IEEE ICASSP 2022

2022 IEEE International Conference on Acoustics, Speech and Signal Processing

AUD-27.6

Peer Collaborative Learning for Polyphonic Sound Event Detection

Hayato Endo and **Hiromitsu Nishizaki**

endo@alps-lab.org, hnishi@yamanashi.ac.jp

Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences,
University of Yamanashi, *JAPAN*

UNIVERSITY
OF
YAMANASHI

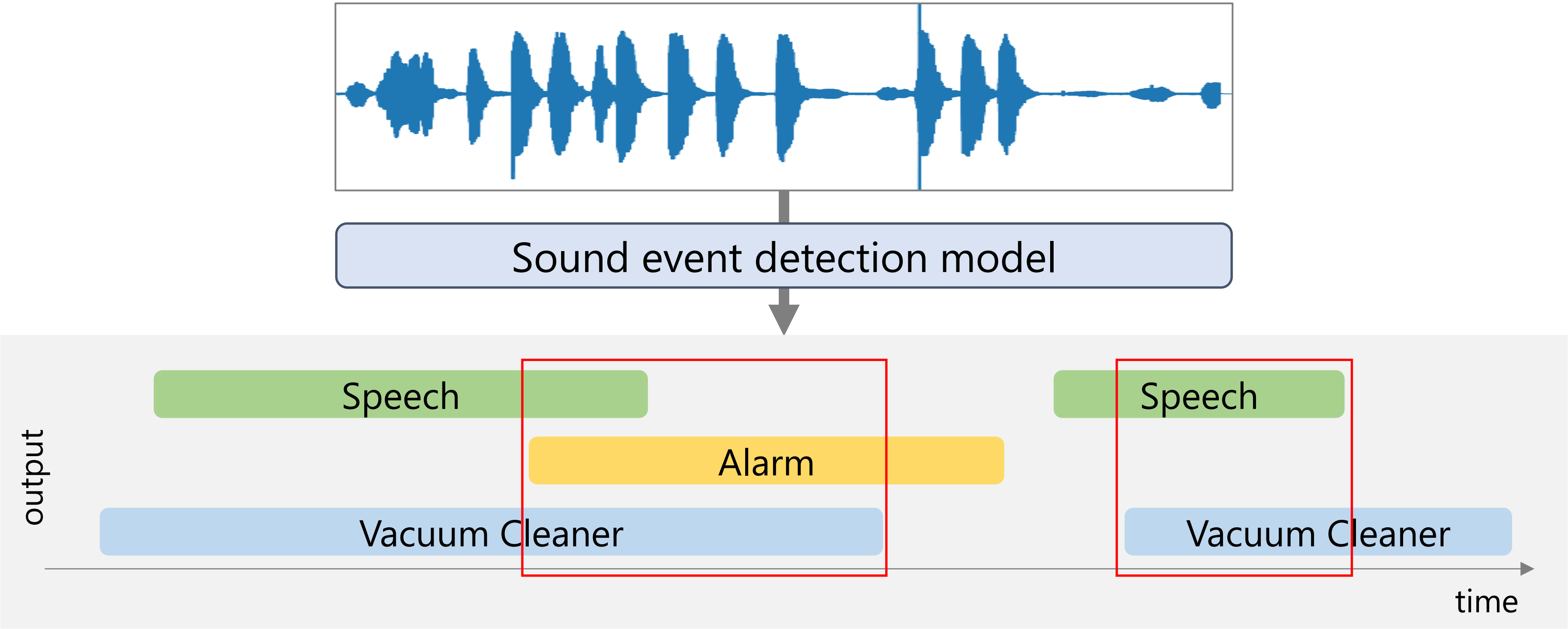
Regional Core
&
Global Professionals

Polyphonic Sound Event Detection Task

- DCASE2019・2020 Task 4 [1, 2] ※

※ DCASE : Detection and Classification of Acoustic Scenes and Events

  - Task Definition: detection of multiple sound event intervals in acoustic data for domestic environments



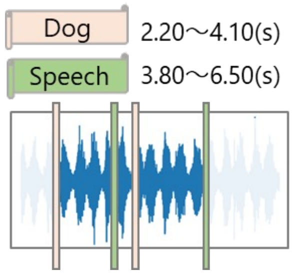
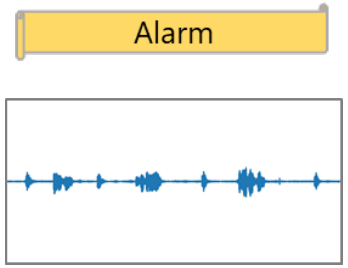
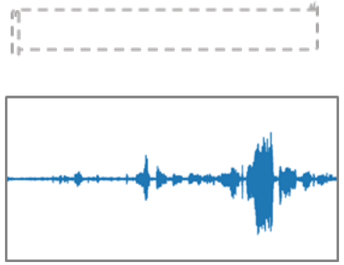
Goal

Improvement of detection accuracy of sound event intervals in practical environment situations

[1] N. Turpault, et al., "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," Proc. of DCASE2019, pp.253–257, 2019.
[2] N. Turpault, et al., "Training sound event detection on a heterogeneous dataset," Proc. of DCASE2020, pp. 200–204, 2020.

