Domain Generalization for Small Self-supervised Speech Processing Models

Kuan-Po Huang, Yu-Kuan Fu, Tsu-Yuan Hsu
Prof. Hung-yi Lee
National Taiwan University

Outline

- Problems of speech models: 1. domain mismatch   2. huge model size
- Domain generalization
    - MLDG, MASF, Augmentation
- Knowledge distillation – DistilHuBERT
- Robustness of DistilHuBERT

Problems of Speech Models

- **Domain mismatch** problem → **Domain generalization**
  - Performance degrades when distortions are introduced.
  - Training data for downstream maybe clean, but distorted during testing.
  - Enhance generalizability of speech representations to avoid having to deal with generalizability whenever we switch to a new downstream task.
- **Models are too large** to deploy → **Knowledge Distillation**
  - HuBERT Base 1.1GB, Large 3.5GB, X-Large 10.8GB
  - However, compressed models usually have poor generalizability.

Our goal

To **reduce model size** while having **domain generalizability**.

HuBERT: https://arxiv.org/abs/2106.07447

Domain generalization

To have robustness on out-of-domain data without having knowledge of it during training.

<u>Domain generalization methods:</u>

- **Learn strategies**
  - Ensemble learning
  - Meta-learning: **MLDG, MASF**
- **Representation learning**
  - Domain-invariant representation learning
  - Feature disentanglement
- **Data manipulation**
  - **Augmentation**

https://www.ijcai.org/proceedings/2021/0628.pdf

MLDG & MASF

Task: Intent Classification   Model: HuBERT base

IC	upstream finetune	WHAM!	DNS	FSD50K
Deep all	X	92.33%	61.20%	91.22%
MLDG	X	91.86%	61.17%	90.88%
Deep all	final proj	92.78%	62.40%	90.75%
MLDG	final proj	92.30%	61.93%	90.35%
MASF	final proj	93.22%	61.90%	90.17%
Deep all	last trans layer + final proj	99.10%	81.36%	97.70%
MLDG	last trans layer + final proj	98.95%	79.20%	97.50%
MASF	last trans layer + final proj	98.44%	78.54%	97.81%
Deep all	all	98.86%	88.08%	98.31%

MLDG & MASF

Task: Intent Classification Model: HuBERT base

Limited performance improvement

IC	upstream finetune	WHAM!	DNS	FSD50K
Deep all	X	92.33%	61.20%	91.22%
MLDG	X	91.86%	61.17%	90.88%
Deep all	final proj	92.78%	62.40%	90.75%
MLDG	final proj	92.30%	61.93%	90.35%
MASF	final proj	93.22%	61.90%	90.17%
Deep all	last trans layer + final proj	99.10%	81.36%	97.70%
MLDG	last trans layer + final proj	98.95%	79.20%	97.50%
MASF	last trans layer + final proj	98.44%	78.54%	97.81%
Deep all	all	98.86%	88.08%	98.31%

MLDG & MASF

Task: Intent Classification   Model: HuBERT base

IC	upstream finetune	WHAM!	DNS	FSD50K
Deep all	X	92.33%	61.20%	91.22%
MLDG	X	91.86%	61.17%	90.88%
Deep all	final proj	92.78%	62.40%	90.75%
MLDG	final proj	92.30%	61.93%	90.35%
MASF	final proj	93.22%	61.90%	90.17%
Deep all	last trans layer + final proj	99.10%	81.36%	97.70%
MLDG	last trans layer + final proj	98.95%	79.20%	97.50%
MASF	last trans layer + final proj	98.44%	78.54%	97.81%
Deep all	all	98.86%	88.08%	98.31%

Large amount of memory requirement, cannot set whole upstream model trainable

Domain generalization

To have robustness on out-of-domain data without having knowledge of it during training.

Domain generalization methods:

- **Learn strategies**
  - Ensemble learning
  - Meta-learning: **MLDG, MASF**
- **Representation learning**
  - Domain-invariant representation learning
  - Feature disentanglement
- **Data manipulation**
  - **Augmentation**

