Information Maximized Variational Domain Adversarial Learning for Speaker Verification

Youzhi TU[1], Man-Wai MAK[1] and Jen-Tzung CHIEN[2]

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

[2]Dept. of Electrical and Computer Engineering,
National Chiao Tung University, Taiwan

ICASSP'20
4-8 May 2020

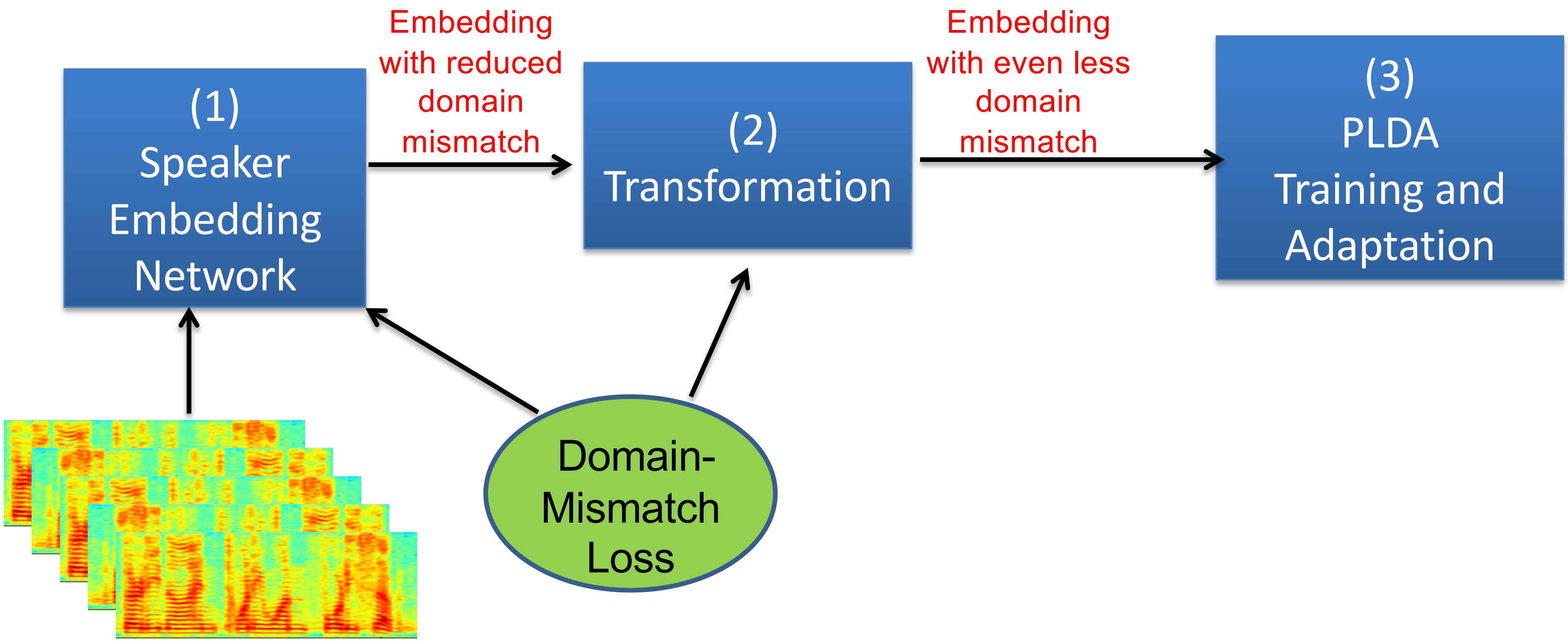
Contents

Domain Mismatch

- When training data and test data of speaker recognition systems have a severe mismatch, the performance degrades rapidly.
- The mismatch can be caused by languages, channels, noises, genders, etc.
- Collecting a large amount of **in-domain labeled** data to retrain the system is time-consuming and costly.
- We need to **adapt the existing system** to new environments or create a **domain-invariant** feature space.