Label Information on the Task

- Three sorts of label types are included in the dataset

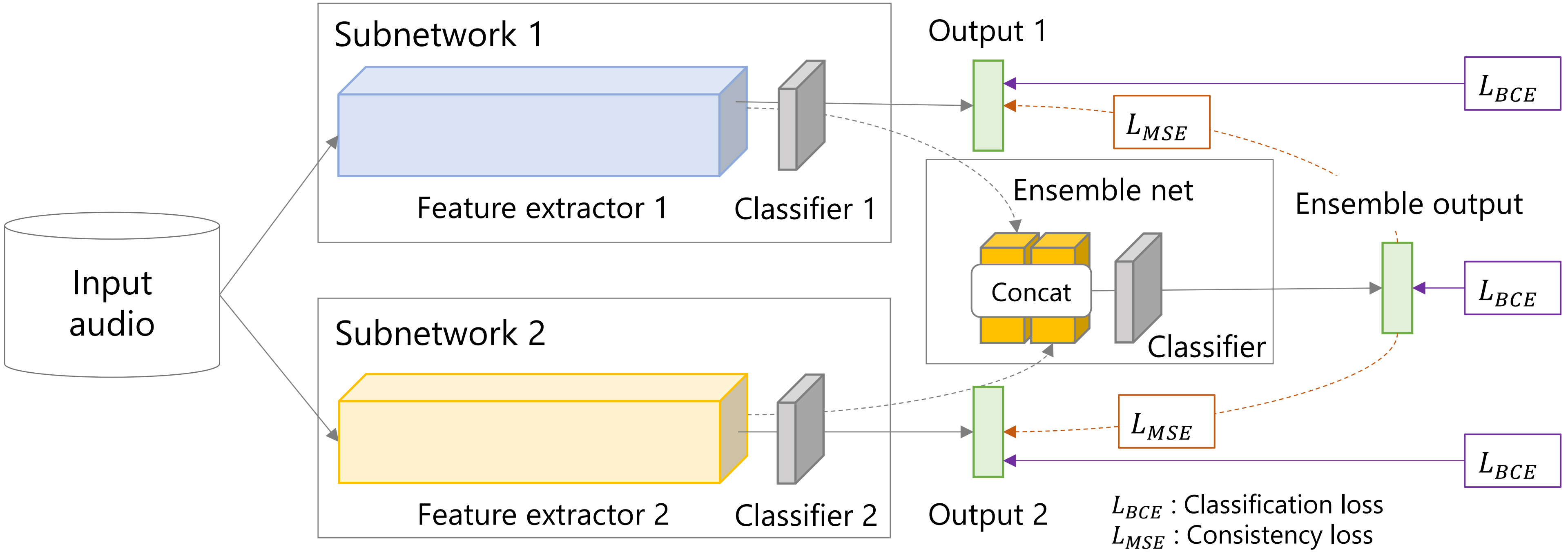
		Hard label	Soft label	Unlabeled
Label image				
Label	class	○	○	×
	interval	○	×	×
Amount of data		small	small	large
Difficulty of collection		high	middle	low

Because collecting hard-labeled data is very costly, soft-labeled or unlabeled data should be utilized

This study proposes a model structure that can utilize soft-labeled and unlabeled data

Related Work (1/2)

- Online Knowledge Distillation [3]
  - Considering the output of the ensemble net as a reference, each subnetwork extracts powerful features for classification



