Domain Robust Deep Embedding Learning for Speaker Recognition

Hang-Rui Hu[1], *Yan Song*[1], *Ying Liu*[1], *Li-Rong Dai*[1], *Ian McLoughlin*[1,2], *Lin Liu*[3]

[1]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China.
[2]ICT Cluster, Singapore Institute of Technology, Singapore.
[3]iFLYTEK Research, iFLYTEK CO. LTD., Hefei, China.

Video Presentation for ICASSP 2022
Paper ID: #4020 Presenter: Hang-Rui Hu

Introduction

- **Speaker Recognition（SRE）** is the task of automatically determining whether a speech utterance belongs to a certain speaker identity.

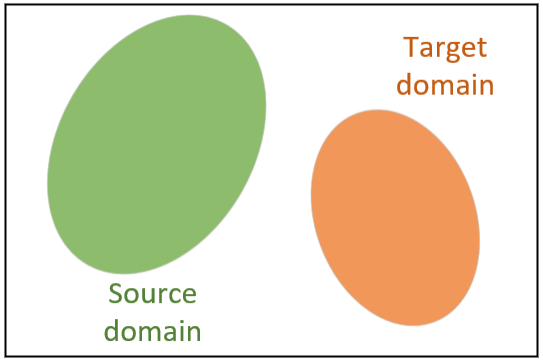
- **Domain Shift**: However, the performance degrades significantly when applied to a new target domain.
  - **Domain Mismatch**: distribution discrepancy between source and target domain.

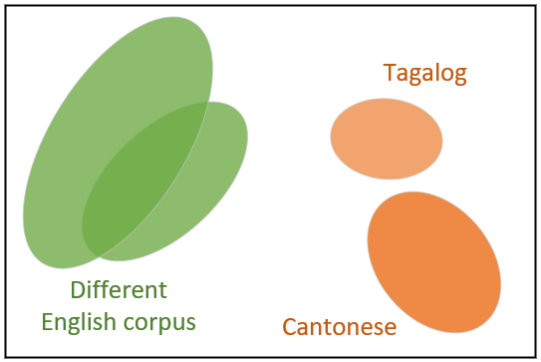
    i.e. in NIST SRE16:
    Source domain: English utterances
    Target domain: Non-English utterances

  - **Latent Sub-Domain Mismatch**: each domain could be composed of multiple mismatched sub-domain

    i.e. in NIST SRE16:
    Source domain is collected from several datasets
    Target domain contains two different languages

  - **Lack of target domain labels**



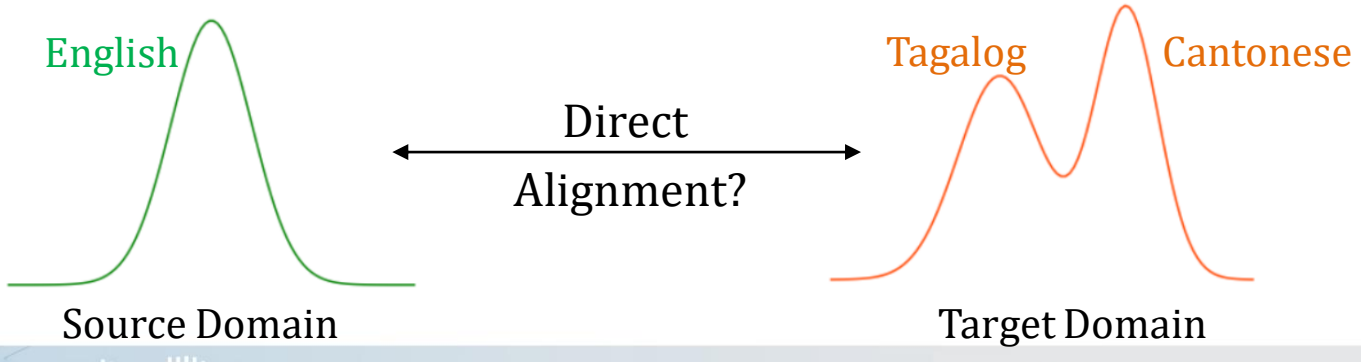


Introduction

- **Unsupervised Domain Adaptation (UDA)** addresses the situation where there are labeled source data and unlabeled target data available for use during training.

