Autoencoder based Domain Adaptation for Speaker Recognition under Insufficient Channel Information

Suwon Shon, Seongkyu Mun[*], Wooil Kim[**], Hanseok Ko[*]

MIT Computer Science and Artificial Intelligence Laboratory,
Cambridge, MA, USA

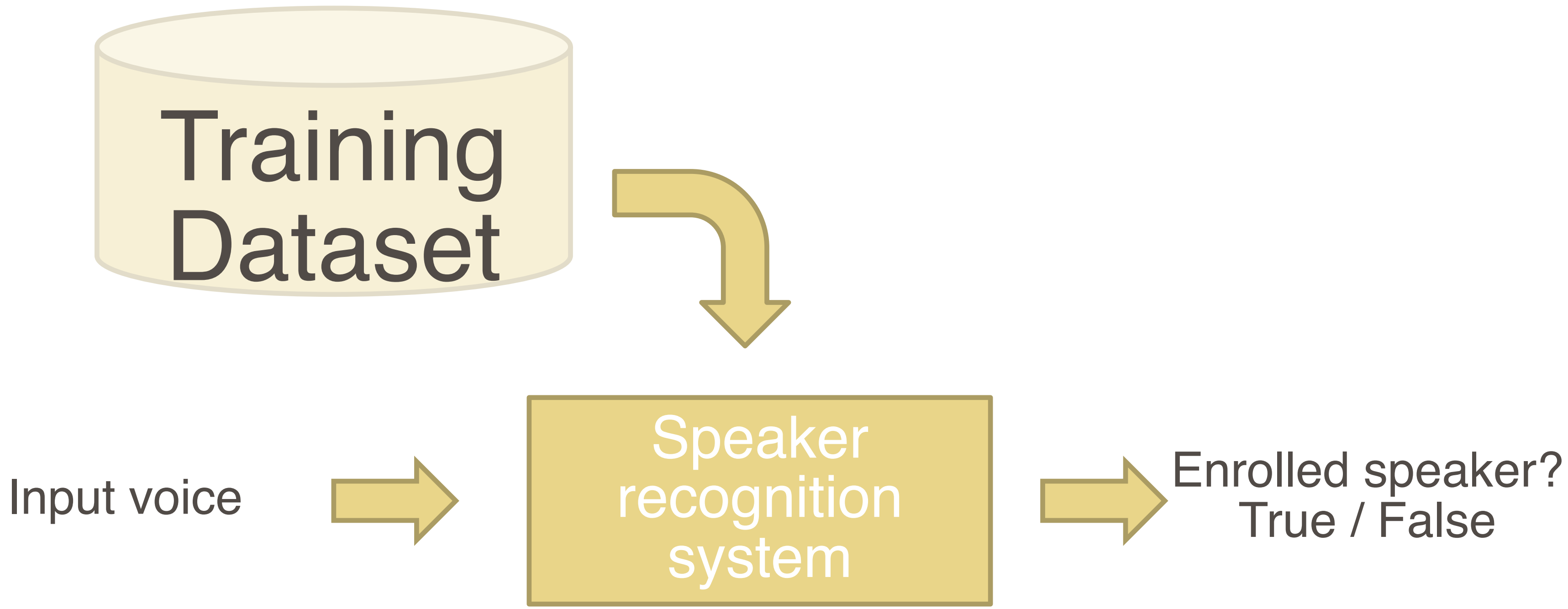
Korea University, Seoul, South Korea[*]

Incheon National University, South Korea[**]

