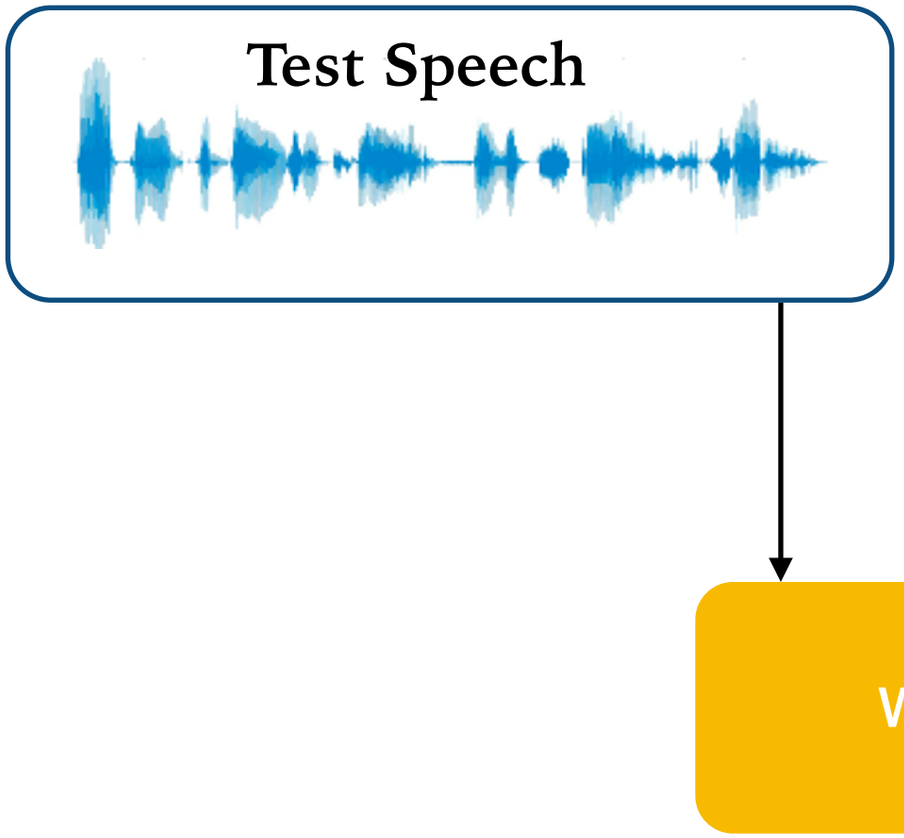
Adversarial Learning for Robust Speaker Verification

Yanmin Qian
SpeechLab, Shanghai Jiao Tong University

Speaker recognition and verification



Speaker recognition: recognize the speaker identity of the test speech given some known enrolled speakers.

Speaker recognition and verification

Test Speech, claimed as Tom

Tom's enrolled speech

Same speaker?

Speaker verification: verify whether the identity of the test speech is the same as the enrolled speech.
Speaker verification is more commonly used in real application.

Text-dependent vs text-independent verification

Text-dependent verification:

- The enrollment and test speech is constrained to the specific phase.
- Can only be used in the condition where the speakers are cooperative.
- Knowledge of spoken text can improve system performance.

Text-independent verification:

- The enrollment and test speech phase is unconstrained.
- More flexible system but also more difficult problem.

Big Progress of Speaker identification with Embedding learning

GMM → (Feature Engineering & Representation Learning) → DNN

I-vector → D-vector (Deep Feature) → X-vector

MIT & CRIM:
i-vector, T-ASLP2011

Speaker:
Google: D-vector, ICASSP2014
SJTU: Deep feature, IS2014/ SpeechCom2015

JHU: X-vector, SLT2016/IS2017/ ICASSP18

Arch: DNN
Aggregate: Frame
Loss: Softmax

→

Better Arch: CNN/TDNN/ResNet
Better Aggr: Seg-level
Better Loss: Triplet/Center/Angular/Focal/ Additive Margin/Additive Angular Margin

Still difficulty and Challenging

- Main Challenges
  - Noise corruption, channel mismatch, multi-domain, multi-speaker, short-utterance, time varying, large scale, anti-spoofing, ……

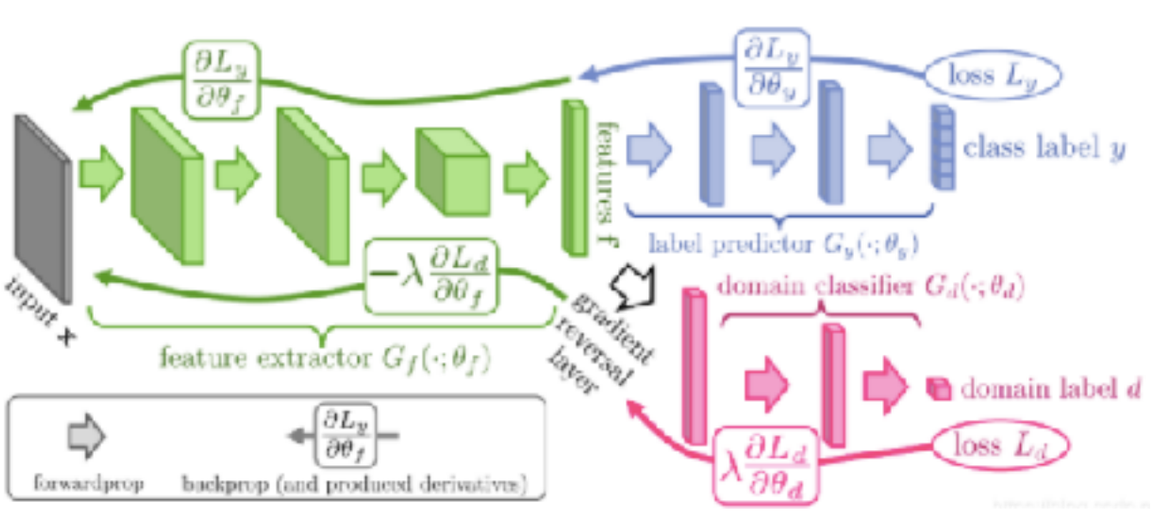
- Variability in many aspects
  - Text mismatch between enrollment and test speech in text-independent verification task.
  - Channel (environment/recording device, etc.) mismatch between enrollment and test speech.
  - Domain (genre/language/age/accent, etc) mismatch between the source domain of the training data in system construction and the target domain in the real system deployment.

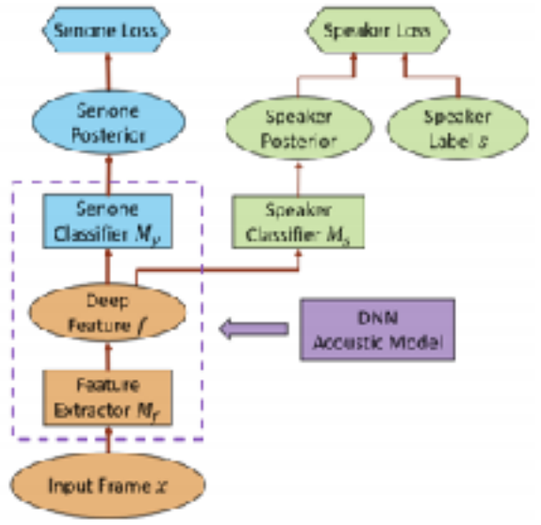
A robust speaker verification system is necessary, which can show the high performance when facing these variabilities !

Adversarial Learning

Adversarial learning is popularly used to remove the nuisance information unrelated to the specific task.

- **For speech recognition, adversarial training has been used to suppress the variety from the speakers or domains.**
- **For speaker anti-spoofing, adversarial training has been used to mitigate the mismatch between different corpus.**





- **The adversarial learning is fully explored for speaker verification in our work, to reduce the mismatch from the variability in text, channel and domain.**

Adversarial learning for text variation in robust speaker verification

Work#1: Multi-task and adversarial training using phonetic information.

