THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Disentangled Speaker Embedding for Robust Speaker Verification

Lu YI, Man-Wai MAK

Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong SAR of China

ICASSP'22

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

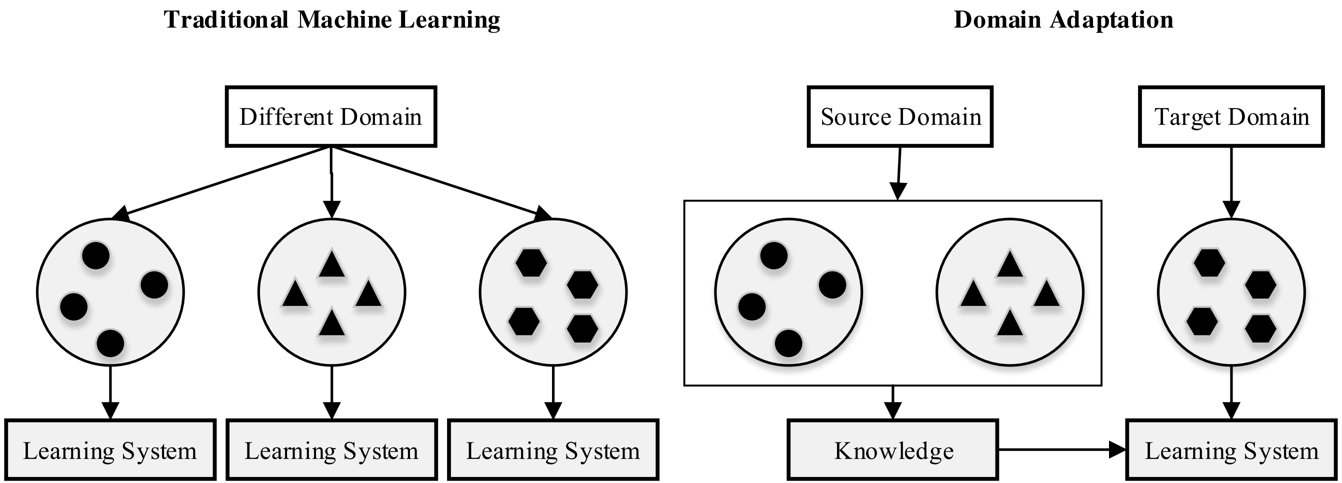
Contents

Domain Mismatch

- Domain mismatch occurs when speech is collected from different acoustic environments.
- For example, there is a domain mismatch between near-field microphone speech and far-field microphone speech due to the difference in microphone characteristics.
- This mismatch can make a speaker verification system trained on near-field microphone speech perform poorly on far-field microphone speech.
- Collecting more data to retrain the system is time-consuming and computationally-expensive.

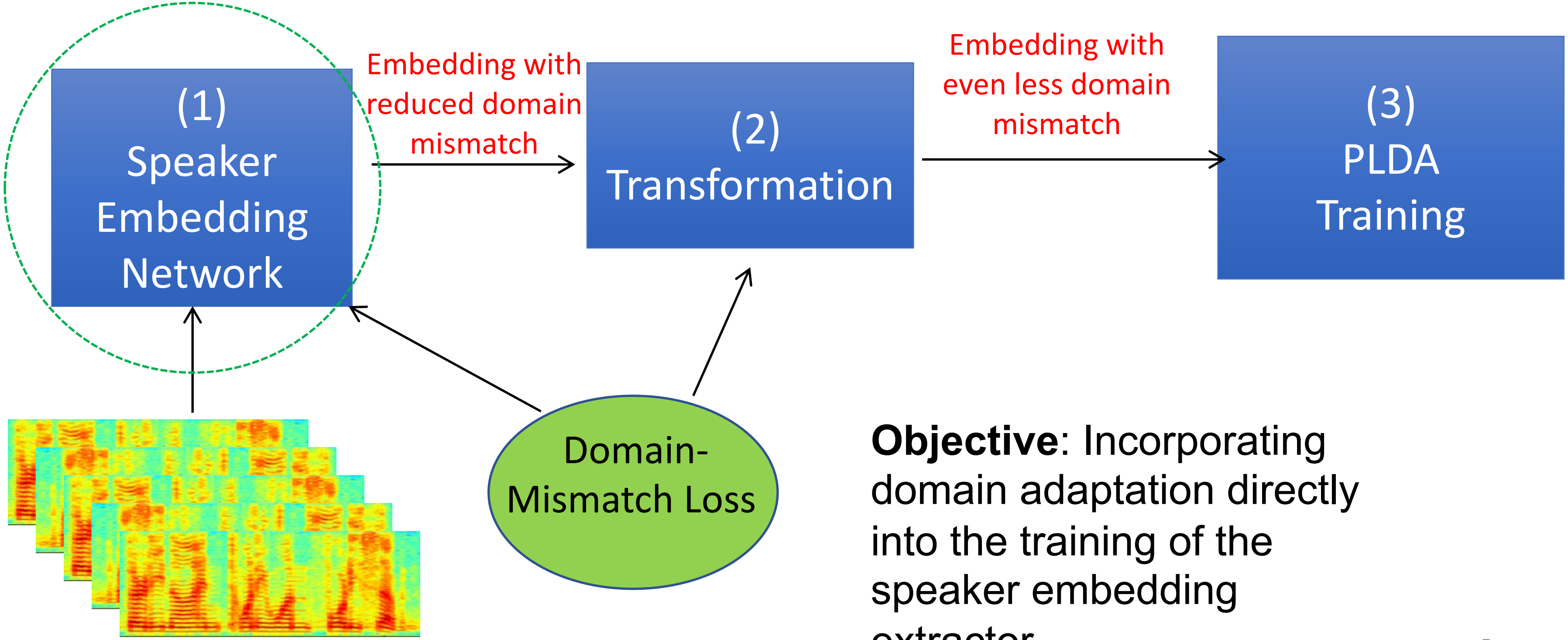
THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