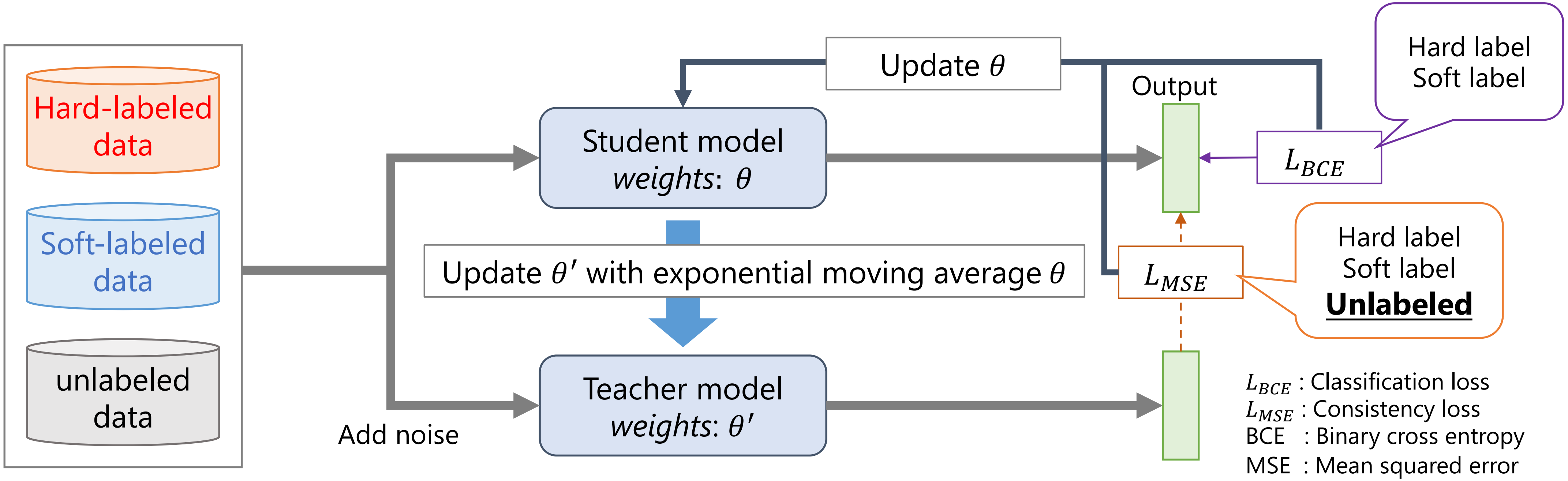
L_{BCE} : Classification loss
L_{MSE} : Consistency loss

Improved performance of each subnetwork ➡ Improved overall performance

[3] J. Kim, M. Hyun, I. Chung, N. Kwak, "Feature Fusion for Online Mutual Knowledge Distillation," Proc. of the 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4619-4625, 2020.

Related Work (2/2)

- Mean-Teacher model (the baseline model of the DCASE 2019・2020 Task4)
  - Student model (For training and evaluation) : Use the recent weights for classification
  - Teacher model (For training only) : Use the past to recent weights of the student model



L_{BCE} : Classification loss
L_{MSE} : Consistency loss
BCE : Binary cross entropy
MSE : Mean squared error

Guiding the student model training, effective use of unlabeled data

Summary of Our Research

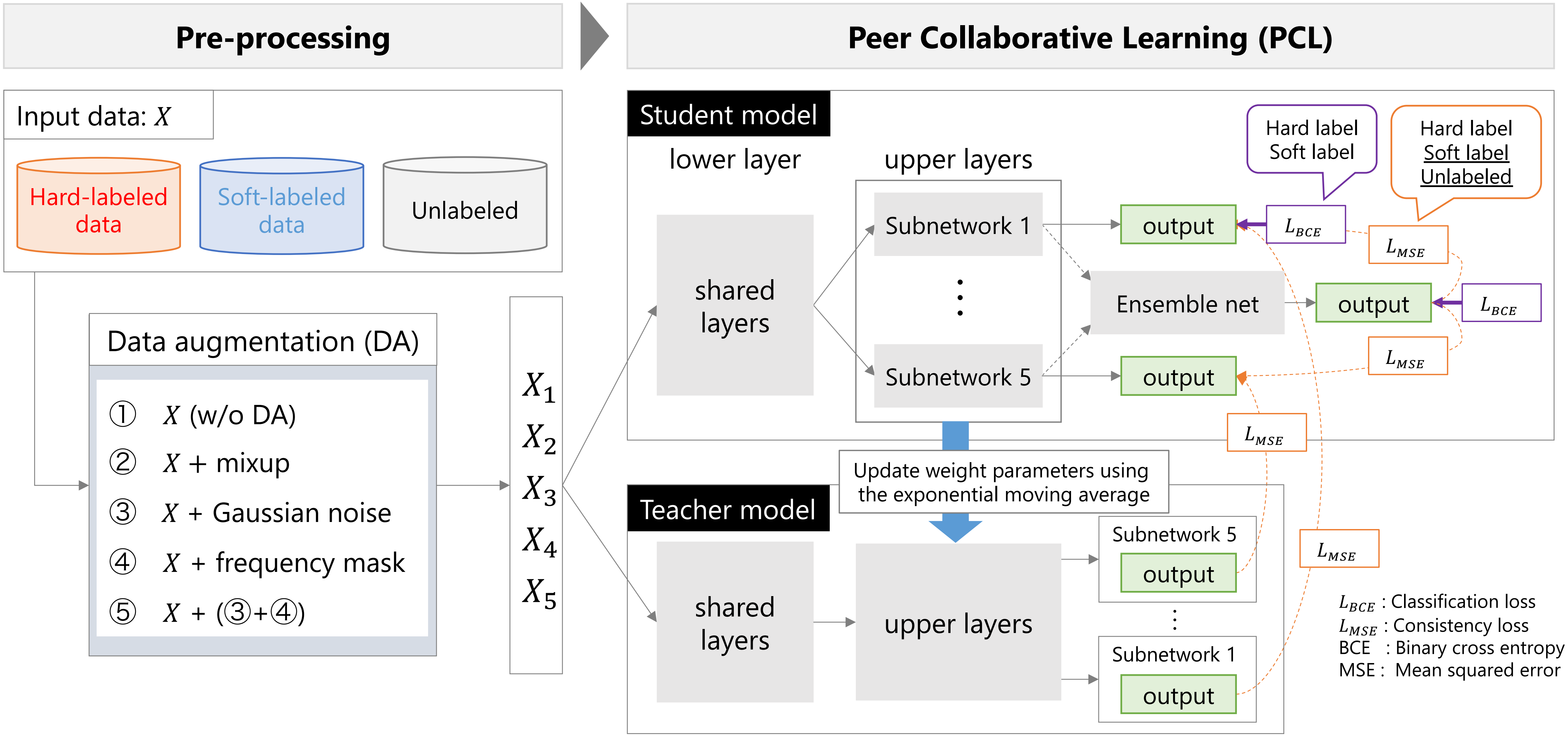
- **Goal** of our research
  - Improvement of accuracy of sound event detection on the DCASE Task 4
- **Proposed approach**
  - Use Peer Collaborative Learning (PCL) [4] , an integration and development of online knowledge distillation and mean-teacher approaches
  - Propose an effective combination of PCL and acoustic data augmentation

