Challenge No.6
The 2020 CORSMAL Challenge
Multi-modal fusion and learning for robotics

Audio-Visual Hybrid Approach for Filling Mass Estimation

Hyper Vision Research Laboratory

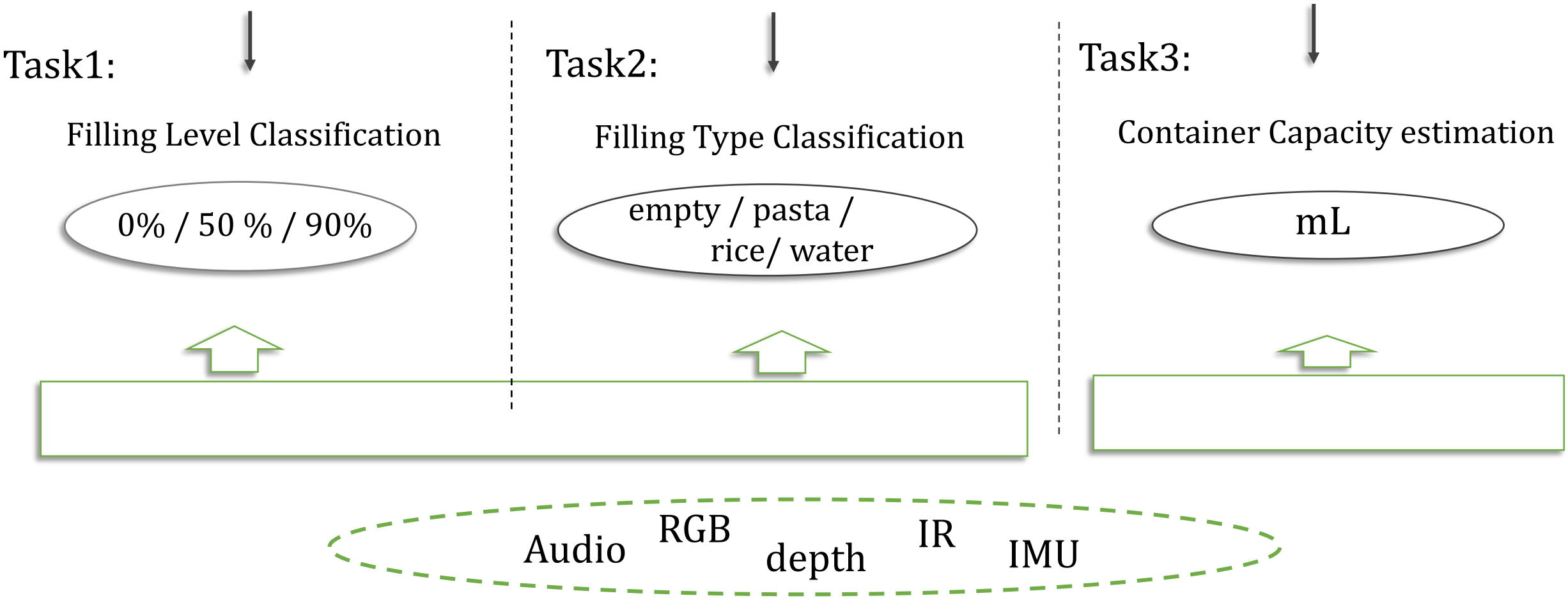
Reina Ishikawa*, Yuichi Nagao*, Ryo Hachiuma, Hideo Saito

{reina-ishikawa, soccerbass03, ryo-hachiuma, hs}@keio.jp

* indicates equal contribution

Multi modal fusion and learning for robotics

Estimating the capacity and mass of unseen containers



Solution for Task2 (Filling type classification)