Interspeech 2017

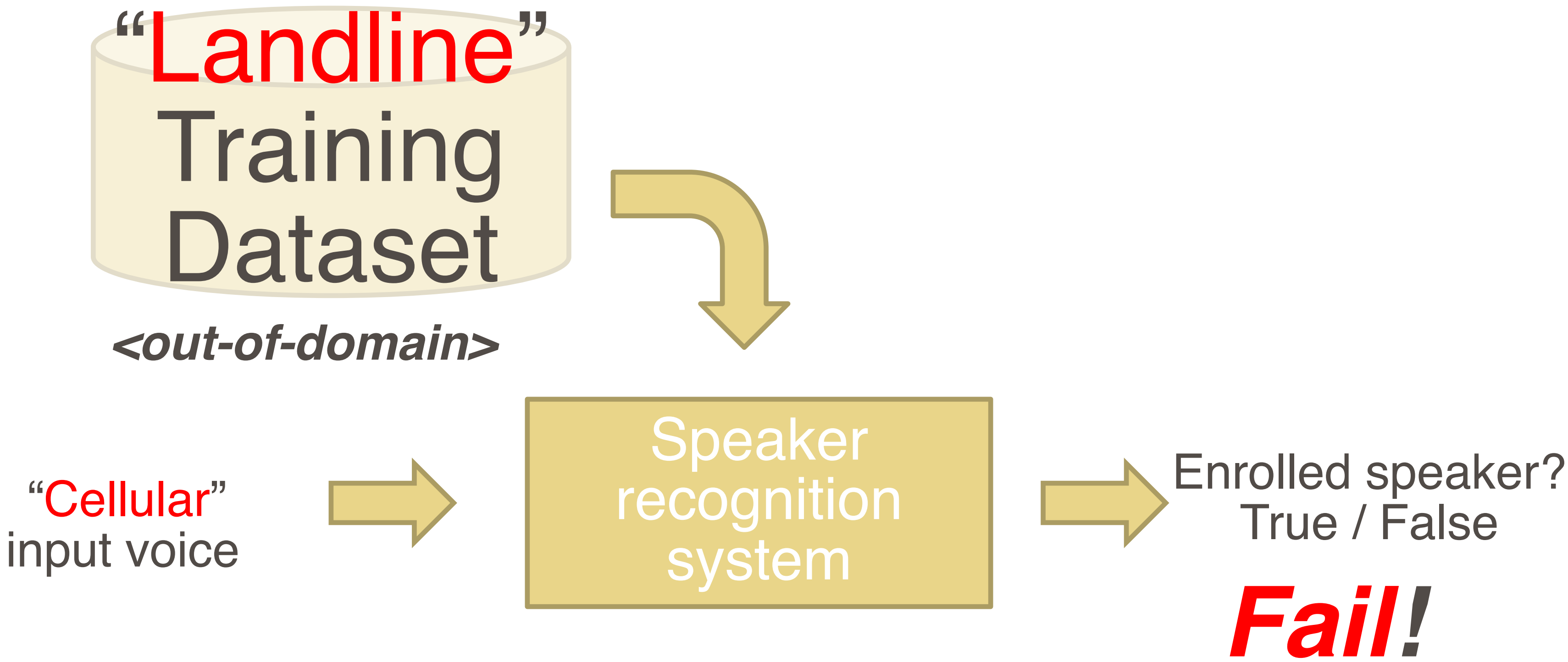
Introduction

- **Speaker recognition task**



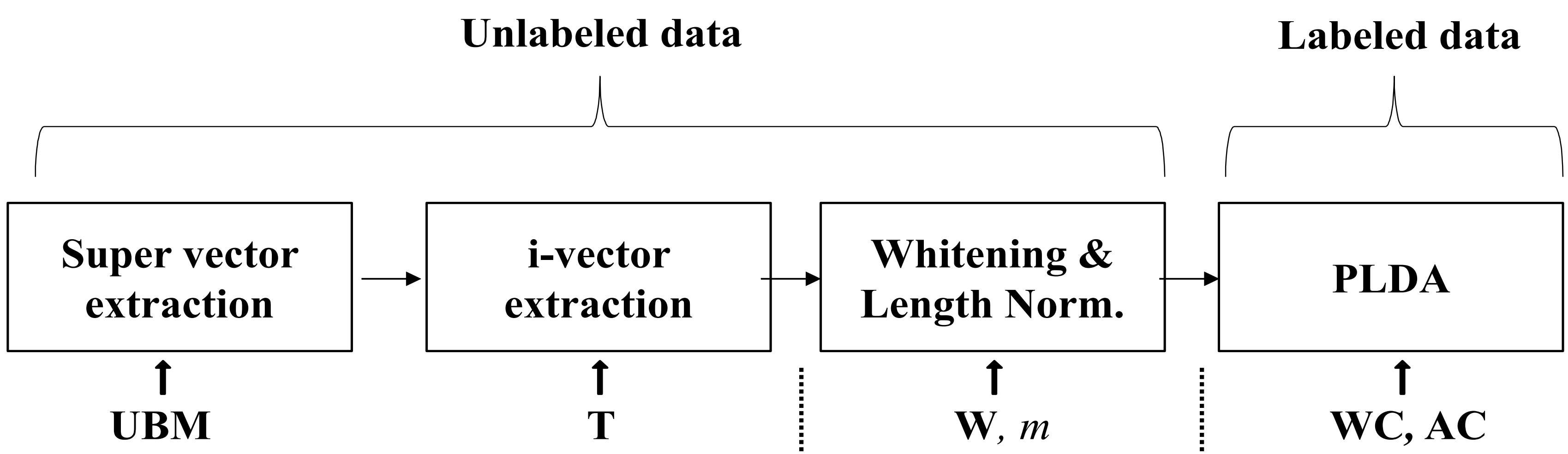
Introduction

- **Channel domain mismatched condition**



Introduction

- **Domain adaptation challenge 2013 @ JHU workshop**
  - SRE10 (evaluation) collected in 2010 (mostly cellular)
    - **7,169 target and 408,950 non-target trials**
  - SWB collected from 1992-2000 (mostly landline), mismatched
  - SRE collected from 2004-2008 (mostly cellular), matched
    - **Suppose we don't have labels on SRE**



Introduction

- **Domain adaptation challenge 2013 @ JHU workshop**
  - SRE10 (evaluation) collected in 2010 (mostly cellular)
    - **7,169 target and 408,950 non-target trials**
  - SWB collected from 1992-2000 (mostly landline), mismatched
  - SRE collected from 2004-2008 (mostly cellular), matched
    - **Suppose we don't have labels on SRE**

	Unlabeled data		Labeled data	
System #	UBM, T	W,m	WC,AC	EER
0*	SRE	SRE	SRE	2.43
1	SWB	SRE	SRE	2.33
2	SWB	SRE	SWB	5.70
3*	SWB	SWB	SWB	6.92

Domain matched benchmark (0*)

Domain mismatched (3*)

Table 2: *SRE10 Test using DAC13 i-vector set.*

S. Shum, D. a. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised Clustering Approaches for Domain Adaptation in Speaker Recognition Systems," Proc. Odyssey 2014 - Speak. Lang. Recog. Work., no. June, pp. 265–272, 2014.

Motivation

- **Insufficient Channel Information**

	SWB	SRE	SRE-1phn
#spkrs	3114	3790	3787
#calls	33039	36470	25640
Avg. #calls/spkrs	10.6	9.6	6.77
Avg. #phone_num/spkr	3.8	2.8	1

<Statistics in DAC 13 i-vector Dataset>

System #	**UBM, T**	**W**,*m*	**WC,AC**	EER
1	**SWB**	**SRE**	**SRE**	2.33
2	**SWB**	**SRE**	**SWB**	5.70
3	**SWB**	**SRE-1phn**	**SRE-1phn**	$\boxed{9.34}$
4	**SWB**	**SRE-1phn**	**SWB**	5.66