Domain Adaptation



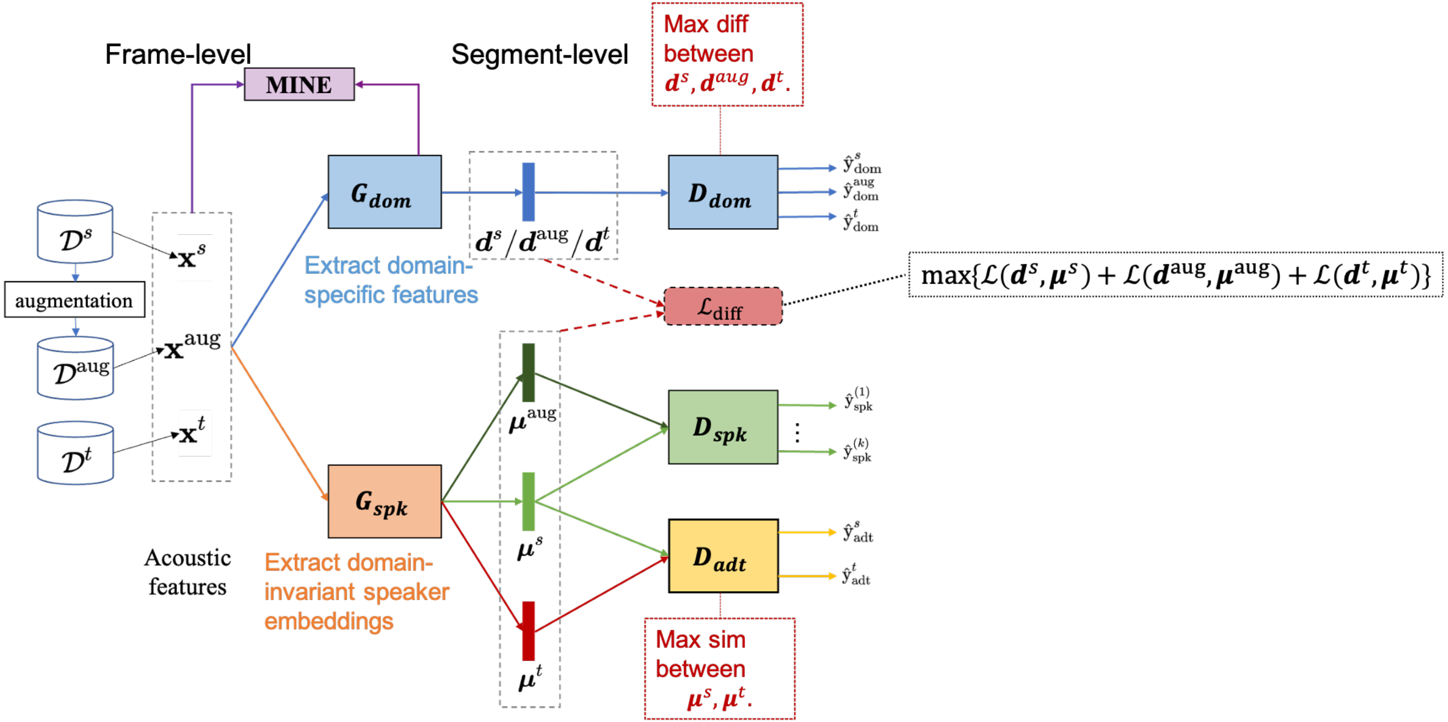
Source: M. Wang, & W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, *312*, 2018, pages 135-153.

Domain Adaptation

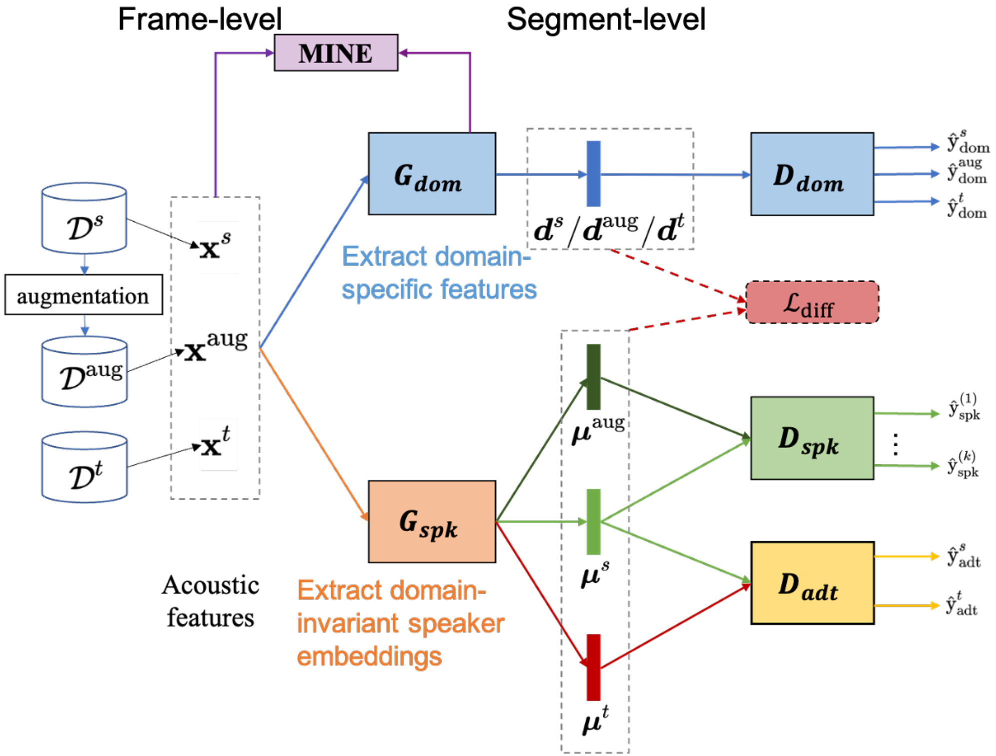


Objective: Incorporating domain adaptation directly into the training of the speaker embedding extractor.