※ F1-score was used as evaluation measure

RESULT: Baseline (31.1%※) ▶ Proposed (44.2%※)

[4] Guile Wu and Shaogang Gong, "Peer collaborative learning for online knowledge distillation," Proc. of AAAI 2021, vol. 35, no. 12, pp. 10302–10310, May 2021.

【Proposed】PCL with Data Augmentation



Pre-processing

Peer Collaborative Learning (PCL)

Input data: X

Hard-labeled data

Soft-labeled data

Unlabeled

Data augmentation (DA)

① X (w/o DA)

② X + mixup

③ X + Gaussian noise

④ X + frequency mask

⑤ X + (③+④)

X_1 X_2 X_3 X_4 X_5

Student model

lower layer

upper layers

shared layers

Subnetwork 1

⋮

Subnetwork 5

output

Ensemble net

output

output

Hard label
Soft label

Hard label
Soft label
Unlabeled

L_{BCE}

L_{MSE}

L_{BCE}

L_{MSE}

L_{MSE}

L_{MSE}

Update weight parameters using the exponential moving average

Teacher model

shared layers

upper layers

Subnetwork 5

output

⋮

Subnetwork 1

output

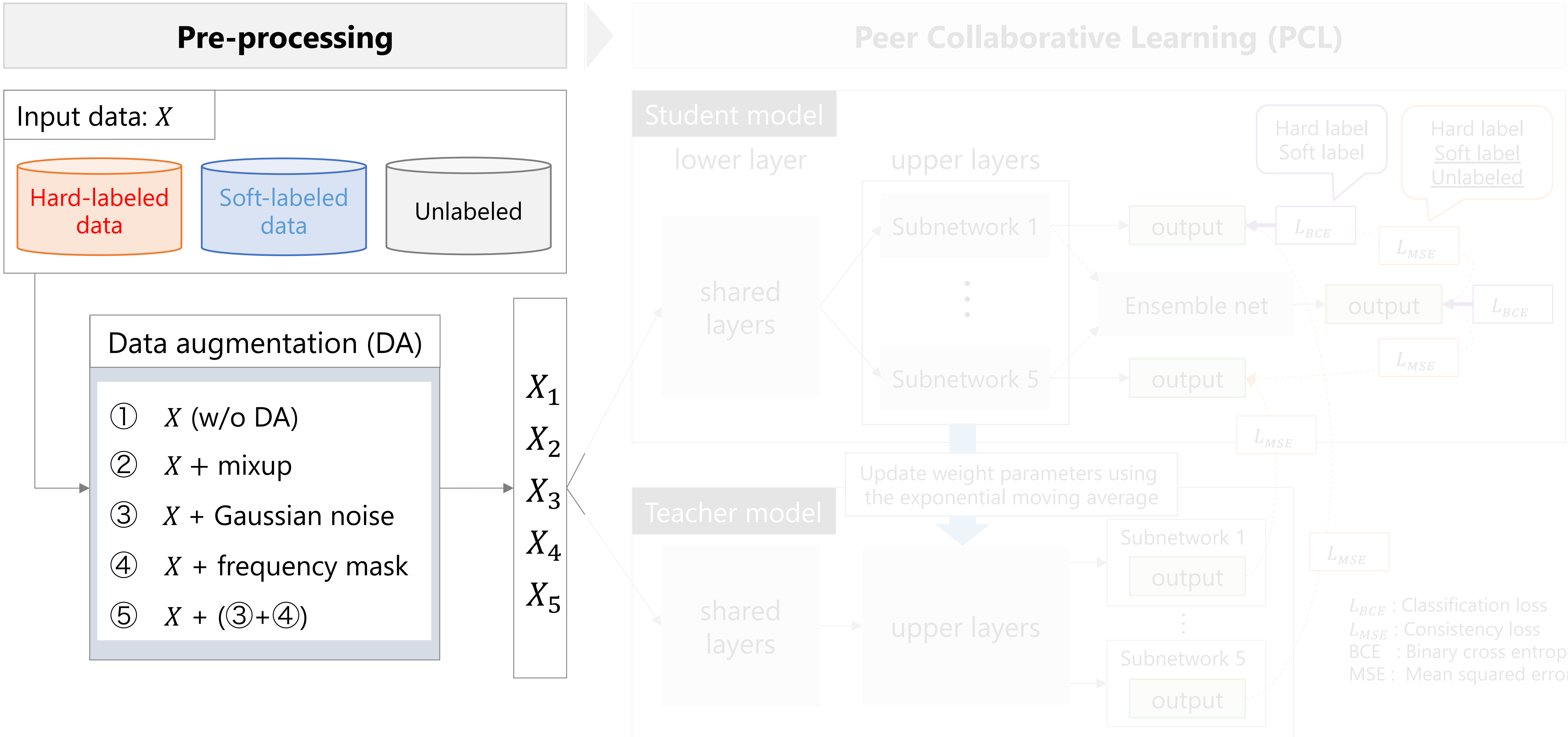
L_{BCE} : Classification loss

L_{MSE} : Consistency loss

BCE : Binary cross entropy

MSE : Mean squared error

Data Pre-Processing



Pre-processing

Peer Collaborative Learning (PCL)

Input data: X

Hard-labeled data

Soft-labeled data

Unlabeled

Data augmentation (DA)

① X (w/o DA)