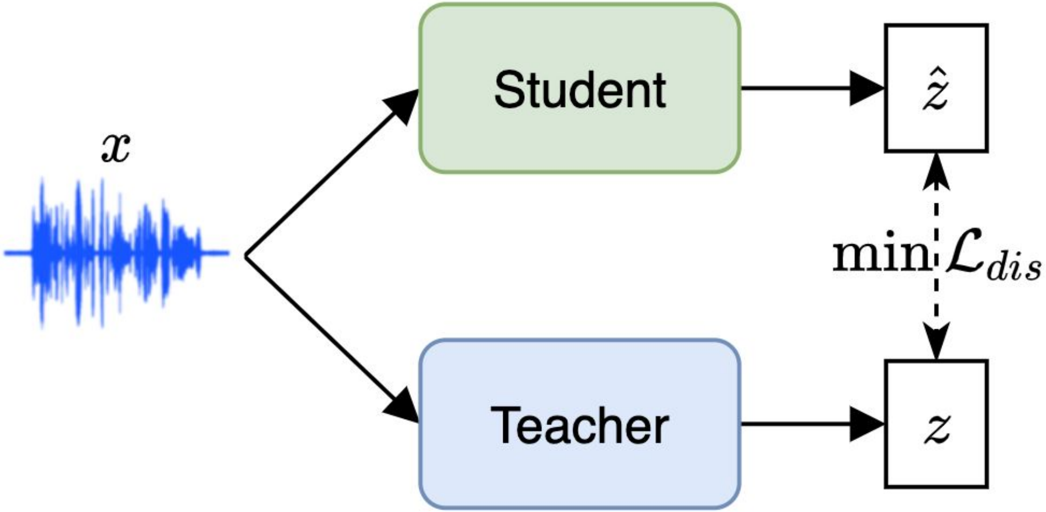
https://www.ijcai.org/proceedings/2021/0628.pdf

Augmentation

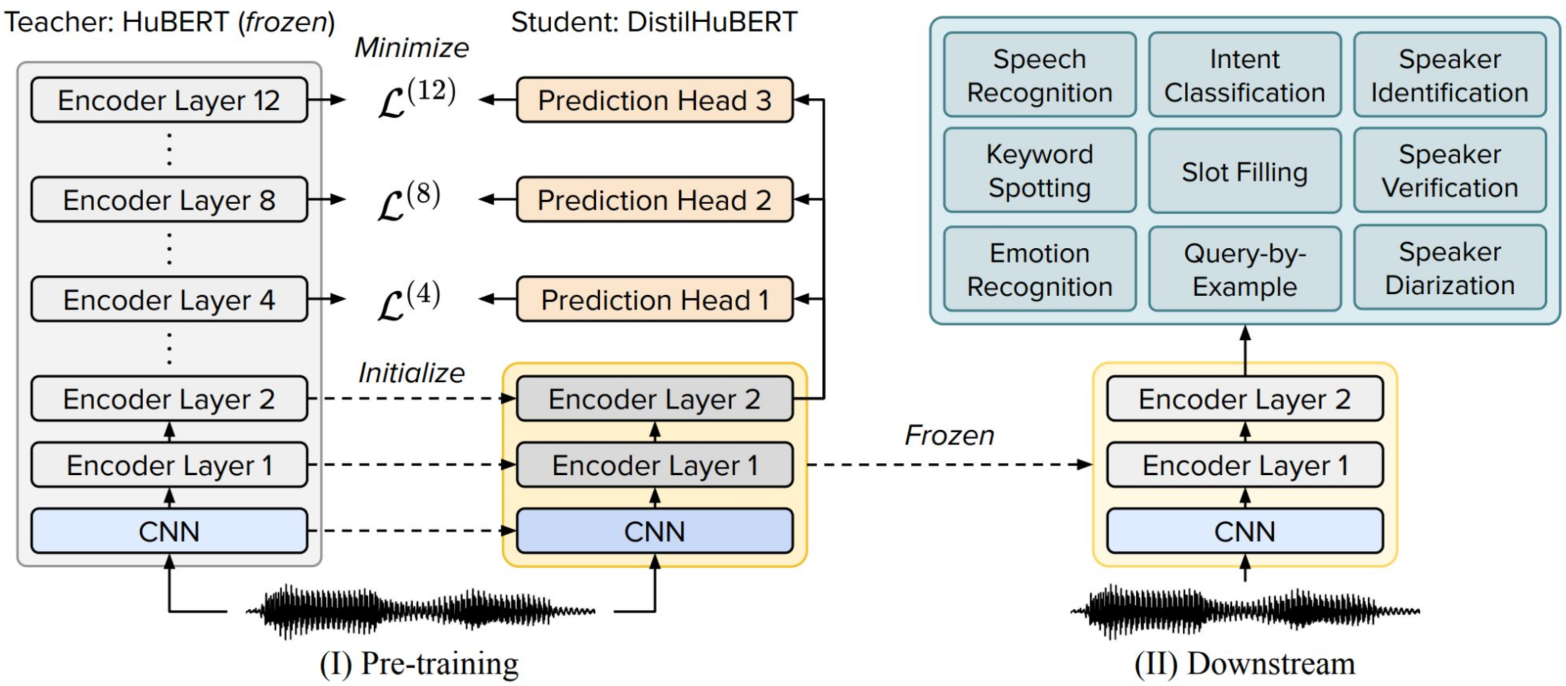
- Pre-defined augmentations (distortions)
  - additive noises: Gaussian noise, Musan noise, …
  - reverberation
  - time-frequency masks
  - speaking rate
  - …
- Trainable augmenter
  - Mixup of different distortions with trainable SNR weights.