InfoMax Domain Separation and Adaptation Network



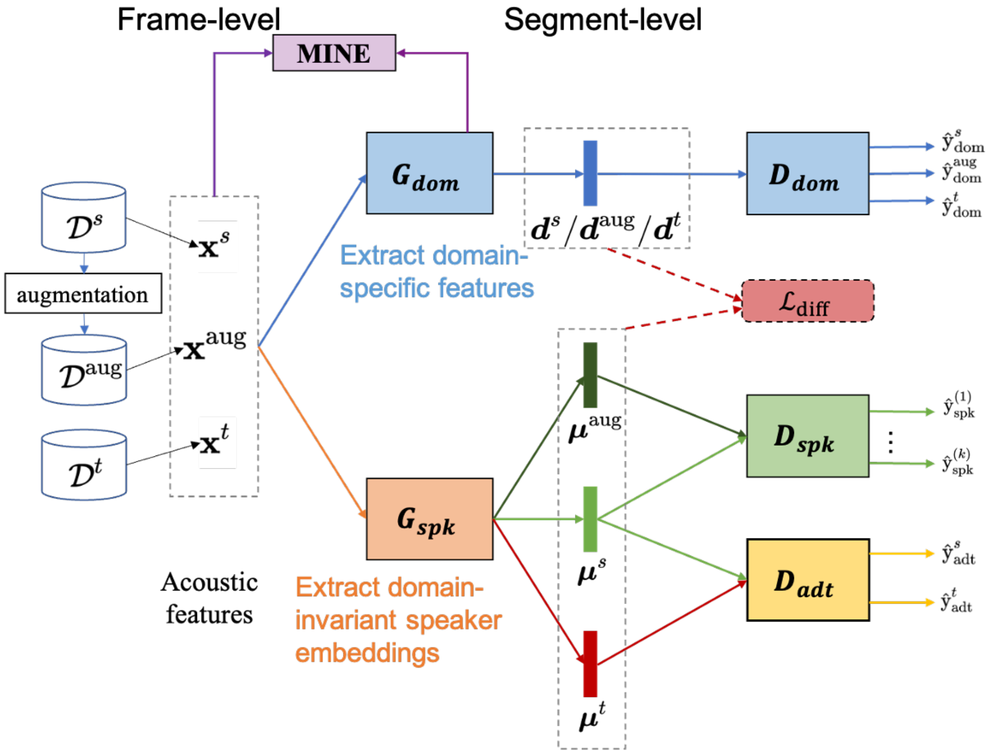
InfoMax Domain Separation and Adaptation Network



$$
\begin{aligned}
\mathcal{L}_{\text{dom}} = \; & \mathbb{E}_{\mathbf{x}^t \sim \mathcal{D}^t}\left[-\log D_{\text{dom}}\left(G_{\text{dom}}\left(\mathbf{x}^t\right)\right)_0\right] \\
& + \mathbb{E}_{\mathbf{x}^s \sim \mathcal{D}^s}\left[-\log D_{\text{dom}}\left(G_{\text{dom}}\left(\mathbf{x}^s\right)\right)_1\right] \\
& + \mathbb{E}_{\mathbf{x}^{\text{aug}} \sim \mathcal{D}^{\text{aug}}}\left[-\log D_{\text{dom}}\left(G_{\text{dom}}\left(\mathbf{x}^{\text{aug}}\right)\right)_2\right].
\end{aligned}
$$

Subscripts 0, 1, and 2 correspond to the target, source, and augmented-source, respectively.

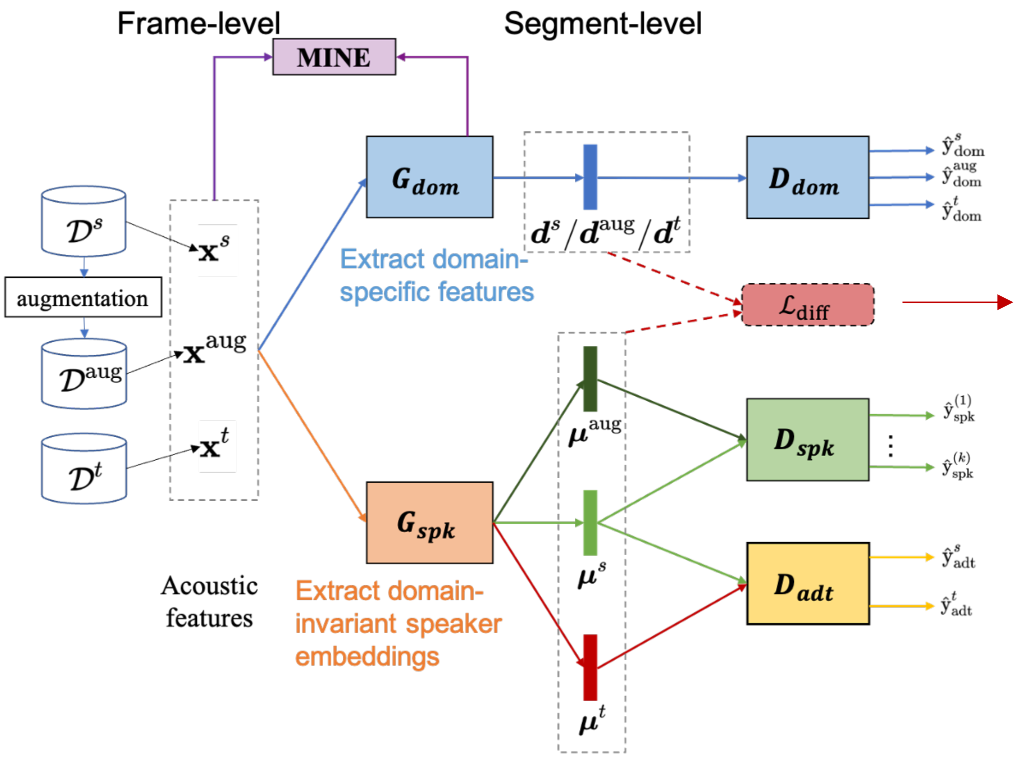
InfoMax Domain Separation and Adaptation Network