better

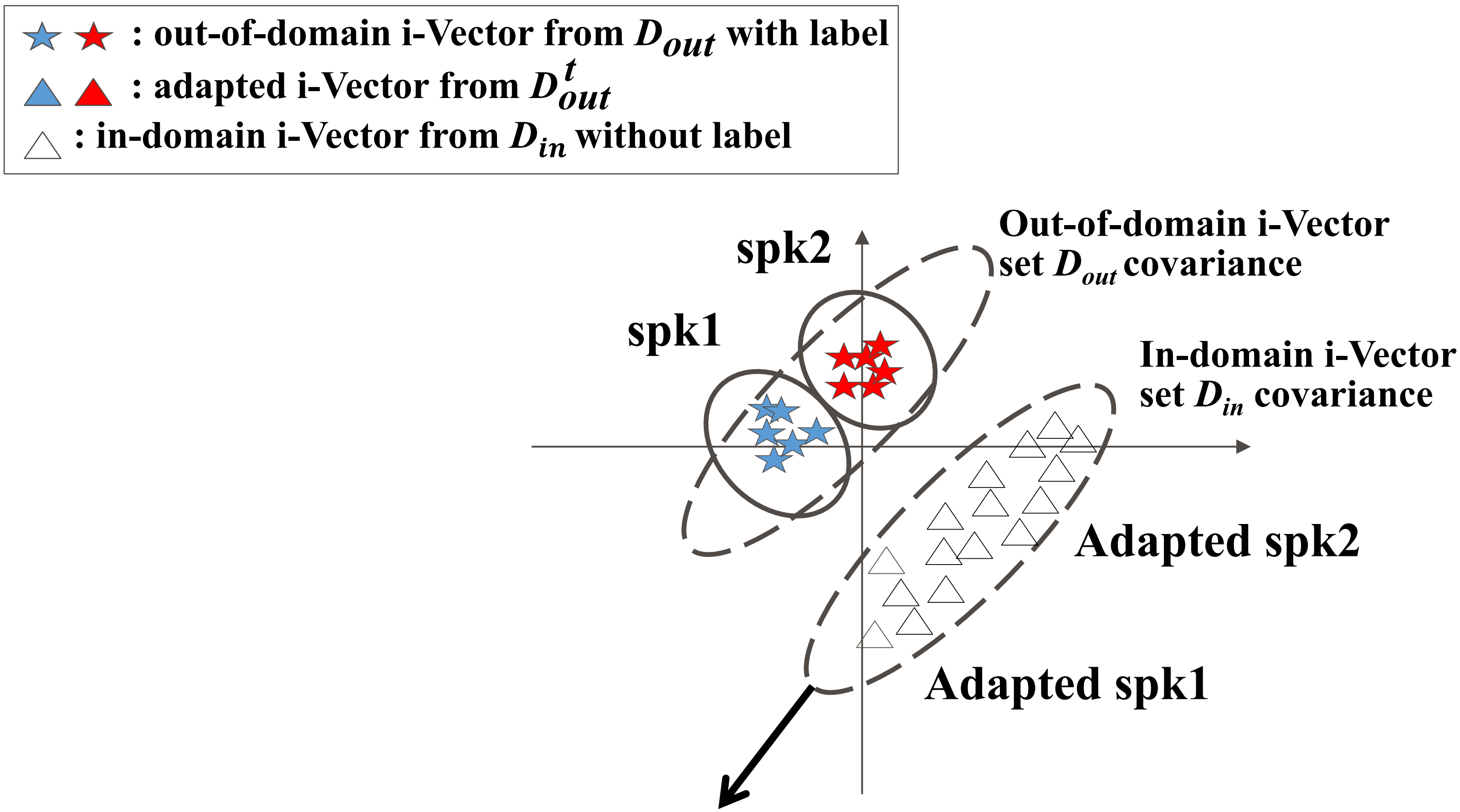
worse

<SRE10 Test using DAC13 i-vector set>

Performance degraded by Insufficient channel information although it is matched domain dataset

Proposed Approach

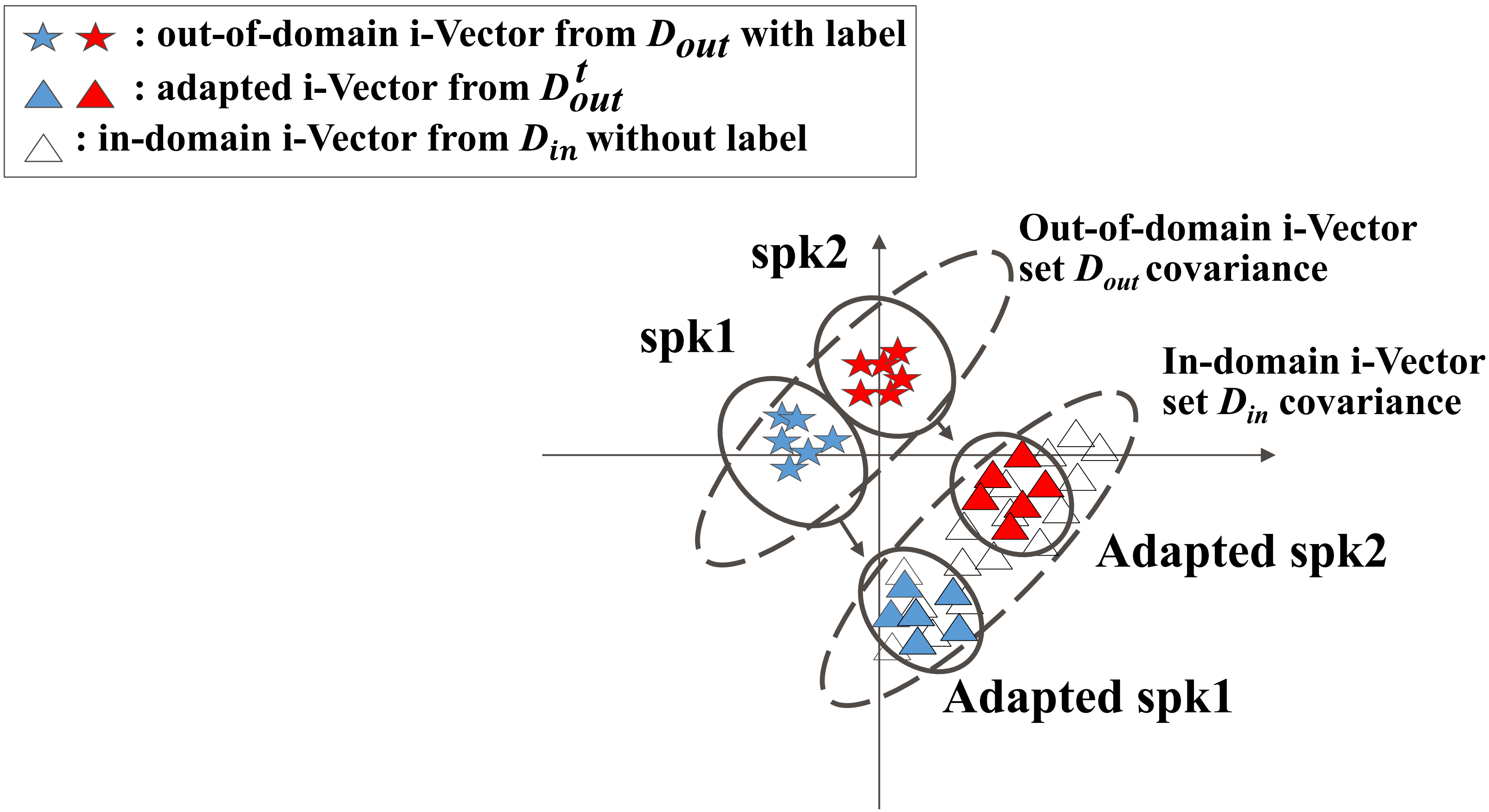
- **Auto-encoder based Domain Adaptation (AEDA)**