② X + mixup

③ X + Gaussian noise

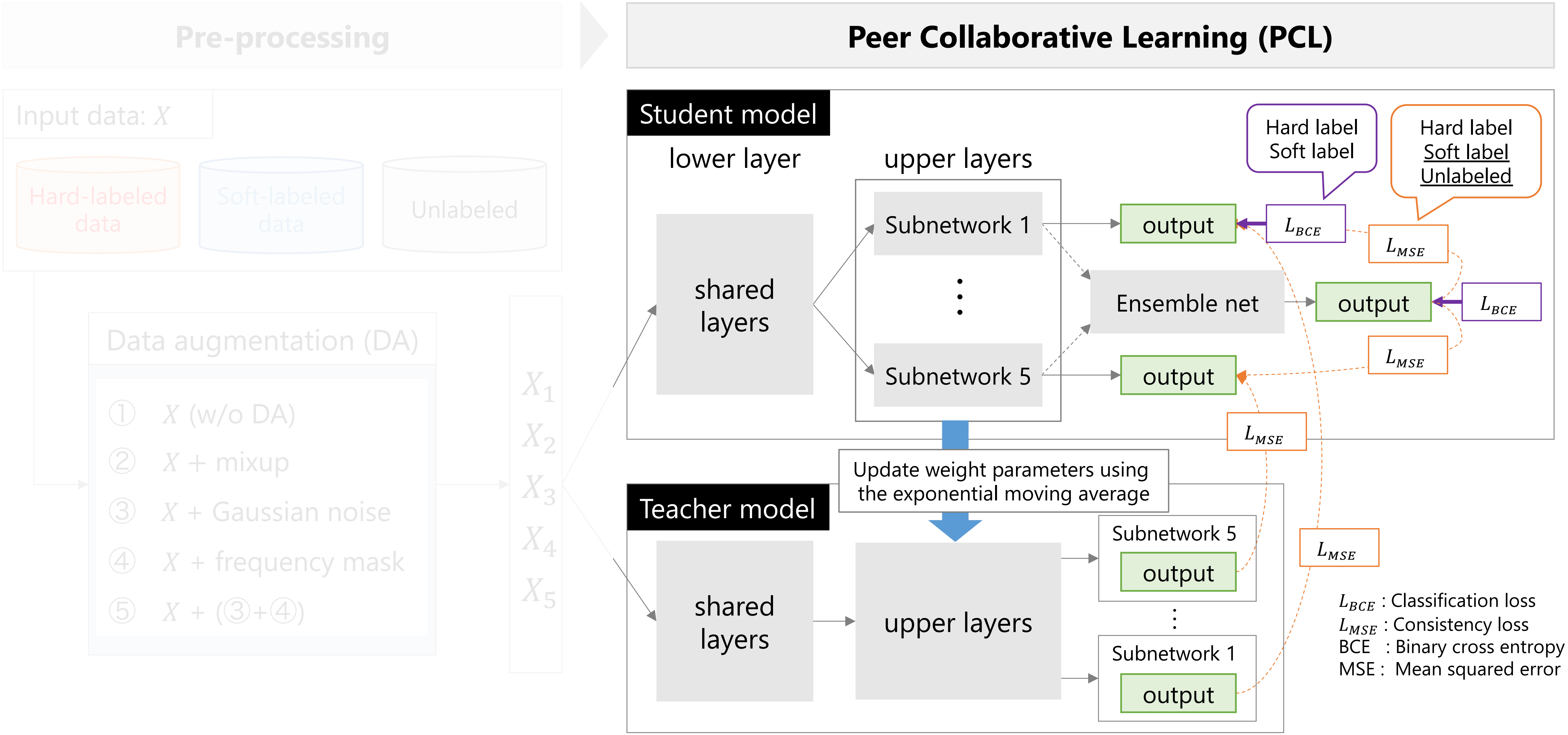
④ X + frequency mask

⑤ X + (③+④)

X_1 X_2 X_3 X_4 X_5

L_{BCE} : Classification loss
L_{MSE} : Consistency loss
BCE : Binary cross entropy
MSE : Mean squared error

Peer Collaborative Learning



Pre-processing

Input data: X

Hard-labeled data

Soft-labeled data

Unlabeled

Data augmentation (DA)

① X (w/o DA)

② X + mixup

③ X + Gaussian noise

④ X + frequency mask

⑤ X + (③+④)

X_1 X_2 X_3 X_4 X_5

Peer Collaborative Learning (PCL)

Student model