Reducing model size – Knowledge distillation

Knowledge distillation: teacher-student learning



DistilHuBERT



$$\mathcal{L}^{i}_{Distil} = ||z_i - z'_i||_1 - \lambda \log \sigma\Big(cossim(z_i, z'_i)\Big)$$

Robustness of DistilHuBERT

Does student model have **robustness**?

Check by adding **distortions** to testing data.

	DistilHuBERT		HuBERT base	
	clean	**distorted**	**clean**	**distorted**
KS	0.9604	0.8984	0.9714	0.9338
IC	0.9478	0.6641	0.9947	0.9694
SID	0.7302	0.4042	0.8497	0.6551
ER	0.6387	0.5392	0.6396	0.5733
ASR	13.77	37.59	6.72	10.16

huge performance drop !

Proposal: Enhance Robustness of DistilHuBERT

Training HuBERT student with **knowledge distillation**.

Problem: Models are not robust to distortions.

- **Add distortions** to the input of the student model.

Problem: Teacher models may not have robustness to distortions.

- **Continually train** the teacher model.

Problem: Teacher and student representations are not alike.

Representations are not domain-invariant.

- Perform **adversarial training** when distilling models.

Experiment settings

Dataset for distillation: LibriSpeech 960 hr

Distortions: Musan noise, Gaussian noise, Reverberation (Maybe more in the future)

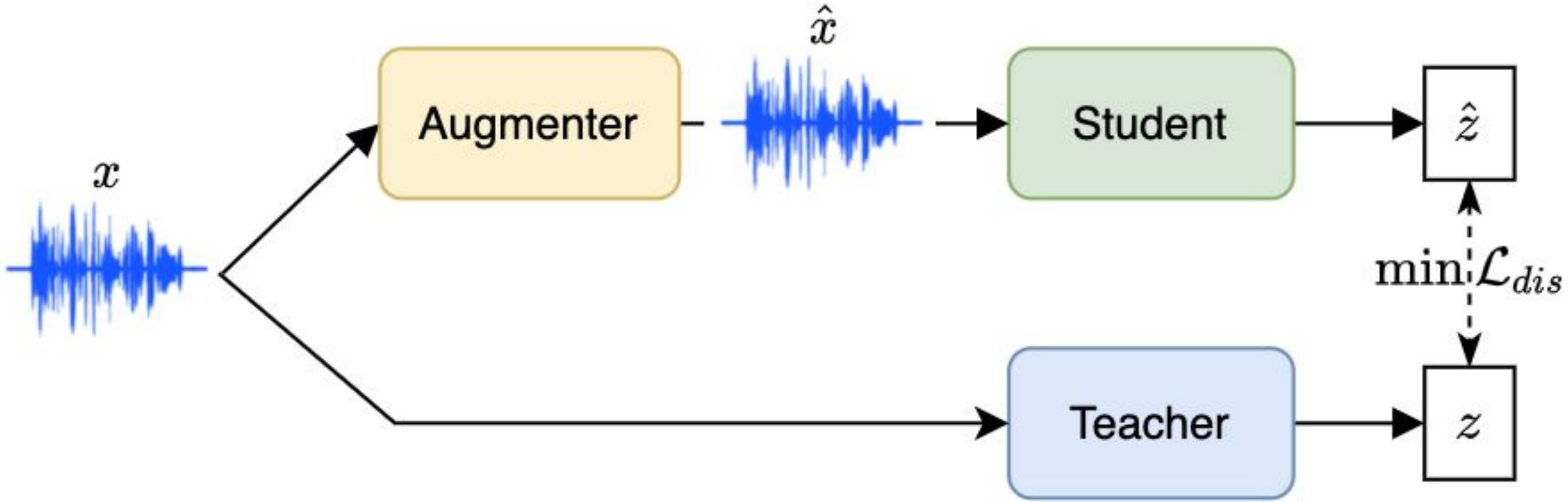
Teacher: pre-trained HuBERT base (or continually trained with distorted data)

Student: fewer transformer layers than the teacher

Add distortions

Add distortions to the input of the student model during distilling.