Useless because of insufficient channel information

Proposed Approach

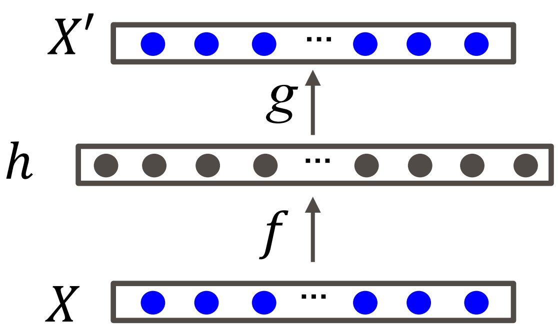
- **Auto-encoder based Domain Adaptation (AEDA)**

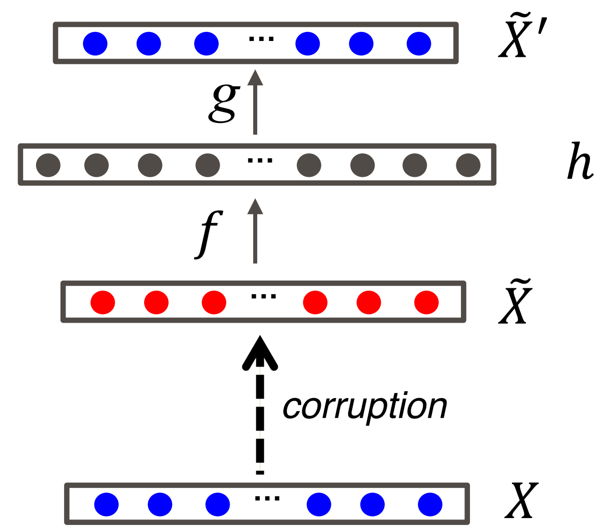


Transferring labeled out-of-domain dataset to in-domain

Proposed Approach

- **Autoencoder and Denoising Autoencoder**





$$\mathcal{L}(X, X') = ||X - X'||^2$$

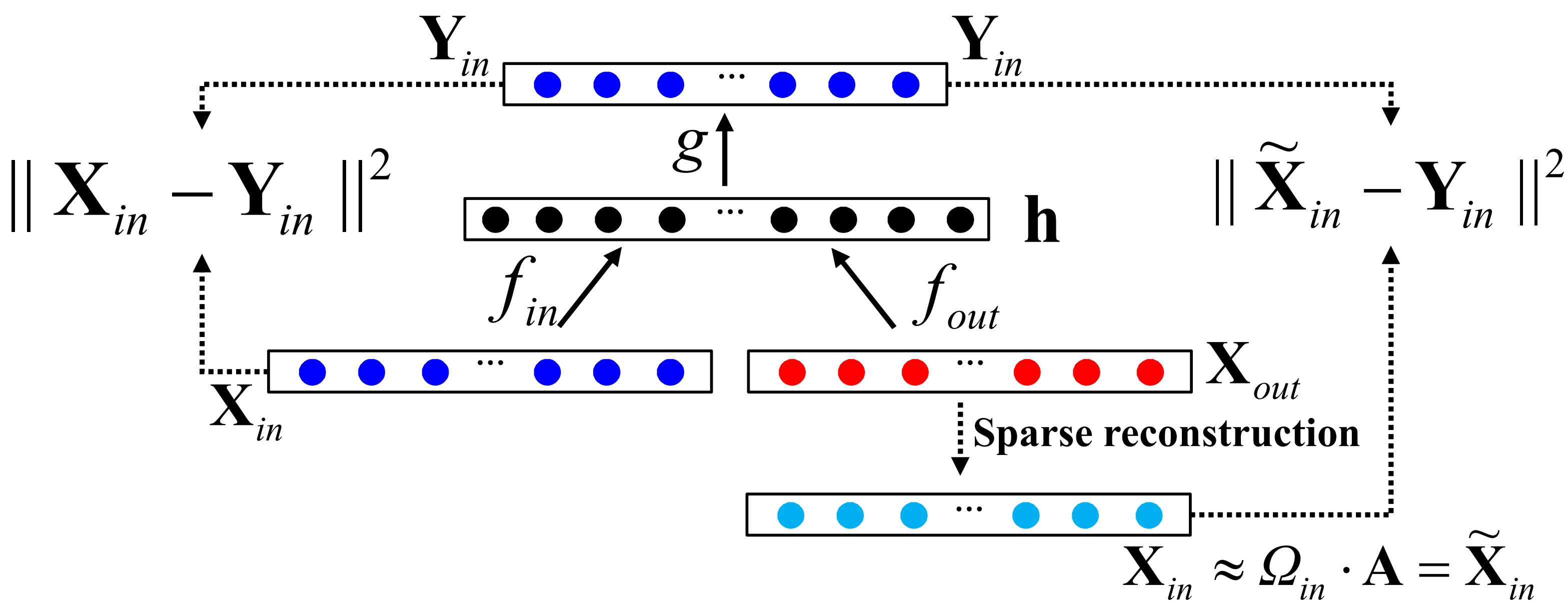