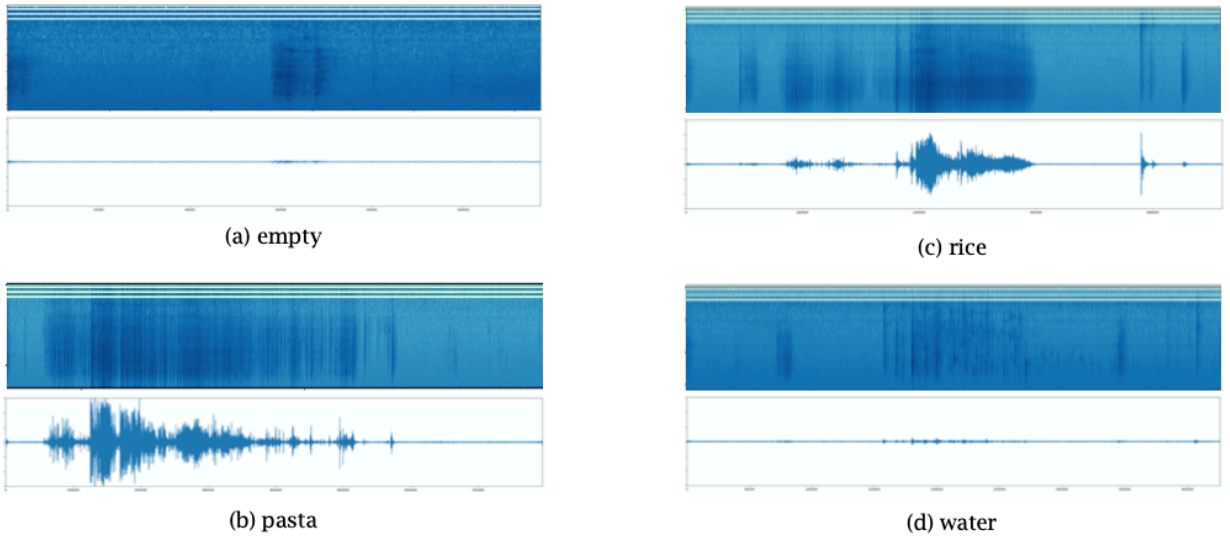
- We used audio data

Quite clear sound

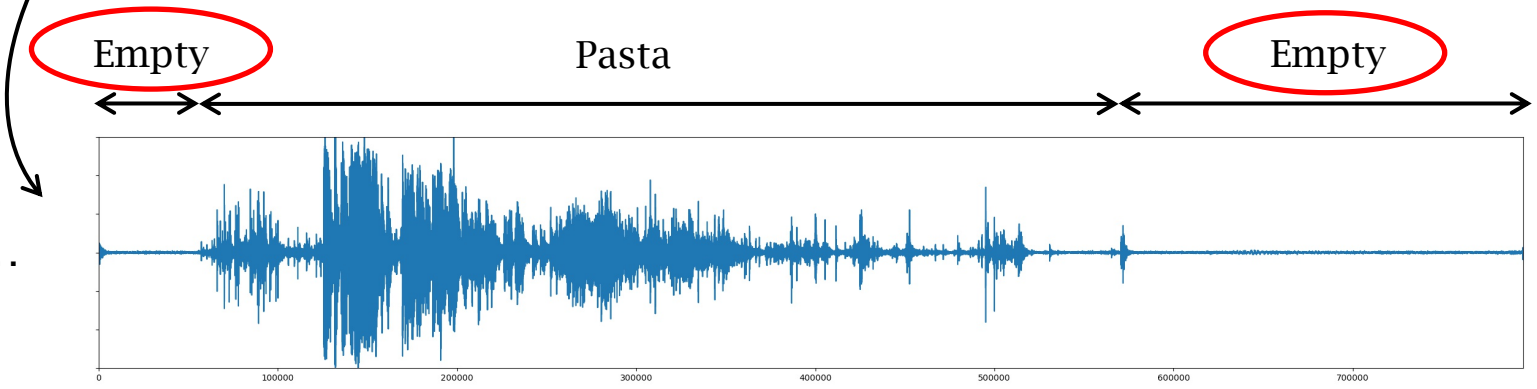
- Build CNN with spectrogram?