- **Existing UDA methods** usually aim to minimize the **distribution discrepancy**:
  - Domain distribution alignment in the back-end
    i.e. Kaldi APLDA, CORAL …
  - Learning domain-invariant features in the front-end
    i.e. Adversarial learning, Deep CORAL, MMD …

- **Issues:**
  - Ignoring the speaker information of the target domain.
  - Difficulty handling latent sub-domain mismatch.
    i.e. Direct alignment of unimodal and multimodal Gaussian distributions may not be optimal.



Introduction

- Some other UDA methods leveraged potential **label information** from target domain.
  - Unsupervised Clustering
  - Self-supervised Contrastive Learning[1]

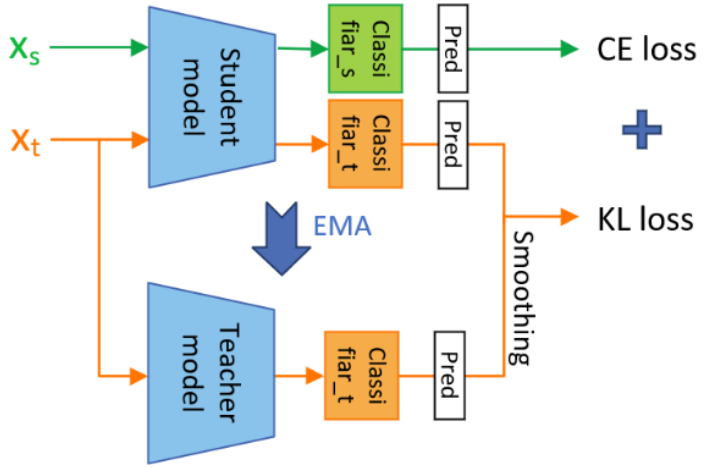
- Issues:
  - Clustering:
    - the number of speakers is difficult to determine
    - the clustering results may not be accurate enough in practice
  - Contrastive learning:
    - It is difficult to construct positive and negative pairs correctly and effectively

Overview of the proposed framework

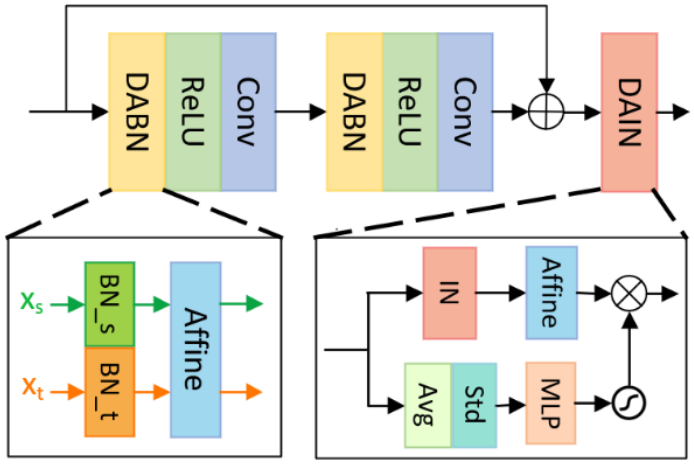
a) We proposed Smoothed Knowledge Distillation (SKD) to reduce the label noise and exploit the latent structural information of target domain.

- Source labeled data is sent to the student model and supervised by cross-entropy loss.
- Target domain data is assigned initial pseudo-labels and supervised by SKD loss- the KL divergence between the student output and the smoothed teacher output.

b) We designed two domain-robust modules to address the Domain Mismatch and Sub-domain Mismatch issues, respectively.