Domain Adaptation

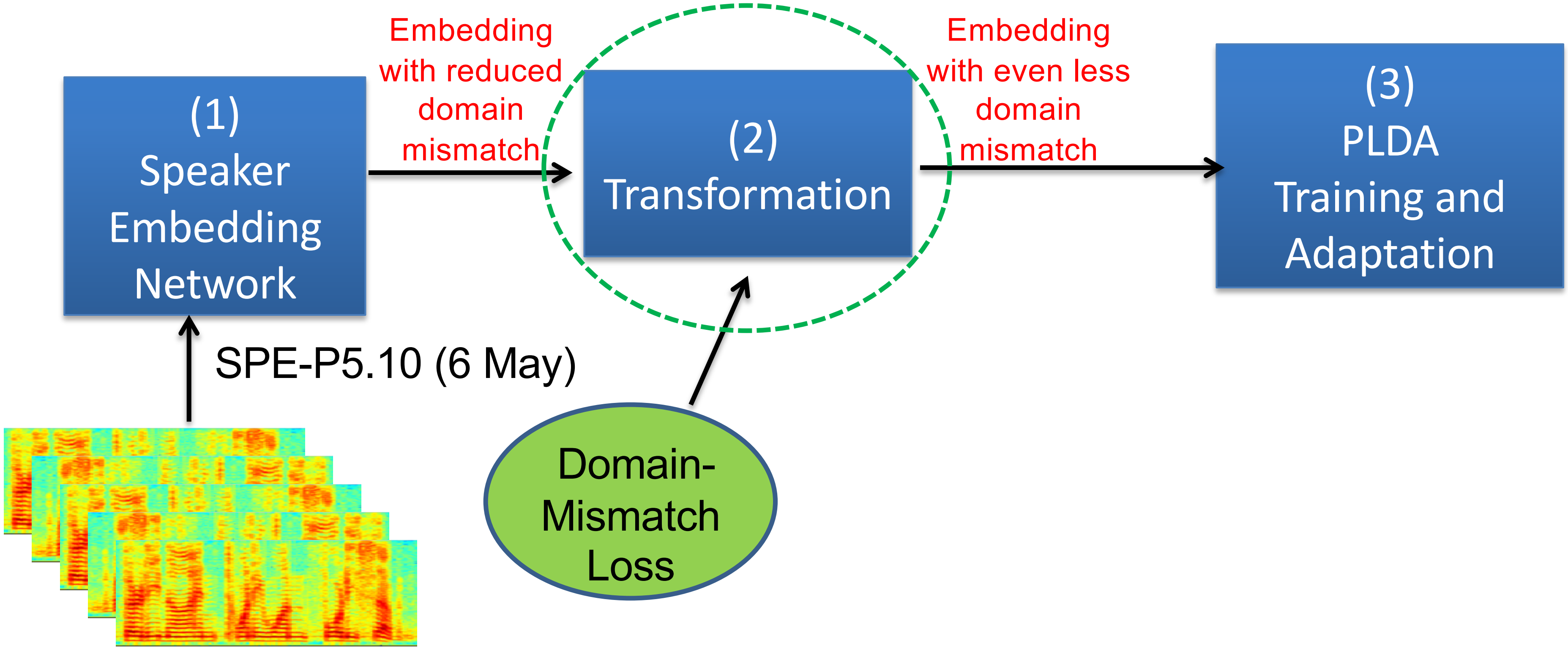
- Can be performed during system training by
  1. making the speaker embedding network domain-invariant
  2. transforming the speaker embedding to domain-invariant space
  3. adapting the PLDA model



Speech from multiple domains

Domain Adaptation

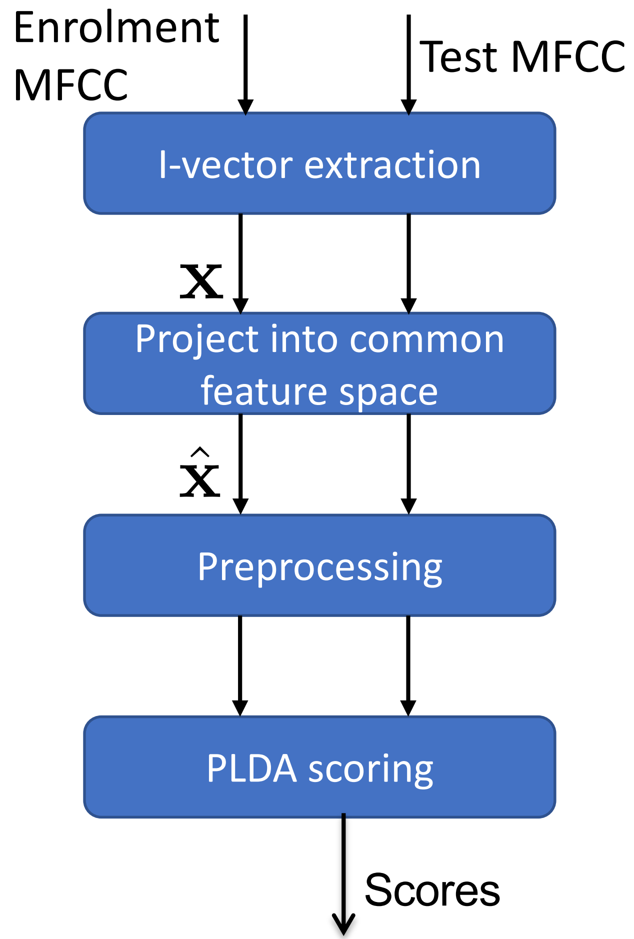
- Can be performed during system training by
  1. making the speaker embedding network domain-invariant
  2. transforming the speaker embedding to domain-invariant space
  3. adapting the PLDA model