$$\mathcal{L}_{\text{spk}} = \mathbb{E}_{\mathbf{x}^s\sim\mathcal{D}^s}\left[-\sum_{k=1}^{K} y^{(k)}_{\text{spk}} \log D_{\text{spk}}\left(G_{\text{spk}}\left(\mathbf{x}^s\right)\right)_k\right] + \mathbb{E}_{\mathbf{x}^{\text{aug}}\sim\mathcal{D}^{\text{aug}}}\left[-\sum_{k=1}^{K} y^{(k)}_{\text{spk}} \log D_{\text{spk}}\left(G_{\text{spk}}\left(\mathbf{x}^{\text{aug}}\right)\right)_k\right]$$

$$\min_{G_{\text{spk}}}\max_{D_{\text{adt}}} \mathcal{L}_{\text{adt}} = \mathbb{E}_{\mathbf{x}^s\sim\mathcal{D}^s}\left[\log D_{\text{adt}}\left(G_{\text{spk}}\left(\mathbf{x}^s\right)\right)\right] + \mathbb{E}_{\mathbf{x}^t\sim\mathcal{D}^t}\left[\log\left[1 - D_{\text{adt}}\left(G_{\text{spk}}\left(\mathbf{x}^t\right)\right)\right]\right].$$

InfoMax Domain Separation and Adaptation Network



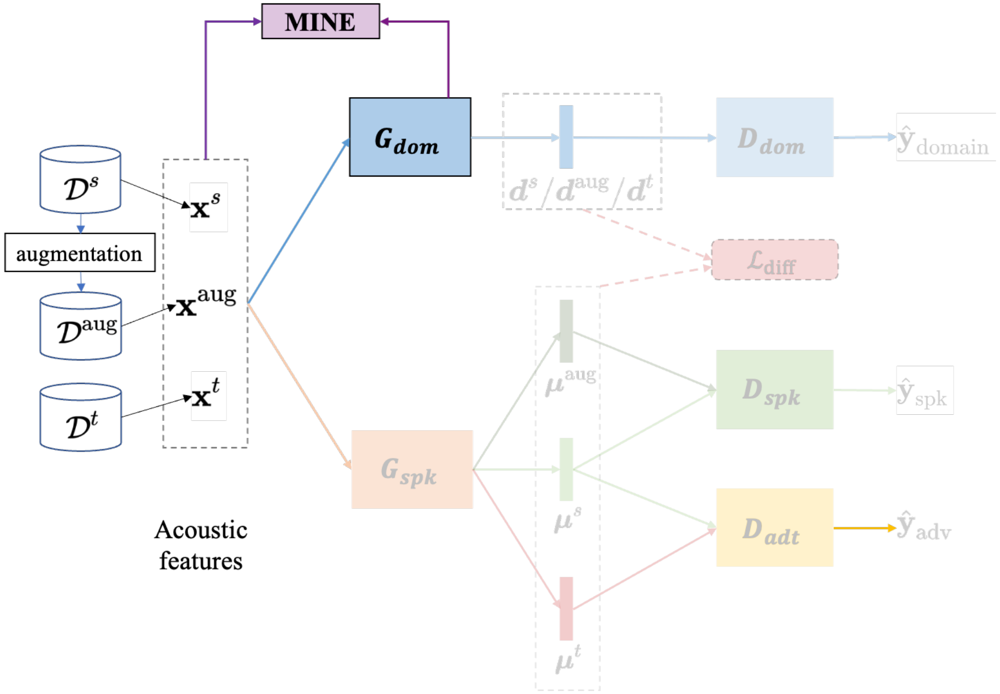
Orthogonality-based:

$$\mathcal{L}_{\text{diff}} = \sum_i \left| (\boldsymbol{\mu}_i^s)^\top \mathbf{d}_i^s \right| + \sum_j \left| \left(\boldsymbol{\mu}_j^{\text{aug}}\right)^\top \mathbf{d}_j^{\text{aug}} \right| + \sum_k \left| (\boldsymbol{\mu}_k^t)^\top \mathbf{d}_k^t \right|$$

MI-based:

$$\mathcal{L}_{\text{diff}} = -I_\Theta^{\text{JS}}(\boldsymbol{\mu}^s, \mathbf{d}^s) - I_\Theta^{\text{JS}}(\boldsymbol{\mu}^{\text{aug}}, \mathbf{d}^{\text{aug}}) - I_\Theta^{\text{JS}}(\boldsymbol{\mu}^t, \mathbf{d}^t)$$

Mutual Information Neural Estimator



Extract informative features.

Consistency

Effectiveness

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Mutual Information Neural Estimator

- Mutual information neural estimator utilizes a deep neural network with parameters $\theta \epsilon \Theta$ to find a lower bound of the mutual information:

$$I(X;Z) \geq I_{\Theta}(X,Z),$$

$$I_{\Theta}(X,Z) = \sup_{\theta \in \Theta} \underbrace{\mathbb{E}_{\mathbb{P}_{XZ}}}_{\text{joint distribution}} [T_\theta] - \log\left(\underbrace{\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}}_{\text{product of the marginal distributions}}\left[e^{T_\theta}\right]\right)$$

$$\boxed{I(X,Z) = D_{KL}(\mathbb{P}_{XZ} \| \mathbb{P}_X \otimes \mathbb{P}_Z)}$$

Source: M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, D. Hjelm, and A. Courville, “Mutual information neural estimation,” in *International Conference on Machine Learning*, 2018, pp. 530–539.

