

# Reducing Domain Mismatch by Maximum Mean Discrepancy Based Autoencoders

Wei-wei Lin, Man-Wai Mak, Longxin Li

*The Hong Kong Polytechnic University*

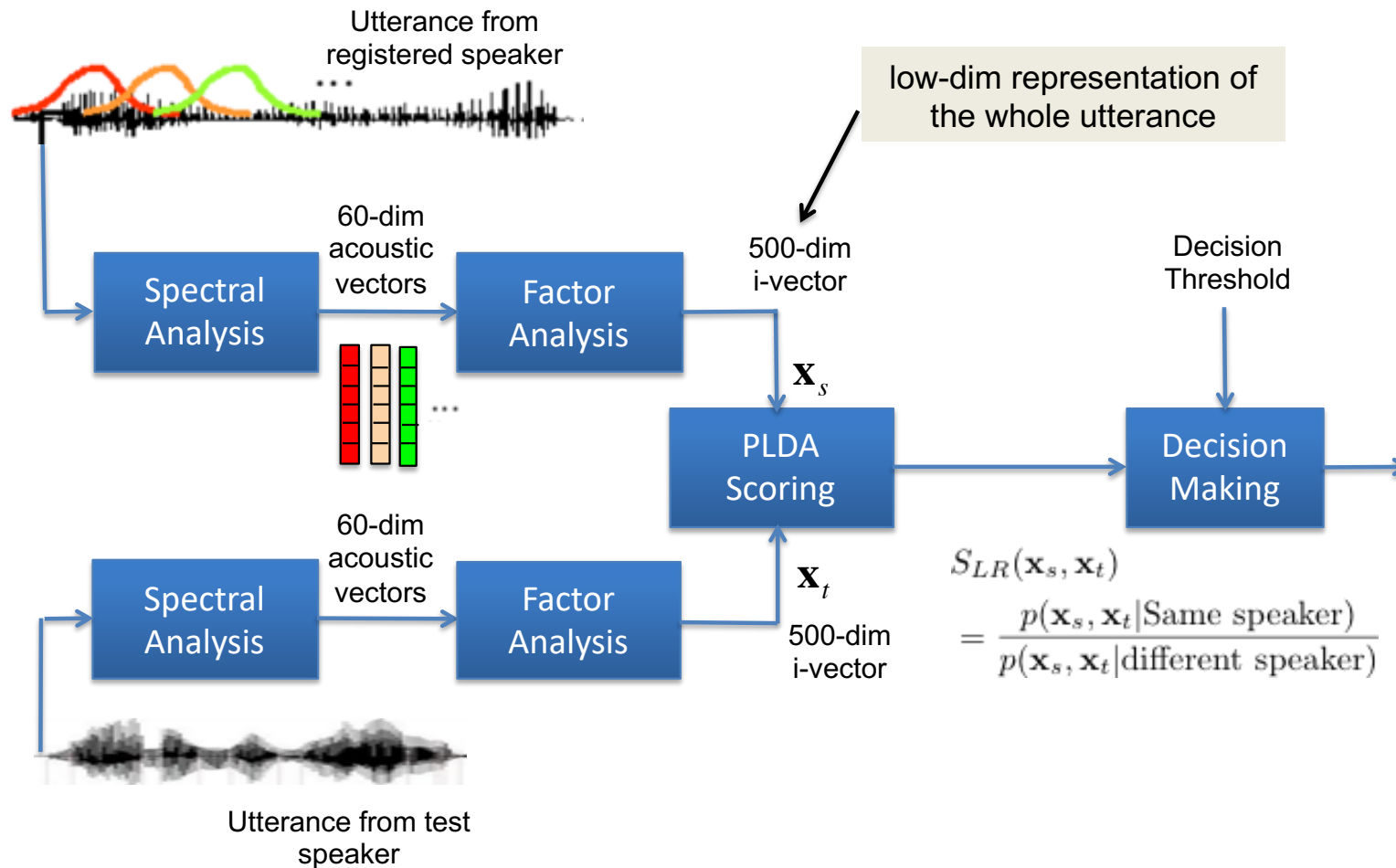
Jen-Tzung Chien

*National Chiao Tung University*

# Contributions

- We show how maximum mean discrepancy (MMD) can be generalized to measure the discrepancies among multiple distributions.
- We propose a new domain adaptation method based on MMD and demonstrated that it can greatly reduce multi-source variability.

# Process of Speaker Verification



# I-Vectors

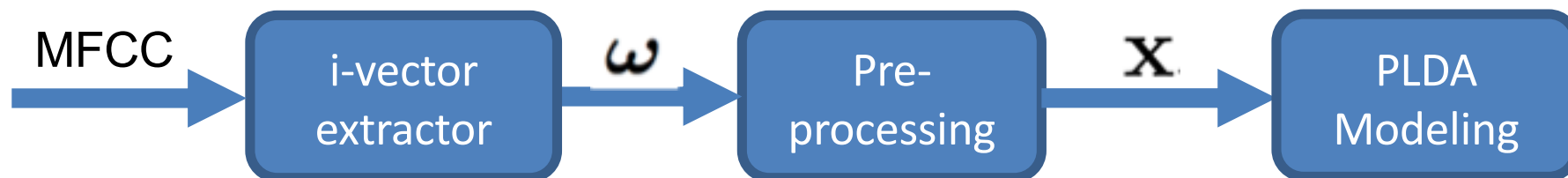
$$\beta = \mathbf{m} + \mathbf{T}\eta \quad \eta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Speaker supervector  $\beta$   
UBM Supervector  $\mathbf{m}$   
Total variability matrix  $\mathbf{T}$   
Total variability factor  $\eta$

- **I-vector** is the maximum-a-posteriori (MAP) estimate of  $\eta$ , which we denote as  $\omega$ .
- Instead of using high-dimension supervector  $\beta$  to represent a speaker, we use more compact (low-dimension) i-vector  $\omega$  to represent a speaker.
- $\mathbf{T}$  represents the subspace where i-vectors vary.