Input: The entire spectrogram??
→Not effective

- We made “additional annotations”…

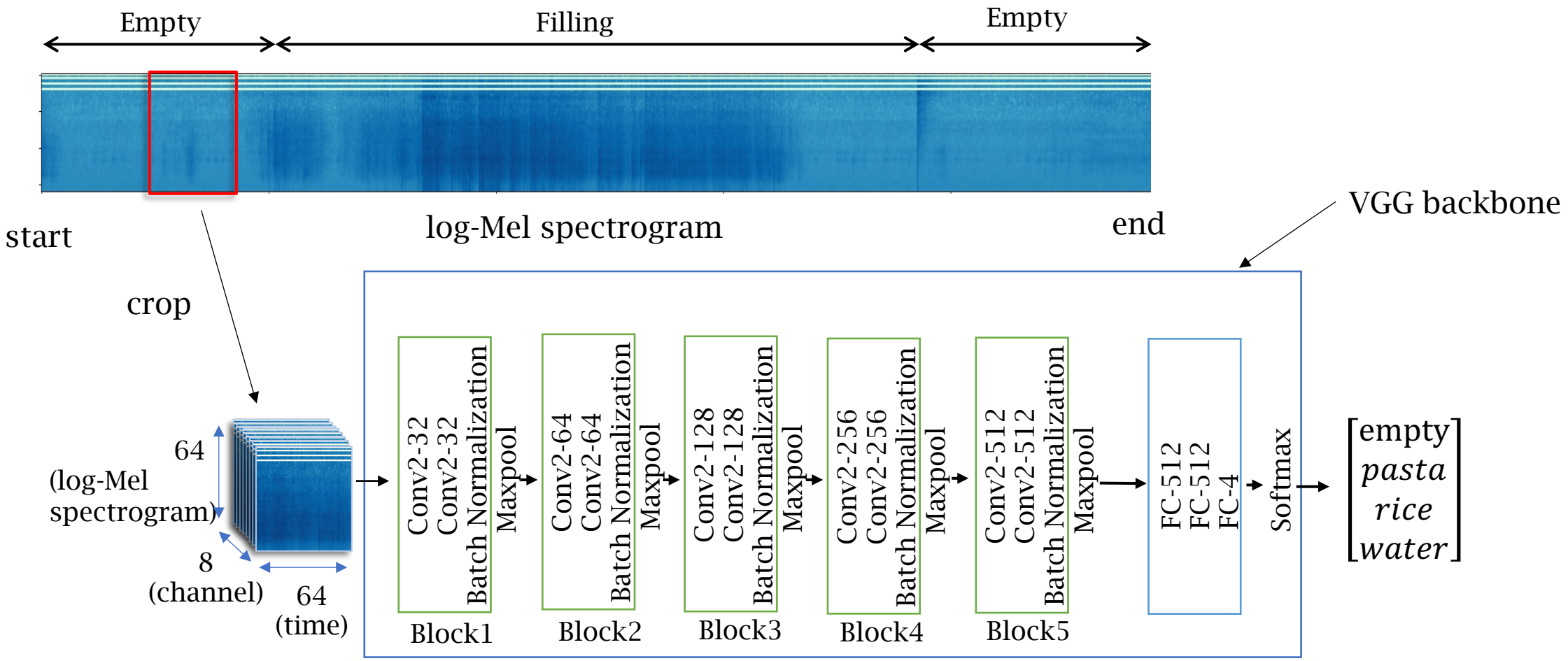


Comparison of raw wave and log-Mel spectrogram



Additional annotation

Solution for Task2 (Filling type classification)