Maps **distorted inputs** to **teacher's clean representations.**

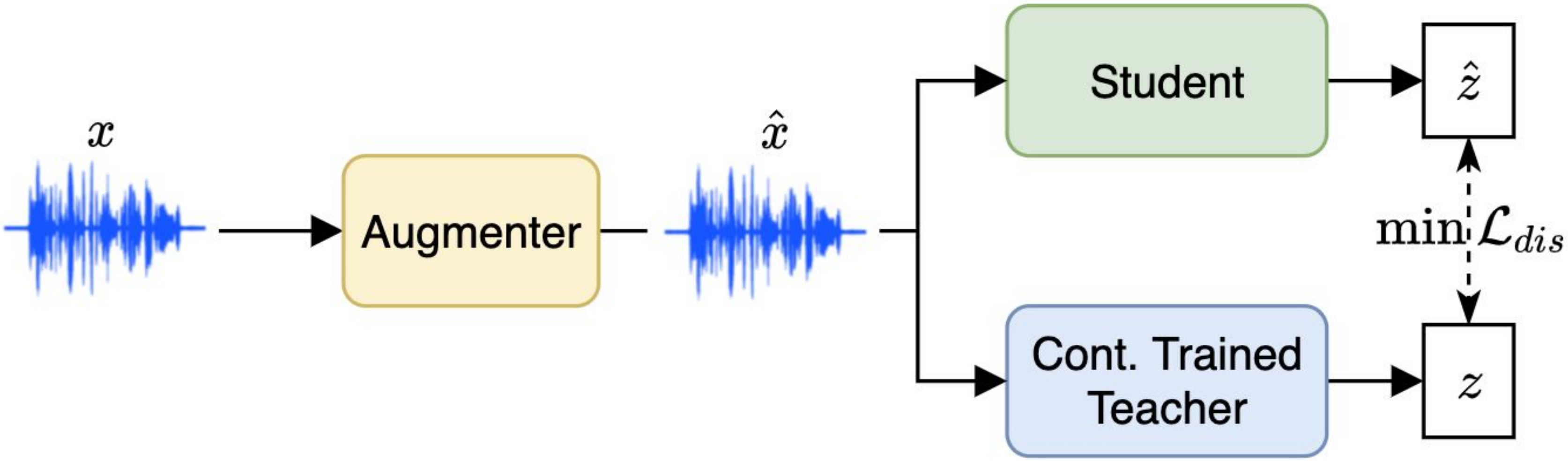


$$\mathcal{L}_{dis} = ||z - \hat{z}||_1 + \lambda \log \sigma(cossim(z, \hat{z}))$$

Continually trained teacher with distorted input

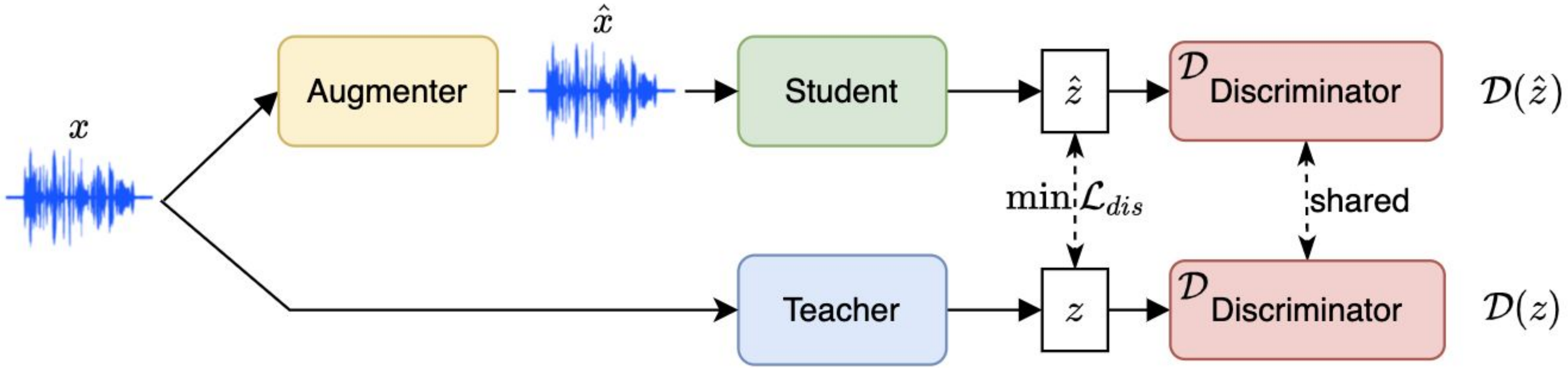
***Stage 1*: Continually train** the teacher model with distorted data.