# I-Vector/PLDA

- Procedure of i-vector/PLDA:



- In Gaussian PLDA, a preprocessed i-vector  $\mathbf{x}_{ij}$  from the  $j$ -th session of speaker  $i$  is considered generated from a factor analysis model:

$$\mathbf{x}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{z}_i + \epsilon_{ij}$$

Diagram illustrating the components of the Gaussian PLDA model equation:

- $\mathbf{x}_{ij}$ : Pre-processed i-vector
- $\mathbf{m}$ : Mean of i-vectors in training set
- $\mathbf{V}$ : Speaker subspace
- $\mathbf{z}_i$ : Speaker factor
- $\epsilon_{ij}$ : Residue

# I-Vector/PLDA

- Given a test i-vector  $\mathbf{x}_s$  and target-speaker's i-vectors  $\mathbf{x}_t$ , the verification score is the log-likelihood ratio between two hypotheses:

$$\begin{aligned}\log S_{LR}(\mathbf{x}_s, \mathbf{x}_t) &= \log \frac{p(\mathbf{x}_s, \mathbf{x}_t | \text{Same speaker})}{p(\mathbf{x}_s, \mathbf{x}_t | \text{different speaker})} \\ &= \frac{1}{2} \mathbf{x}_s^T \mathbf{Q} \mathbf{x}_s + \mathbf{x}_s^T \mathbf{P} \mathbf{x}_t + \frac{1}{2} \mathbf{x}_t^T \mathbf{Q} \mathbf{x}_t + \text{const}\end{aligned}$$

where

$$\mathbf{Q} = \Sigma_{tot}^{-1} - (\Sigma_{tot} - \Sigma_{ac} \Sigma_{tot}^{-1} \Sigma_{ac})^{-1}$$

$$\Sigma_{ac} = \mathbf{V} \mathbf{V}^T$$

$$\mathbf{P} = \Sigma_{tot}^{-1} \Sigma_{ac} (\Sigma_{tot} - \Sigma_{ac} \Sigma_{tot}^{-1} \Sigma_{ac})^{-1}$$

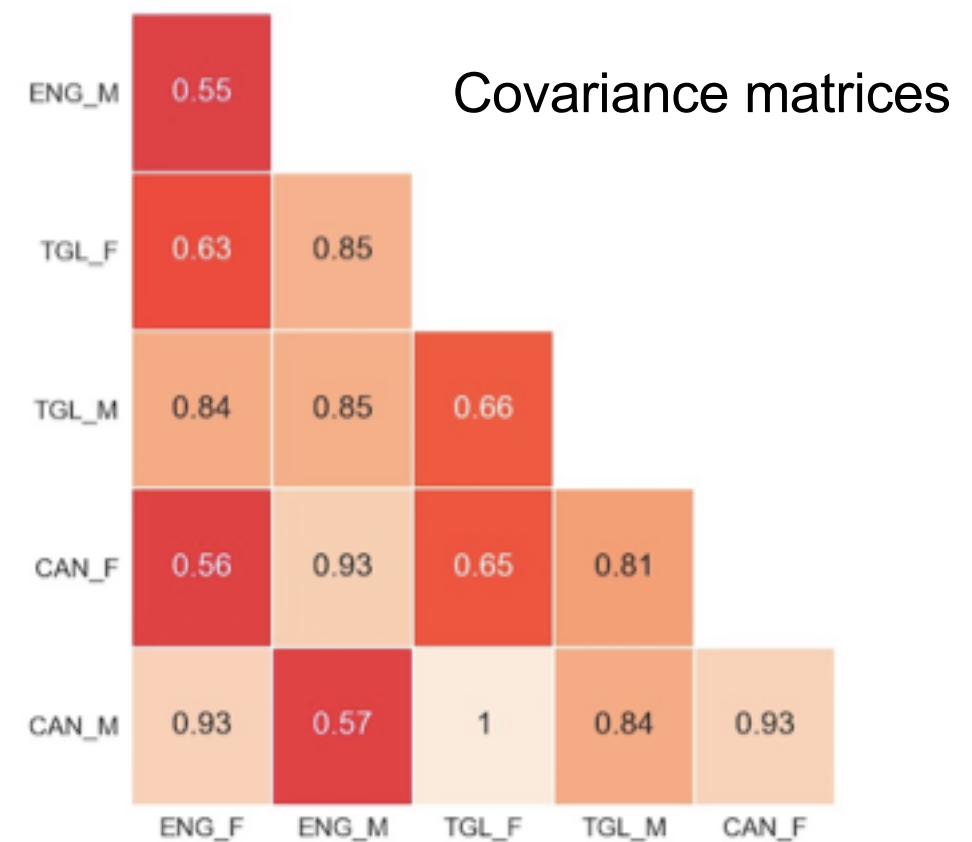
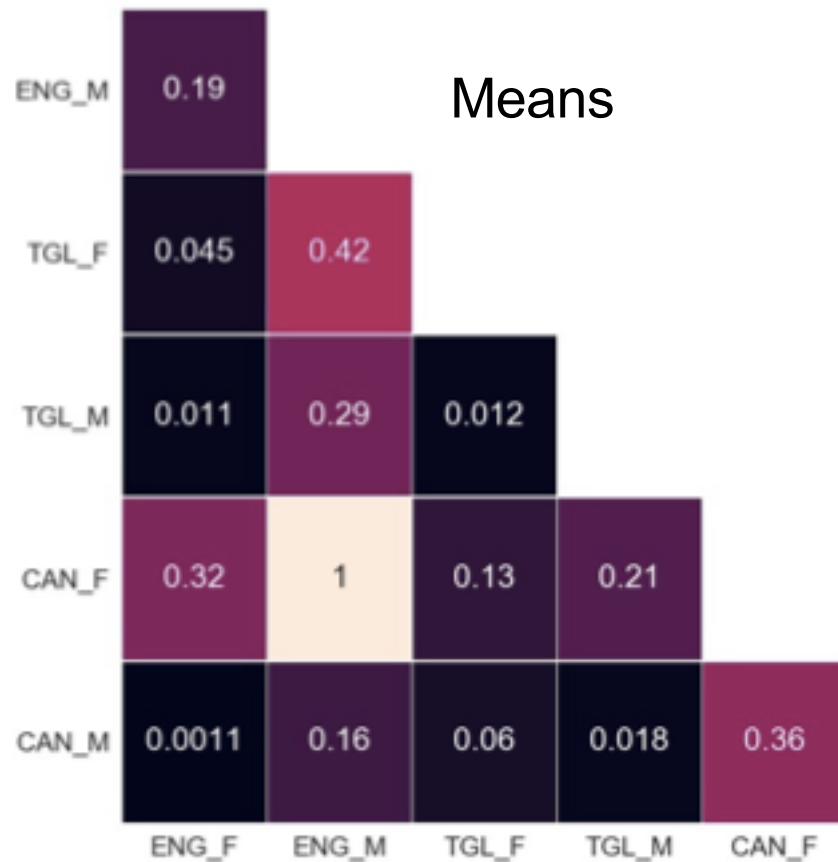
$$\Sigma_{tot} = \mathbf{V} \mathbf{V}^T + \Sigma$$