Mutual Information Neural Estimator

- Mutual information neural estimator utilizes a deep neural network with parameters $\theta \epsilon \Theta$ to find a lower bound of the mutual information:

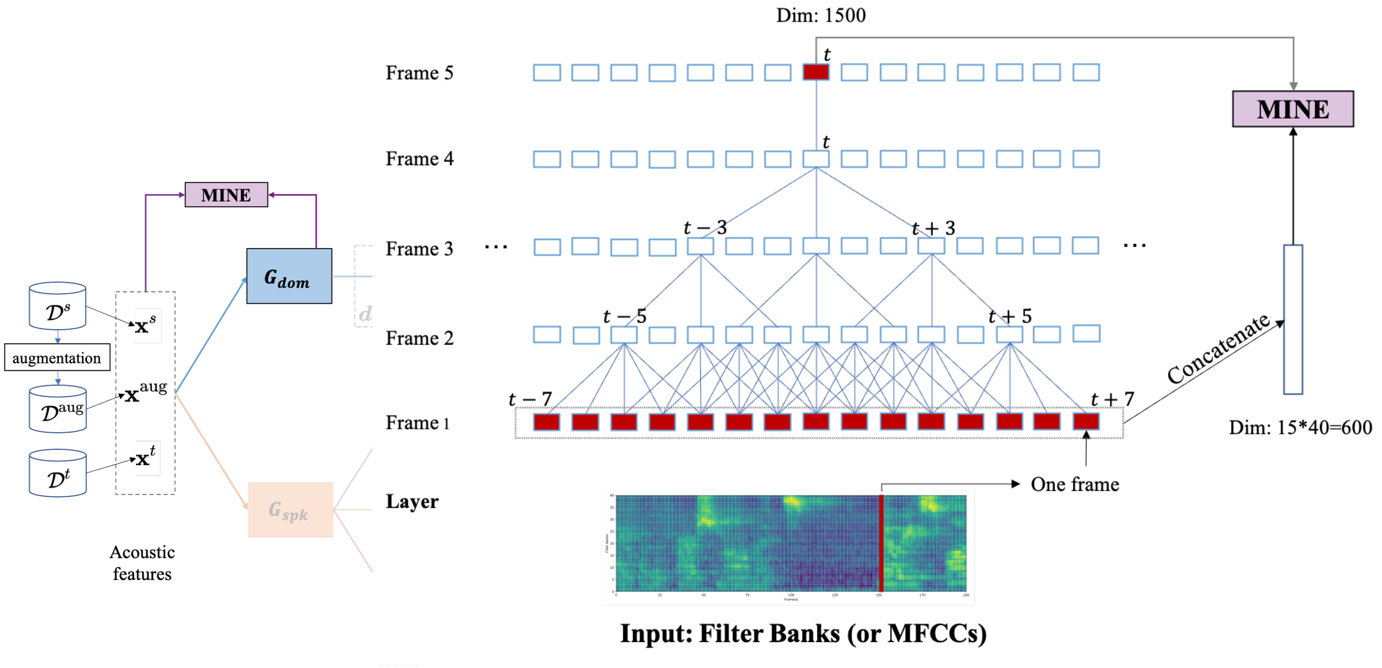
$$\boxed{I(X, Z) = D_{KL}(\mathbb{P}_{XZ} || \mathbb{P}_X \otimes \mathbb{P}_Z)}$$

$$\Big\uparrow \text{Approximate}$$

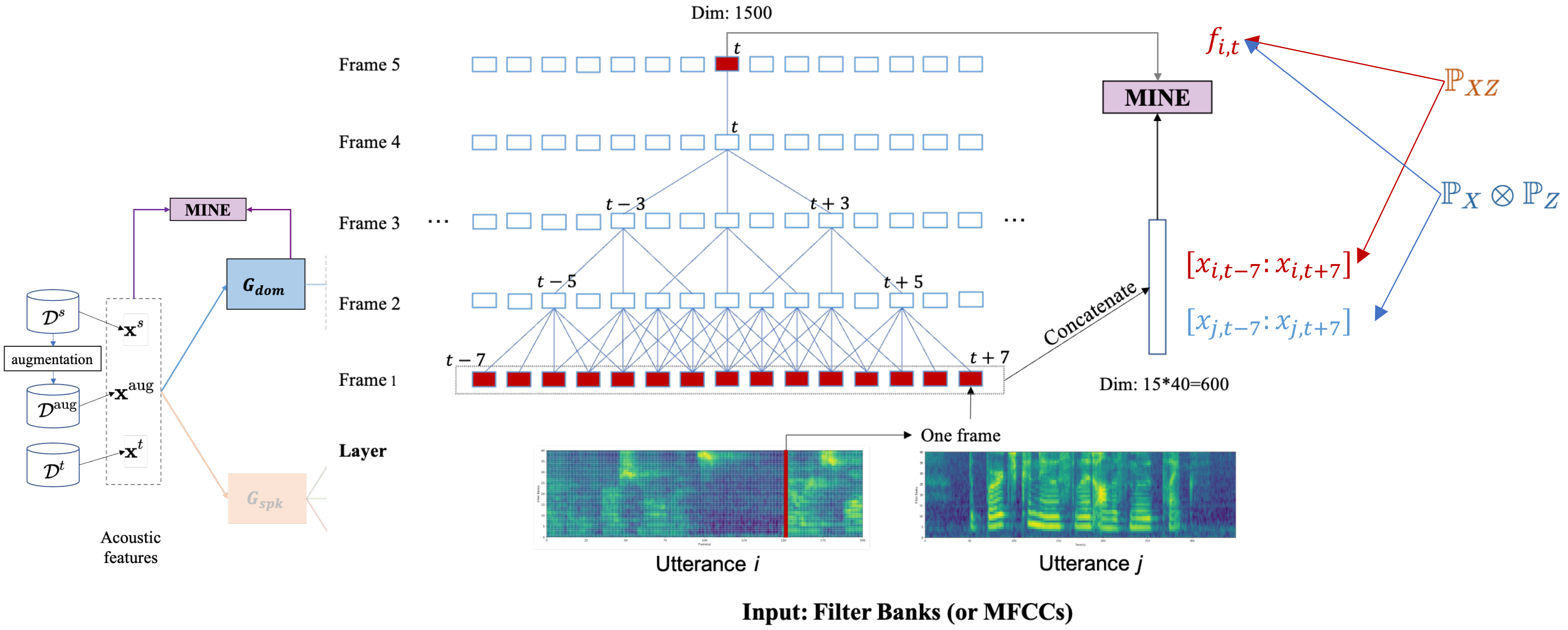
$$I_{\Theta}^{\text{JS}}(X, Z) = \mathbb{E}_{\mathbb{P}_{XZ}}[-\operatorname{sp}(-T_\theta)] - \mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}[\operatorname{sp}(T_\theta)]$$