MOTIVATION: Previous work shows that encouraging phonetic information learning at frame level with multi-task learning is helpful for both the text-dependent and text-independent verification task. However, intuitively, it's doubted that spoken content should matter for speaker embedding.

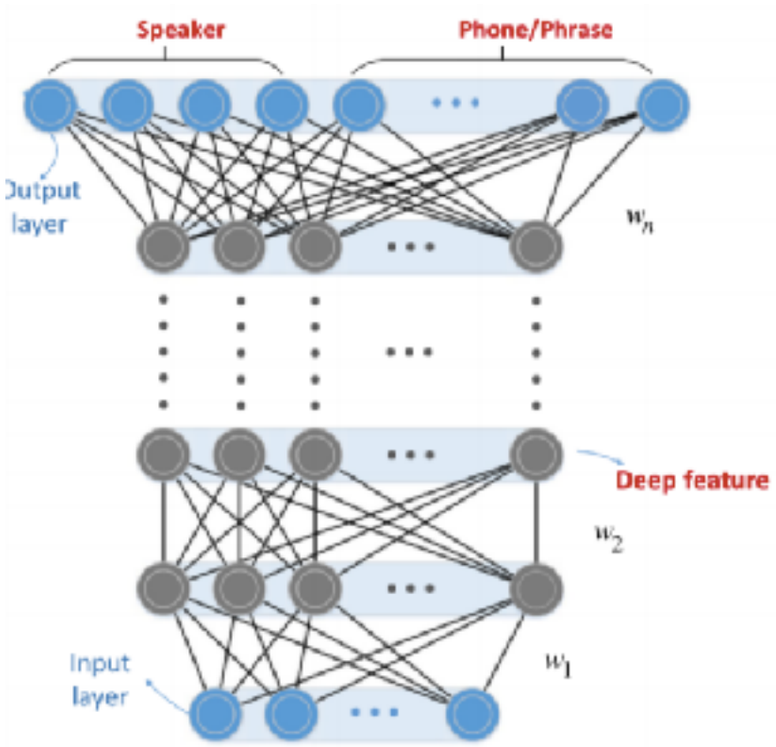
QUESTION: Whether to suppress (adversarial training) phonetic information or encourage (multi-task training) phonetic information learning for robust text-independent verification task?

TASK: Detail analysis about the multi-task training and adversarial training using phonetic information is necessary.

Shuai Wang, Yanmin Qian and Kai Yu. On the Usage of Phonetic Information for Text-independent Speaker Embedding Extraction. InterSpeech 2019

Previous related works

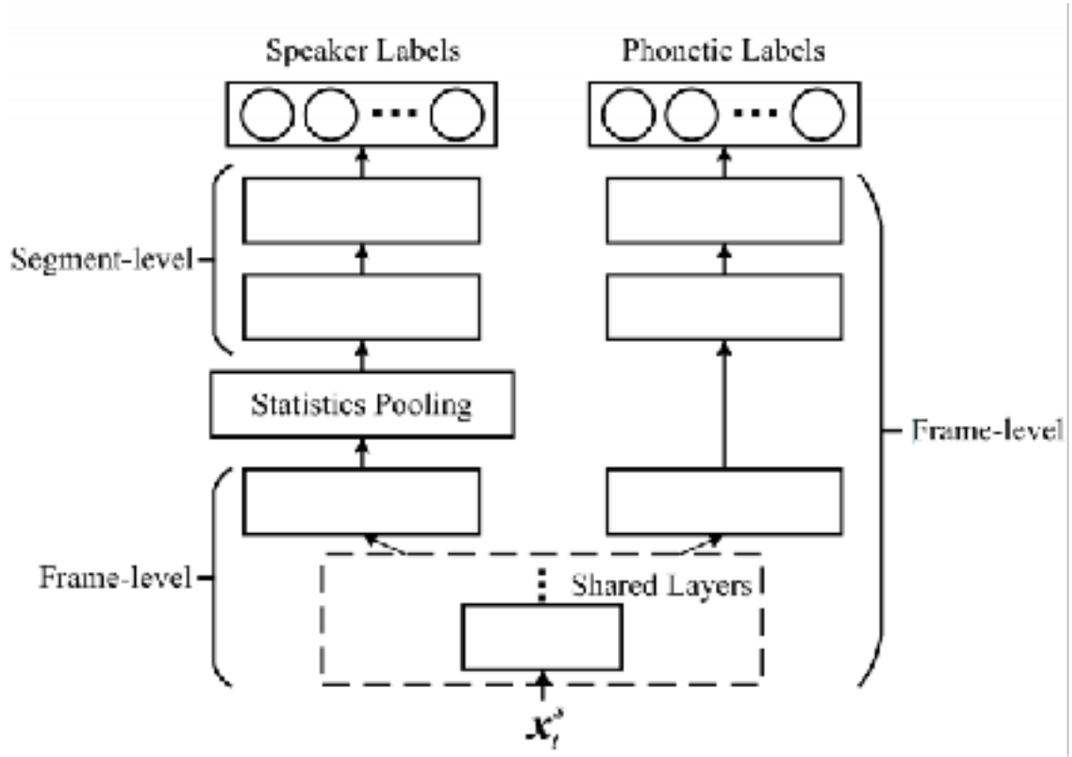
multi-task learning for text-dependent SID task



Explicitly modeling phonetic information helps the text-dependent speaker verification, which is intuitive.

Yuan Liu, et al. Deep feature for text-dependent speaker verification. Speech Communication, 2015

multi-task learning for text-independent SID task



Why explicitly learning phonetic information helps the text-independent speaker verification task? Is it counter-intuitive?

Yi Liu, et al. Speaker Embedding Extraction with Phonetic Information. Interspeech 2018

Frame level multi-task and adversarial training

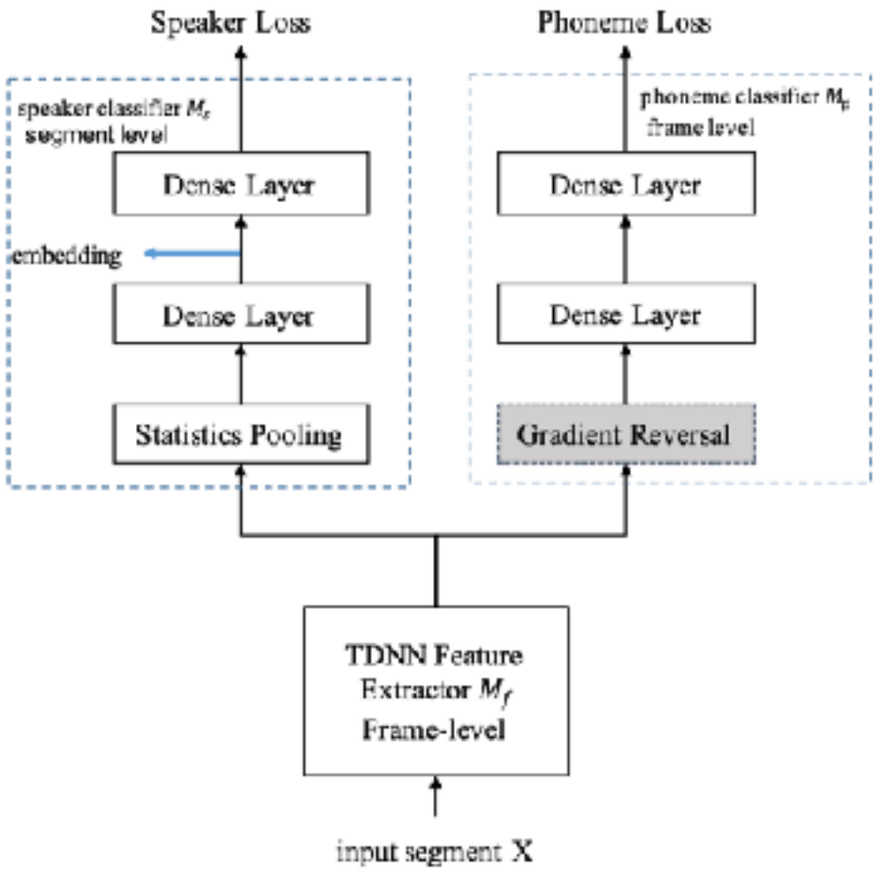


Figure 1: *Structure of the frame-level multitask (without the gradient reversal layer) and adversarial learning for x-vector*

Multi-task training:
Supervised by speaker classification loss and phoneme loss. Evaluated on Voxceleb1.

