Multi-level Deep Neural Network Adaptation for Speaker Verification Using MMD and Consistency Regularization

Weiwei Lin[1], Man-Wai MAK[1], Na Li[2], Dan Su[2] and Dong Yu[2]

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

[2]Tencent AI Lab, China

ICASSP'20
4-8 May 2020

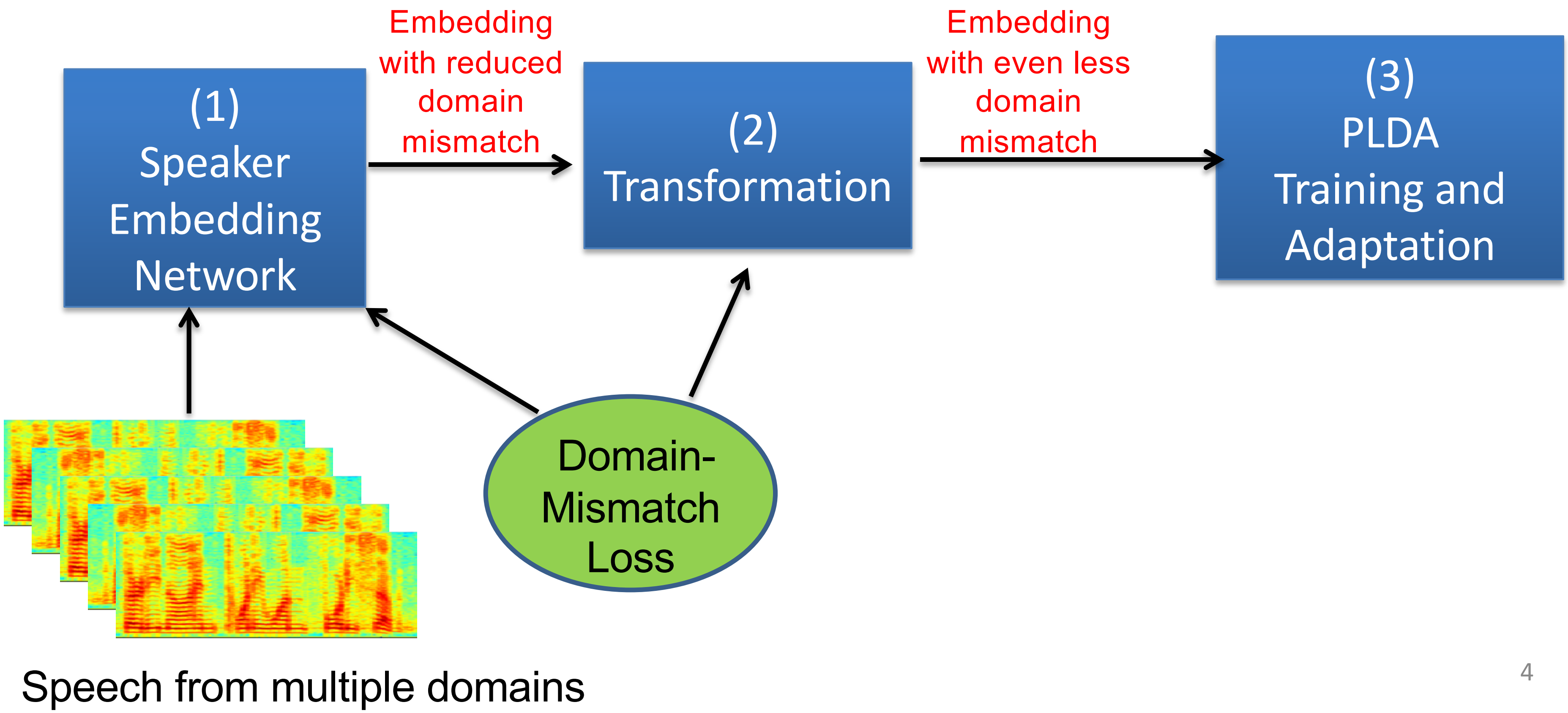
Contents

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, and genders.
- Collecting more data to retrain the system is time-consuming and computationally-expensive.
- We need to **adapt existing systems** to new environments or create a **domain-invariant** feature space.