Source: R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.

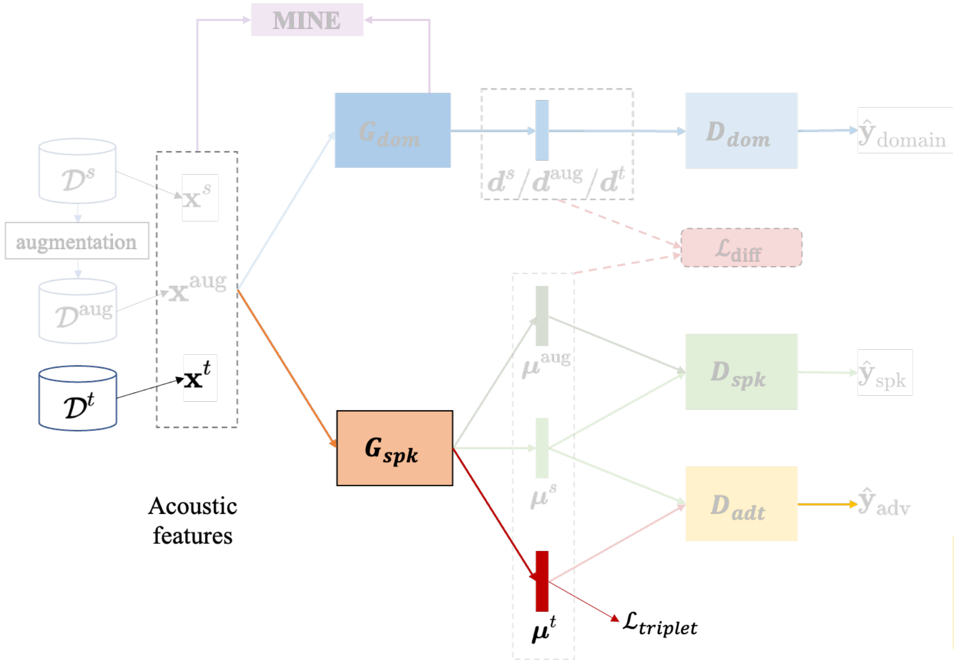
Frame-based Mutual Information Neural Estimation



Frame-based Mutual Information Neural Estimation



Self-supervised Learning



Objective: To minimize the distance between an anchor and a positive sample and maximize the distance between an anchor and a negative sample

Positive pair: segments from the same utterance

Negative pair: segments from different utterances (find the closest segment for each anchor within a batch)

Positive pair: x_i^t(frame 0-200), x_i^t(frame 200-400)
Negative pair: x_i^t (frame 0-200), $x_j^{s/aug}$

Experiments

- **Source domain data $\mathcal{D}^s$:**
  - VoxCeleb1 dev, VoxCeleb2 dev & test
  - ~2.2M utterances spoken by 7,323 speakers

- **Augmented source domain data $\mathcal{D}^{aug}$:**
  - by adding noise, babble, and music from MUSAN and reverberation from the RIR dataset to speech in $\mathcal{D}^s$

- **Target domain data (unlabeled) $\mathcal{D}^t$:**
  - VOiCES Challenge 2019 development set

- **Evaluation set:**
  - VOiCES Challenge 2019 development and evaluation set

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Results

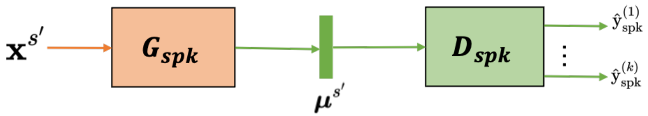
Row	System	Source domain	Target domain	MINE	$\mathcal{L}_{\text{triplet}}$	D_{adt}	$\mathcal{L}_{\text{diff}}$	VOiCES dev.		VOiCES eval.	
								EER(%)	minDCF	EER(%)	minDCF
1	TDNN [4]	$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	×	×	✓	×	3.18	-	7.15	-
2	TDNN (G_{spk})	$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	×	×	×	×	×	2.35	0.2596	6.42	0.4398
3		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	×	×	2.16	0.2627	6.32	0.4305
4		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	×	✓	×	×	×	2.29	0.2453	5.98	0.4351
5		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	✓	×	2.31	0.2645	6.29	0.4399
6		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	✓	✓	×	2.54	0.2671	6.23	0.4379
7	InfoMax–DSAN (G_{dom}&G_{spk})	$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	✓	×	✓	MI	2.17	0.2418	5.93	0.4265
8		$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	✓	×	✓	ort	2.33	0.2577	5.98	0.4131
9		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	×	✓	MI	3.29	0.3278	6.75	0.4614
10		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	✓	MI	2.31	0.2490	6.02	0.4161
11		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	✓	✓	MI	**2.06**	**0.2375**	**5.69**	**0.4127**

[4]: J. Huang and T. Bocklet, “Intel far-field speaker recognition system for VOiCES challenge 2019,” in *Proc. Interspeech*, 2019, pp. 2473–2477.