Table 1: *Systems combining frame-level phonetic information, FRM-MT and FRM-ADV denote two systems described in Section 2, trained using multitask or adversarial objectives, with or without the gradient reversal layer, respectively*

System	EER(%)	$\text{minDCF}_{0.01}$	$\text{minDCF}_{0.1}$
x-vector baseline	3.73	0.389	0.192
FRM-MT	3.38	0.357	0.180
FRM-ADV	5.24	0.502	0.269

Frame-level multi-task training improve the performance.

Shuai Wang, Yanmin Qian and Kai Yu. On the Usage of Phonetic Information for Text-independent Speaker Embedding Extraction. InterSpeech 2019

Frame level multi-task and adversarial training

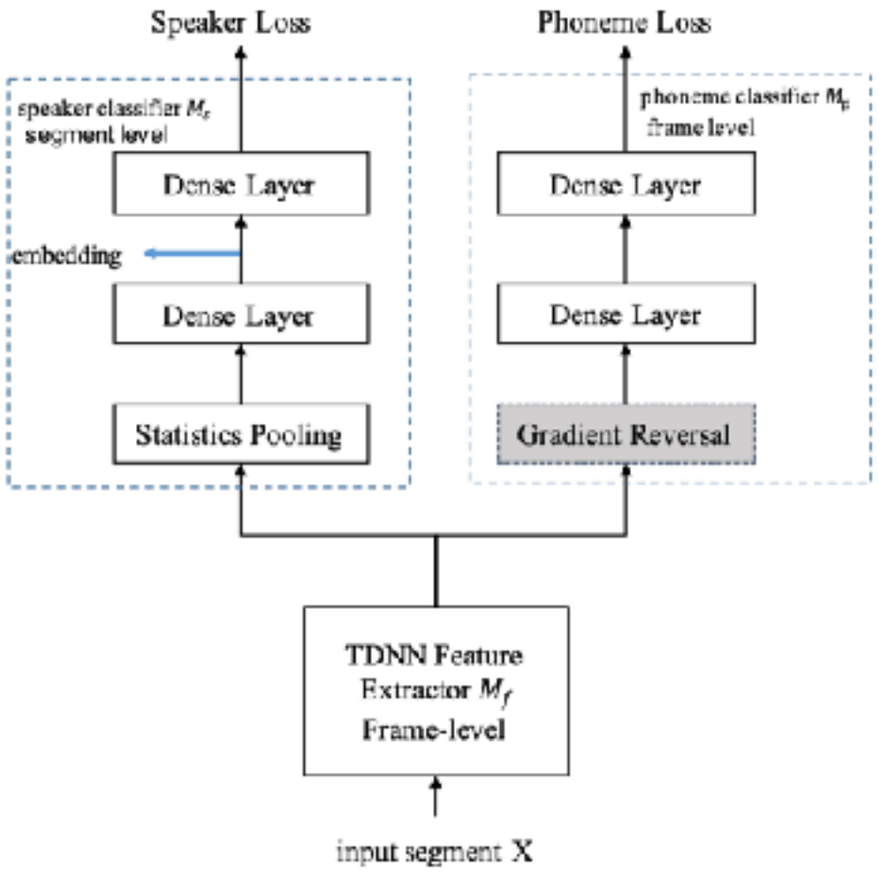


Figure 1: *Structure of the frame-level multitask (without the gradient reversal layer) and adversarial learning for x-vector*

Adversarial training:
Gradient reversal layer is inserted to remove the phonetic information.

Table 1: *Systems combining frame-level phonetic information, FRM-MT and FRM-ADV denote two systems described in Section 2, trained using multitask or adversarial objectives, with or without the gradient reversal layer, respectively*

System	EER(%)	$\text{minDCF}_{0.01}$	$\text{minDCF}_{0.1}$
x-vector baseline	3.73	0.389	0.192
FRM-MT	3.38	0.357	0.180
FRM-ADV	5.24	0.502	0.269

However frame-level adversarial training hurts the performance.

Shuai Wang, Yanmin Qian and Kai Yu. On the Usage of Phonetic Information for Text-independent Speaker Embedding Extraction. InterSpeech 2019

Segment level multi-task and adversarial training

How to construct segment level phoneme label?

Phoneme soft label construct according to the phoneme statistic in a segment.

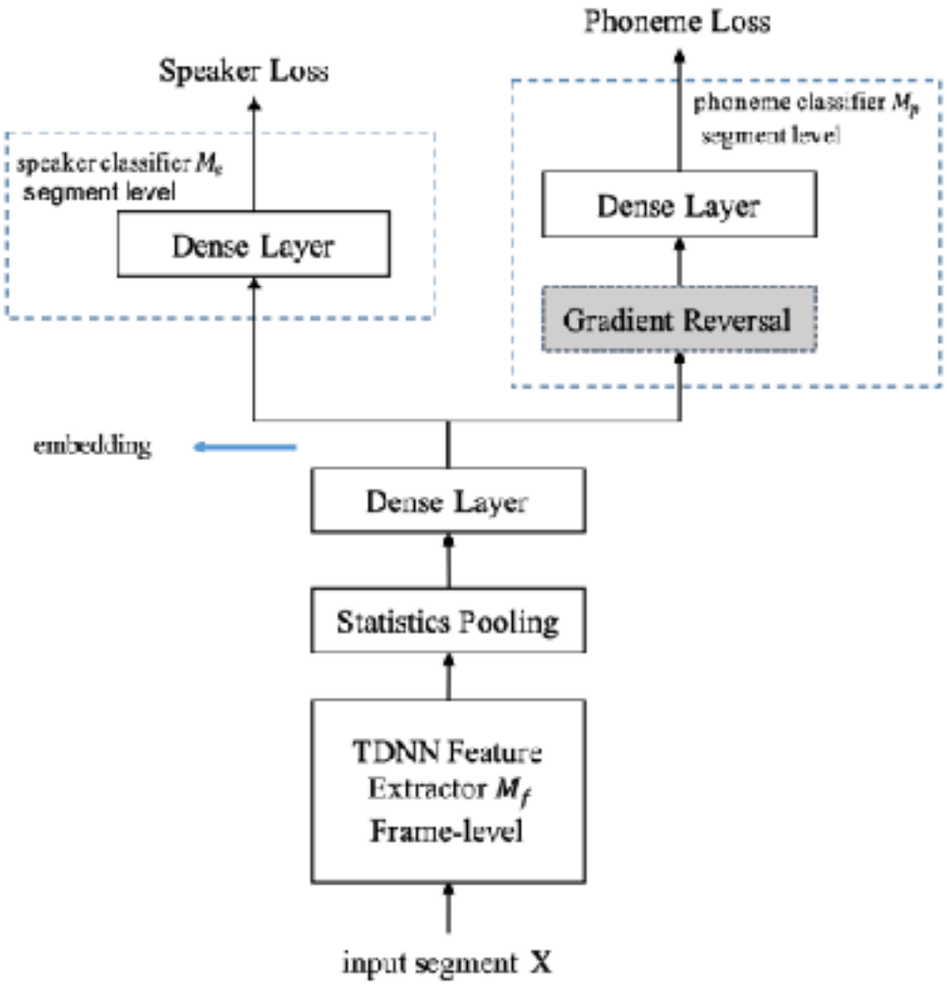
$$\mathbf{y}^p = y_1, \ldots, y_C, \; y_c = \frac{N_c}{N}$$

Such soft label replaces the hard label in cross entropy loss.

$$\mathcal{L}_p = \mathrm{CE}(M_p(M_f(\mathbf{x}_i)), \mathbf{y}^p)$$

Shuai Wang, Yanmin Qian and Kai Yu. On the Usage of Phonetic Information for Text-independent Speaker Embedding Extraction. InterSpeech 2019

Segment level multi-task and adversarial training



The phoneme classification branch is inserted at the embedding level rather frame level.