lower layer

upper layers

shared layers

Subnetwork 1

Subnetwork 5

Ensemble net

output

output

output

L_{BCE}

L_{BCE}

L_{MSE}

L_{MSE}

L_{MSE}

L_{MSE}

Hard label
Soft label

Hard label
Soft label
Unlabeled

Update weight parameters using the exponential moving average

Teacher model

shared layers

upper layers

Subnetwork 5

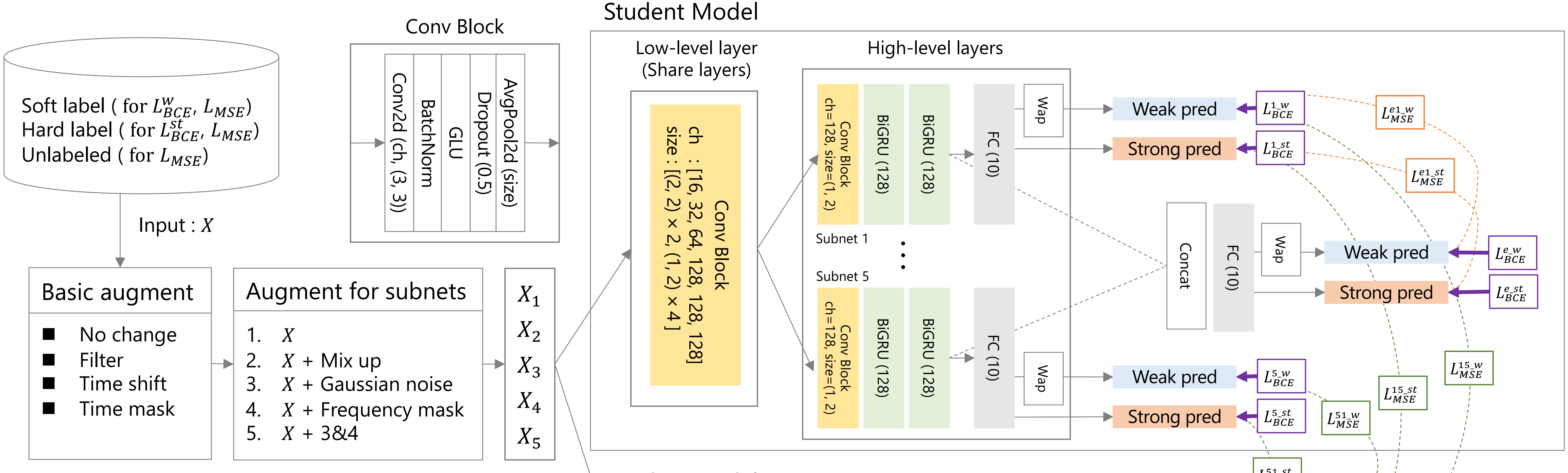
output

Subnetwork 1

output

L_{BCE} : Classification loss
L_{MSE} : Consistency loss
BCE : Binary cross entropy
MSE : Mean squared error