<Autoencoder>

$$\mathcal{L}(X, X') = ||X - \tilde{X}'||^2$$

<Denoising Autoencoder>

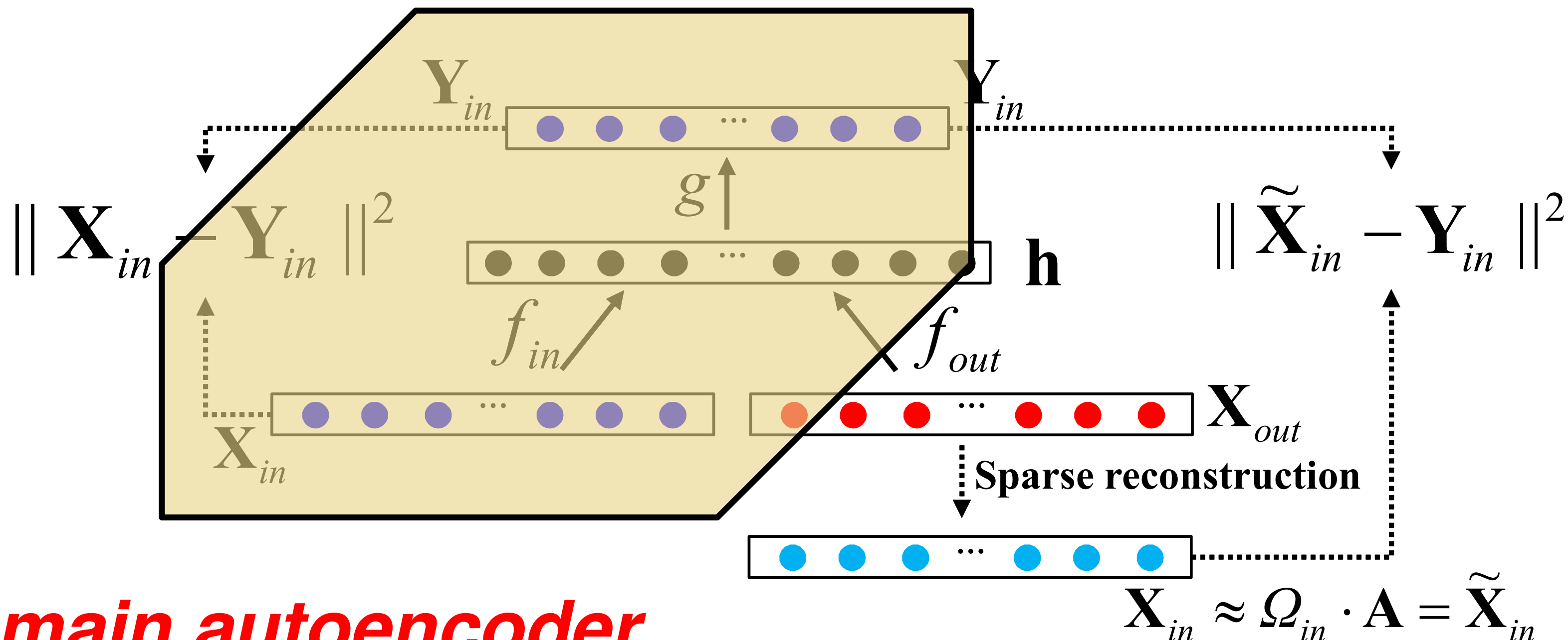
Proposed Approach

- **Auto-encoder based Domain Adaptation (AEDA)**



Proposed Approach

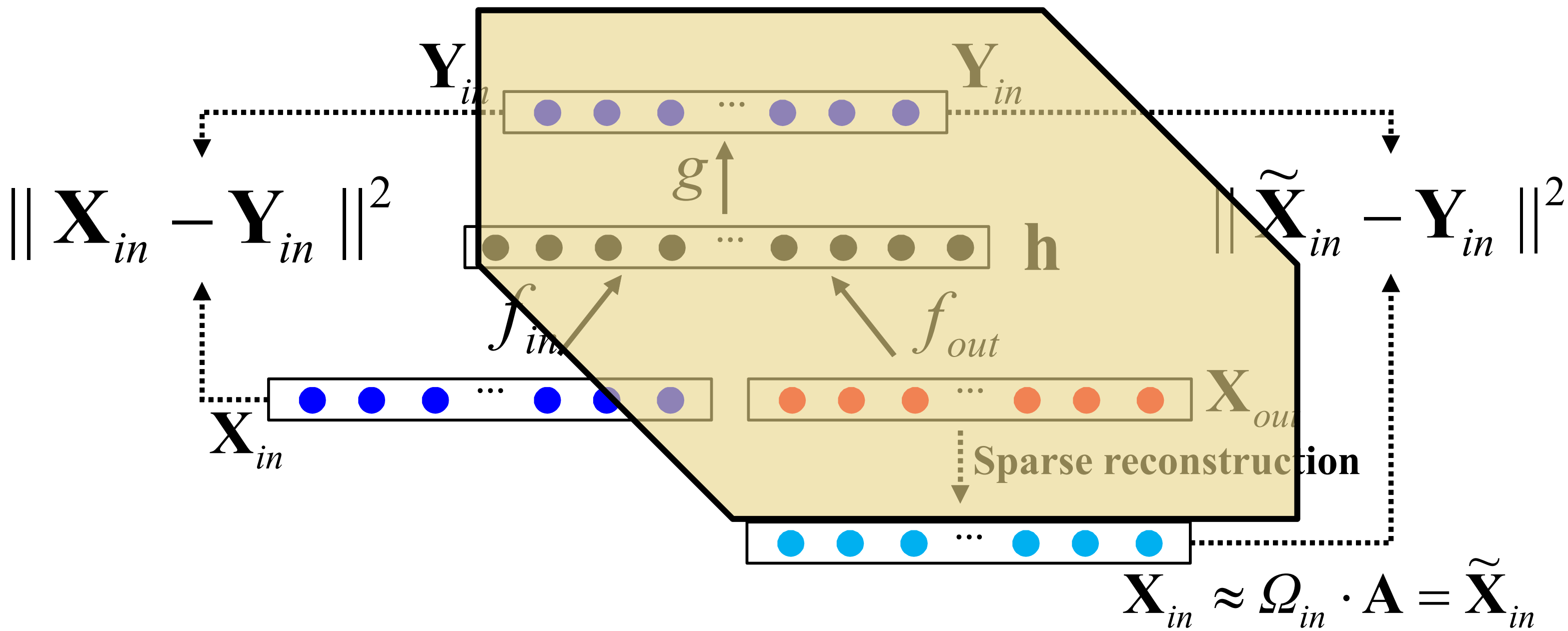
- **Auto-encoder based Domain Adaptation (AEDA)**



In-domain autoencoder
(using unlabeled in-domain dataset)