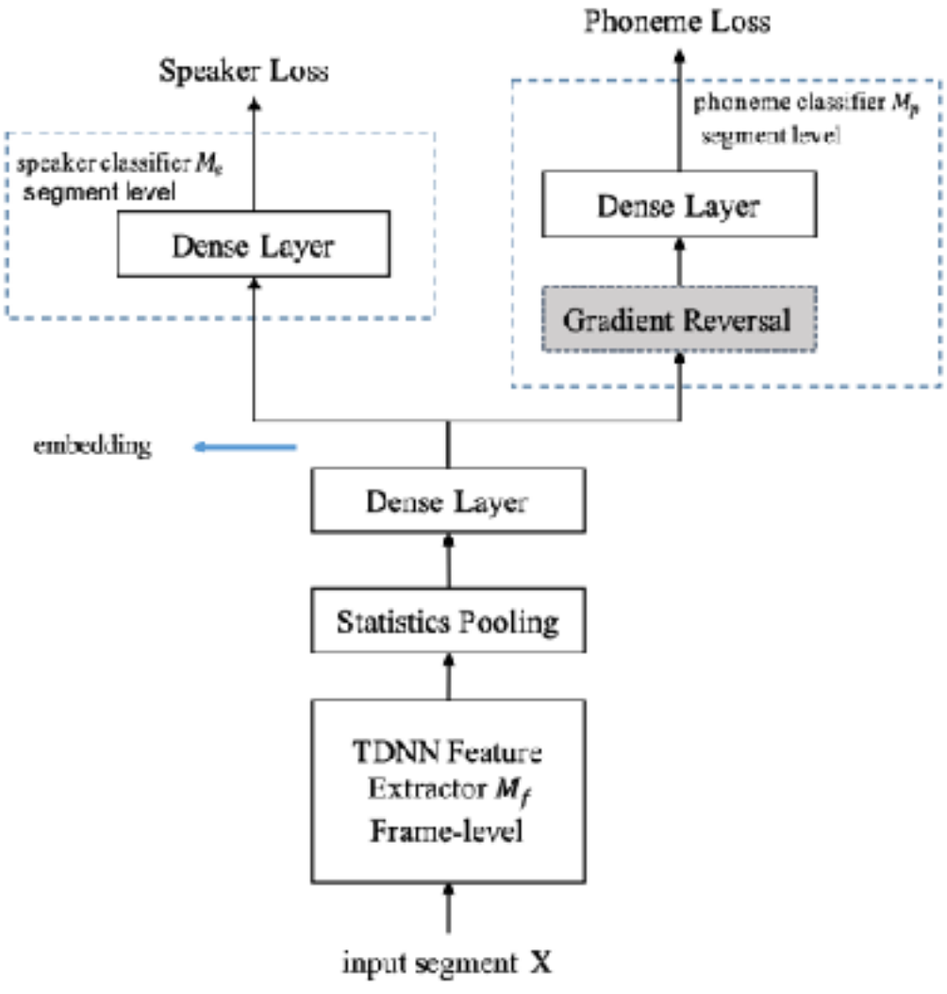
Table 2: *Systems combining segment-level phonetic information, SEG-MT and SEG-ADV denote two systems described in Section 3, trained using multitask or adversarial objectives, with or without the gradient reversal layer, respectively*

System	EER(%)	$\text{minDCF}_{0.01}$	$\text{minDCF}_{0.1}$
x-vector baseline	3.73	0.389	0.192
SEG-MT	3.71	0.327	0.175
SEG-ADV	3.35	0.332	0.159

Segment level multi-task training obtains no obvious improvement.

Shuai Wang, Yanmin Qian and Kai Yu. On the Usage of Phonetic Information for Text-independent Speaker Embedding Extraction. InterSpeech 2019

Segment level multi-task and adversarial training



The phoneme classification branch is inserted at the embedding level rather frame level.

Table 2: *Systems combining segment-level phonetic information, SEG-MT and SEG-ADV denote two systems described in Section 3, trained using multitask or adversarial objectives, with or without the gradient reversal layer, respectively*

System	EER(%)	$\text{minDCF}_{0.01}$	$\text{minDCF}_{0.1}$
x-vector baseline	3.73	0.389	0.192
SEG-MT	3.71	0.327	0.175
SEG-ADV	3.35	0.332	0.159

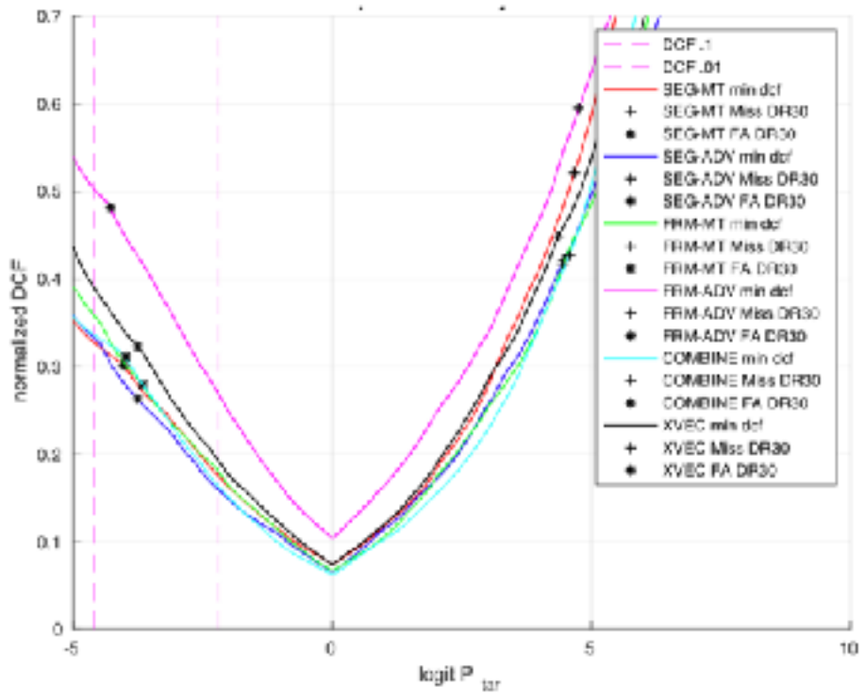
Segment level adversarial training significantly improves the performance.

Shuai Wang, Yanmin Qian and Kai Yu. On the Usage of Phonetic Information for Text-independent Speaker Embedding Extraction. InterSpeech 2019

Multi-task and adversarial training using phonetic information

Table 3: *Systems combining frame-level multitask and segment-level adversarial learning, COMBINE denotes the architecture which performs both strategies*

System	EER(%)	$\text{minDCF}_{0.01}$	$\text{minDCF}_{0.1}$
x-vector baseline	3.73	0.389	0.192
FRM-MT	3.38	0.357	0.180
SEG-ADV	3.35	0.332	0.159
COMBINE	3.17	0.336	0.163



- Encourage the phonetic information at the frame level for generic feature learning.
- Suppress the phonetic variety at the segment level to obtain the more speaker-discriminant embeddings.

Shuai Wang, Yanmin Qian and Kai Yu. On the Usage of Phonetic Information for Text-independent Speaker Embedding Extraction. InterSpeech 2019

Conclusion

- Frame-level phonetic information is helpful for generic feature learning and can be encouraged with multi-task learning.
- Phonetic information should be suppressed at the embedding level with adversarial learning for extracting text-independent speaker embedding.
- Combining both multi-task and adversarial learning can take both advantages on phonetic information usage, and it achieves the best system performance for text-independent speaker verification.