PCL Model Details



$$L = L^{n\_w}_{BCE} + L^{n\_st}_{BCE} + L^{e\_w}_{BCE} + L^{e\_st}_{BCE} + L^{nm\_w}_{MSE} + L^{nm\_st}_{MSE} + L^{en\_w}_{MSE} + L^{en\_st}_{MSE}$$

n : Subnet Number

m : Subnet number, excluding n

L^{n}_{BCE} : BCE for each subnet's output

L^{e}_{BCE} : BCE for ensemble output

L^{nm}_{MSE} : MSE of each subnet output of the student model and each subnet output of the teacher model except n

L^{en}_{MSE} : MSE with each subnet output and ensemble output

Wap : Weighted average pooling
BCE : Binary Cross Entropy
MSE : Mean Squared Error

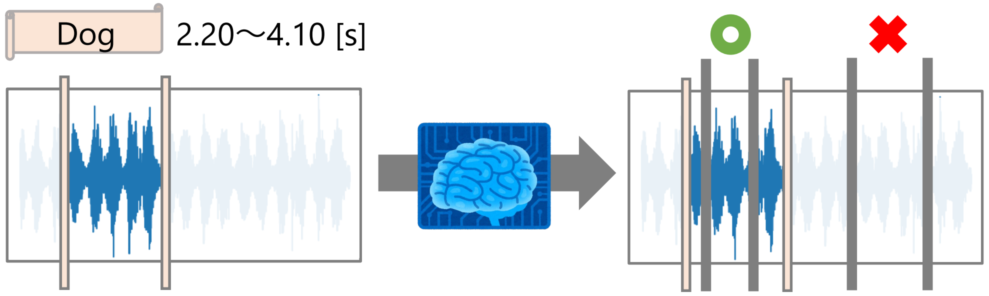
Experimental Setup

- Dataset
  - DCASE 2019 Task4 [1]
  - Sounds expected to occur in home environment (1 file = 10 seconds duration)

	Label type	# of data [/file]	Remarks
Training	Hard label	2,045	Known event intervals
	Soft label	1,578	Unknown event intervals
	Unlabeled	14,412	
Validation	Hard label	1,168	Known event intervals
Evaluation		692	

Num. of event classes: 10	
Alarm/bell/ringing	Electric shaver/ toothbrush
Blender	Frying
Cat	Running water
Dishes	Speech
Dog	Vacuum cleaner

- Evaluation measure
  - F1-score [%] based on the interval of sound event occurrence
    - ➢ The student model is used for evaluation



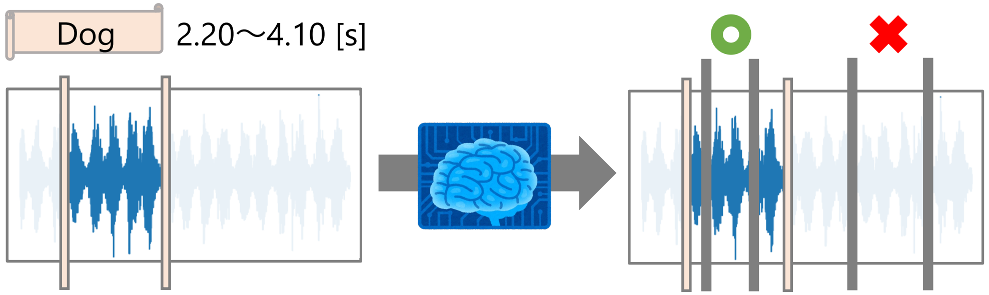
Experimental Setup

- Dataset
  - DCASE 2019 Task4 [1]
  - Sounds expected to occur in home environment (1 file = 10 seconds duration)

	Label type	# of data [/file]
Training	Hard	2,045
	Soft	1,578
	Unlabeled	14,412
Validation	Hard	1,168
Evaluation		692