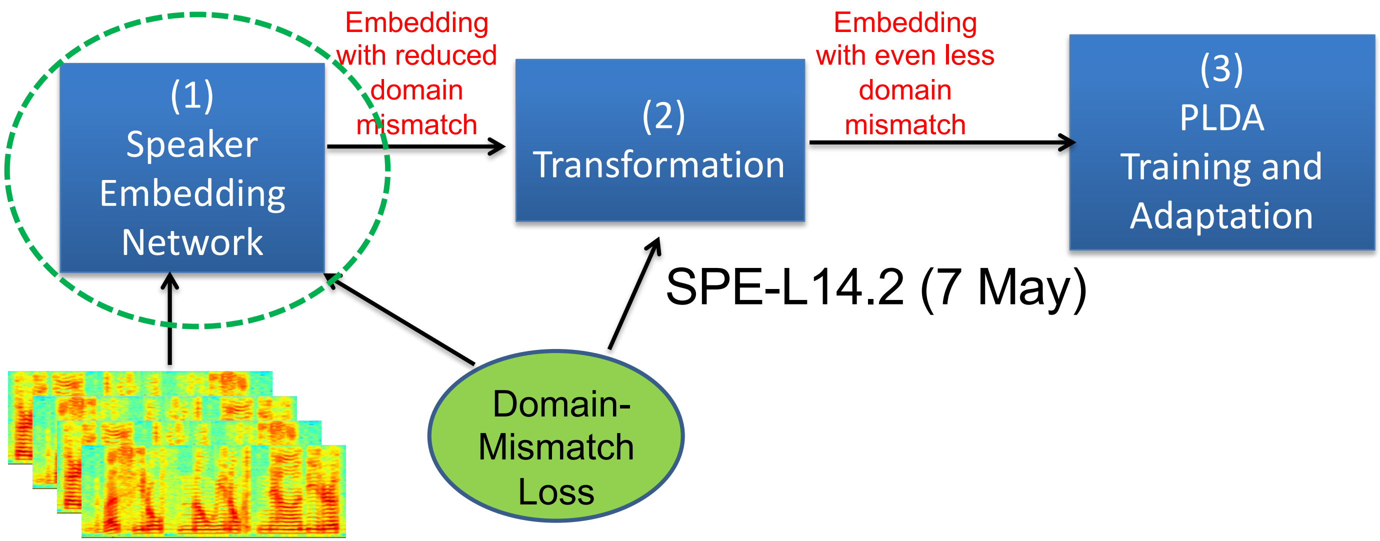
Domain Adaptation

- 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

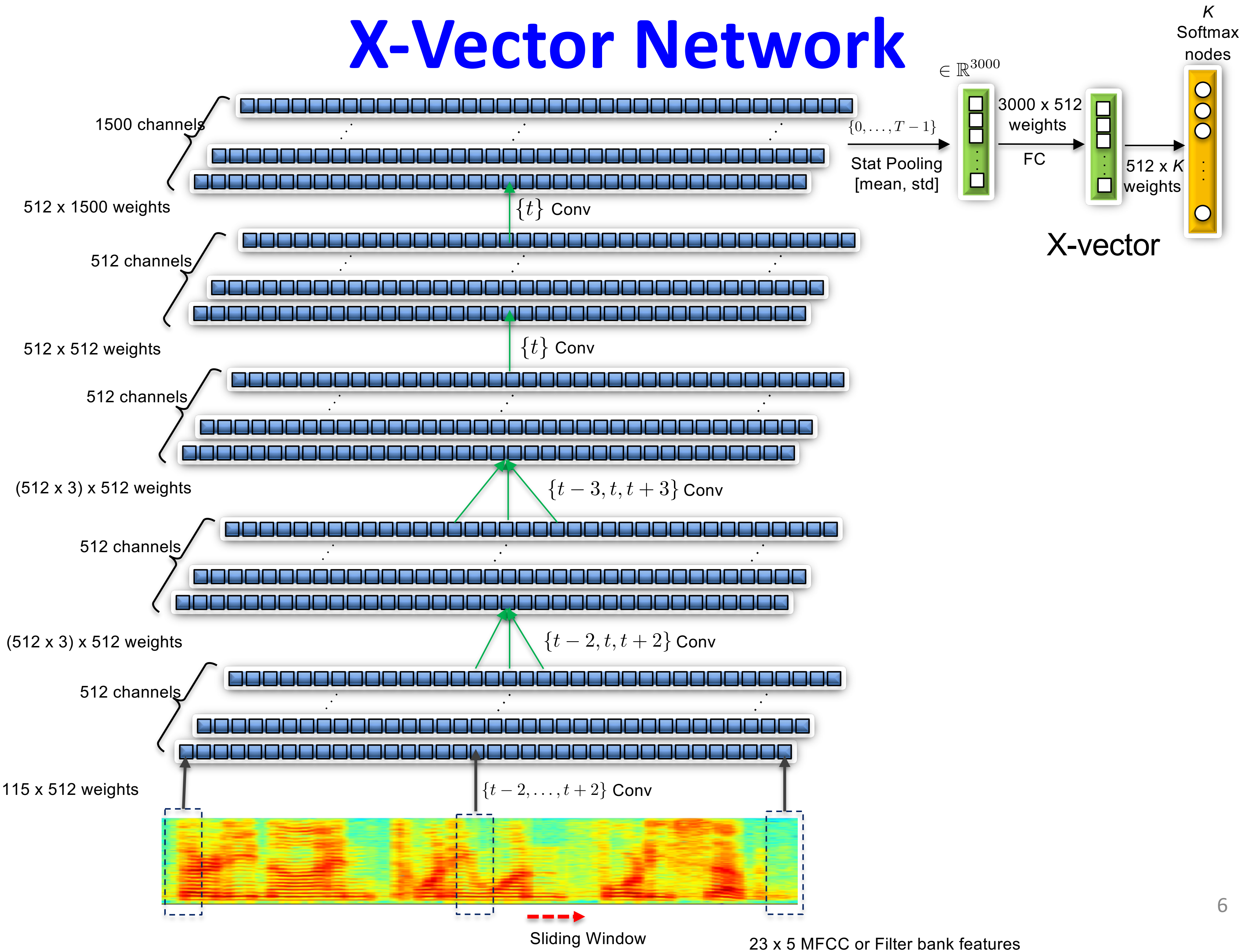


Speaker-embedding Adaptation

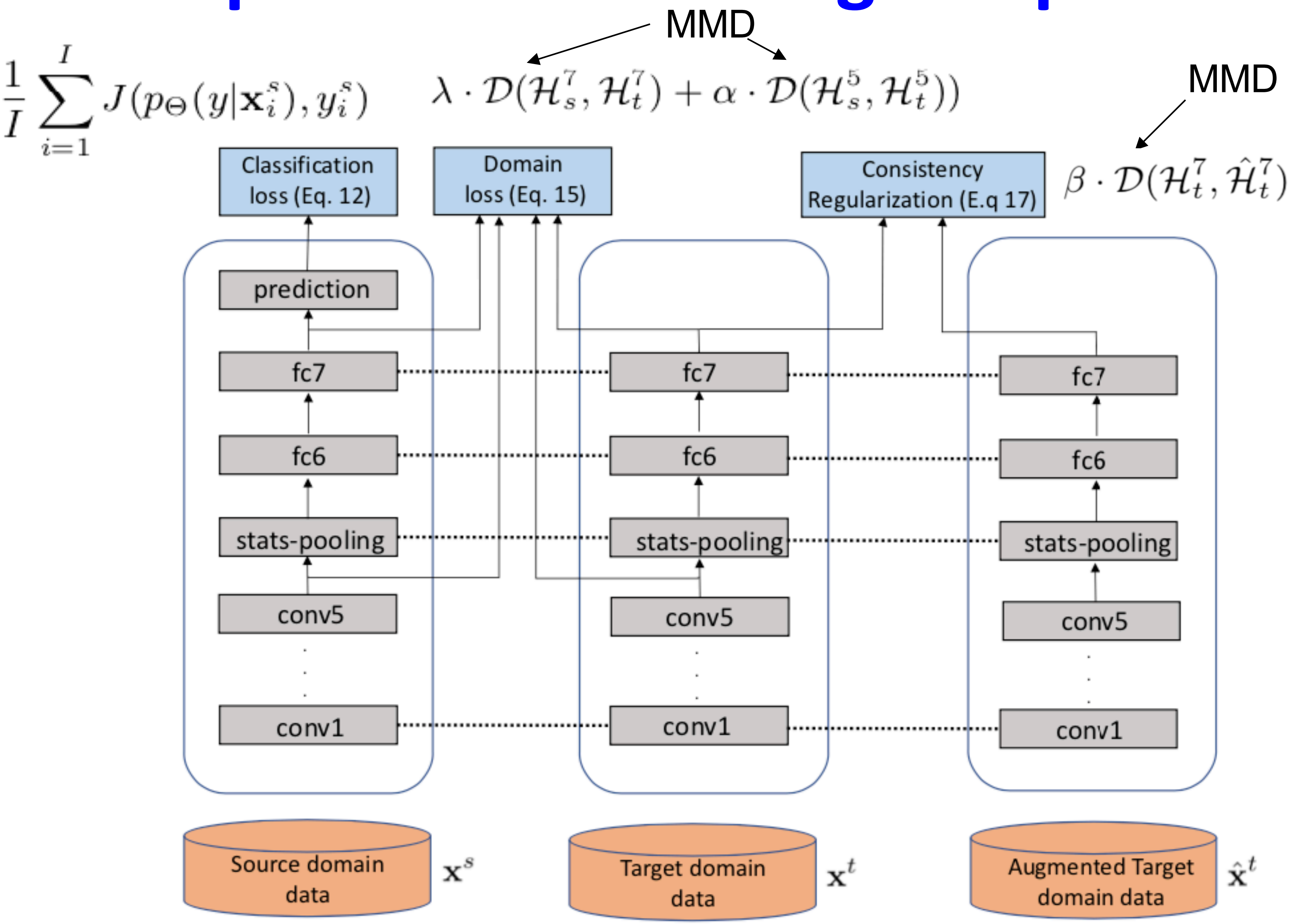
- ***Goal:*** Train the speaker embedding network to produce domain-invariant feature vectors.
- Minimize domain discrepancy at both frame-level and utterance-level
- Apply consistency regularization to leverage unlabeled target-domain data.