Shuai Wang, Yanmin Qian and Kai Yu. On the Usage of Phonetic Information for Text-independent Speaker Embedding Extraction. InterSpeech 2019

Adversarial learning for channel variation in robust speaker verification

Work#2: Multi-task and adversarial training using channel information.

MOTIVATION: Speaker verification system degrades dramatically due to the channel mismatch (environment, device, etc) between enrollment and test speech. Speaker embedding should be channel-invariant.

QUESTION: Should we simply suppress the channel information or better channel information usage method should be explored?

INSPIRATION: From work#1, simply suppressing the phonetic information is not the best choice for text-independent speaker verification task. Should we use channel information in a similar way?

DATASET: A wake-up word text-dependent dataset.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Channel Invariant Speaker Embedding Learning with Joint Multi-Task and Adversarial Training. ICASSP 2020

Multi-task and adversarial training using channel information

- Channel information resides at the segment level.
- We explored to do multi-task or adversarial training after the statistic layer (lower level) or at the embedding layer.

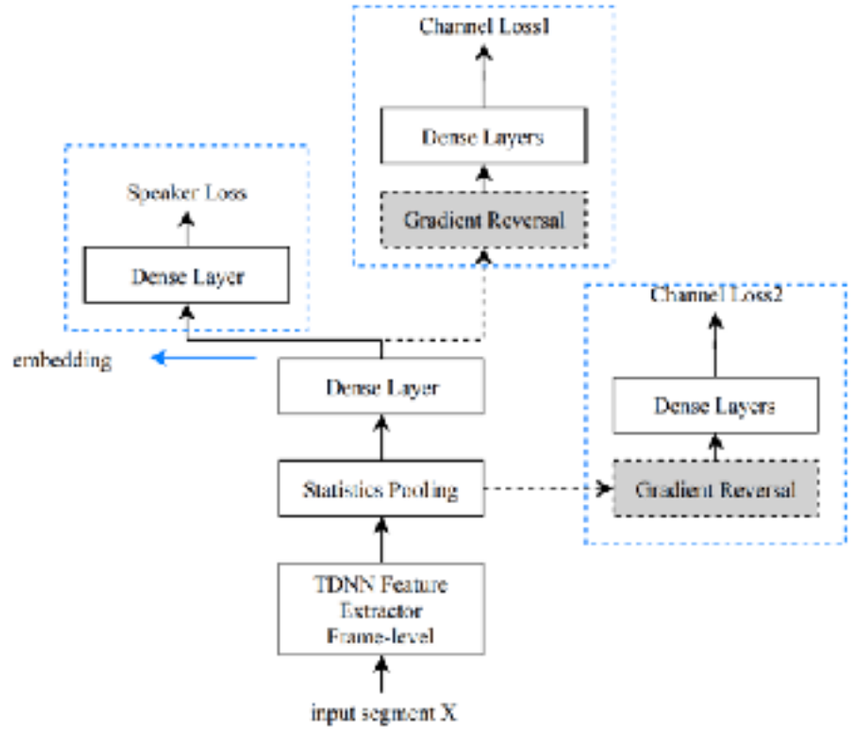


Fig. 2. The proposed structure of applying channel-level multi-task and adversarial training at the different positions of the model.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Channel Invariant Speaker Embedding Learning with Joint Multi-Task and Adversarial Training. ICASSP 2020

Multi-task and adversarial training using channel information

- Recording environment (6 different recording conditions) is considered as channel information here.
- Multi-task training should be performed at the statistic level (low level) to encourage the generic acoustic feature learning.
- Adversarial training should be performed at the embedding level to extract channel invariant embedding.

Table 2. Comparison of multitask or adversarial training results at different position of the model using environment information, STA-MT and STA-ADV denote multitask or adversarial training at the statistics pooling layer, while EMB-MT and EMB-ADV denote the related learning is performed at the embedding layer.

System	Dataset (EER(%))		
	Device1	Device2	Average
baseline	6.12	8.26	7.24
EMB-MT	6.61	7.93	7.27
STA-MT	**6.07**	**7.86**	**6.97**
EMB-ADV	**5.91**	**7.78**	**6.85**
STA-ADV	6.08	8.18	7.13

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Channel Invariant Speaker Embedding Learning with Joint Multi-Task and Adversarial Training. ICASSP 2020

Multi-task and adversarial training using channel information

Table 3. Comparison of two training strategies using environment information for the proposed architecture, JOINT denotes joint multitask-adversarial training mode and PROGRESSIVE denotes the progressive multitask-adversarial training mode for the proposed architecture.

System	Dataset (EER(%))		
	Device1	Device2	Average
baseline	6.12	8.26	7.24
EMB-ADV	5.91	7.78	6.85
STA-MT	6.07	7.86	6.97
JOINT	5.83	7.41	6.62
PROGRESSIVE	**5.57**	**7.36**	**6.46**

- JOINT: Perform EMB-ADV and STA-MT during the whole training period.
- PROGRESSIVE: STA-MT training is only applied at the first half training period and then we do EMB-ADV training.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Channel Invariant Speaker Embedding Learning with Joint Multi-Task and Adversarial Training. ICASSP 2020

Multi-task and adversarial training using channel information

Table 4. Comparison of different systems using device information.

System	EER(%)
baseline	4.27
EMB-MT	4.12
STA-MT	4.09
EMB-ADV	4.10
STA-ADV	4.23
JOINT	3.93
PROGRESSIVE	**3.87**

When using recording device (5 devices) as channel information, we get the same conclusion.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Channel Invariant Speaker Embedding Learning with Joint Multi-Task and Adversarial Training. ICASSP 2020

Conclusion

- Channel information should be encouraged at the lower level and then be suppressed at the embedding level.
- A joint multi-task and adversarial learning can better utilize the channel information than the individual one, which can make the system more robust on different channels.
- The progressive mode of joint learning with multi-task and adversarial learning can better optimize the systems.
- It's better to do lower level multi-task learning at the early training stage to encourage generic acoustic feature learning. Then, system can focus on learning channel-invariant embedding by doing embedding level adversarial training.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Channel Invariant Speaker Embedding Learning with Joint Multi-Task and Adversarial Training. ICASSP 2020

Adversarial learning for domain variation in robust speaker verification

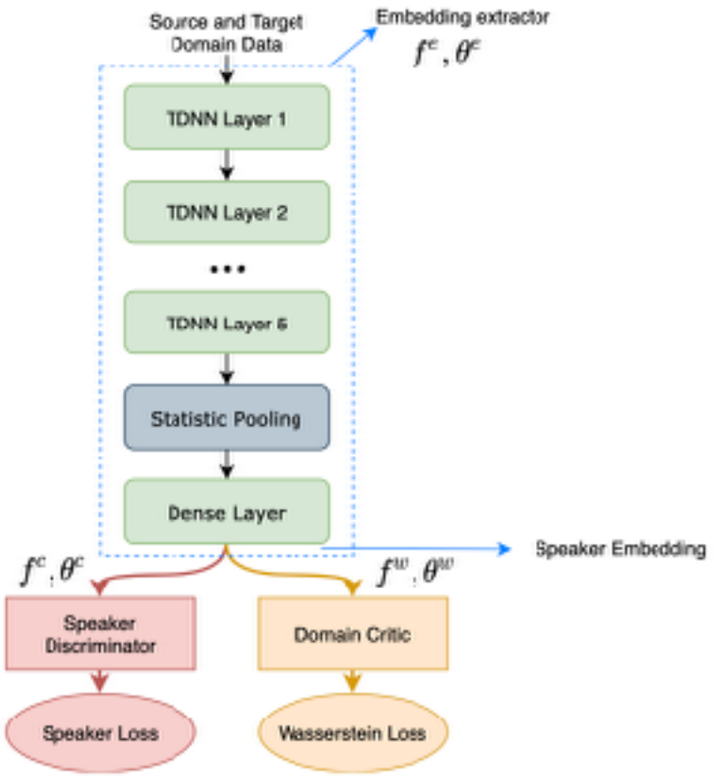
Work#3: The adversarial architecture for unsupervised domain adaptation.