(a)Smoothed Knowledge Distillation based learning method

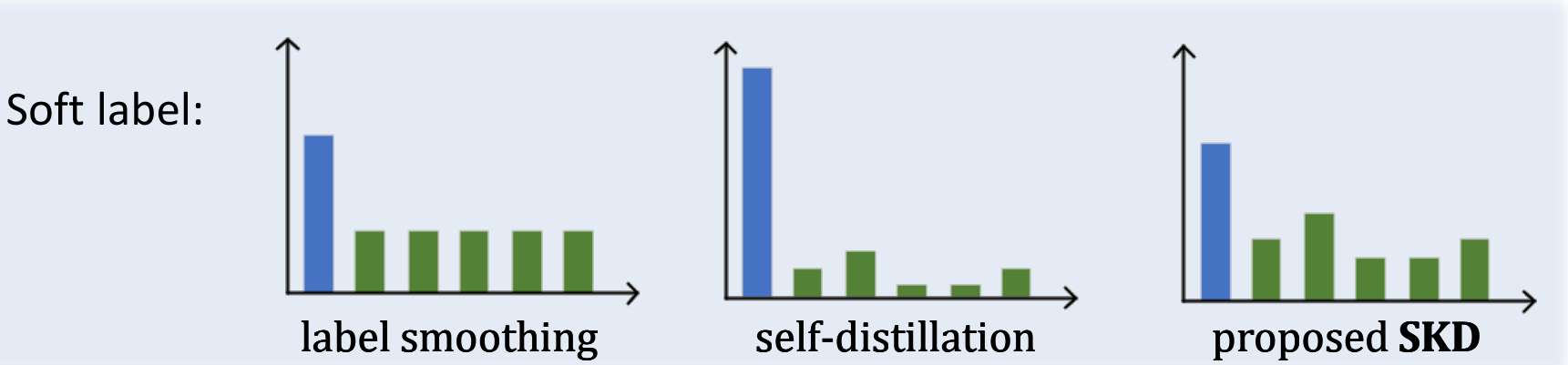


(b) Residual block with domain-robust modules

Methods - 1/4

(a) Smoothed Knowledge Distillation (SKD)

- Noisy label information in the target domain may hinder any performance improvement.



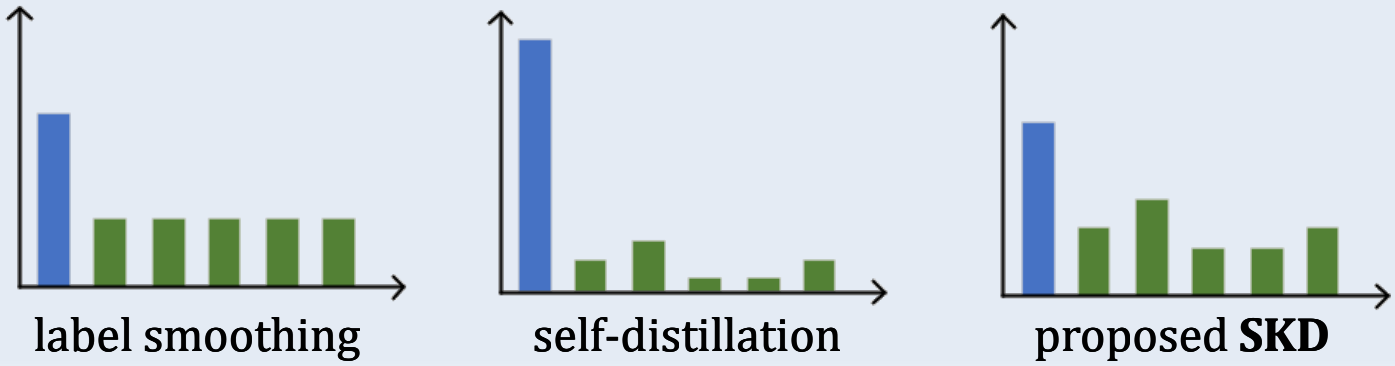
- Inspired by
  - Label smoothing:
    - ✓ reduces the confidence of the hard label,
    - ✘ but divides it equally among all negative classes.
  - Self-knowledge distillation:
    - ✓ learns inter-class relationships captured by the teacher model,
    - ✘ but cannot solve overfitting of noisy labels.

Methods – 2/4

(a) Smoothed Knowledge Distillation (SKD)

- SKD loss - learning the smoothed teacher output.