Speech from multiple domains

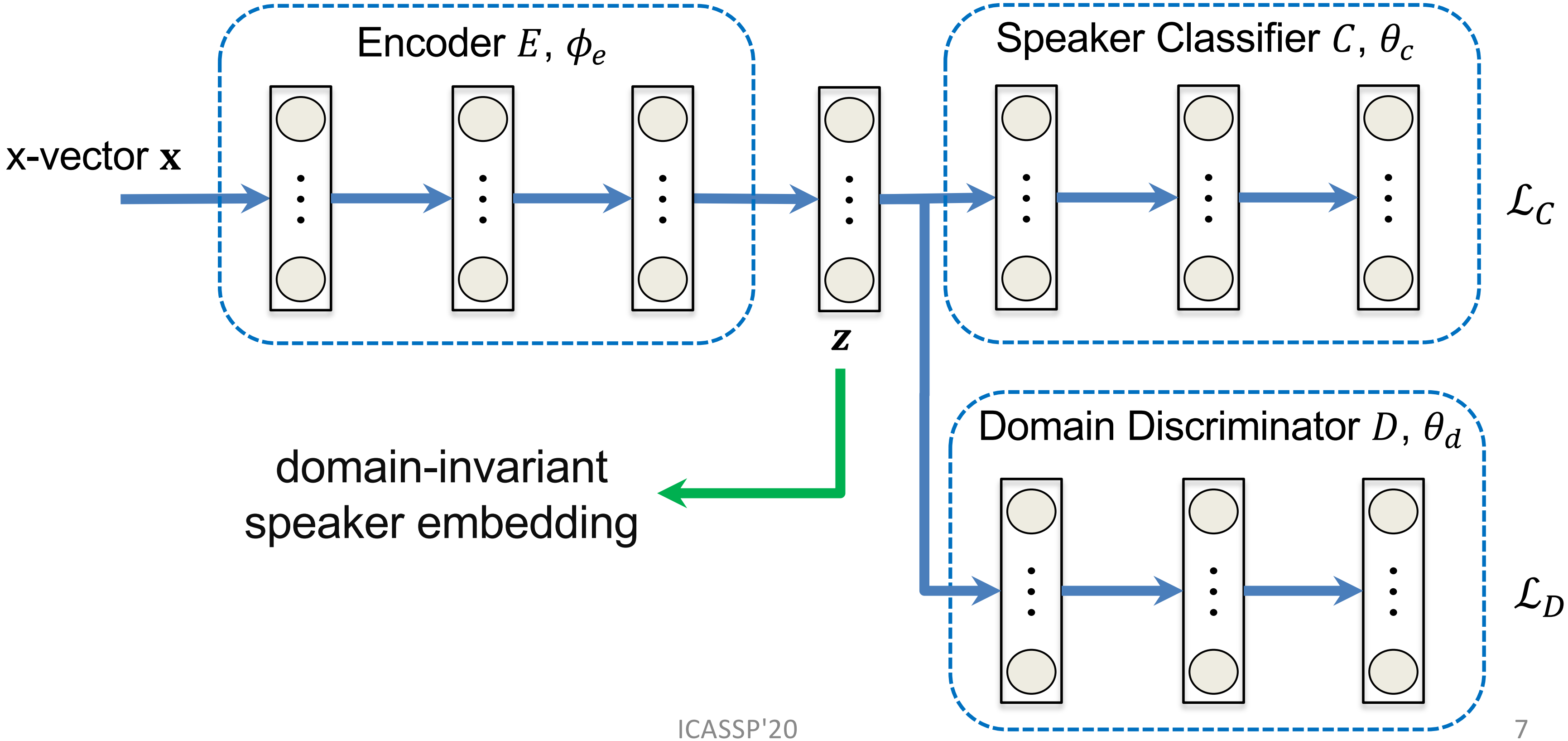
Transformation of i/x-vectors

- **Fix** the i-vector extractor or speaker embedding network
- **Transform** the i/x-vectors to a domain-invariant space, followed by PLDA scoring
- Transformation can be performed by a **domain adversarial neural network**