Proposed Approach

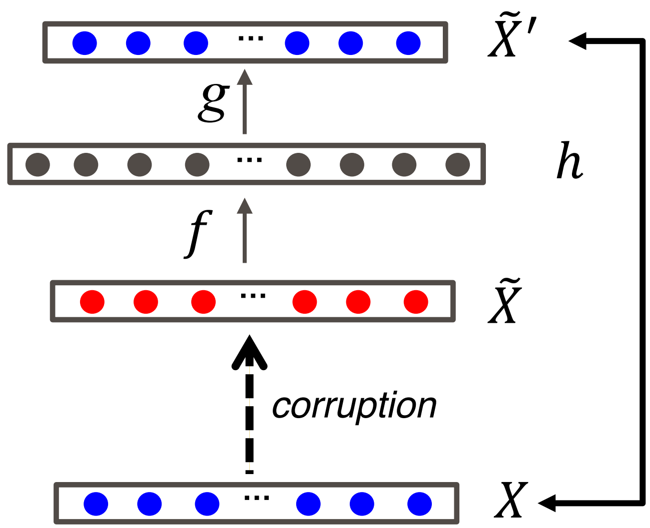
- **Auto-encoder based Domain Adaptation (AEDA)**

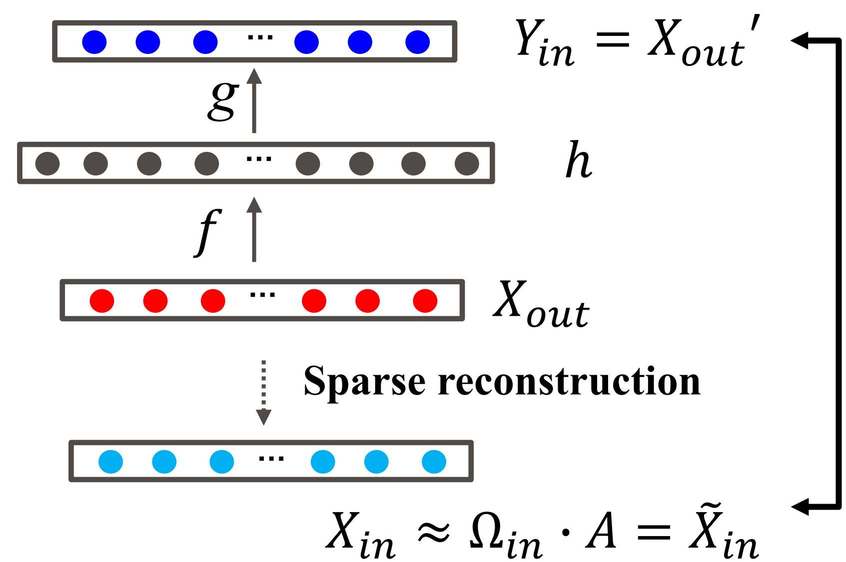


domain transferring autoencoder (using labeled out-of-domain dataset)

Proposed Approach

- **Sparse reconstruction**





Objective function : $\min_{\alpha_j} \| \Omega_{in}\alpha_j - \mathbf{y}_j^{in} \|^2 + \gamma | \alpha_j |^2$

$$\mathcal{L}(X, X') = ||X - \tilde{X}'||^2$$

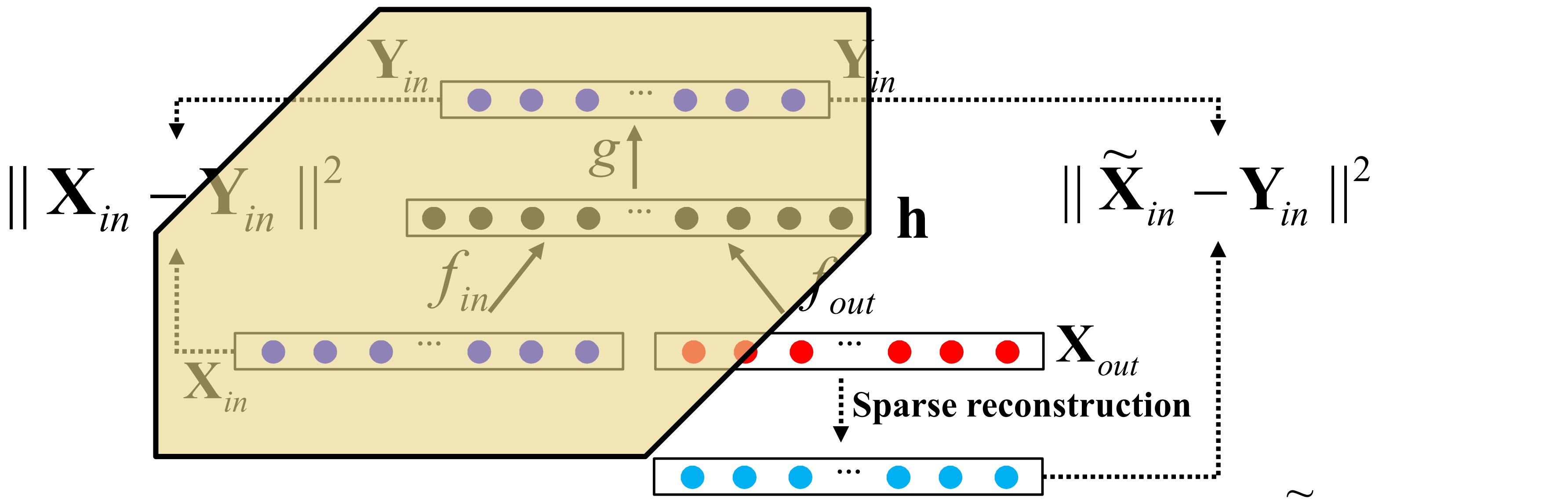
<Denoising Autoencoder>

$$\mathcal{L}(X_{in}, Y_{in}) = ||X_{in} - Y_{in}||^2 = ||\tilde{X}_{in} - Y_{in}||^2$$