# Domain Mismatch

- NIST SRE16 is a multilingual dataset for speaker verification.
- Test data include Cantonese and Tagalog speakers. But both Cantonese and Tagalog speech in the development set are unlabeled and small in number (2344 segments).

Dataset	Category	Language
Dev	Unlabeled	Cantonese and Tagalog
Dev	Unlabeled	Mandarin and Cebuano
Dev	Labeled	Mandarin and Cebuano
Eval	Enrolment	Cantonese and Tagalog
Eval	Test	Cantonese and Tagalog

# Domain Mismatch



Pairwise normalized distance between different languages and genders

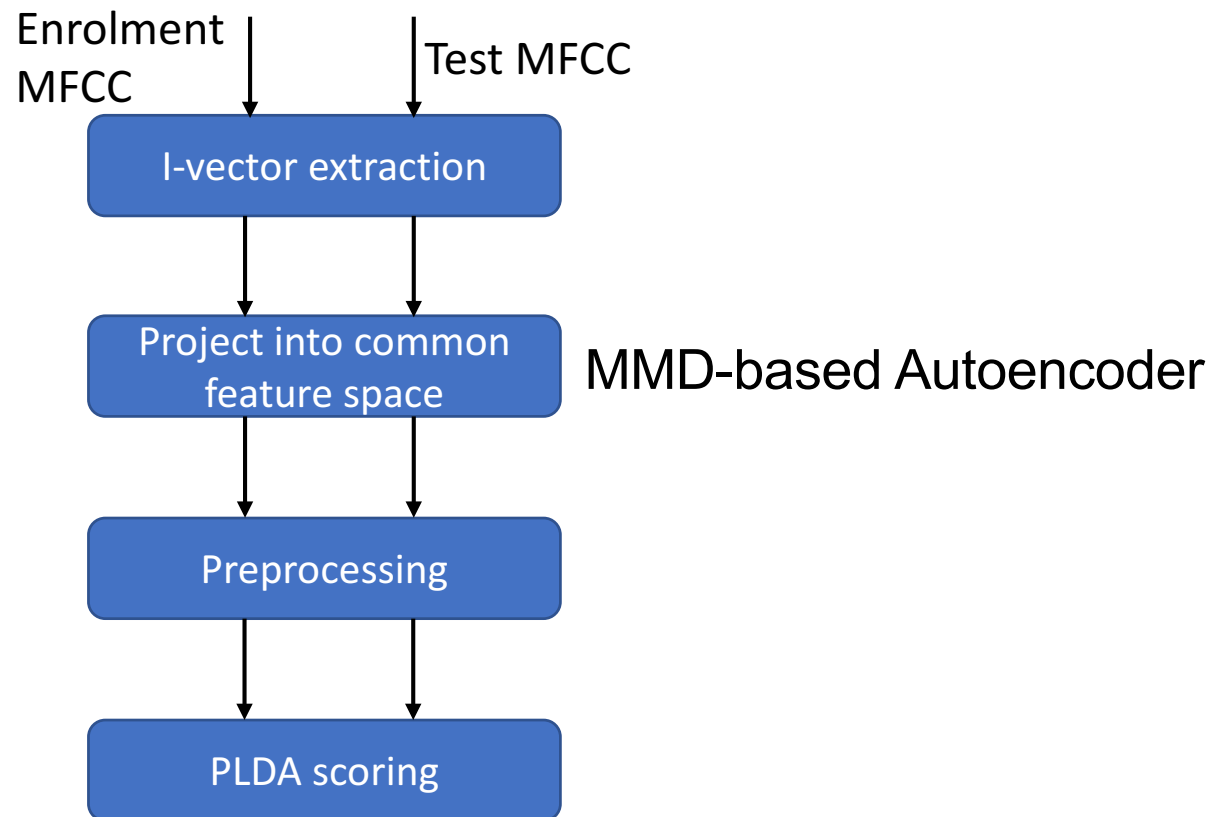


# Domain Mismatch

- We have English corpora from previous SREs and SWB, which are large in number and have speaker labels. But the language mismatch in SRE16 makes these corpora less useful.
- We aim to adapt the i-vectors of English speech to look more like the i-vectors of Cantonese and Tagalog.
- Then, we use the adapted English i-vectors to train a PLDA model for scoring Cantonese and Tagalog i-vectors.

# Domain Adaptation

- I-vector based domain adaptation:



# IDVC

- Inter-dataset variability compensation (IDVC) is a popular domain adaptation technique for speaker verification.
- IDVC aims to remove the subspace that causes most of the inter-dataset variability:

$$\hat{\mathbf{x}} = (\mathbf{I} - \mathbf{W}\mathbf{W}^T)\mathbf{x}$$

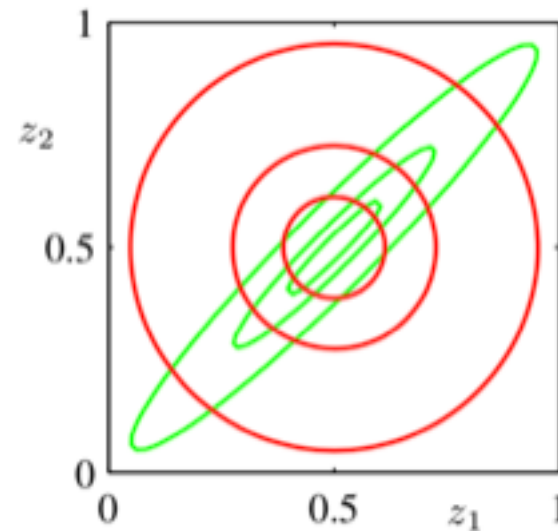
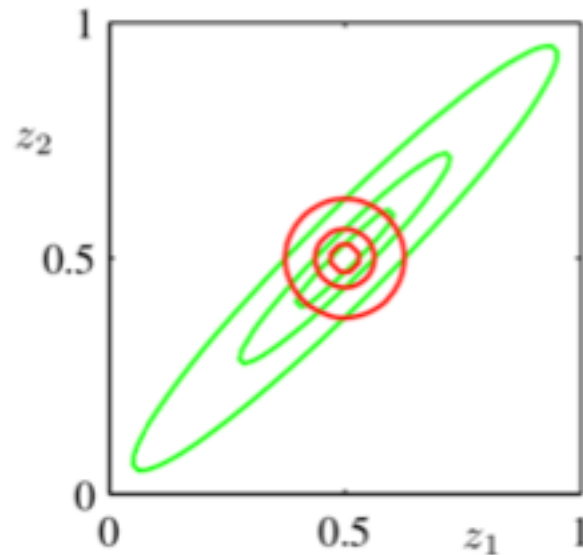
where  $\mathbf{x}$  is an i-vector, the columns of  $\mathbf{W}$  comprise the eigenvectors of the covariance matrix of the **domain means**.

# Motivations of Our Work

- A drawback of IDVC is that the domain mismatch is entirely defined by the domain means.
- From the perspective of reducing the divergence between probabilistic distributions, this is not enough.

# Motivations of Our Work

- Means are the first moment of probabilistic distributions only.
- Even if two distributions have exactly the same mean, they could still be very different, due to the difference in the higher order statistics.



# Maximum Mean Discrepancy

- The theoretical work in domain adaptation suggests that it is important to have a good measurement of the divergence between the data distributions of different domains.
- Maximum mean discrepancy (MMD) is a distance measure in the space of probability.
- Given two datasets, MMD computes the mean squared difference between the statistics of the two datasets:

$$\mathcal{D}_{\text{MMD}} = \left\| \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{x}_i) - \frac{1}{M} \sum_{j=1}^M \phi(\mathbf{y}_j) \right\|^2$$

# Maximum Mean Discrepancy


$$\begin{aligned}\mathcal{D}_{\text{MMD}} &= \left\| \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{x}_i) - \frac{1}{M} \sum_{j=1}^M \phi(\mathbf{y}_j) \right\|^2 \\ &= \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_{i'}) - \frac{2}{NM} \sum_{i=1}^N \sum_{j=1}^M \phi(\mathbf{x}_i)^\top \phi(\mathbf{y}_j) + \frac{1}{M^2} \sum_{j=1}^M \sum_{j'=1}^M \phi(\mathbf{y}_j)^\top \phi(\mathbf{y}_{j'}) \\ &= \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N k(\mathbf{x}_i, \mathbf{x}_{i'}) - \frac{2}{NM} \sum_{i=1}^N \sum_{j=1}^M k(\mathbf{x}_i, \mathbf{y}_j) + \frac{1}{M^2} \sum_{j=1}^M \sum_{j'=1}^M k(\mathbf{y}_j, \mathbf{y}_{j'})\end{aligned}$$



Kernel function

# Maximum Mean Discrepancy

- Assume that we have  $D$  sets of data  $\{\mathbf{h}_i^d\}_{i=1}^{N_d}$ , where  $d = 1, 2, \dots, D$ .
- We can generalize MMD to measure the discrepancies among multiple domains:

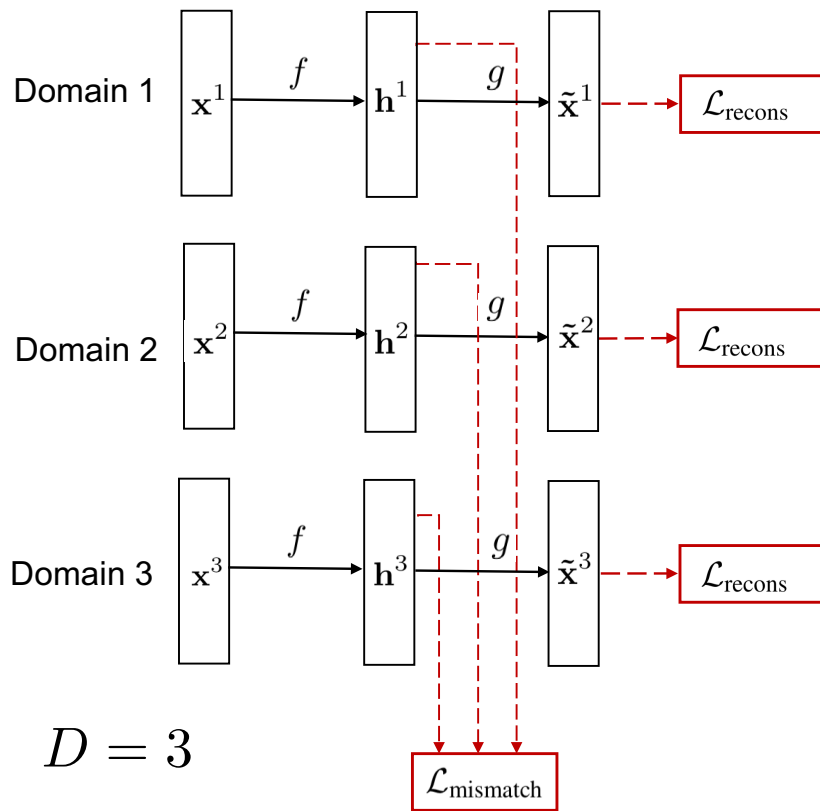
$$\mathcal{L}_{\text{mismatch}} = \sum_{d=1}^D \sum_{\substack{d'=1 \\ d' \neq d}}^D \left( \frac{1}{N_d^2} \sum_{i=1}^{N_d} \sum_{i'=1}^{N_d} k(\mathbf{h}_i^d, \mathbf{h}_{i'}^d) \right. \\ \left. - \frac{2}{N_d N_{d'}} \sum_{i=1}^{N_d} \sum_{j=1}^{N_{d'}} k(\mathbf{h}_i^d, \mathbf{h}_j^{d'}) + \frac{1}{N_{d'}^2} \sum_{j=1}^{N_{d'}} \sum_{j'=1}^{N_{d'}} k(\mathbf{h}_j^{d'}, \mathbf{h}_{j'}^{d'}) \right)$$