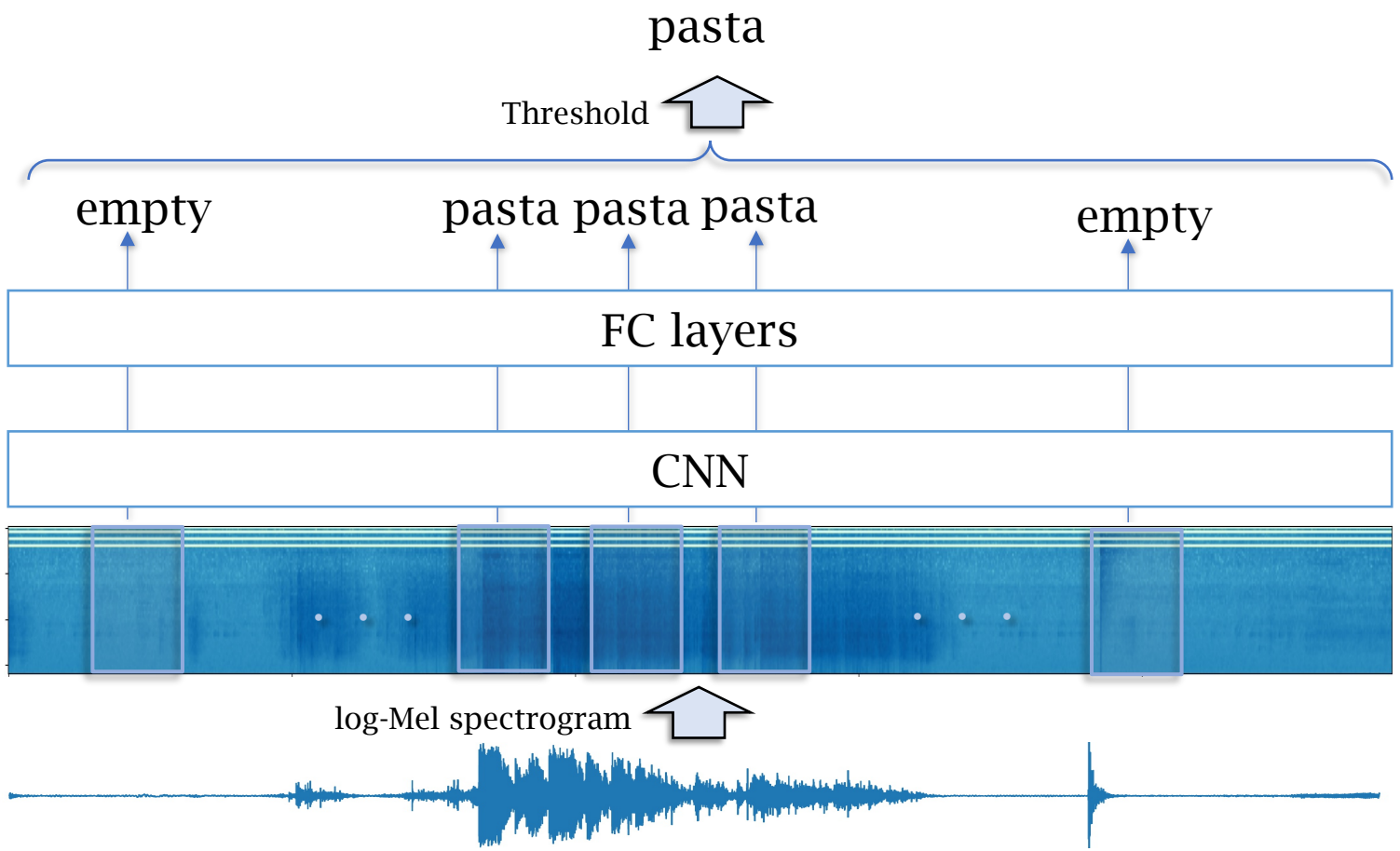


Model structure for Task2

Solution for Task2 (Filling type classification)