Soft label:



- Given the pseudo label c, and the output predictions for category k of the student and teacher model are denoted as p(k) and q(k), respectively, then the SKD loss is:

$$\mathcal{L}_{SKD}(p,q,c) = \mathcal{L}_{KL}(p,q')$$

$$q'(k) = \begin{cases} \gamma \cdot q(k) + \beta, & k = c \\ A \cdot q(k)^{1/t}, & k \neq c \end{cases}$$

Where A is the normalization coefficient

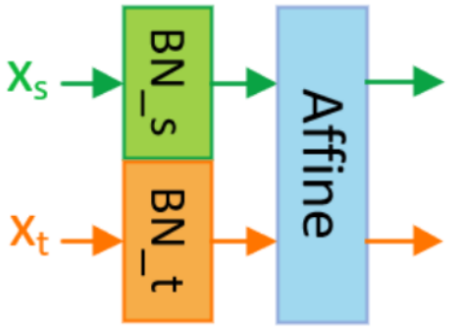
A linear transformation to suppress the main class confidence

An exponential transformation to smooth negative classes confidence

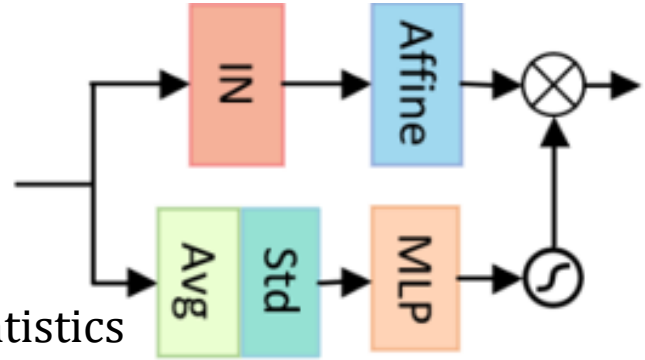
Methods – 3/4

(b) Domain-robust module

- **Domain-Aware BatchNorm (DABN)**
  - ➢ **Mismatch of source and target domains**
  - – Separate the statistics of each domain
    Sharing the statistics of BN is inappropriate when the domain shift is significant
  - – Sharing the affine parameters
    Transforming different domains to the same distribution



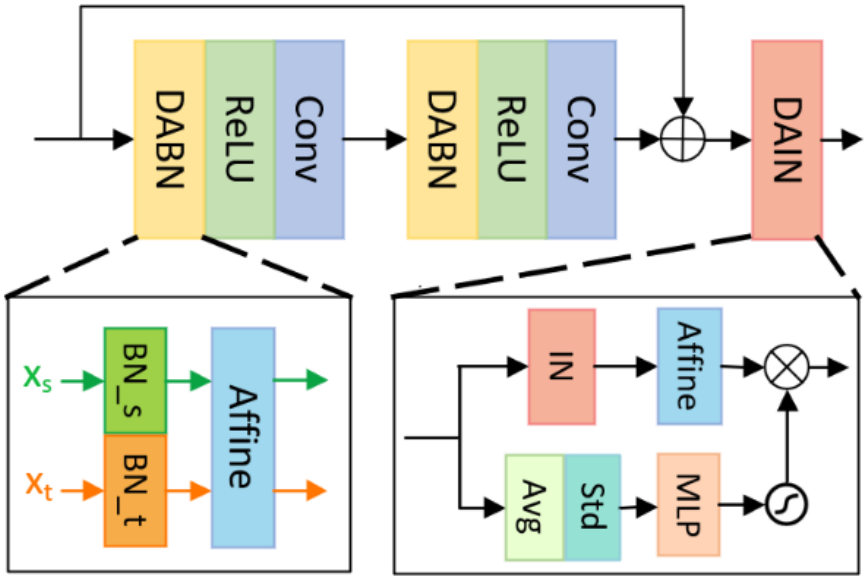
- **Domain-Agnostic InstanceNorm (DAIN)**
  - ➢ **Mismatch of latent sub-domains**
  - – IN alleviates the latent sub-domains discrepancy
    Inspired by style transfer or style normalization
  - – Recovering useful information from the removed statistics
    guarantee the feature discriminative capability



Methods – 4/4

(b) Domain-robust module

- Position:
  - We replaced all the BNs in the network with DABNs modules.
  - We inserted a DAIN module at the end of each residual block.



