

TABLE V: Performance EER(%) with semi-supervised MCT on cross-channel test.

Enroll-Test	BASE	0%	20%	40%	60%	80%	100%
AN-AN	0.797	-	-	-	-	-	-
AN-Mic	2.146	3.329	1.410	1.273	**1.066**	1.259	1.151
AN-iOS	1.425	1.642	1.104	**0.930**	1.029	1.170	1.161
Mic-AN	2.175	3.675	1.953	1.746	1.184	1.307	**1.161**
Mic-Mic	0.778	-	-	-	-	-	-
Mic-iOS	2.251	3.732	1.883	1.675	**1.255**	1.349	1.293
iOS-AN	1.599	2.024	1.472	1.241	**1.156**	1.274	**1.156**
iOS-Mic	2.216	3.697	1.651	1.476	**1.061**	1.236	1.137
iOS-iOS	0.920	-	-	-	-	-	-

MCT needs more global label data

TABLE VI: Performance EER(%) with pure-supervised MCT on cross-channel test.

Enroll-Test	BASE	20%	40%	60%	80%	100%
AN-AN	0.797	-	-	-	-	-
AN-Mic	2.146	5.093	1.896	1.264	1.283	**1.151**
AN-iOS	1.425	5.185	1.628	1.250	1.250	**1.161**
Mic-AN	2.175	5.586	2.284	1.274	1.194	**1.161**
Mic-Mic	0.778	-	-	-	-	-
Mic-iOS	2.251	5.732	2.213	1.491	1.420	**1.293**
iOS-AN	1.599	5.213	1.807	1.236	1.151	**1.156**
iOS-Mic	2.216	5.296	1.900	1.165	1.165	**1.137**
iOS-iOS	0.920	-	-	-	-	-

MLE needs less global label data

TABLE VIII: Performance EER(%) with pure-supervised MLE-B on cross-channel test.

Enroll-Test	BASE	20%	40%	60%	80%	100%
AN-AN	0.797	-	-	-	-	-
AN-Mic	2.146	2.980	2.113	0.868	1.080	0.981
AN-iOS	1.425	1.241	1.085	0.524	0.651	0.623
Mic-AN	2.175	5.921	5.091	0.967	0.750	0.712
Mic-Mic	0.778	-	-	-	-	-
Mic-iOS	2.251	5.582	4.902	1.066	0.830	0.812
iOS-AN	1.599	1.835	1.628	0.642	0.816	0.760
iOS-Mic	2.216	2.740	1.882	0.821	1.094	1.052
iOS-iOS	0.920	-	-	-	-	-

Conclusions

- The NL-based scoring form can be used to address enroll-test mismatch.
- The proposed MLE-B can be regarded as a principle solution, and obtain the best performance.