Results

Row	System	Source domain	Target domain	MINE	$\mathcal{L}_{\text{triplet}}$	D_{adt}	$\mathcal{L}_{\text{diff}}$	VOiCES dev. EER(%)	VOiCES dev. minDCF	VOiCES eval. EER(%)	VOiCES eval. minDCF
1	TDNN [4]	$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	×	×	✓	×	3.18	-	7.15	-
2		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	×	×	×	×	×	2.35	0.2596	6.42	0.4398
3		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	×	×	2.16	0.2627	6.32	0.4305
4	TDNN (G_{spk})	$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	×	✓	×	×	×	2.29	0.2453	5.98	0.4351

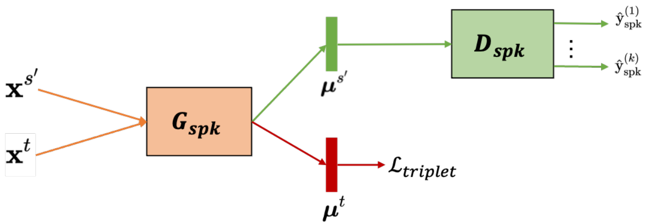
2



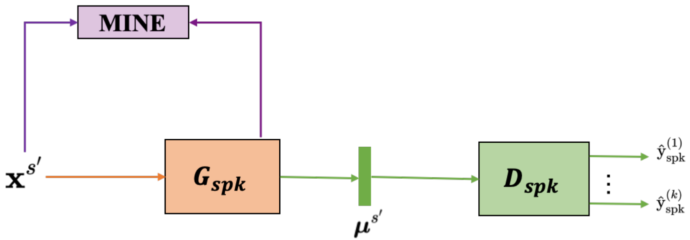
$\mathbf{x}^{s'} \in \{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$

$\mathbf{x}^{t} \in \mathcal{D}^t$

3

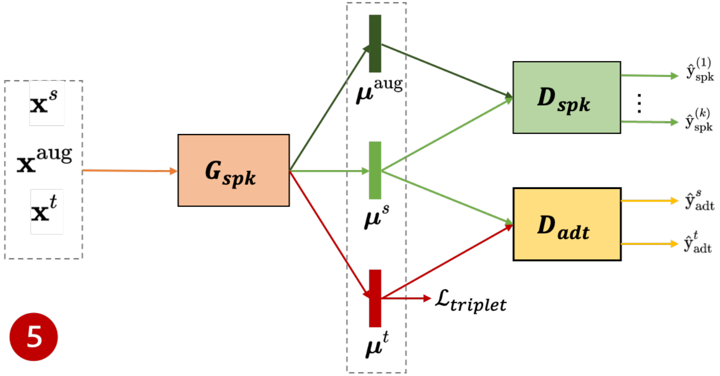


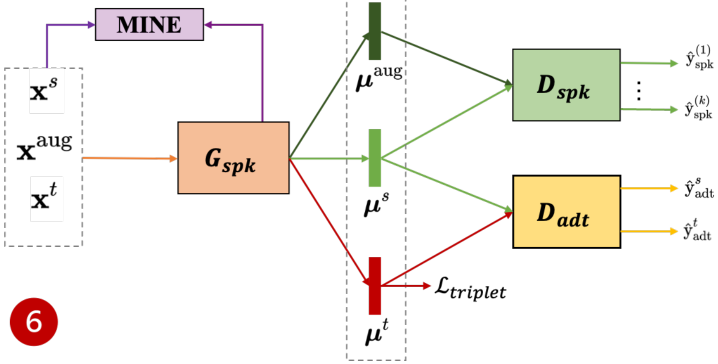
4



Results

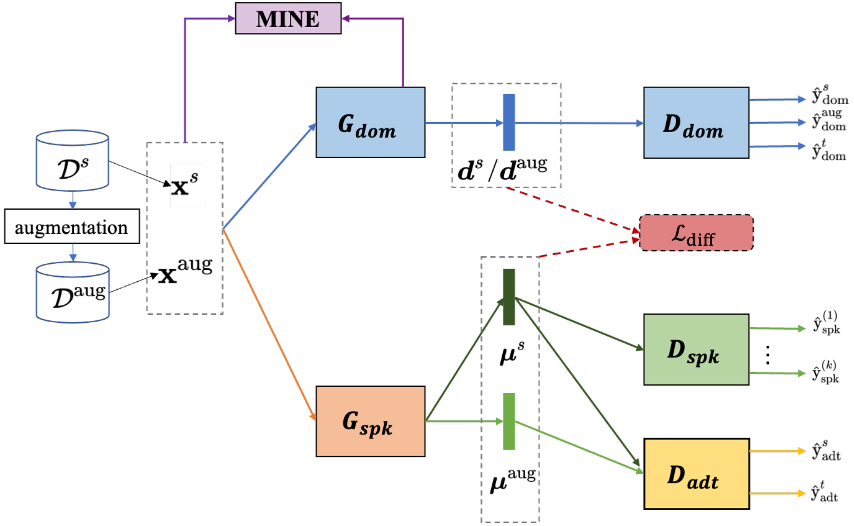
Row	System	Source domain	Target domain	MINE	$\mathcal{L}_{\text{triplet}}$	D_{adt}	$\mathcal{L}_{\text{diff}}$	VOiCES dev.		VOiCES eval.	
								EER(%)	minDCF	EER(%)	minDCF
1	TDNN [4]	$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	×	×	✓	×	3.18	-	7.15	-
2	TDNN (G_{spk})	$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	×	×	×	×	×	2.35	0.2596	6.42	0.4398
3		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	×	×	2.16	0.2627	6.32	0.4305
4		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	×	✓	×	×	×	2.29	0.2453	5.98	0.4351
5		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	✓	×	2.31	0.2645	6.29	0.4399
6		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	✓	✓	×	2.54	0.2671	6.23	0.4379