(b) Residual block with domain-robust modules

Experiment – 1/2

EER(%) comparison with existing adaptation systems on NIST SRE16 evaluation

Systems	Cosine			PLDA			APLDA		
	pooled	Tagalog	Cantonese	pooled	Tagalog	Cantonese	pooled	Tagalog	Cantonese
Baseline(Resnet18)	13.85	18.95	8.73	10.79	15.76	5.49	7.50	11.26	3.60
Our method (Resnet18)	11.22	15.38	7.16	9.81	14.37	5.26	6.95	10.13	3.64
Baseline (Resnet34)	12.64	17.63	7.77	9.36	14.17	4.66	6.85	10.57	3.27
Our method (Resnet34)	**9.62**	**13.41**	**5.76**	**7.87**	**11.46**	**4.29**	**5.77**	**8.92**	**2.59**
LSGAN[2]	11.74	15.63	7.90	/	/	/	/	/	/
FuseGAN [2]	10.88	14.84	6.93	/	/	/	/	/	/
Multi MMD[3]	/	/	/	9.03	/	/	8.29	/	/
PSN[4]	/	/	/	8.98	12.90	5.18	/	/	/
SpecAug[5]	/	/	/	/	/	/	/	11.49	3.72
CGAN[6]	/	/	/	/	/	/	/	9.86	2.99
CVAE[6]	/	/	/	/	/	/	/	10.07	2.97
NDM[7]	/	/	/	/	/	/	/	9.89	2.80
VB-MAP[8]	/	/	/	/	/	/	/	9.96	3.18

Experiment – 2/2

Cosine EER (%) results of **Ablation** experiments on NIST SRE16 evaluation performed on ResNet18

DABN	DAIN	SKD	Pooled	Tagalog	Cantonese
			13.85	18.95	8.73
✓			13.09	17.37	8.36
✓	✓		12.25	16.81	7.86
✓		✓	11.94	16.45	7.43
✓	✓	✓	11.22	15.38	7.16

Conclusions

- We proposed a **smoothed knowledge distillation (SKD)** based self-supervised learning method to exploit latent structural information from the unlabeled target domain.

- We designed two domain-robust modules:
  - **Domain-Aware BatchNorm (DABN)** module aims to reduce the cross-domain distribution discrepancy
  - **Domain-Agnostic InstanceNorm (DAIN)** module aim to learn features that are robust to within-domain variance.

- Our proposed method can be flexibly combined with many other adaption methods.

References

[1] Zhengyang Chen, Shuai Wang, and Yanmin Qian, ”Self-supervised learning based domain adaptation for robust speaker verification”, ICASSP 2020

[2] Gautam Bhattacharya, Joao Monteiro, Patrick Kenny et al., “Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification,” ICASSP 2019

[3] Weiwei Lin, Man-Mai Mak, Na Li et al., “Multi-Level Deep Neural Network Adaptation for Speaker Verification Using MMD and Consistency Regularization”, ICASSP 2020

[4] Zhengyang Chen, Shuai Wang and Yanmin Qian , “Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network”, Interspeech 2020

[5] Shuai Wang, Johan Rohdin, Oldrich Plchot el al., “Investigation of Specaugment for Deep Speaker Embedding Learning”, ICASSP 2020

[6] Shuai Wang, Yexin Yang, Zhanghao Wu, Yanmin Qian and Kai Yu, “Data Augmentation using Deep Generative Models for Embedding based Speaker Recognition”, IEEE/ACM Transactions on Audio, Speech, and Language Processing 2020

[7] Xun Gong, Zhengyang Chen, Yexin Yang, Shuai Wang, Lan Wang and Yanmin Qian, “Speaker Embedding Augmentation with Noise Distribution Matching”, ISCSLP2021

[8] Bengt J. Borgström, “Unsupervised Bayesian Adaptation of PLDA for Speaker Verification”, Interspeech 2021

Thank you for listening!