Kernel function



# Domain-Invariant Autoencoders

- The **domain-invariant autoencoder (DAE)** directly encodes the features that minimize the multi-source mismatch:



$$\mathcal{L}_{\text{recons}} = \frac{1}{2} \sum_{d=1}^D \sum_{i=1}^{N_d} \left\| \mathbf{x}_i^d - \tilde{\mathbf{x}}_i^d \right\|^2$$

$$\mathcal{L}_{\text{mismatch}} = \sum_{d=1}^D \sum_{\substack{d'=1 \\ d' \neq d}}^D \left( \frac{1}{N_d^2} \sum_{i=1}^{N_d} \sum_{i'=1}^{N_d} k(\mathbf{h}_i^d, \mathbf{h}_{i'}^{d'}) \right. \\ \left. - \frac{2}{N_d N_{d'}} \sum_{i=1}^{N_d} \sum_{j=1}^{N_{d'}} k(\mathbf{h}_i^d, \mathbf{h}_j^{d'}) + \frac{1}{N_{d'}^2} \sum_{j=1}^{N_{d'}} \sum_{j'=1}^{N_{d'}} k(\mathbf{h}_j^{d'}, \mathbf{h}_{j'}^{d'}) \right)$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mismatch}} + \lambda \mathcal{L}_{\text{recons}}$$

# Nuisance-Attribute Autoencoders

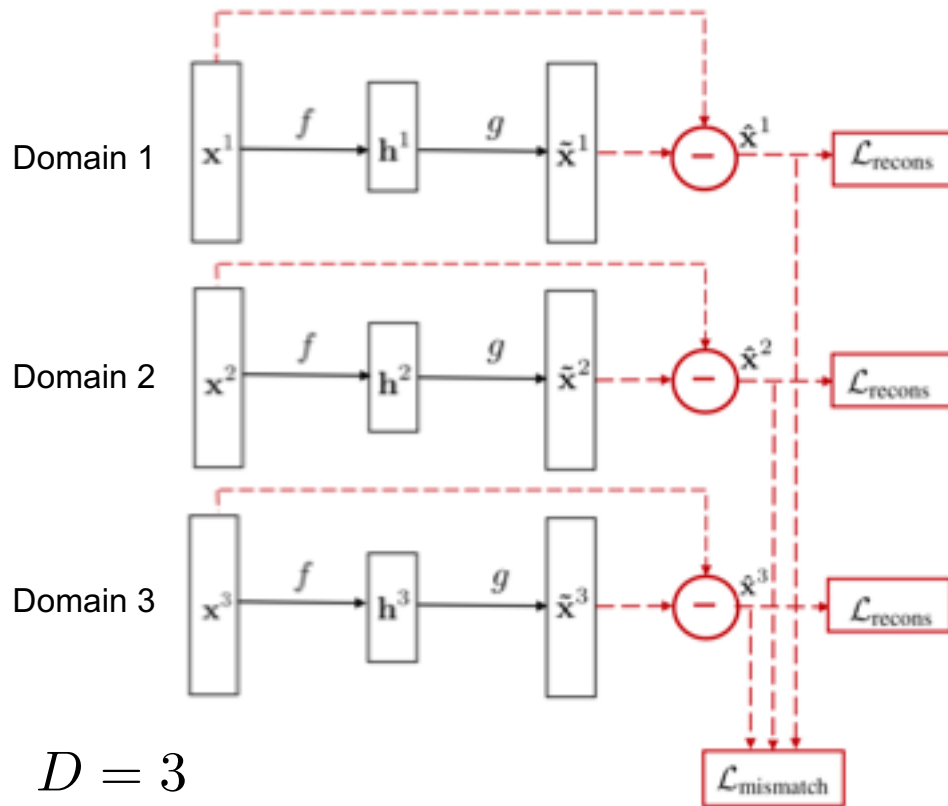
- The **n**uisance-attribute **a**utoencoder (NAE) borrows the idea of IDVC in that it removes the domain specific features using:

$$\hat{\mathbf{x}} = \mathbf{x} - g(f(\mathbf{x}))$$

where  $g(f(\mathbf{x}))$  should contain all of the domain-specific info.

- Therefore,  $\hat{\mathbf{x}}$  will become domain-indistinguishable.
- $g(f(\mathbf{x}))$  is realized by an autoencoder called NAE, which encodes the features that cause most of the multi-source mismatch.

# Nuisance-Attribute Autoencoders

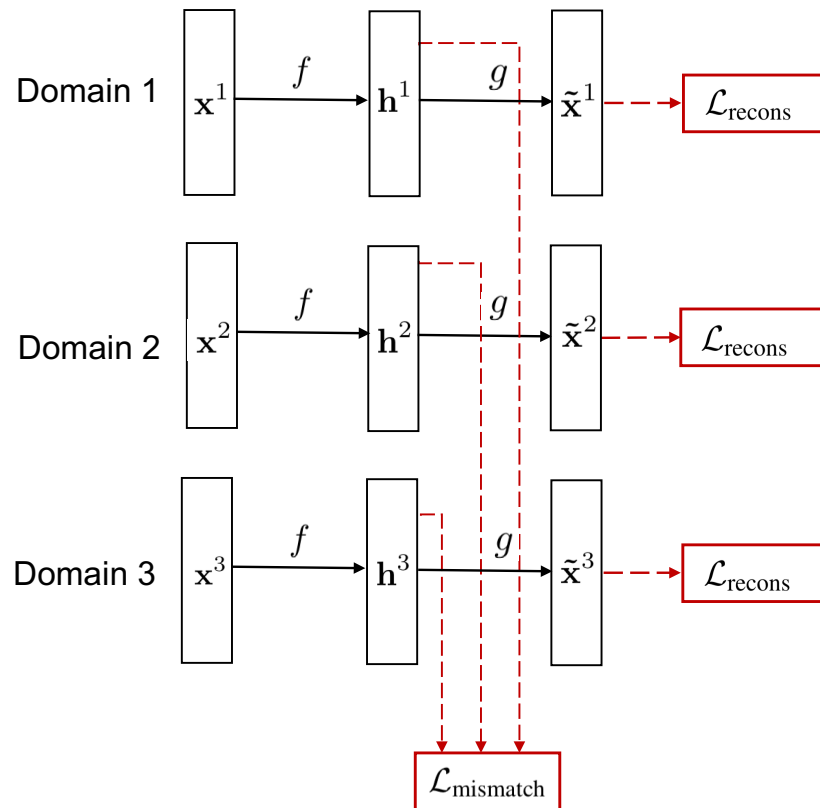


$$\mathcal{L}_{\text{recons}} = \frac{1}{2} \sum_{d=1}^D \sum_{i=1}^{N_d} \left\| \mathbf{x}_i^d - \tilde{\mathbf{x}}_i^d \right\|^2$$