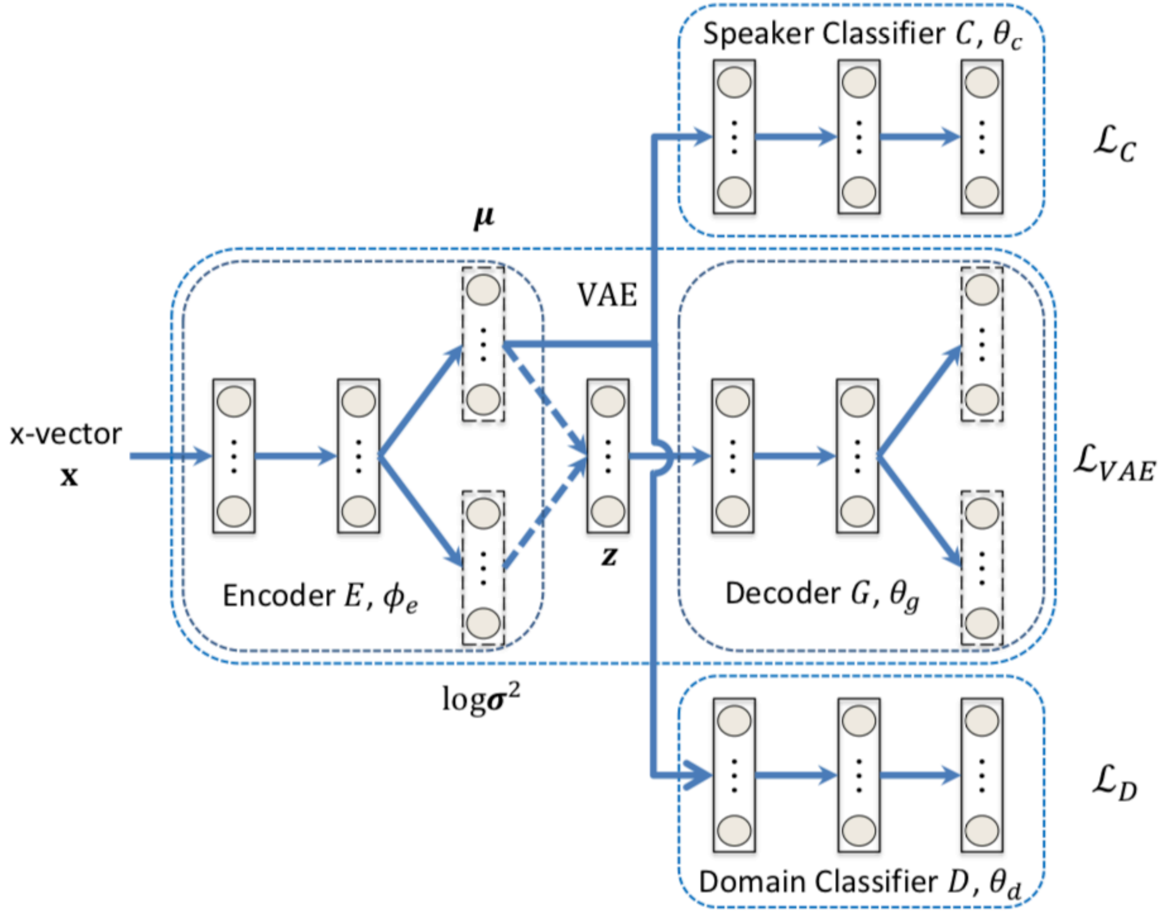
Domain Adversarial Neural Network

- A feature encoder, a speaker classifier, and a domain discriminator are trained with contradictory objectives
- After training, the encoder produces domain-invariant feature vectors



Variational DANN

- Variational domain adversarial neural network (VDANN) incorporates a variational autoencoder (VAE) into domain adversarial training



VAE aims to perform variational regularization so that **z** follows a Gaussian distribution

Y.Z. Tu, M.W. Mak, and J.T. Chien, "Variational Domain Adversarial Learning for Speaker Verification", Interspeech'19.

Limitations of VDANN

- **Posterior collapse**: When the decoder is too flexible, a VAE can produce non-informative representations **z** independent of the input **x**

$$\mathrm{ELBO}_{\mathrm{VAE}} = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[-\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right]$$

$$\propto -\mathbb{E}_{q_\phi(\mathbf{z})}[\mathrm{KL}(q_\phi(\mathbf{x}|\mathbf{z})||\, p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$$

$p_\theta(\mathbf{x}|\mathbf{z})$: Reconstruction likelihood

. $q_\phi(\mathbf{z}|\mathbf{x})$: Variational posterior

$p(\mathbf{z})$: Latent prior

. $q_\phi(\mathbf{z}) = \int_{\mathbf{x}} p_{\mathcal{D}}(\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x}) d\mathbf{x}$: Aggregated posterior

S.J Zhao, J.M. Song, S. Ermon. InfoVAE: Balancing Learning and Inference in Variational Autoencoders. AAAI-19.

InfoVAE

- InfoVAEs overcome the limitations of VAE by
  - incorporating a term $\eta I_q(\mathbf{x}, \mathbf{z})$ that explicitly preserves high mutual information between $\mathbf{x}$ and $\mathbf{z}$
  - adding a scalar λ to balance variational inference and data reconstruction

$$\mathrm{ELBO}_{\mathrm{VAE}} \propto -\mathbb{E}_{q_\phi(\mathbf{z})}\left[\mathrm{KL}(q_\phi(\mathbf{x}|\mathbf{z}) || p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$$