X-Vector Network

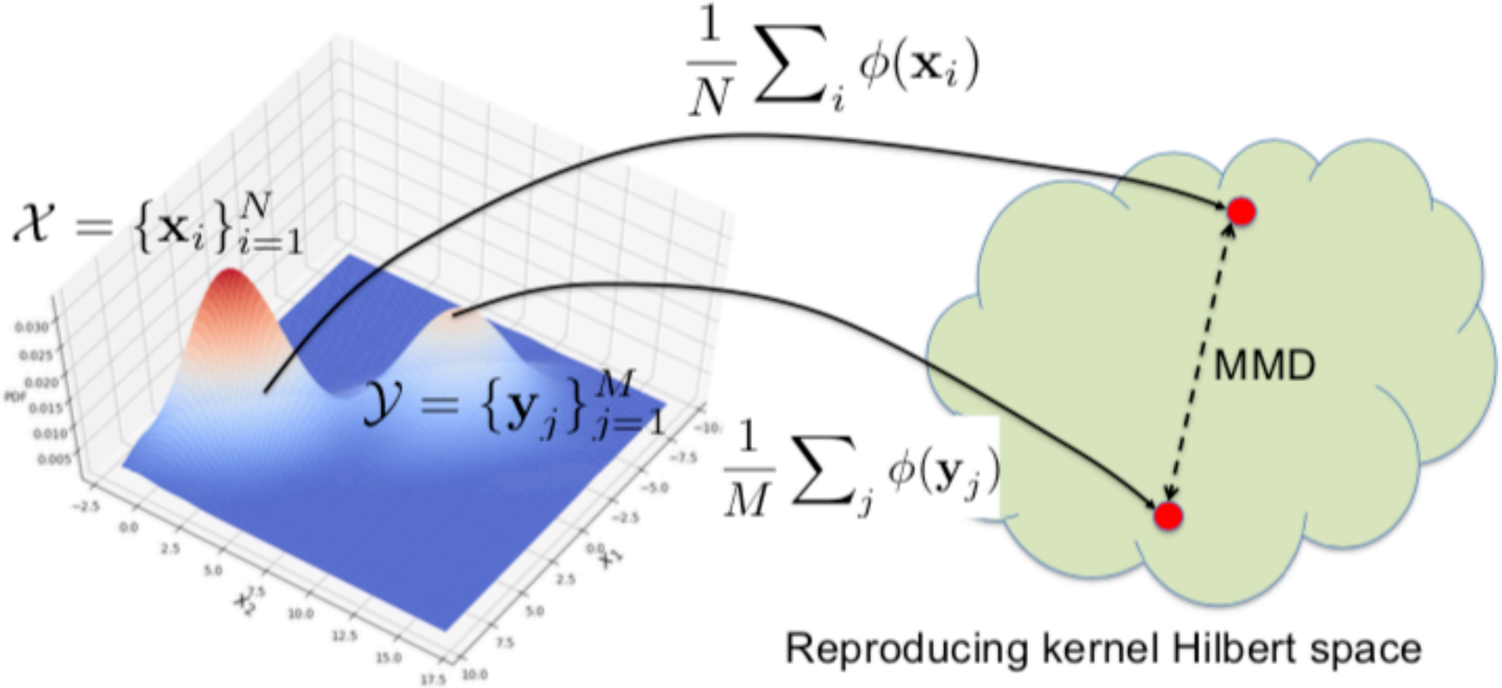


Speaker Embedding Adaptation



Maximum Mean Discrepancy (MMD)

- MMD is a nonparametric approach to measuring the distance between two distributions.
- The basic idea is to non-linearly map the input to an RKHS and compute the distance between the means of the two distributions in that space.



Maximum Mean Discrepancy (MMD)

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

$$= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \phi(\mathbf{x}_i)^\mathsf{T} \phi(\mathbf{x}_{i'}) - \frac{2}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \phi(\mathbf{x}_i)^\mathsf{T} \phi(\mathbf{y}_j)$$

$$+ \frac{1}{M^2} \sum_{j=1}^{M} \sum_{j'=1}^{M} \phi(\mathbf{y}_j)^\mathsf{T} \phi(\mathbf{y}_{j'}).$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} k(\mathbf{x}_i, \mathbf{x}_{i'}) - \frac{2}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} k(\mathbf{x}_i, \mathbf{y}_j) + \frac{1}{M^2} \sum_{j=1}^{M} \sum_{j'=1}^{M} k(\mathbf{y}_j, \mathbf{y}_{j'})$$

Maximum Mean Discrepancy (MMD)

- Quadratic kernel:

$$k(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x})^{\mathsf{T}}\phi(\mathbf{y}) = (\mathbf{x}^{\mathsf{T}}\mathbf{y} + c)^2.$$

$$\mathcal{D}_{\text{MMD}} = 2c\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i - \frac{1}{M}\sum_{j=1}^{M}\mathbf{y}_j\right\|^2 + \left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}} - \frac{1}{M}\sum_{j=1}^{M}\mathbf{y}_j\mathbf{y}_j^{\mathsf{T}}\right\|_F^2$$

- With a quadratic kernel, MMD can measure the distance between two distributions up to their second order stats.
- Multi-RBF kernels:

$$k(\mathbf{x},\mathbf{y}) = \sum_{q=1}^{K}\exp\left(-\frac{1}{2\sigma_q^2}\|\mathbf{x}-\mathbf{y}\|^2\right)$$

Consistency Regularization

- Exploit the unlabeled data for domain adaptation by applying data augmentation on them.
- Consistency training is to regularize a network such that the predictions are consistent even if the network's input is subjected to noise perturbation.
- Achieved by minimizing the KL divergence