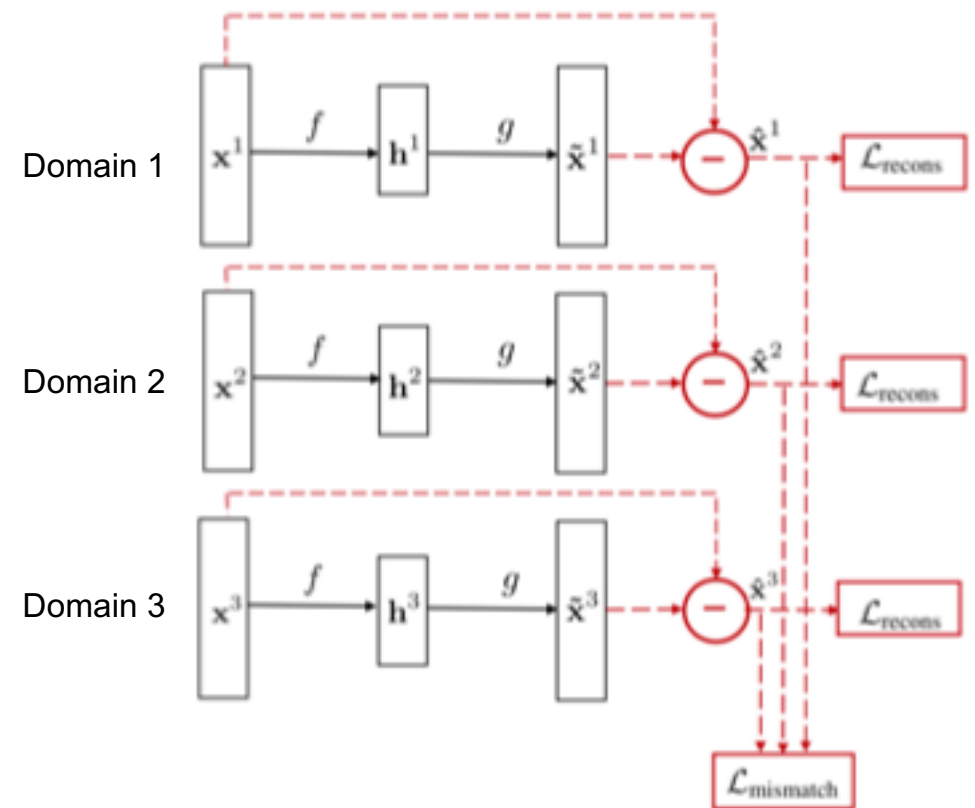
$$\mathcal{L}_{\text{mismatch}} = \sum_{d=1}^D \sum_{\substack{d'=1 \\ d' \neq d}}^D \left( \frac{1}{N_d^2} \sum_{i=1}^{N_d} \sum_{i'=1}^{N_d} k(\hat{\mathbf{x}}_i^d, \hat{\mathbf{x}}_{i'}^{d'}) \right. \\ \left. - \frac{2}{N_d N_{d'}} \sum_{i=1}^{N_d} \sum_{j=1}^{N_{d'}} k(\hat{\mathbf{x}}_i^d, \hat{\mathbf{x}}_j^{d'}) + \frac{1}{N_{d'}^2} \sum_{j=1}^{N_{d'}} \sum_{j'=1}^{N_{d'}} k(\hat{\mathbf{x}}_j^{d'}, \hat{\mathbf{x}}_{j'}^{d'}) \right)$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mismatch}} + \lambda \mathcal{L}_{\text{recons}}$$

# MMD-Baesda Autoencoders

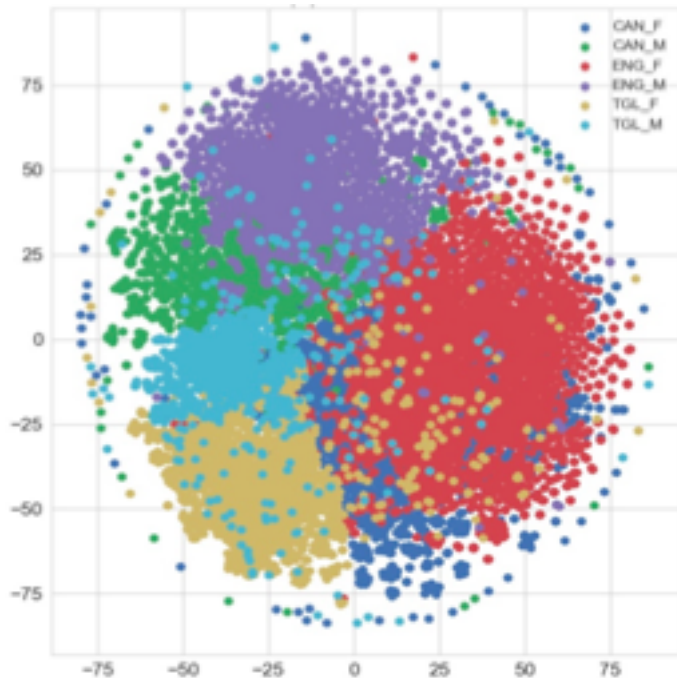


Domain-Invariant Autoencoder (DAE)