MOTIVATION: Data from different domain always has different distributions, and a well trained system in on domain may perform badly in another domain. Adversarial learning is usually used to reduce the domain mismatch, however the normal fully shared neural network for different domain data in domain adaptation is inappropriate.

SOLUTION: revise the normal fully shared NNs to partially shared ones in adversarial domain adaptation.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network. InterSpeech 2020

Review of unsupervised adversarial adaptation

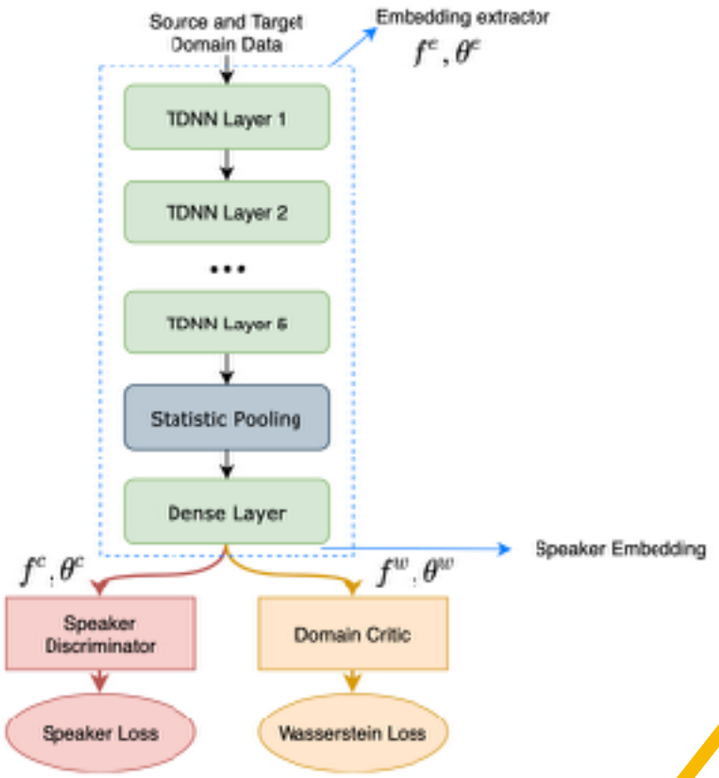


Source domain: large amount of well-labeled data available.
Target domain: only small amount of un-labeled data available.

- Speaker loss: train a speaker discriminant embedding extractor based on the labeled source domain data.
- Adversarial loss: reduce the mismatch between the source and target domain embedding distribution, and make the new speaker embeddings domain-invariant for different domains.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network. InterSpeech 2020

Problem of unsupervised adversarial adaptation



Performance improves or degrades?

Adversarial training using GAN based loss.

Compensate domain mismatch

Improve

Target domain performance

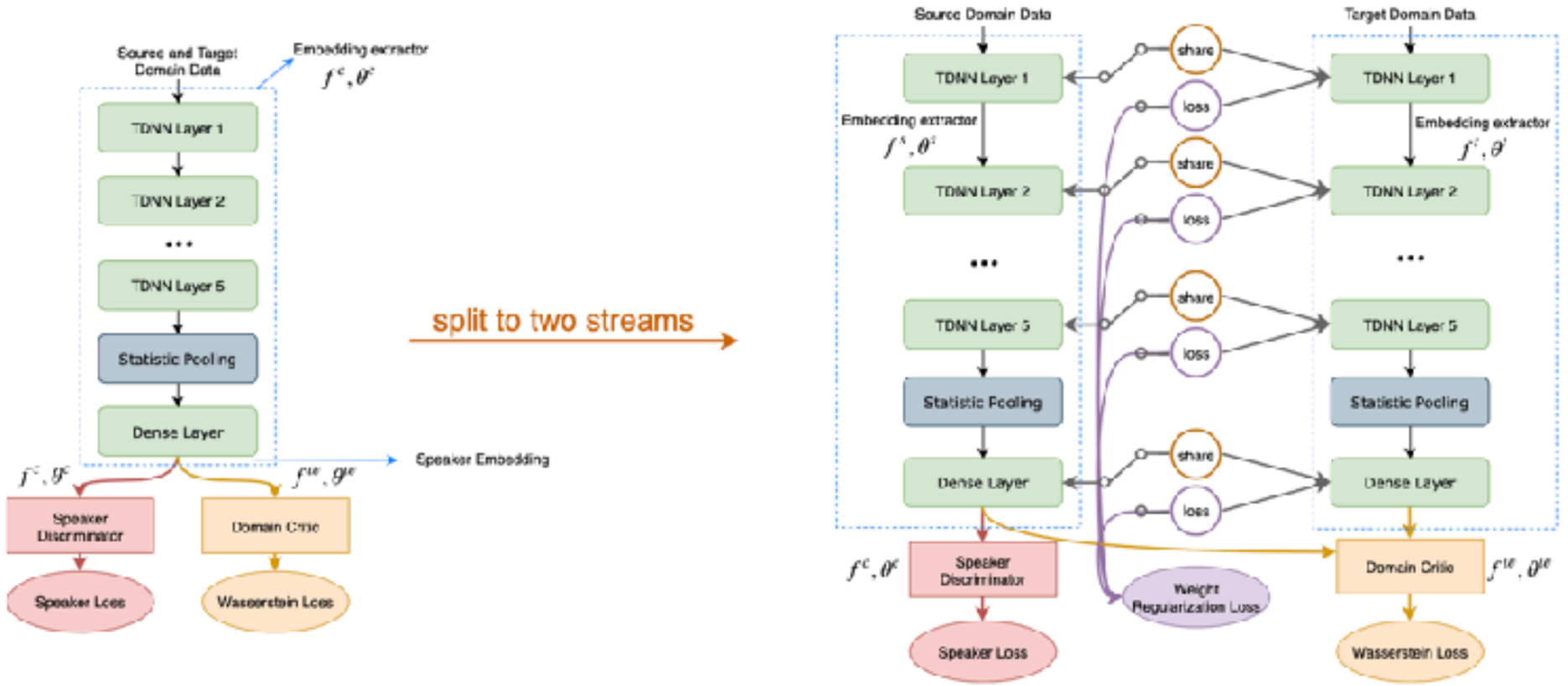
Degrade speaker discrimination

Hurt

- Different domain data has too large data distribution mismatch.
- Pushing the whole model to compensate such mismatch may hurt the speaker discrimination at the same time.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network. InterSpeech 2020

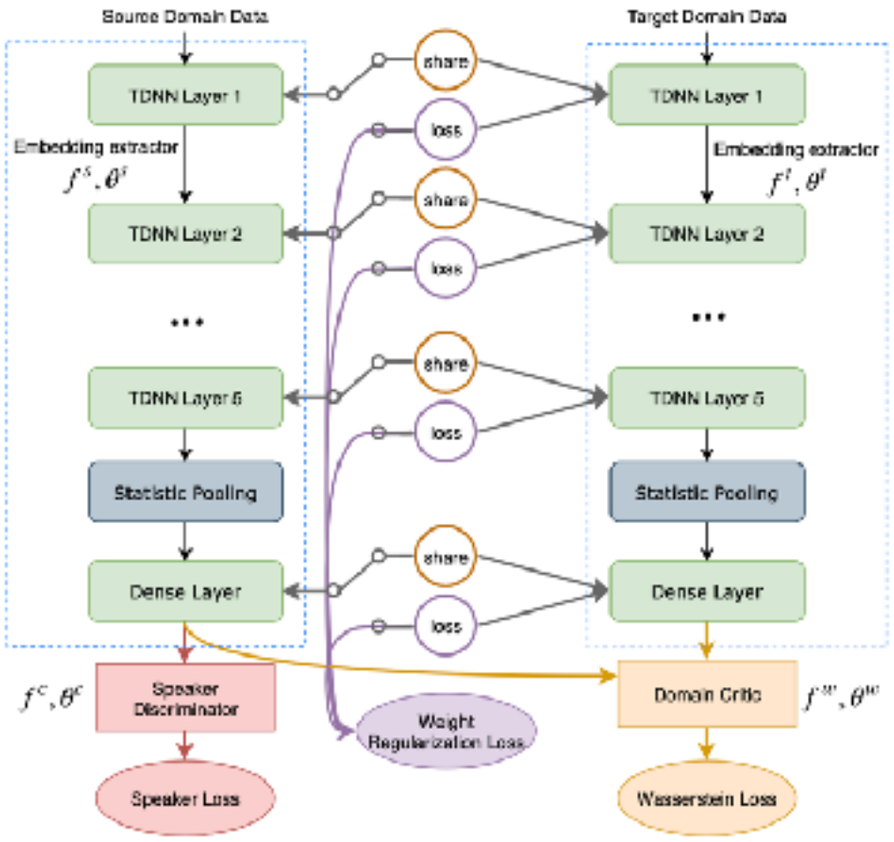
Partially shared network for unsupervised adversarial adaptation