Input: × The entire audio clip
○ Cropped frame

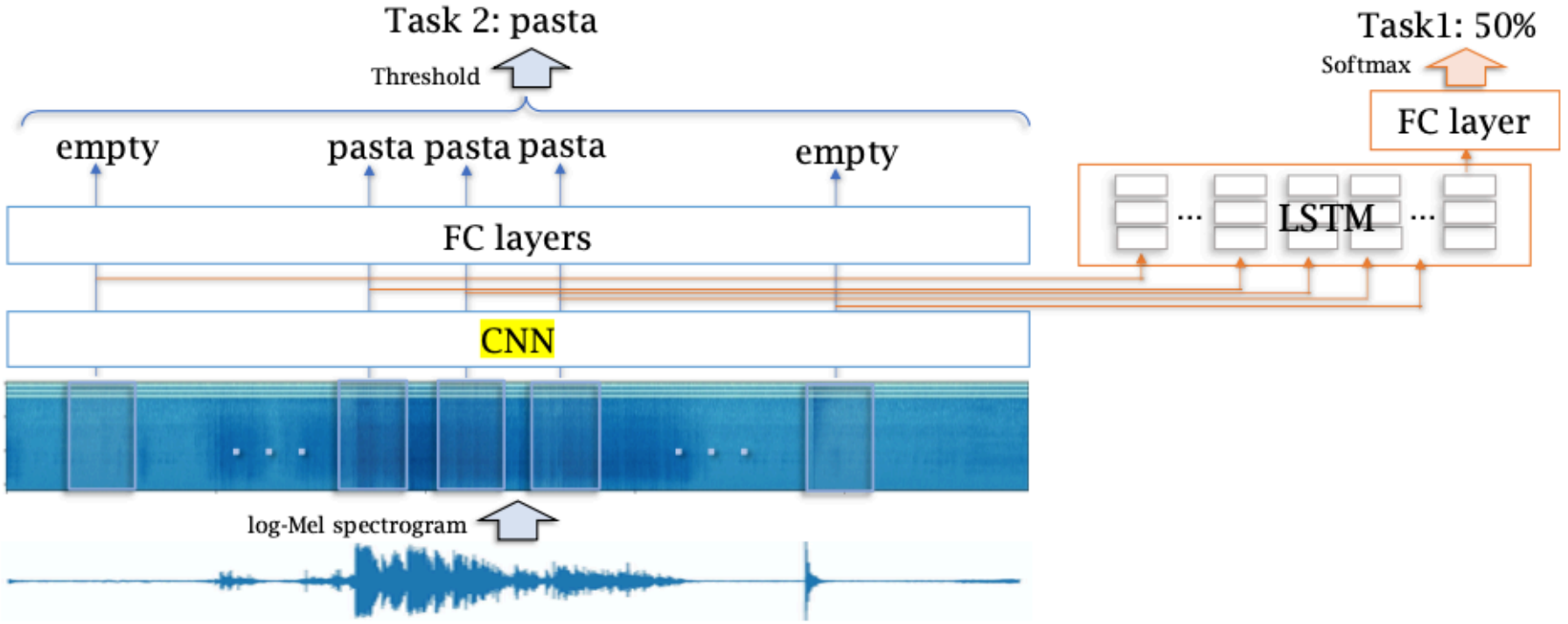
If Number of predictions of
"a filling type" ≥ Threshold
then "that filling type"



Overview of our framework for Task2

Solution for Task1 (Filling level classification)

- Reuse features of Task2
- Use LSTM



Overview of our framework for Task1 and Task2

Results of Task1 and Task2

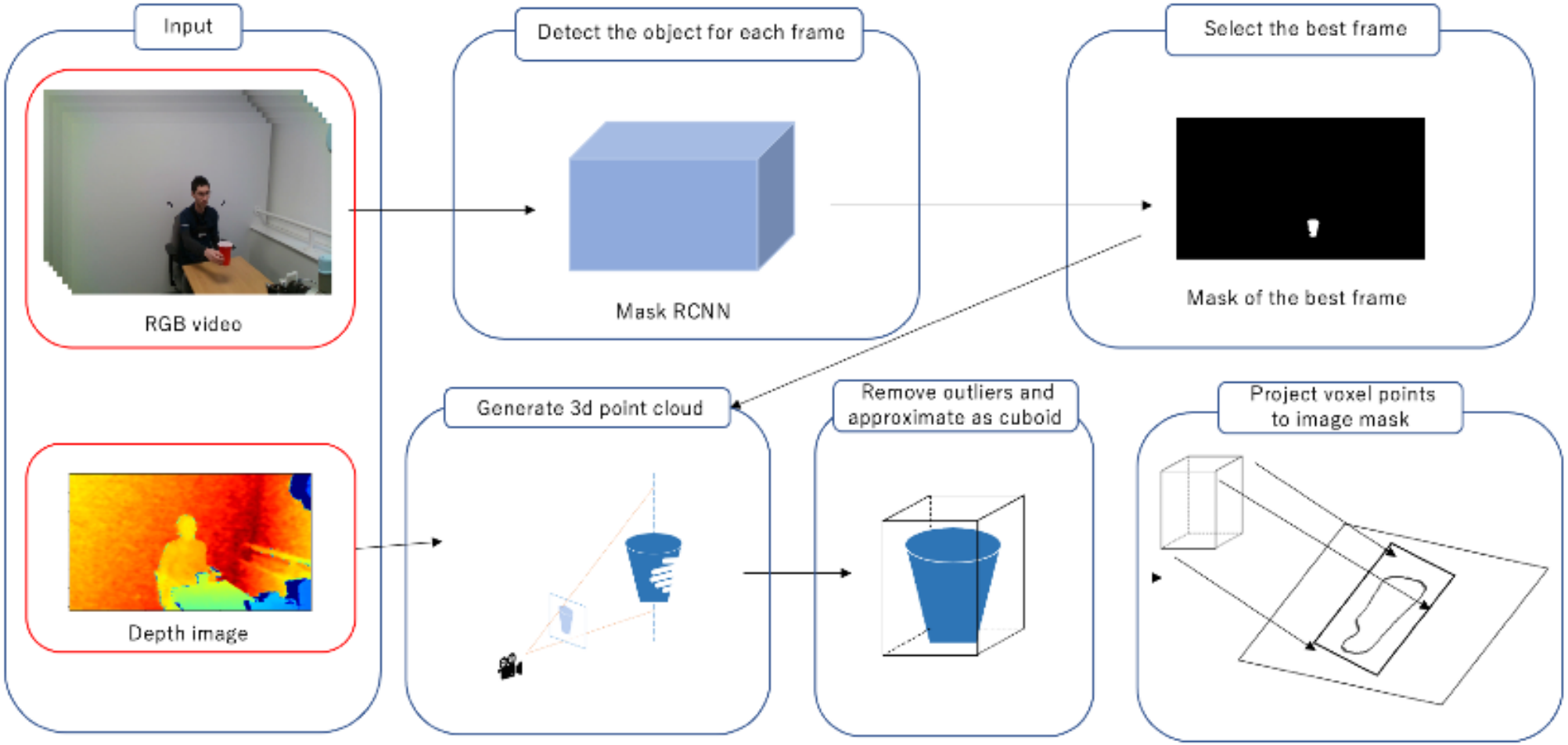
- High accuracy for Task2
- Lower accuracy for Task1