<Out-of-domain transferring autoencoder>

Proposed Approach

- **Structure of Autoencoder which sharing hidden layer h**



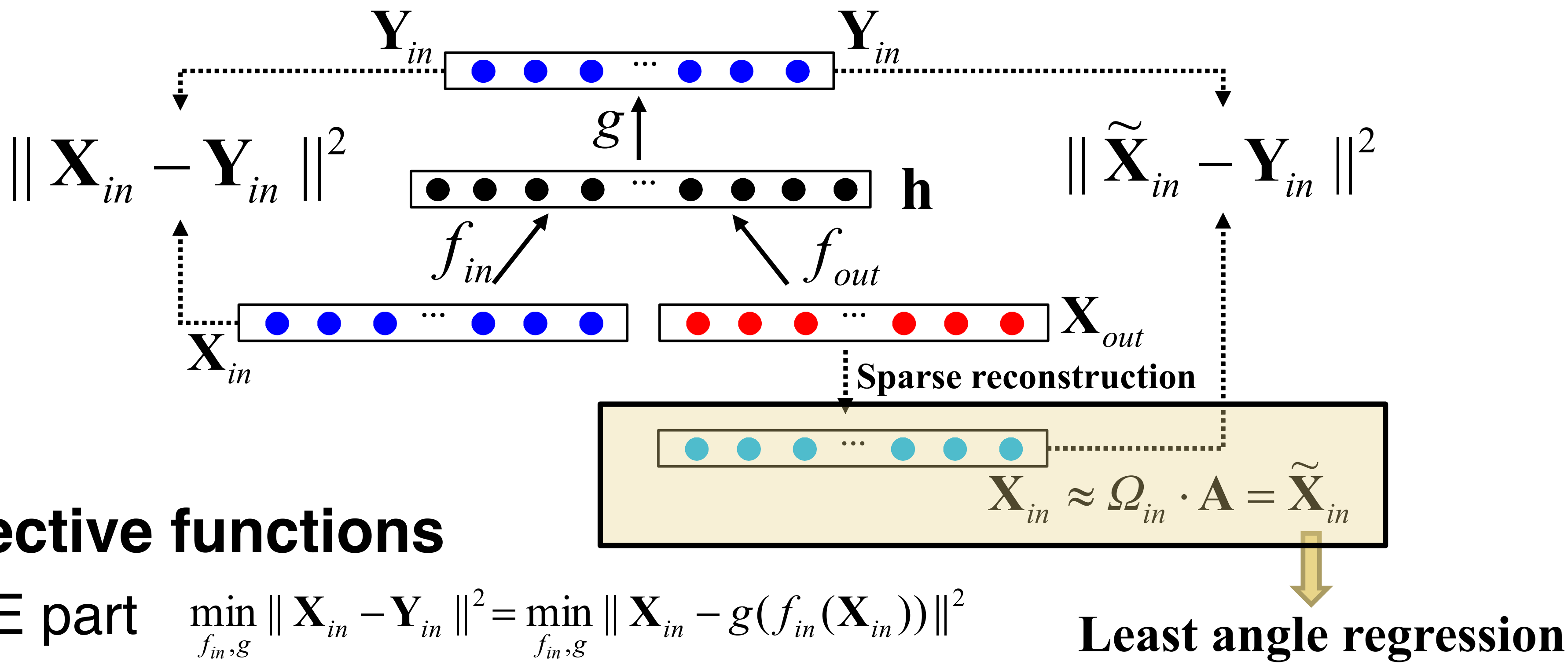
$$\mathbf{X}_{in} \approx \Omega_{in} \cdot \mathbf{A} = \widetilde{\mathbf{X}}_{in}$$

Least angle regression

- **Objective functions**
  - AE part $\min_{f_{in}, g} \| \mathbf{X}_{in} - \mathbf{Y}_{in} \|^2 = \min_{f_{in}, g} \| \mathbf{X}_{in} - g(f_{in}(\mathbf{X}_{in})) \|^2$

Proposed Approach

- **Structure of Autoencoder which sharing hidden layer h**

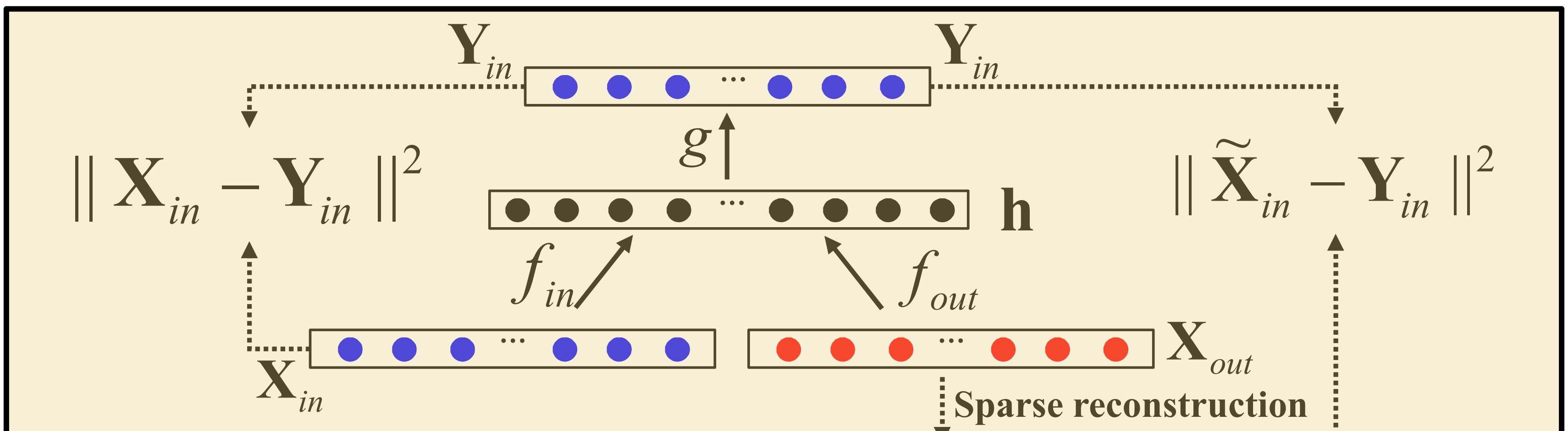


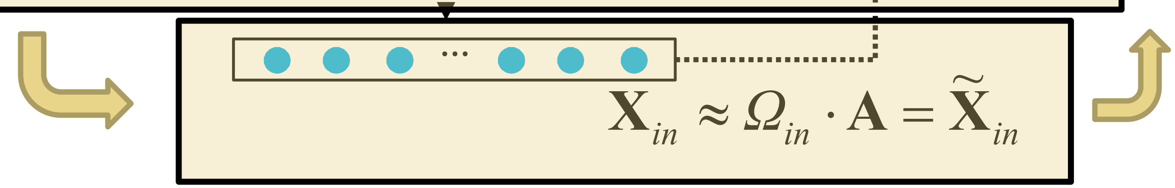
- **Objective functions**
  - AE part $\min_{f_{in},g} \| \mathbf{X}_{in} - \mathbf{Y}_{in} \|^2 = \min_{f_{in},g} \| \mathbf{X}_{in} - g(f_{in}(\mathbf{X}_{in})) \|^2$
  - Least angle regression for sparse reconstruction

$$\min_{\alpha_j} \| \Omega_{in} \alpha_j - \mathbf{y}_j^{in} \|^2 + \gamma | \alpha_j |^2$$