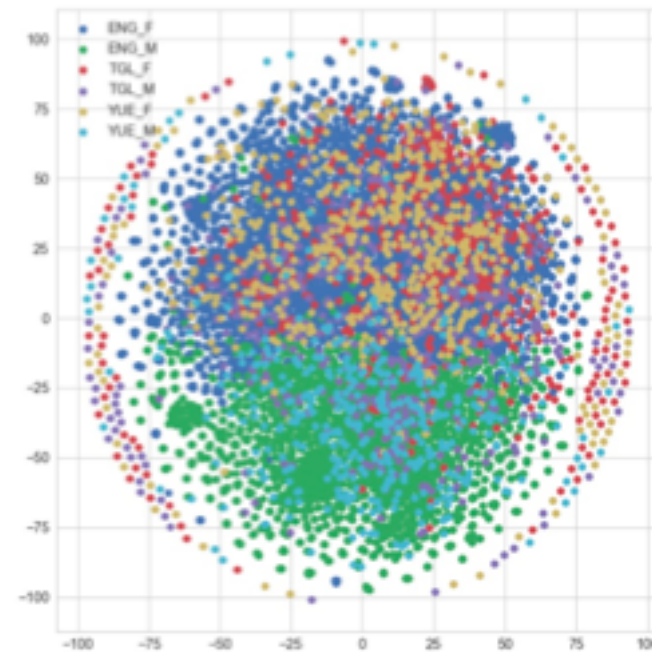


Nuisance-Attribute Autoencoder (NAE)

# t-SNE Visualizations of Learned Features



Before DAE Transformation



After DAE Transformation

The t-SNE plot of the hidden activations of DAE has less domain-clustering effect than that of the i-vectors, which shows that the DAE indeed learns a domain-invariant representation.

# Experimental Setup

- **Parameterization:** 19 MFCCs together with energy plus their 1<sup>st</sup> and 2<sup>nd</sup> derivatives → 60-Dim
- **UBM:** gender-dependent, 512 mixtures, trained by SRE16-dev
- **Total Variability Matrix:** gender-independent, 300 total factors, trained by SRE16-dev
- **DAE- and NAE-transformed vectors:** 300-dim
- **I-Vector Preprocessing:** PCA to 200-dim followed by length normalization
- **PLDA:** 200 latent factors

# Experimental Setup

- We have conducted two sets of experiments
  1. domain adaptation experiment
  2. domain robustness experiment.
- In the domain adaptation experiment, i-vectors derived from SRE04--SRE10 and SRE16-dev were used for training the DAE, the NAE and the projection matrices in IDVC.
- I-vectors derived from SRE16-eval were used for testing.

# Domain Adaptation Experiment

Method	EER	mCprim	aCprim
No Adapt	15.84	0.89	0.93
IDVC	13.08	0.86	0.93
DAE	12.79	0.85	0.91
NAE	12.81	0.85	0.91

Pooling genders and languages

- All of the domain adaptation methods improve system performance significantly.
- Both DAE and NAE outperform IDVC by a small margin.

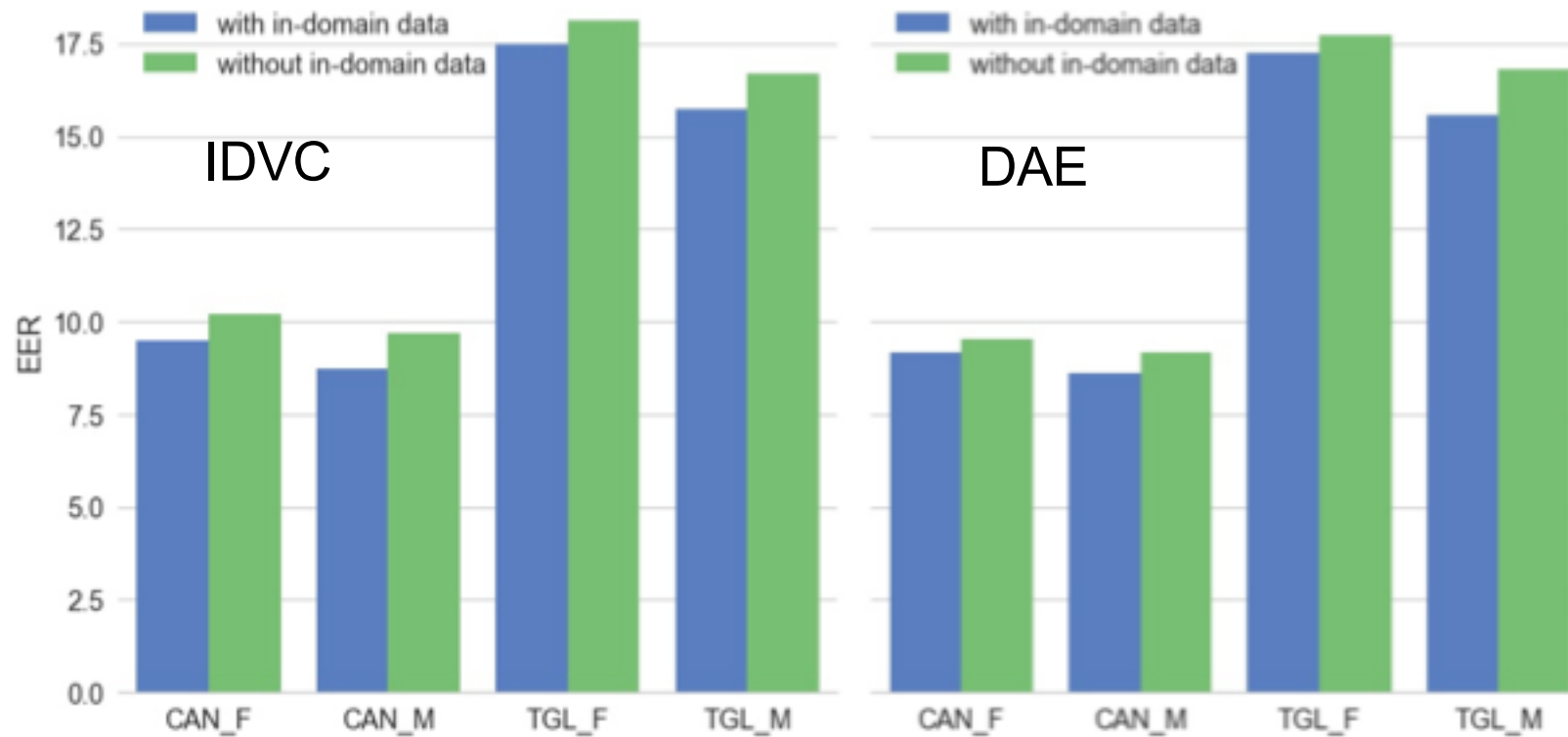


# Domain Robustness Experiment

- In the domain robustness experiment, for each gender and language (TGL/CAN) in test sessions, we exclude the speech of the same gender who speak that language from training.