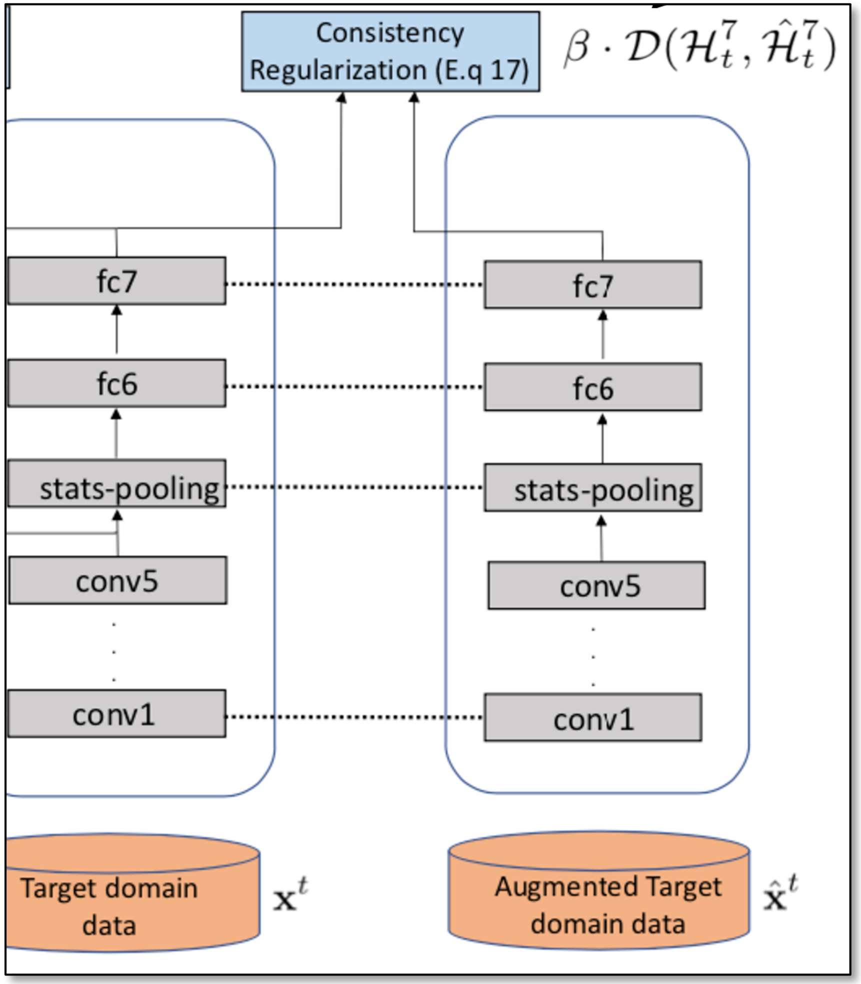
$$\mathop{\mathbb{E}}_{q(\hat{\mathbf{x}}_{\text{unlab}}|\mathbf{x}_{\text{unlab}})}\left[\mathrm{KL}\left(p_{\Theta}(y|\mathbf{x}_{\text{unlab}})\,||\,p_{\Theta}(y|\hat{\mathbf{x}}_{\text{unlab}})\right)\right]$$

  where *q*( ) is a data augmentation transformation, e.g., adding noise or reverb effect.

- We propose minimizing the discrepancy between the embeddings produced by the clean data and the embeddings produced by the augmented data.

Consistency Regularization

- Achieved by minimizing the MMD between target-domain data and unlabeled augmented data:

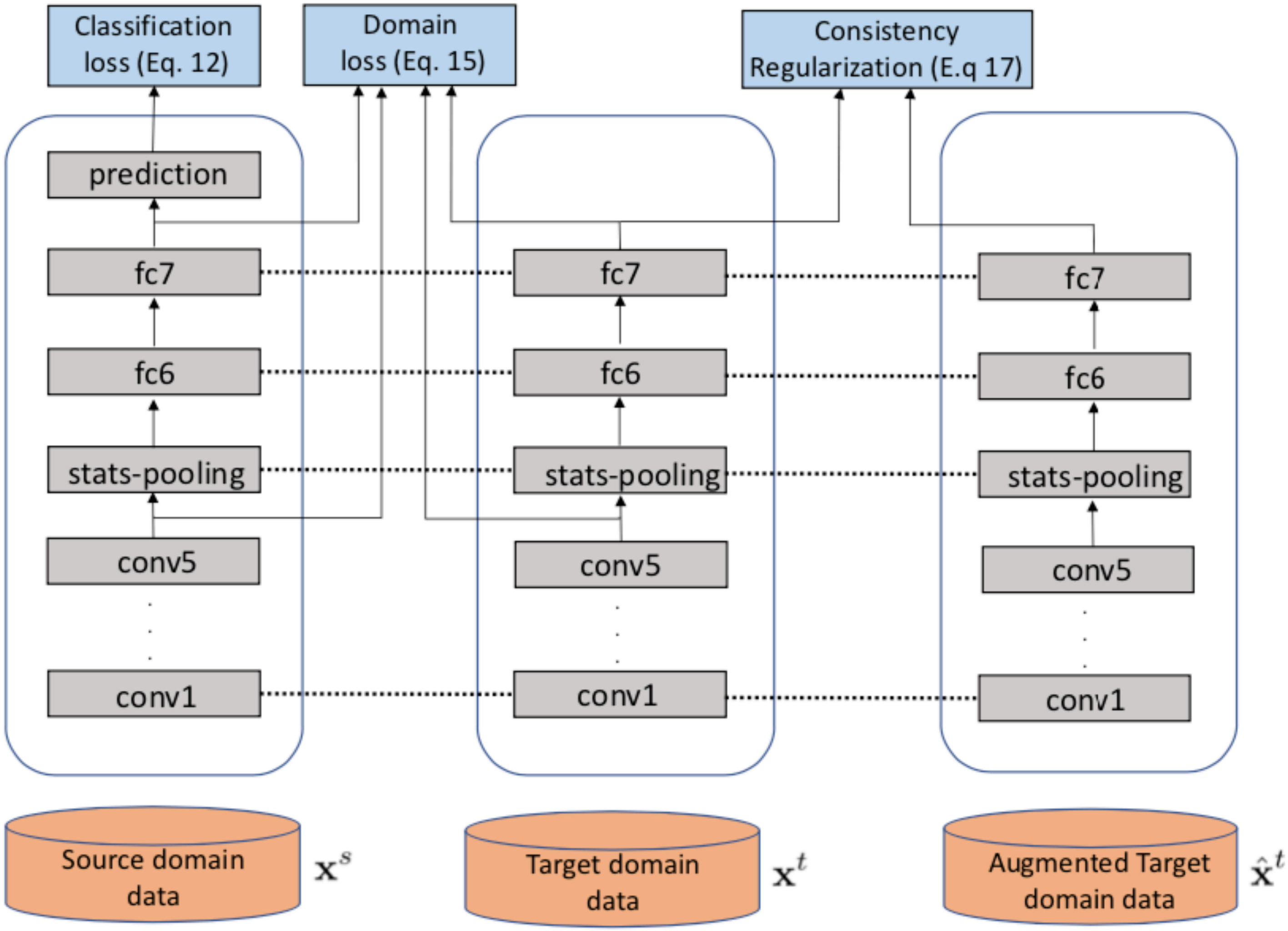


$$\mathcal{D}(\mathcal{H}_t^7, \hat{\mathcal{H}}_t^7) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N} k(\mathbf{h}_i^7, \mathbf{h}_{i'}^7) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} k(\mathbf{h}_i^7, \hat{\mathbf{h}}_j^7) + \frac{1}{M^2}\sum_{j=1}^{M}\sum_{j'=1}^{M} k(\hat{\mathbf{h}}_j^7, \hat{\mathbf{h}}_{j'}^7)$$

Experiments

- **Training data for DNN and PLDA**: 4808 speakers from SRE04-10 and Switchboard
- **Consistency Regularization**: SRE16 and SRE18 unlabeled
- **Test data**: SRE16-eval and SRE18-eval-cmn2
- **Kernel of MMD**: 19 RBF kernels with width ranges from $2^{-8}\sigma_m$ to $2^8\sigma_m$, where σ_m is the median pairwise distance from training data.
- **Acoustic vectors**: 23-dim MFCC with mean norm
- **VAD**: Kaldi's energy-based VAD
- **PLDA adaptation and CORAL**: SRE16 and SRE18 unlabeled
- **Hyperparameters for DNN Objective**: $\alpha = \beta = \lambda = 1$

Experiments