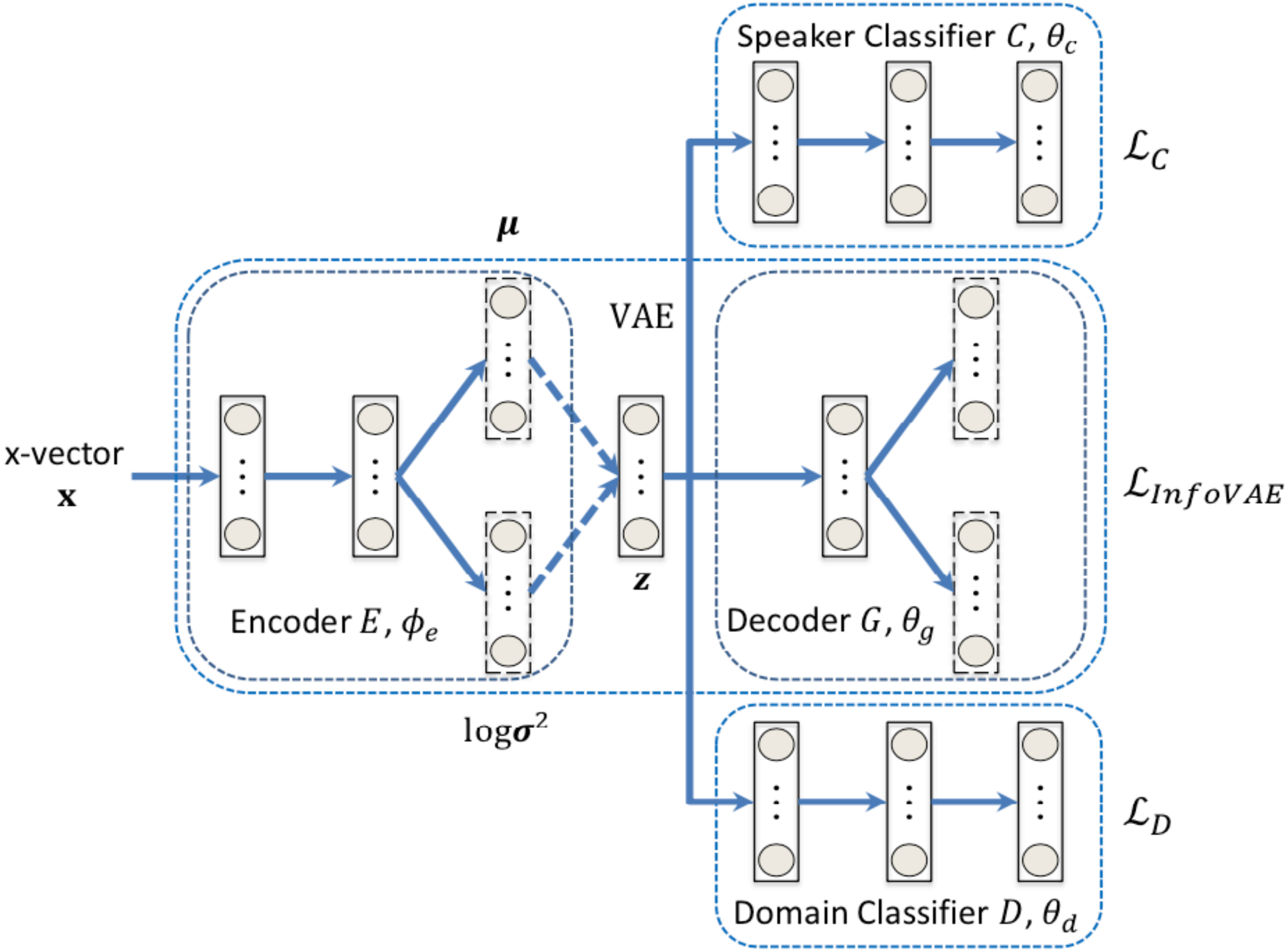
$$\begin{aligned}\mathrm{ELBO}_{\text{InfoVAE}} &\equiv -\mathbb{E}_{q_\phi(\mathbf{z})}\left[\mathrm{KL}(q_\phi(\mathbf{x}|\mathbf{z}) || p_\theta(\mathbf{x}|\mathbf{z})\right] - \lambda \mathrm{KL}(q_\phi(\mathbf{z})||p(\mathbf{z})) + \eta I_q(\mathbf{x}, \mathbf{z}) \\ &\propto \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right] \\ &\qquad -(1-\eta)\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))\right] \\ &\qquad -(\lambda+\eta-1)\mathrm{D}_g(q_\phi(\mathbf{z})||p(\mathbf{z}))\end{aligned}$$

$I_q(\mathbf{x}, \mathbf{z})$: Mutual information between $\mathbf{x}$ and $\mathbf{z}$ under $q_\phi(\mathbf{z}, \mathbf{x})$

$\mathrm{D}_g(\cdot \,||\, \cdot)$: Generalized divergence, e.g., maximum mean discrepancy (MMD) and adversarial training

InfoVDANN

- Information-maximized variational DANN (InfoVDANN) incorporates an **InfoVAE** into domain adversarial training



InfoVAE aims to

- Perform variational regularization so that **z** follows a Gaussian distribution
- Preserve the mutual information between **z** and **x**.

InfoVDANN

- Loss function:

$$\mathcal{L}_{\text{InfoVDANN}}\left(\theta_c, \theta_d, \phi_e, \theta_g\right) = \mathcal{L}_C(\theta_c, \phi_e) - \alpha\mathcal{L}_D(\theta_d, \phi_e) + \beta\mathcal{L}_{\text{InfoVAE}}\left(\phi_e, \theta_g\right)$$

$$\mathcal{L}_C(\theta_c, \phi_e) = \sum_{r=1}^{R} \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x}^{(r)})}\left\{-\sum_{k=1}^{K} y_k^{(r)} \log C\left(E\left(\mathbf{x}^{(r)}\right)\right)_k\right\}$$

$$\mathcal{L}_D(\theta_d, \phi_e) = \sum_{r=1}^{R} \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x}^{(r)})}\left\{-\log D\left(E\left(\mathbf{x}^{(r)}\right)\right)_r\right\}$$

r indexes domain
k indexes speaker
j indexes the dim of $\mathbf{z}$

$$\begin{aligned}\mathcal{L}_{\text{InfoVAE}}(\theta_g, \phi_e) \triangleq &- \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right] + (1-\eta)\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}[\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] \\ &+ (\lambda + \eta - 1)\mathrm{D}_g(q_\phi(\mathbf{z})||p(\mathbf{z})) \\ = &- \sum_{r=1}^{R}\sum_{i=1}^{N_r}\left\{\log p_\theta\left(\mathbf{x}_i^{(r)}|\mathbf{z}_i^{(r)}\right) + \frac{1-\eta}{2}\sum_{j=1}^{J}\left[1 + \log\left(\sigma_{ij}^{(r)}\right)^2 - \left(\mu_{ij}^{(r)}\right)^2 - \left(\sigma_{ij}^{(r)}\right)^2\right]\right\} \\ &+ (\lambda + \eta - 1)\mathrm{D}_g\left(q_\phi(\mathbf{z})\|p(\mathbf{z})\right)\end{aligned}$$