***Stage 2*:** Knowledge distillation



Augmented student input with DAT (Binary domain setting)

DAT with Binary domain (**teacher / student**) setting

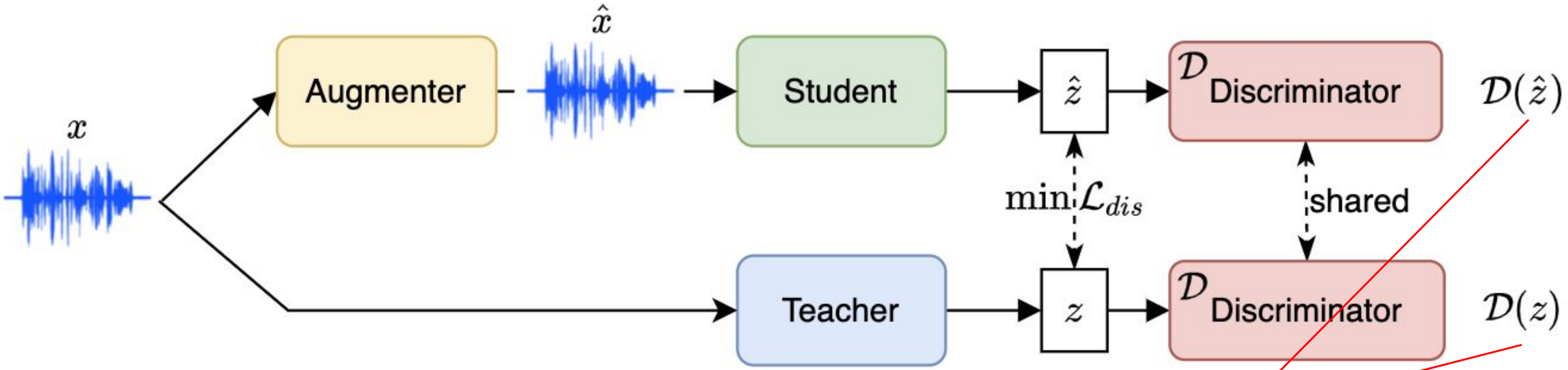


$$\mathcal{L}_{dis} = ||z - \hat{z}||_1 - \lambda \log \sigma(cossim(z, \hat{z}))$$

$$\sum_{i=0}^{B-1} \mathcal{L}_{dis}^i - \alpha \left[\sum_{i=0}^{\frac{B}{2}-1} \log D(z_i) + \sum_{i=\frac{B}{2}}^{B-1} \log(1 - D(\hat{z}_i)) \right]$$

Augmented student input with DAT (Binary domain setting)

DAT with Binary domain (**teacher / student**) setting

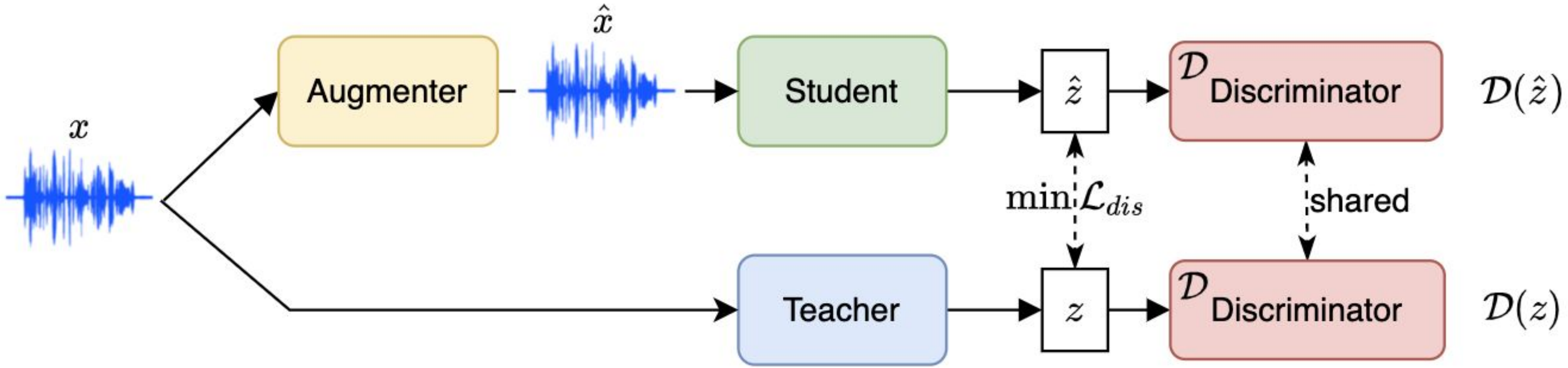


$$\mathcal{L}_{dis} = ||z - \hat{z}||_1 - \lambda \log \sigma(cossim(z, \hat{z}))$$