SRE04-10 + Switchboard

SRE16 or SRE18 unlabeled

SRE16 or SRE18 augmented unlabeled

Experiments

- **DNN Architecture**

Layer	Kernel	Channel_in $\times$ Channel_out
Conv1	5,1,1	23×512
Conv2	3,1,2	512×512
Conv3	3,1,3	512×512
Conv4	1,1,1	512×512
Conv5	1,1,1	512×1536
Statistics pooling		1536×3072
FC6	–	3072×512
FC7	–	512×512
Am-softmax	–	$512 \times N$

$$\mathcal{L}_{AMS} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s\cdot\left(\mathbf{W}_{y_i}^T\mathbf{x}_i - m\right)}}{e^{s\cdot\left(\mathbf{W}_{y_i}^T\mathbf{x}_i - m\right)} + \sum_{j=1,j\neq y_i}^{c} e^{s\mathbf{W}_j^T\mathbf{x}_i}},$$

Results

	SRE16		SRE18	
Adapt Method	EER (%)	minDCF	EER(%)	minDCF
WGAN [12]	13.25	0.899	9.59	0.652
Sup. WGAN [12]	9.59	0.746	8.88	0.619
LSGAN [21]	11.74	-	-	-
Our DNN Adapt.	**9.03**	**0.585**	**8.33**	**0.520**

- All the results are without backend adaptation.
- Our DNN adaptation performs significantly better than the previously proposed methods.