Results

Row	System	Source domain	Target domain	MINE	$\mathcal{L}_{\text{triplet}}$	D_{adt}	$\mathcal{L}_{\text{diff}}$	VOiCES dev. EER(%)	VOiCES dev. minDCF	VOiCES eval. EER(%)	VOiCES eval. minDCF
1	TDNN [4]	$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	×	×	✓	×	3.18	-	7.15	-
2		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	×	×	×	×	×	2.35	0.2596	6.42	0.4398
6		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	✓	✓	×	2.54	0.2671	6.23	0.4379
7	InfoMax–DSAN	$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	✓	×	✓	MI	2.17	0.2418	5.93	0.4265
8		$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	✓	×	✓	ort	2.33	0.2577	5.98	0.4131

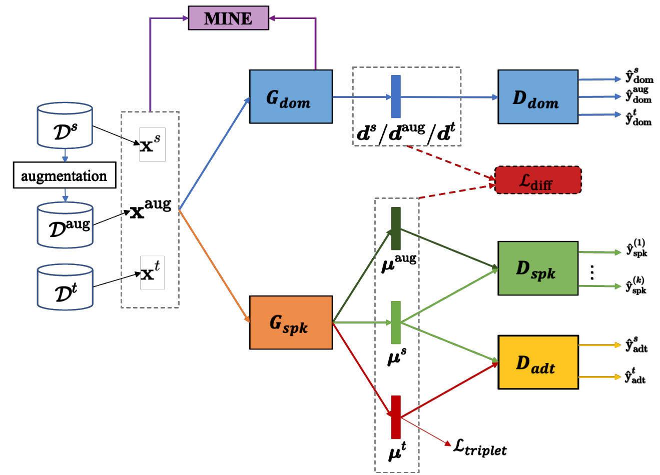


THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Results

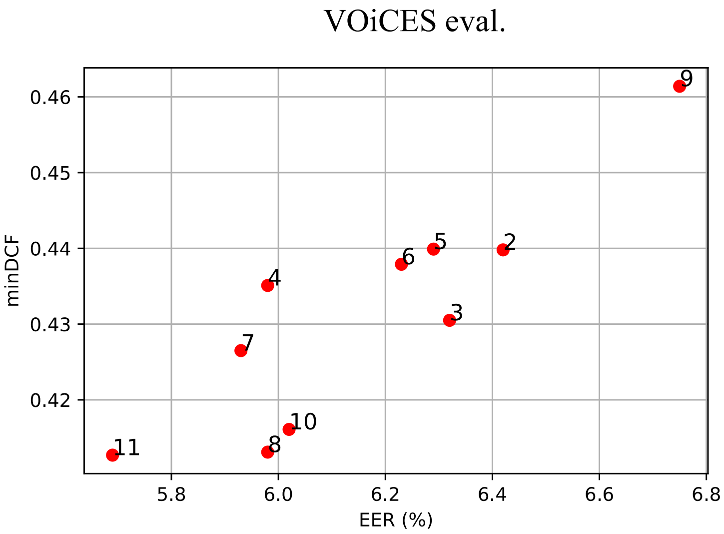
Row	System	Source domain	Target domain	MINE	$\mathcal{L}_{\text{triplet}}$	D_{adt}	$\mathcal{L}_{\text{diff}}$	VOiCES dev. EER(%)	VOiCES dev. minDCF	VOiCES eval. EER(%)	VOiCES eval. minDCF
1	TDNN [4]	$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	×	×	✓	×	3.18	-	7.15	-
2		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	×	×	×	×	×	2.35	0.2596	6.42	0.4398
9	$(G_{\text{dom}}\&G_{\text{spk}})$	$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	×	✓	MI	3.29	0.3278	6.75	0.4614
10		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	✓	MI	2.31	0.2490	6.02	0.4161
11		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	✓	✓	MI	**2.06**	**0.2375**	**5.69**	**0.4127**

11



Results

Row	System	Source domain	Target domain	MINE	$\mathcal{L}_{\text{triplet}}$	D_{adt}	$\mathcal{L}_{\text{diff}}$
1	TDNN [4]	$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	×	×	✓	×
2	TDNN (G_{spk})	$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	×	×	×	×	×
3		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	×	×
4		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	×	✓	×	×	×
5		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	✓	×
6		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	✓	✓	×
7	InfoMax–DSAN (G_{dom}&G_{spk})	$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	✓	×	✓	MI
8		$\mathcal{D}^s$	$\mathcal{D}^{\text{aug}}$	✓	×	✓	ort
9		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	×	✓	MI
10		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	×	✓	✓	MI
11		$\{\mathcal{D}^s, \mathcal{D}^{\text{aug}}\}$	$\mathcal{D}^t$	✓	✓	✓	MI



Conclusions

- InfoMax-DSAN can enforce the shared encoder to disentangle the domain-invariant features from the domain-specific properties, which can help to **address domain mismatch**.

- The frame-based MINE can effectively help **extract informative** features.

- Self-supervised learning can help **mitigate the label mismatch** problem for domain adaptation.