$$\sum_{i=0}^{B-1} \mathcal{L}_{dis}^{i} - \alpha \boxed{\left[\sum_{i=0}^{\frac{B}{2}-1} \log D(z_i) + \sum_{i=\frac{B}{2}}^{B-1} \log(1 - D(\hat{z}_i)) \right]}$$

Augmented student input with DAT (Binary domain setting)

DAT with Binary domain (**teacher / student**) setting



$$\mathcal{L}_{dis} = ||z - \hat{z}||_1 - \lambda \log \sigma(cossim(z, \hat{z}))$$

$$\min \boxed{\sum_{i=0}^{B-1} \mathcal{L}_{dis}^{i} - \alpha \overbrace{\left[\sum_{i=0}^{\frac{B}{2}-1} \log D(z_i) + \sum_{i=\frac{B}{2}}^{B-1} \log(1 - D(\hat{z}_i)) \right]}^{\max}}$$

Augmented student input with DAT – training method

Train the student model and the discriminator in turn in a loop as follows:

Step 1: Set the discriminator trainable

Step 2: Train the discriminator with the teacher's and student's output representations to classify the outputs of the teacher and student.

$$\left[\sum_{i=0}^{\frac{B}{2}-1} \log D(z_i) + \sum_{i=\frac{B}{2}}^{B-1} \log(1 - D(\hat{z}_i))\right]$$

Step 3: Set the discriminator non-trainable

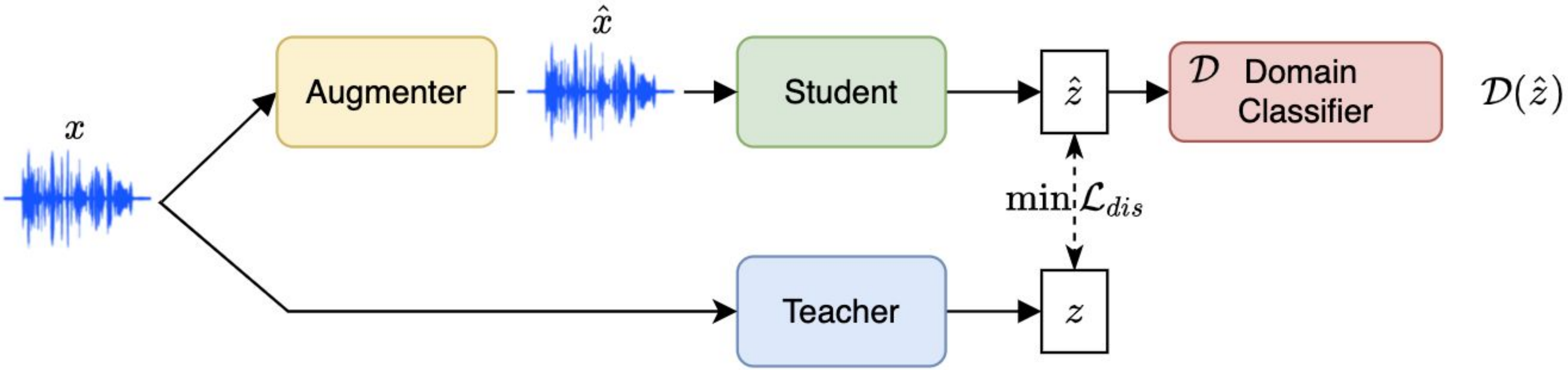
Step 4: Train the student model with the distillation loss and the domain adversarial loss.

$$\mathcal{L}_{dis} = ||z - \hat{z}||_1 - \lambda \log \sigma(cossim(z, \hat{z}))$$

$$\min \sum_{i=0}^{B-1} \mathcal{L}_{dis}^i - \alpha \left[\sum_{i=0}^{\frac{B}{2}-1} \log D(z_i) + \sum_{i=\frac{B}{2}}^{B-1} \log(1 - D(\hat{z}_i))\right]$$

Augmented student input with DAT (Multi-domain setting)