Our model doesn't take into account the capacity of the container.
Necessary to use visual data?

container	Task2		Task1	
	ACC [%]	WAFs [%]	ACC [%]	WAFs [%]
Red cup	98.80	98.80	48.80	35.23
Small white cup	88.09	88.27	75.00	73.16
Small transparent cup	97.61	97.65	78.57	78.05
Green glass	96.42	96.38	64.28	58.93
Wine glass	100.00	100.00	55.95	41.97
Champagne flute glass	100.00	100.00	82.14	82.05
Cereal box	95.00	95.11	56.66	55.58
Biscuit box	98.33	98.34	53.33	44.16
Tea box	95.00	95.11	61.66	52.59

The quantitative results of Task1 and Task2 using **leave one container cross-validation out scheme**. Each task is evaluated with accuracy, weighted F1-score

Solution for Task3 (Container capacity estimation)



Overview of our framework for Task3

Results of Task3 (Container capacity estimation)

- Bad prediction?

  ✓ We generated a 3D point cloud using only one point of view camera
  ✓ Cuboid approximation was a very rough approximation

Container	Score
Red cup	0.524
Small white cup	0.503
Small transparent cup	0.495
Green glass	0.598
Wine glass	0.649
Champagne flute glass	0.373
Cereal box	0.507
Biscuit box	0.462
Tea box	0.478

The evaluation of Task3 using **ACS metrics** for each container.

Conclusion

✓ Our solution still has room for improvement
← we only used the data independently for each task

✓ We will need to build a model that fuses multi-modal data (audio, infrared, depth, and RGB of multiple views)

Team	Description	Task 1	Task 2	Task 3	Public	Priv...	Over...
Because it's Tac...	GRU+ Random Forest for filling properties estimation. LoDE with RGB-D-IR data from selected frames in a video for volume estimation.	✓	✓	✓	64.98	65.15	65.06
HVRL	Log-Mel spectrogram-based audio features as input to VGG-based CNN and LSTM for filling properties estimation. Container volume from the shape approximation as cuboid of the 3D point cloud obtained with RGB-D data and object detection with Mask R-CNN.	✓	✓	✓	63.32	61.01	62.16
Concatenation	Multi-modal learning with audio features and prior of container categories through object detection for inferring container capacity and fluid properties.	✓	✓	✓	52.80	54.14	53.47
NTNU-ERC	MFCC features in a 20s-window + neural network to classify filling type. Object detection and selection of the closest contours (up to 700 mm) in the depth data + regression with a CNN for container capacity.		✓	✓	38.56	39.80	39.18
Random	Baseline with random estimations for each task.	✓	✓	✓	38.47	31.65	35.06

The results of overall tasks