Results

	SRE16		SRE18	
Adapt Method	EER(%)	minDCF	EER(%)	minDCF
Our DNN Adapt.	9.03	0.585	8.33	0.520
CORAL Adapt.	8.49	0.560	8.74	0.553
PLDA Adapt.	8.55	0.556	8.88	0.563
Ours+CORAL Adapt.	**8.28**	**0.541**	8.13	0.519
Ours+PLDA Adapt.	8.29	0.546	**8.09**	**0.521**

- Combining the proposed method with backend adaptation further improves the performance.

Results

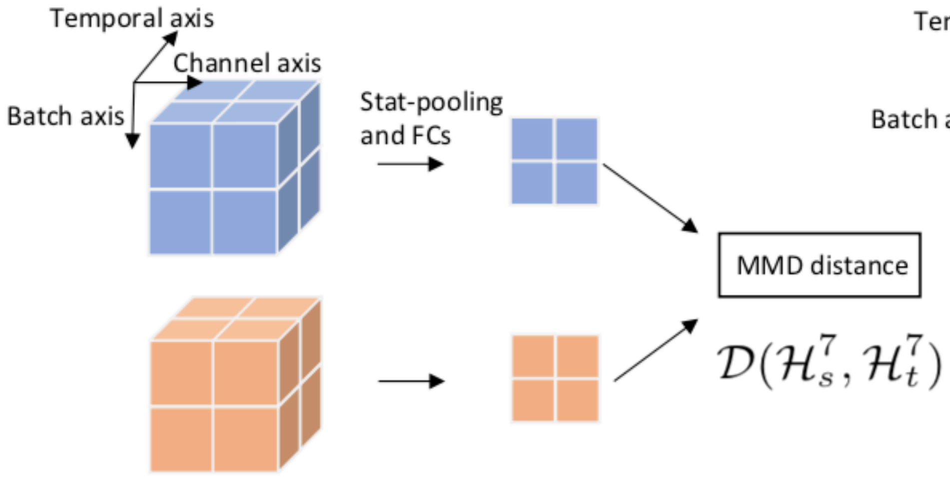
			SRE16		SRE18	
Layer 7	Layer 6	Consis.	EER(%)	DCF	EER(%)	DCF
×	×	×	12.02	0.990	11.59	0.72
✓	×	×	9.79	0.621	9.08	0.580
✓	✓	×	9.63	0.606	8.77	0.555
✓	✓	✓	9.03	0.585	8.33	0.520

- Multi-level adaptation significantly improves the performance in both SRE16 and SRE18.
- Consistency regularization also helps.

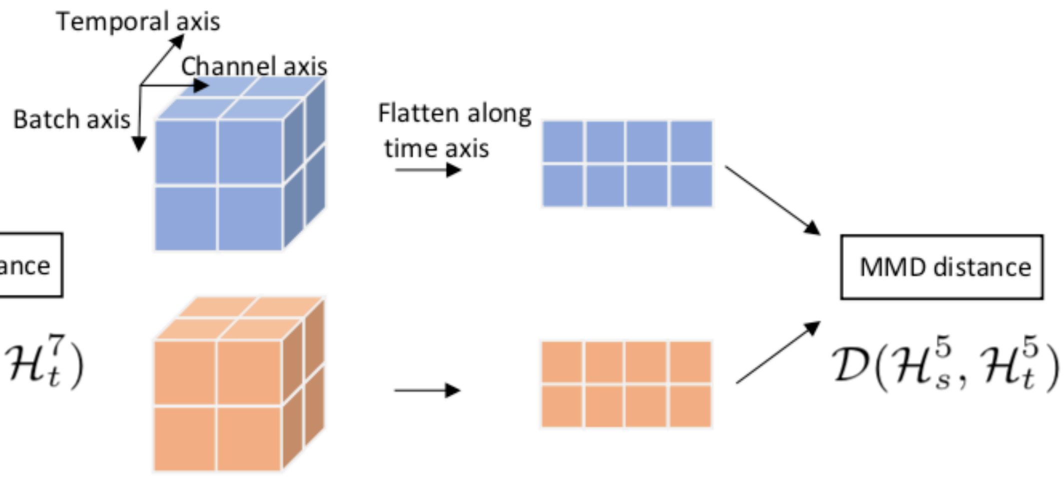
Conclusions

- Domain mismatch loss can be applied at both both frame-level and utterance-level
- Apply MMD at frame level performs significantly better than at utterance-level alone
- Data augmentation can be utilized in the unlabeled target-domain through consistency regularization.

Utterance- and Frame-level MMD



Utterance-level MMD



Frame-level MMD