DAT with Multi-domain (**augmentations**) setting

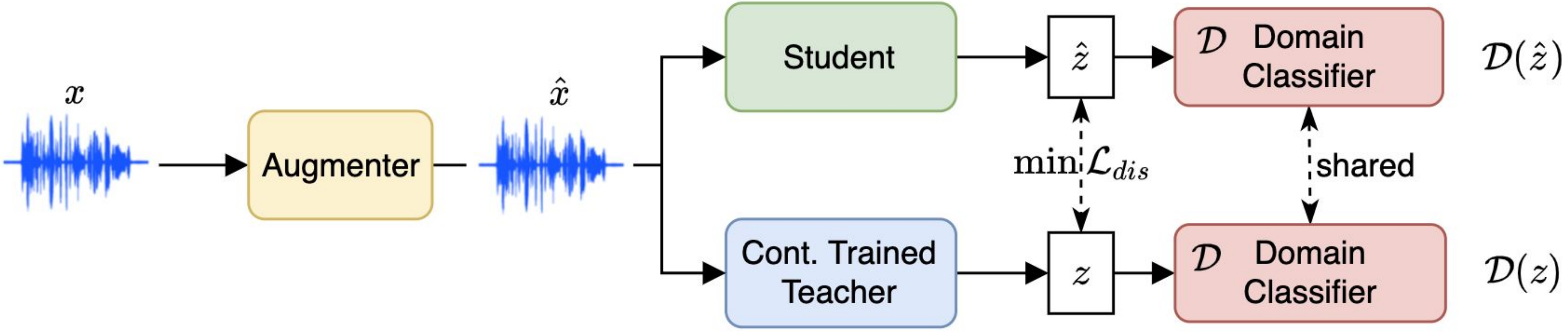


$$\mathcal{L}_{dis} = ||z - \hat{z}||_1 + \lambda \log \sigma(cossim(z, \hat{z}))$$

$$\mathcal{L}_{dis} - \alpha \cdot \mathcal{L}_{CE}(\mathcal{D}(\hat{z}), d_{\hat{z}})$$

Augmented student input with DAT (Multi-domain setting)

DAT with Multi-domain (**augmentations**) setting

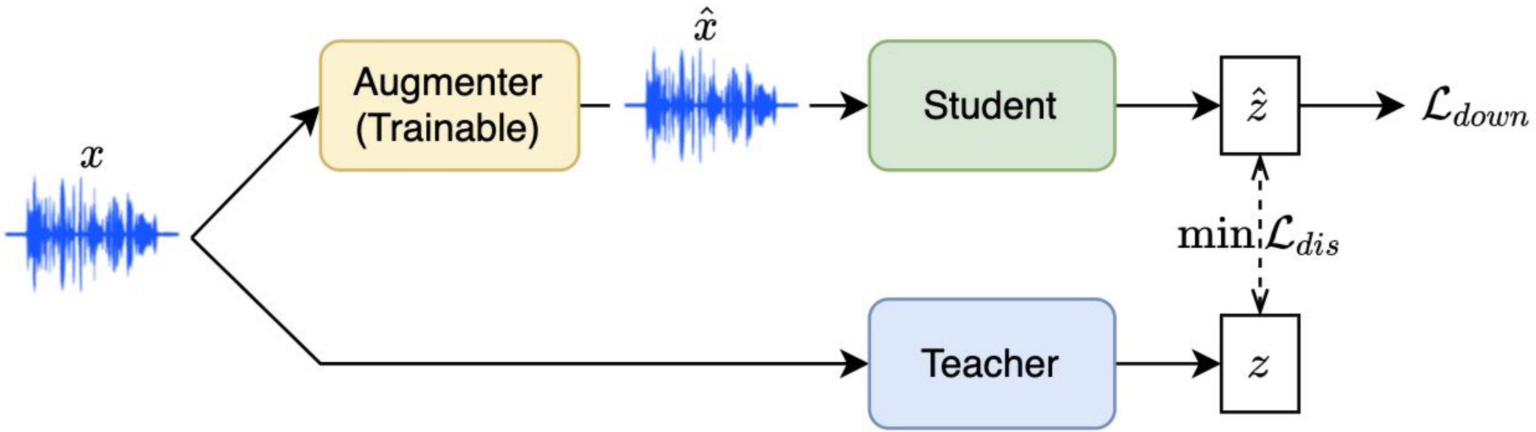


$$\mathcal{L}_{dis} = ||z - \hat{z}||_1 + \lambda \log \sigma(cossim(z, \hat{z}))$$