MMD or adversarial training by introducing a discriminator to distinguish samples drawn from $q_\phi(\mathbf{z})$ and $p(\mathbf{z})$

Gaussian regularization: Push $q_\phi(\mathbf{z}|\mathbf{x})$ towards a Gaussian distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$

InfoVDANN

- Optimization:

$$\mathcal{L}_{\text{InfoVDANN}}\left(\theta_c, \theta_d, \phi_e, \theta_g\right) = \mathcal{L}_C(\theta_c, \phi_e) - \alpha\mathcal{L}_D(\theta_d, \phi_e) + \beta\mathcal{L}_{\text{InfoVAE}}\left(\phi_e, \theta_g\right)$$

$$\hat{\theta}_d = \underset{\theta_d}{\operatorname{argmax}} \mathcal{L}_{\text{InfoVDANN}}\left(\hat{\theta}_c, \theta_d, \hat{\phi}_e, \hat{\theta}_g\right)$$

$$\left(\hat{\theta}_c, \hat{\phi}_e, \hat{\theta}_g\right) = \underset{\theta_c, \phi_e, \theta_g}{\operatorname{argmin}} \mathcal{L}_{\text{InfoVDANN}}\left(\theta_c, \hat{\theta}_d, \phi_e, \theta_g\right)$$

- Special cases
  - $\eta = 0, \lambda = 1$, InfoVDANN → VDANN
  - Remove the decoder and the sampling procedure ($\beta = 0$), InfoVDANN → DANN

Experiments

- **InfoVDANN training:** each dataset corresponds to a domain

dataset	No. of speakers	No. of utterances
SRE04–10	1,806	54,180
Voxceleb1	1,251	37,530
SwitchBoard II	273	6,962
SITW	203	3,700

- **Evaluation data:** SRE16 and SRE18-CMN2
- **PLDA training data:**

  SRE04-10 + augmentation for SRE16

  SRE04-10-mx6 + augmentation for SRE18-CMN2
- **PLDA adaptation**: SRE16 and/or SRE18 unlabeled

Experiments

- **Input:** X-vectors extracted from the pre-trained DNN (512)
- **InfoVDANN:** Transformed to a 400-dimensional latent space

Sub-network	Architecture	Non-linearity
Encoder	1024-1024-400	ReLU + linear (output)
Decoder	2048-512	ReLU + linear (output)
Speaker classifier	1024-1024-3533	Leaky ReLU + softmax (output)
Domain classifier	128-32-4	ReLU + softmax (output)

- **Hyperparameters:** $\alpha = 0.1, \beta = \lambda = 1, \eta = 0.2$
- **Preprocessing:** center + LDA (150) + whitening + length-norm

http://kaldi-asr.org/models/m3

Experiments

- Generalized divergence: $\mathrm{D}_g\left(q_\phi(\mathbf{z})\|p(\mathbf{z})\right)$

$$\mathcal{L}_{\text{InfoVAE}}(\theta_g,\phi_e)=-\sum_{r=1}^{R}\sum_{i=1}^{N_r}\left\{\log p_\theta\left(\mathbf{x}_i^{(r)}|\mathbf{z}_i^{(r)}\right)+\frac{1-\eta}{2}\sum_{j=1}^{J}\left[1+\log\left(\sigma_{ij}^{(r)}\right)^2-\left(\mu_{ij}^{(r)}\right)^2-\left(\sigma_{ij}^{(r)}\right)^2\right]\right\}$$
$$+(\lambda+\eta-1)\mathrm{D}_g\left(q_\phi(\mathbf{z})\|p(\mathbf{z})\right)$$