- Partially shared: parameters from the same-level layer can either be shared or not.
- Domain specific parameters can naturally compensate the domain mismatch without influencing the speaker discrimination ability.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network. InterSpeech 2020

Partially shared network for unsupervised adversarial adaptation



$$\mathcal{L}_{PSN} = \mathcal{L}_c + \lambda_w \mathcal{L}_w + \lambda_r \mathcal{L}_r$$

$$\mathcal{L}_r = \sum_{j \in \Omega} \left[exp\left(\left\| \theta_j^s - \theta_j^t \right\|^2 \right) - 1 \right]$$

Lc : Speaker classification loss
Lw : WGAN based adversarial loss
Lr : Weight regularization loss

- Adversarial loss: reduce the mismatch on the embedding distribution from different domains.
- Weight regularization loss: keeping the weight of target domain extractor not far-away from the source domain.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network. InterSpeech 2020

Partially shared network for unsupervised adversarial adaptation

Source and target domain joint training

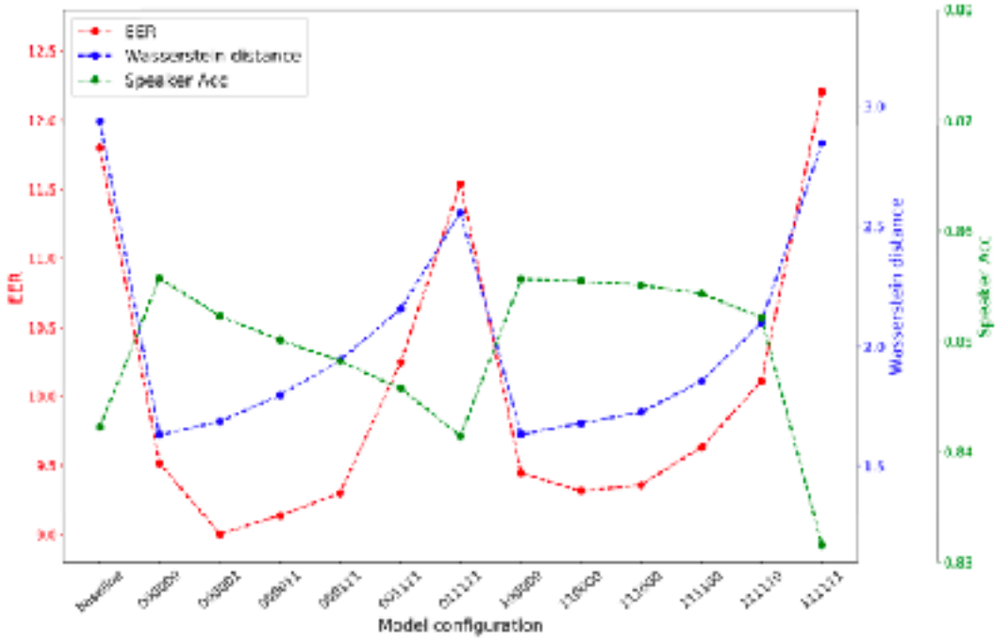


Figure 2: *The results of different weight sharing strategies with jointly training the source and target extractors, and EER (%) denotes the pooled results on SRE16. On the $x-$axis,* 1 *or* 0 *denotes whether or not to share the weights of the corresponding layer (from the lowest to the highest layer, low means close to the input layer), e.g.* 100000 *means only the parameters of the lowest layer is shared.*

Source domain data: NIST SRE04-10 & Switchboard, English (Labeled data)
Target domain data: NIST SRE16, Cantonese/Tagalog (Unlabeled data)

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network. InterSpeech 2020

Partially shared network for unsupervised adversarial adaptation

Source and target domain joint training

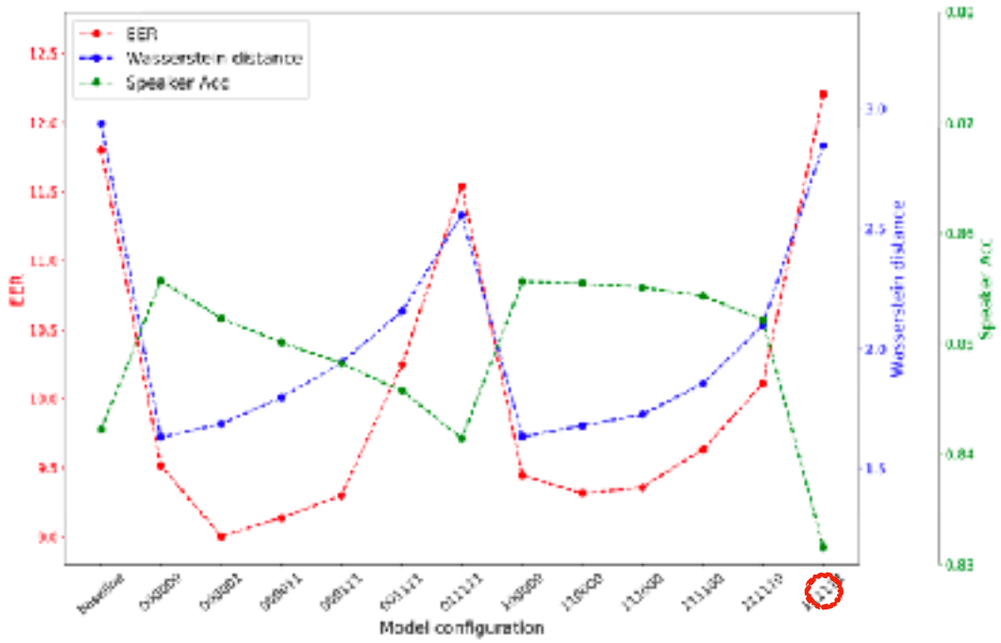


Figure 2: *The results of different weight sharing strategies with jointly training the source and target extractors, and EER (%) denotes the pooled results on SRE16. On the $x-$axis,* 1 *or* 0 *denotes whether or not to share the weights of the corresponding layer (from the lowest to the highest layer, low means close to the input layer), e.g.* 100000 *means only the parameters of the lowest layer is shared.*

Fully shared neural network indeed hurt the speaker discrimination ability.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network. InterSpeech 2020