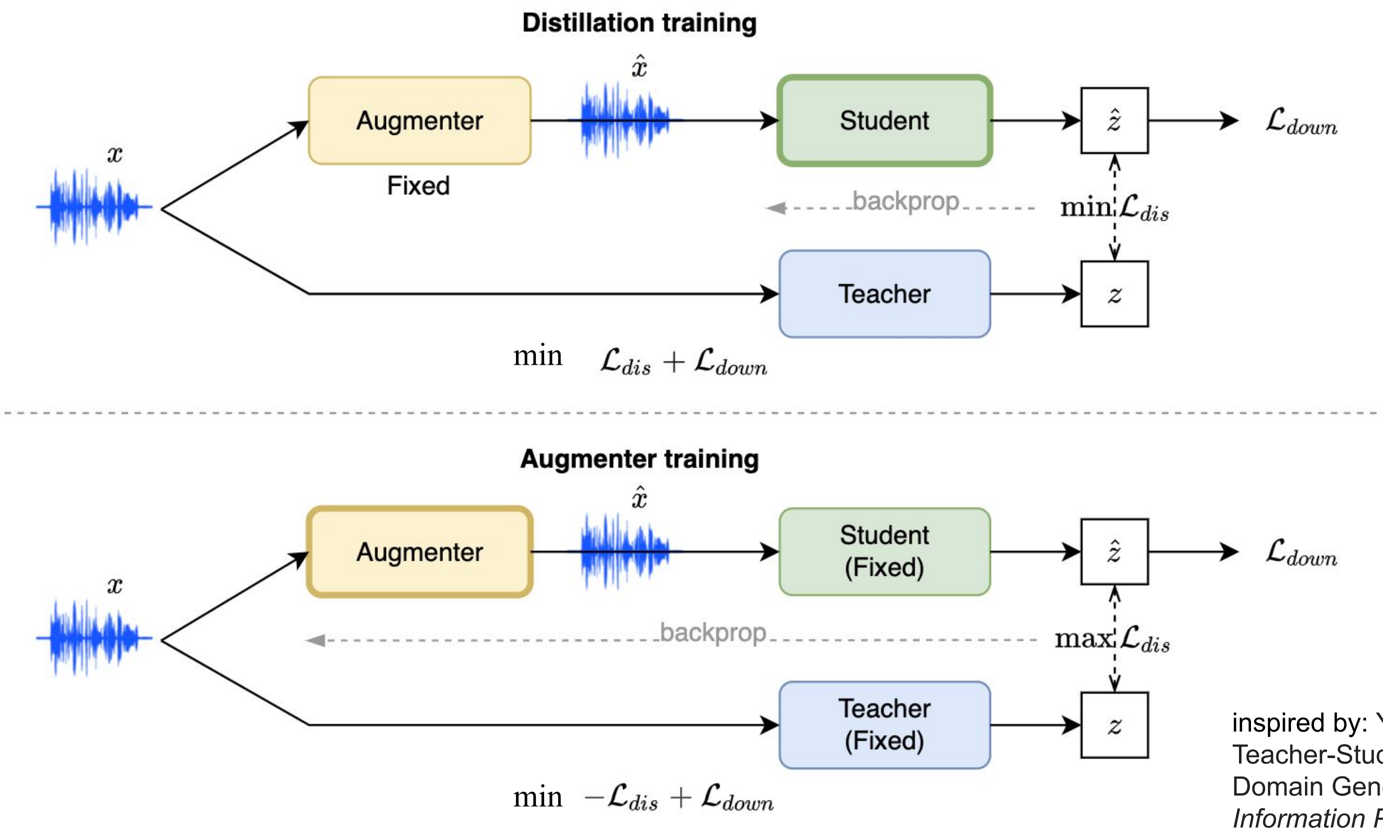
$$\mathcal{L}_{dis} + \alpha \cdot \mathcal{L}_{CE}(\{\mathcal{D}(z), \mathcal{D}(\hat{z})\}, \{d_z, d_{\hat{z}}\})$$

Trainable Augmenter

- Mixup of different distortions with trainable SNR weights.



Trainable Augmenter



inspired by: Yang, Fu-En, et al. "Adversarial Teacher-Student Representation Learning for Domain Generalization." *Advances in Neural Information Processing Systems* 34 (2021).

Long-term goal

- Model agnostic
- Teacher and student models can be the same architecture or different.

Timeline

6/12 - 6/25 Finish experiments of robust DistilHuBERT.

6/26 - 7/16 Write paper and submit to SLT.

6/26 - 7/21 Train a robust HuBERT Large model for public usage.

7/17 - 8/5 Experiments for long-term goals. Preparation for closing presentation

Thanks for listening.