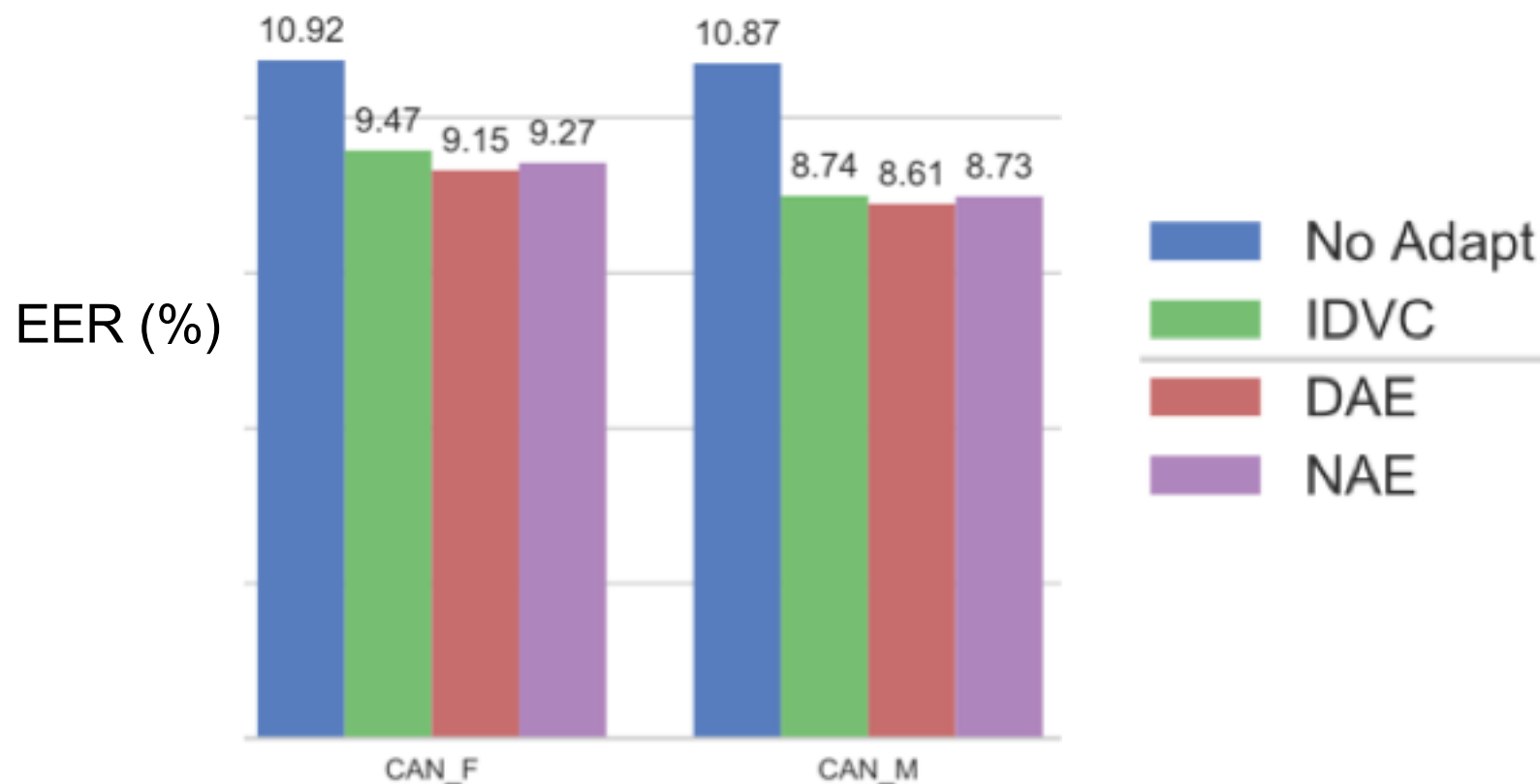
Test Data	Training Data					
	Male			Female		
	ENG	TGL	CAN	ENG	TGL	CAN
Male TGL	✓	✗	✓	✓	✓	✓
Male CAN	✓	✓	✗	✓	✓	✓
Female TGL	✓	✓	✓	✓	✗	✓
Female CAN	✓	✓	✓	✓	✓	✗

# Domain Robustness Experiment



- The performance of DA methods degrades when in-domain data are excluded from training.

# Domain Robustness Experiment



- DAE achieves a relative reduction of 5-6% with respect to IDVC on Cantonese speech. But no gain is found on Tagalog speech.

# I-vector Adaptation + PLDA Interpolation

- I-vectors adaptation can be combined with unsupervised PLDA model interpolation (interpolate the covariance matrices, Garcia-Romero (2014)).

Method	EER	mCprim	aCprim	Method	EER	mCprim	aCprim
No Adapt	15.84	0.89	0.93	No Adapt	13.47	0.86	0.91
IDVC	13.08	0.86	0.93	IDVC	12.88	0.85	0.93
DAE	12.79	0.85	0.91	DAE	12.43	0.84	0.90
NAE	12.81	0.85	0.91	NAE	12.51	0.84	0.91

Without PLDA Interpolation

With PLDA Interpolation,  $\alpha=0.3$

Combining i-vector adaptation and PLDA covariance matrix adaptation and can further improve performance.

# Conclusions

- We proposed two MMD-based autoencoders.
- We show the relative improvement of 11.8% EER in the NIST 2016 SRE compared to PLDA without adaptation.
- We also found that MMD-based autoencoders are more robust to unseen domains.
- In the domain robustness experiments, MMD-based autoencoders show 5.2% and 6.8% improvement over IDVC for male and female Cantonese speakers, respectively.