Partially shared network for unsupervised adversarial adaptation

Source and target domain joint training

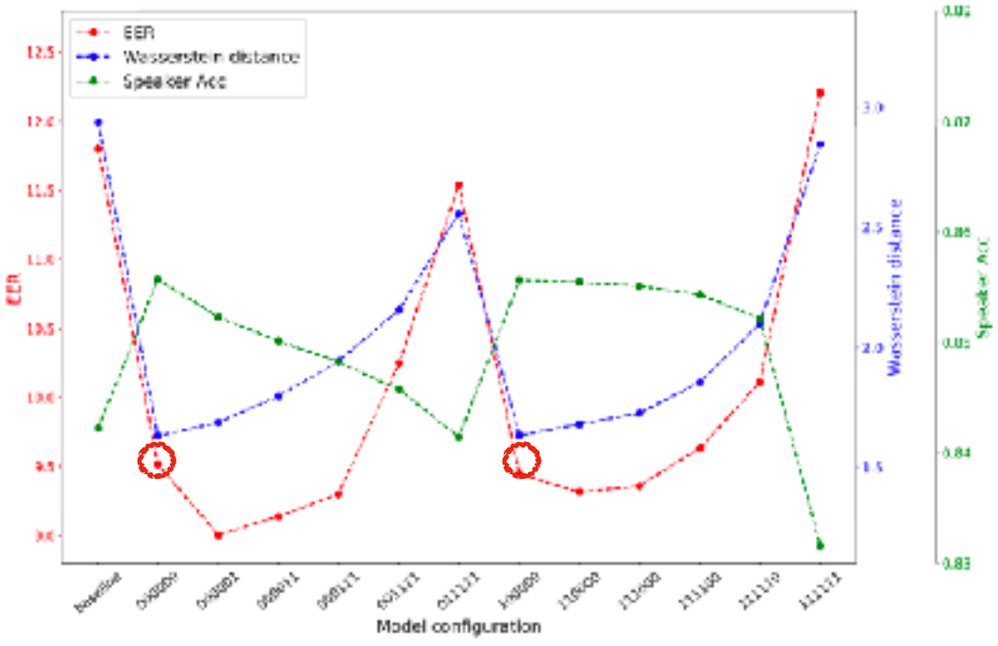


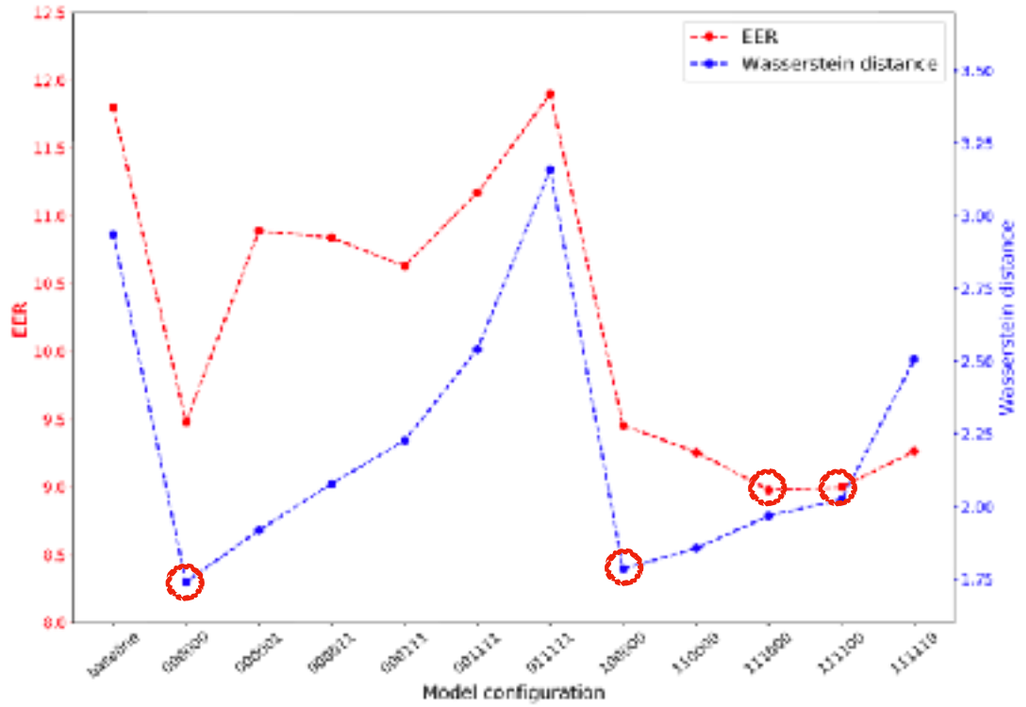
Figure 2: *The results of different weight sharing strategies with jointly training the source and target extractors, and EER (%) denotes the pooled results on SRE16. On the $x-$axis,* 1 *or* 0 *denotes whether or not to share the weights of the corresponding layer (from the lowest to the highest layer, low means close to the input layer), e.g.* 100000 *means only the parameters of the lowest layer is shared.*

Partially shared neural network is more helpful on both reducing distribution mismatch and enhancing speaker discrimination for embeddings.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network. InterSpeech 2020

Partially shared network for unsupervised adversarial adaptation

Fix source domain parameter and find optimal partially shared architecture.



- More unshared layers is helpful to reduce embedding distribution mismatch.
- It is better to share shallow layers for final target domain performance.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network. InterSpeech 2020

Partially shared network for unsupervised adversarial adaptation

Performance comparison between fully shared and partially shared architecture.

Table 2: *Results comparison using different weight sharing strategies.*

System	EER(%)		
	Pooled	Cantonese	Tagalog
baseline	11.81	8.36	15.38
FSN	12.21	7.30	17.20
PSN	**8.98**	**5.18**	**12.90**

Table 3: *Results with or without weight regularization. The model configuration corresponds to 111000 in Fig. 3.*

λ_w	λ_r	EER(%)		
		Pooled	Cantonese	Tagalog
1.0	0.0	26.72	26.60	26.84
	0.01	9.35	5.42	13.36
0.1	0.0	9.08	5.29	13.03
	0.01	**8.98**	**5.18**	**12.90**

- The normal FSN gets no stable improvement on this task with large domain mismatch (different languages), and the proposed PSN has a large improvement for fast domain adaptation.
- The regularization loss is useful in the proposed PSN architecture, and it can make the performance more stable.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network. InterSpeech 2020

Conclusion

- When using adversarial learning for domain adaptation, simply sharing the whole extractor can indeed reduce the embedding distribution mismatch between the source and target domain, but may hurt the speaker discrimination ability.
- Partially shared network naturally has the ability to compensate the domain mismatch without influencing the speaker discrimination ability, and it is better than the fully shared network for domain adaptation.
- It's better to share the lower-level layers (generic acoustic feature learning) and leave the higher-level layers unshared (high-level information learning, language) for domain adaptation.

Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Adversarial Domain Adaptation for Speaker Verification using Partially Shared Network. InterSpeech 2020

Related publications and references

[1] Shuai Wang, Johan Rohdin, Lukáš Burget, Oldřich Plchot, Yanmin Qian, Kai Yu and Jan Černocký. On the Usage of Phonetic Information for Text-independent Speaker Embedding Extraction. In 20th Annual Conference of the International Speech Communication Association (InterSpeech), Graz, Austria, 2019, 1148-1152.

[2] Zhengyang Chen, Shuai Wang, Yanmin Qian and Kai Yu. Channel Invariant Speaker Embedding Learning with Joint Multi-Task and Adversarial Training. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, 6574-6578.

[3] Zhengyang Chen, Shuai Wang and Yanmin Qian. Adversarial Domain Adaptation for Speaker Verification Using Partially Shared Network. In 21st Annual Conference of the International Speech Communication Association (InterSpeech), Shanghai, China, 2020, 3017-3021.

[4] Tsuchiya, Taira, Naohiro Tawara, Testuji Ogawa, and Tetsunori Kobayashi. "Speaker invariant feature extraction for zero-resource languages with adversarial learning." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2381-2385. IEEE, 2018.

[5] Meng, Zhong, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gang, and Biing-Hwang Juang. "Speaker-invariant training via adversarial learning." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5969-5973. IEEE, 2018.

[6] Wang, Hongji, Heinrich Dinkel, Shuai Wang, Yanmin Qian, and Kai Yu. "Cross-Domain Replay Spoofing Attack Detection Using Domain Adversarial Training." In *Interspeech*, pp. 2938-2942. 2019.

[7] Meng, Zhong, Yong Zhao, Jinyu Li, and Yifan Gong. "Adversarial speaker verification." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6216-6220. IEEE, 2019.

The End

Thank You

Q & A