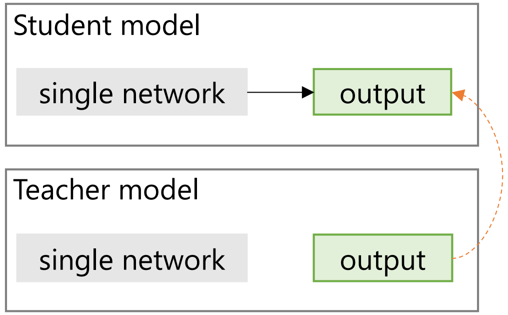
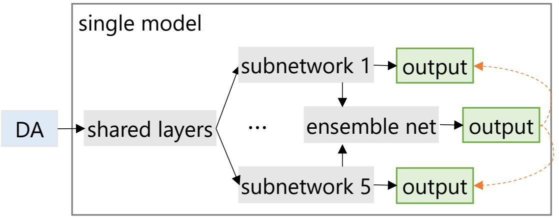
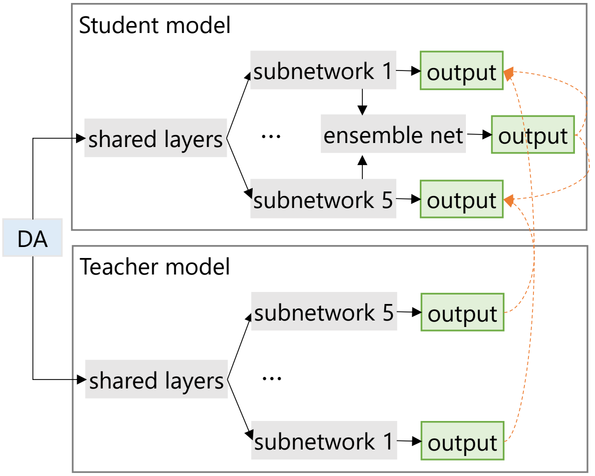
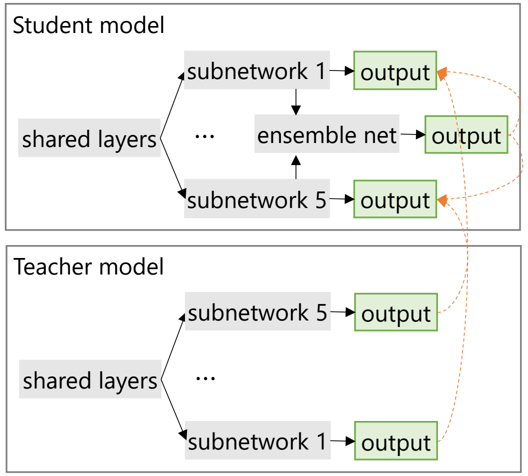
Num. of event classes: 10	
Alarm/bell/ringing	Electric shaver/ toothbrush
Blender	Frying
Cat	Running water
Dishes	Speech
Dog	Vacuum cleaner

- Evaluation measure
  - F1-score [%] based on the interval of sound event occurrence
    - ➢ The student model is used for evaluation



Four Competitive Approaches

Online KD w/ DA: Online knowledge distillation with data augmentation
PCL w/ DA: Peer collaborative learning with data augmentation
PCL w/o DA: Peer collaborative learning without data augmentation

	Baseline (mean-teacher)	Online KD[3] w/ DA	PCL w/ DA (proposed)	PCL w/o DA
Model image	 Student model single network output Teacher model single network output	 single model DA shared layers subnetwork 1 output … ensemble net output subnetwork 5 output	 Student model DA shared layers subnetwork 1 output … ensemble net output subnetwork 5 output Teacher model shared layers subnetwork 5 output … subnetwork 1 output	 Student model shared layers subnetwork 1 output … ensemble net output subnetwork 5 output Teacher model shared layers subnetwork 5 output … subnetwork 1 output

Evaluation Results (F1-score [%])

	Baseline	Online KD w/ DA	PCL w/ DA	PCL w/o DA
Validation	25.9	43.1	**43.8**	41.7
Evaluation	31.1	43.4	**44.2**	42.4

★ Experimental findings

1. PCL, Online KD > Baseline
2. PCL w/ DA, Online KD w/ DA > PCL w/o DA

- **Confirmation of the effectiveness of the PCL model, which evolved from the online knowledge distillation and mean-teacher methods**
- **It is valid to design sub-networks based on the data augmentation process**

Evaluation Results (F1-score [%])

	Baseline	Online KD w/ DA	PCL w/ DA	PCL w/o DA
Validation	25.9	43.1	**43.8**	41.7
Evaluation	31.1	43.4	**44.2**	42.4

★ Experimental findings

1. PCL
Online KD > Baseline
2. PCL w/ DA
Online KD w/ DA > PCL w/o DA

- **Confirmation of the effectiveness of the PCL model, which evolved from the online knowledge distillation and mean-teacher methods**
- **It is valid to design sub-networks based on the data augmentation process**

Conclusions

- Motivation (Goal)
  - Improvement of accuracy of polyphonic sound event detection on the DCASE Task4 task
- Proposed approach
  - **Peer collaborative learning** model, which evolved from the online knowledge distillation and mean-teacher methods with **audio data augmentation**
- Experimental results (F1-score)
  - Baseline (mean-teacher) 31.1% →⇨⇛ PCL with data augmentation **44.2%**
- Future work
  - We will implement and experiment with new knowledge distillation methods, such as collaborating with other knowledge distillation methods