- **MMD:** MMD-VDANN

  Minimize the MMD between $q_\phi(\mathbf{z})$ and $p(\mathbf{z})$

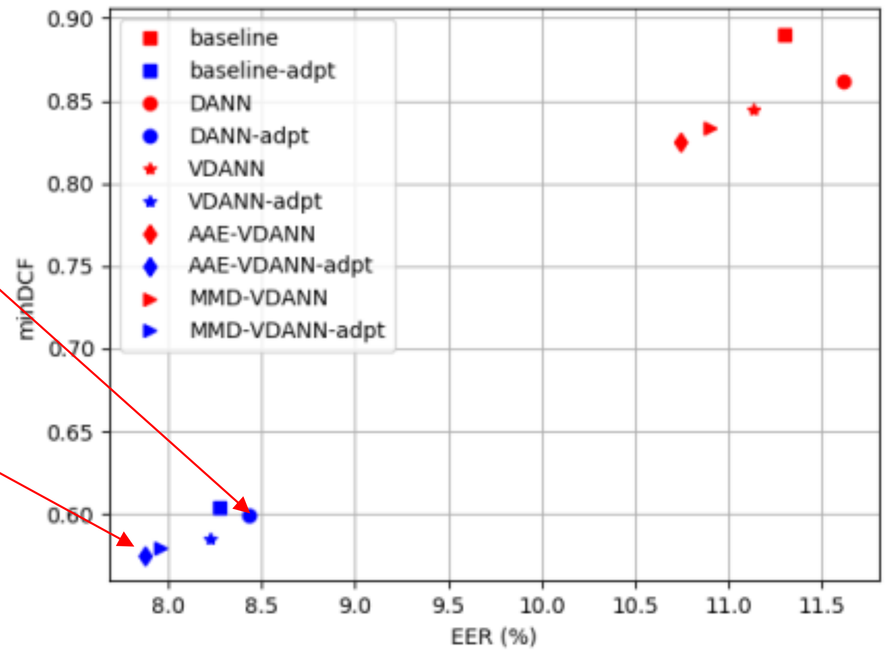
- **Adversarial training**: AAE-VDANN

  Minimize the JS-divergence between $q_\phi(\mathbf{z})$ and $p(\mathbf{z})$

Results

SRE16

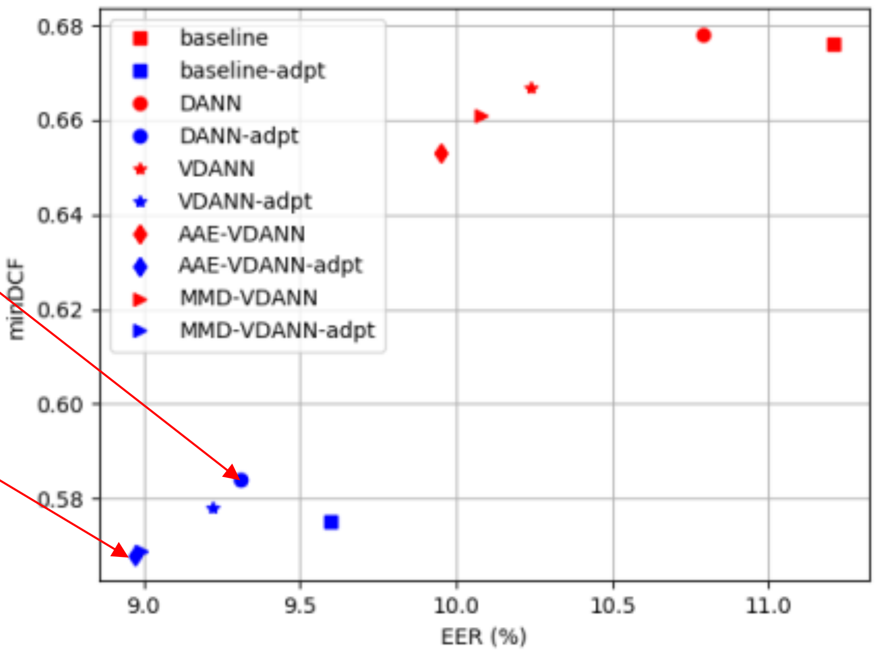
	No PLDA adaptation		PLDA adaptation	
	EER	minDCF	EER	minDCF
Baseline	11.30	0.890	8.27	0.604
DANN	11.62	0.862	8.43	0.599
VDANN	11.13	0.845	8.22	0.585
AAE-VDANN	**10.74**	**0.825**	**7.87**	**0.575**
MMD-VDANN	10.90	0.834	7.96	0.579