Proposed Approach

- **Structure of Autoencoder which sharing hidden layer h**





- **Objective functions**
  - AE part $\min_{f_{in},g} \| \mathbf{X}_{in} - \mathbf{Y}_{in} \|^2 = \min_{f_{in},g} \| \mathbf{X}_{in} - g(f_{in}(\mathbf{X}_{in})) \|^2$
  - DAE part $\min_{f_{out},g} \| \mathbf{X}_{in} - \mathbf{Y}_{in} \|^2 = \min_{f_{out},g} \| \mathbf{X}_{in} - g(f_{out}(\mathbf{X}_{out})) \|^2$
  - AEDA $\min_{f_{in},f_{out},g} \| \mathbf{X}_{in} - g(f_{in}(\mathbf{X}_{in})) \|^2 + \| \tilde{\mathbf{X}}_{in} - g(f_{out}(\mathbf{X}_{out})) \|^2$

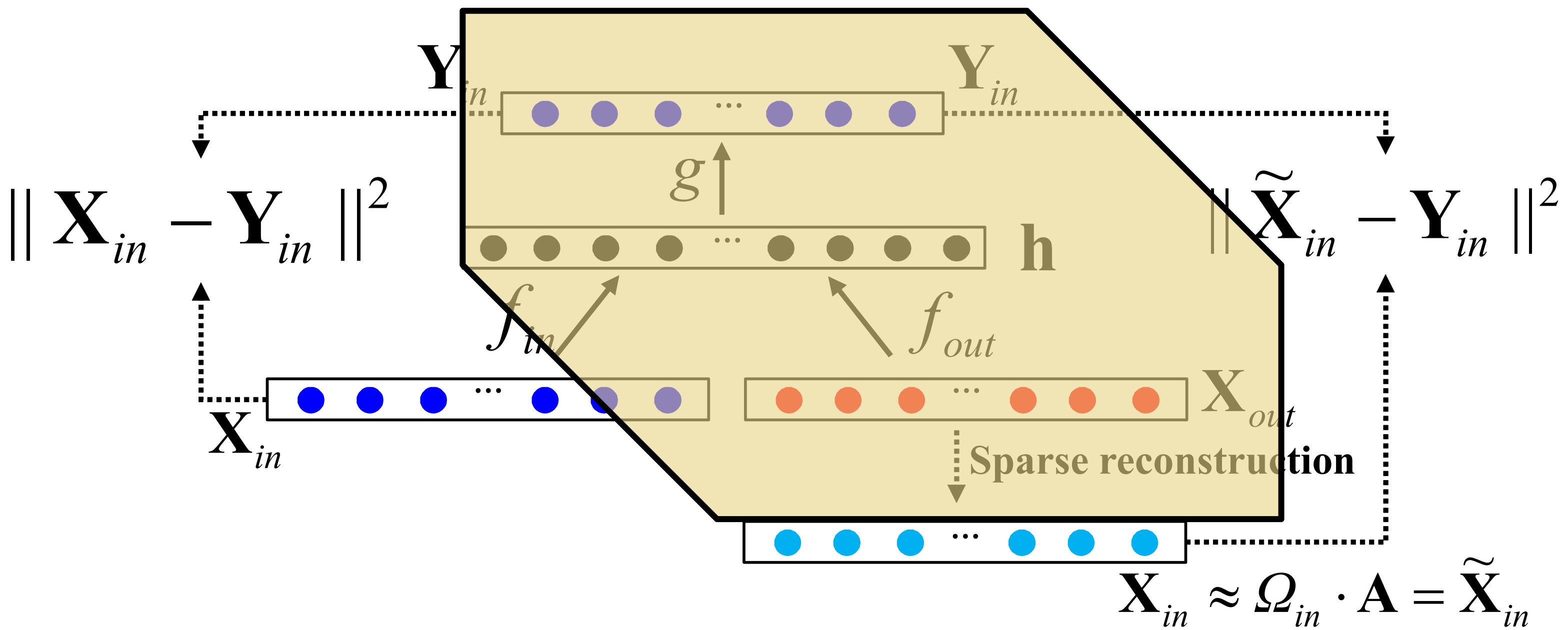
Training network

$$\min_{\alpha_j} \| \Omega_{in} \alpha_j - \mathbf{y}_j^{in} \|^2 + \gamma | \alpha_j |^2$$

Sparse reconstruction

Proposed Approach

- **Structure of Autoencoder which sharing hidden layer h**



- **AEDA**
  - 600 dim i-vector with 1000 hidden node with learning rate 0.005
- **Sparse reconstruction**
  - Least Angle Regression(LARS)
  - Sparsity 0.01
  - Random 1500 spk i-vector for in-domain dictionary Ω_{in}
- **Performance**
  - Using PLDA with 400 eigenvoice after 400 dim LDA transform
  - EER, DCF10, DCF08

Experimental result

- **Auto-encoder based Domain Adaptation (AEDA)**

#	Adaptation & Compensation	**WC,AC**	EER	DCF10	DCF08
3	-	SRE-1phn	9.34	0.721	0.520
4	-	SWB	5.66	0.633	0.426
5	Interpolated [13]	SWB + SRE-1phn	6.55	0.652	0.454
6	IDV [15]	IDV-SWB	6.15	0.676	0.476
7	DICN [16]	DICN-SWB	4.99	0.623	0.416
8	DAE [23]	DAE-SWB	4.81	0.610	0.398
9	**AEDA**	**AEDA-SWB**	**4.50**	**0.589**	**0.362**

<SRE10 evaluation result with DAC 13 Dataset when Unlabeled In-Domain Dataset is Available >

Conclusion

- **Only small subset of unlabeled in-domain dataset is used for domain adaptation**
- **Insufficient channel information dataset is effectively used for transferring knowledge of in-domain**
- **Domain transferring autoencoder part of AEDA can be trained using sparse reconstruction without actual pair of in-domain and out-of-domain**

Q & A

- **Thanks!**

- **Domain related paper :**
  Suwon Shon, Seongkyu Mun and Hanseok Ko,
  “Recursive whitening transformation for speaker recognition on Language Mismatched Condition”
  @ 4.9 Evaluation of Speaker and language identification systems session, Wednesday 10:00~12:00