Results

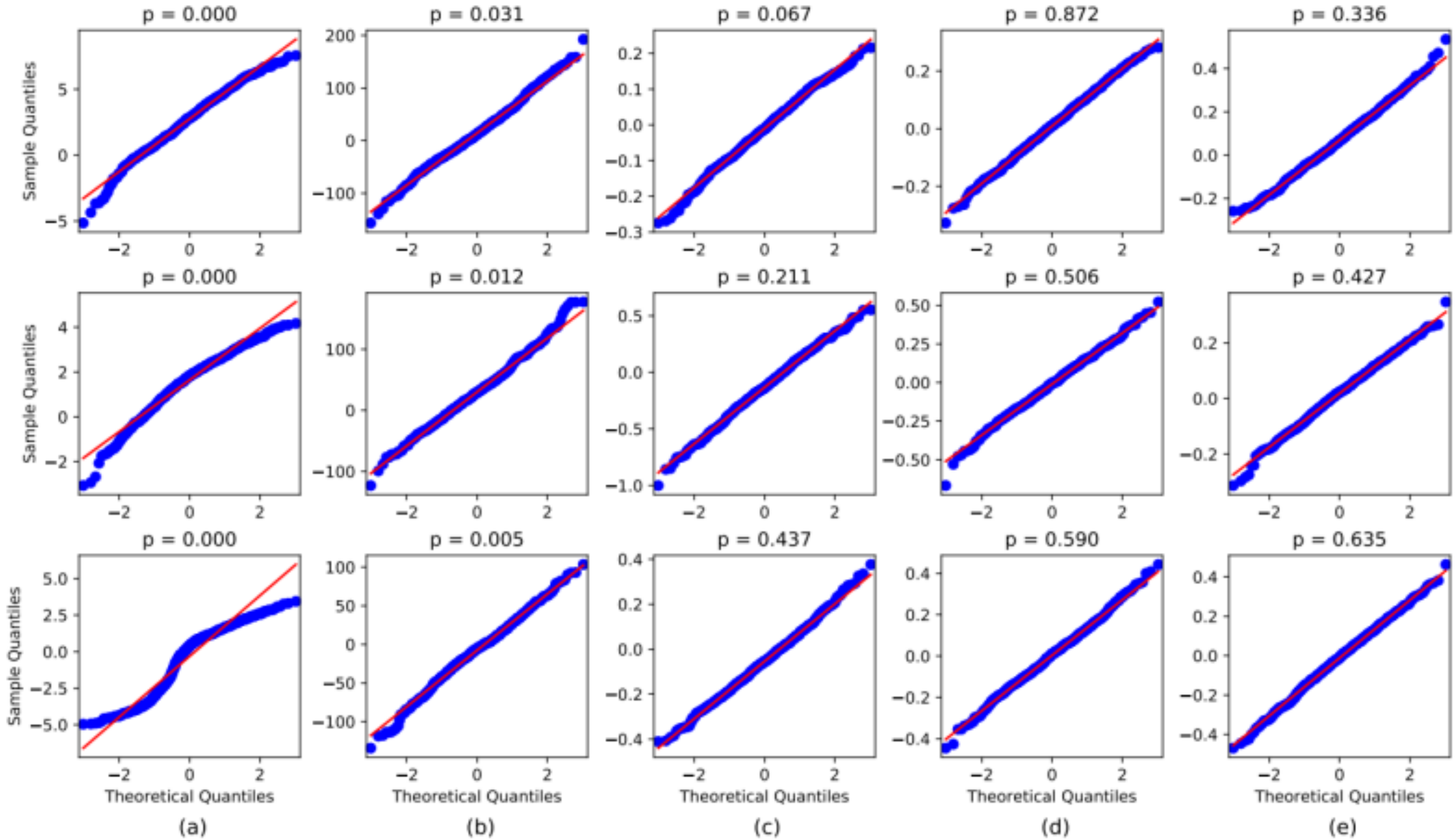
SRE18-CMN2

	No PLDA adaptation		PLDA adaptation	
	EER	minDCF	EER	minDCF
Baseline	11.21	0.676	9.60	0.575
DANN	10.79	0.678	9.31	0.584
VDANN	10.24	0.667	9.22	0.578
AAE-VDANN	**9.95**	**0.653**	**8.97**	**0.568**
MMD-VDANN	10.08	0.661	8.99	0.569



Results

Quantile-quantile (Q–Q) plots and p-values obtained from Shapiro–Wilk tests (The larger the p, the more Gaussian the distribution.)



Raw DANN VDANN MMD-VDANN AAE-VDANN

Conclusions

- InfoVDANN can **reduce domain mismatch** through domain adversarial training.
- InfoVAEs and VAEs are effective in making the transformed x-vectors **more Gaussian**.
- InfoVDANNs are effective for preserving **speaker** information in the latent space to improve the performance of speaker verification.

References

1. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” Journal of Machine Learning Research, vol. 17, no. 1, pp. 2096–2130, 2016.
2. A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Scholkopf, and A. Smola, “A kernel method for the two-sample-problem,” in Proc. NIPS, 2007, pp. 513–520.
3. W. W. Lin, M. W. Mak and J. T. Chien, "Multi-source I-vectors Domain Adaptation using Maximum Mean Discrepancy Based Autoencoders", IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 26, no. 12, pp. 2412-2422, Dec. 2018.
4. Y. Z. Tu, M. W. Mak, and J. T. Chien, "Variational Domain Adversarial Learning for Speaker Verification", Interspeech'19, Graz, Sept. 2019.
5. S. J. Zhao, J. M. Song, and S. Ermon, “InfoVAE: balancing learning and inference in variational autoencoders,” in Proc. AAAI Conference on Artificial Intelligence, 2019, pp. 5885–5892.

Thank you!

Results

- Mutual information estimation

$$\hat{I}_q(\mathbf{x};\mathbf{z}) = \sum_{s=1}^{B}\left\{-\frac{1}{2}\sum_{j=1}^{J}\left[1+\log(2\pi)+\log\sigma_{sj}^2\right]-\log\frac{1}{B}\sum_{b=1}^{B} q_\phi\left(\mathbf{z}_s|\mathbf{x}_b\right)\right\}$$

  - Estimated samples: $B = 1024$
  - 200 runs

	SRE16-eval				SRE18-eval-CMN2			
	Enrollment		Test		Enrollment		Test	
	mean	var	mean	var	mean	var	mean	var
VDANN	4.466	1.092	5.078	1.115	3.922	1.045	4.567	1.077
MMD-VDANN	4.811	1.052	5.770	1.150	**5.357**	1.228	5.028	1.327
AAE-VDANN	**5.114**	1.047	**6.263**	1.151	5.038	1.163	**5.031**	1.248