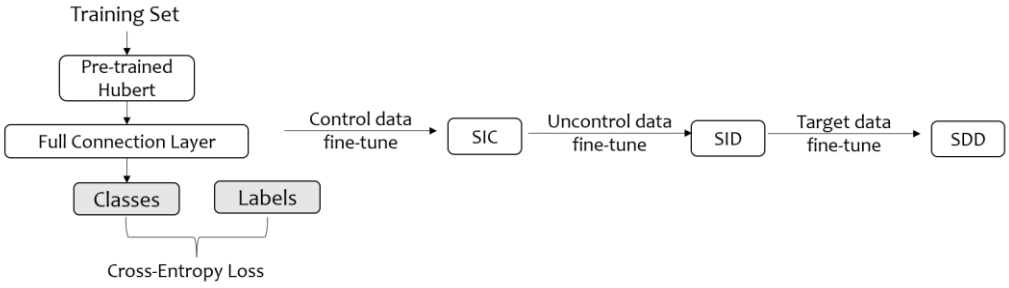
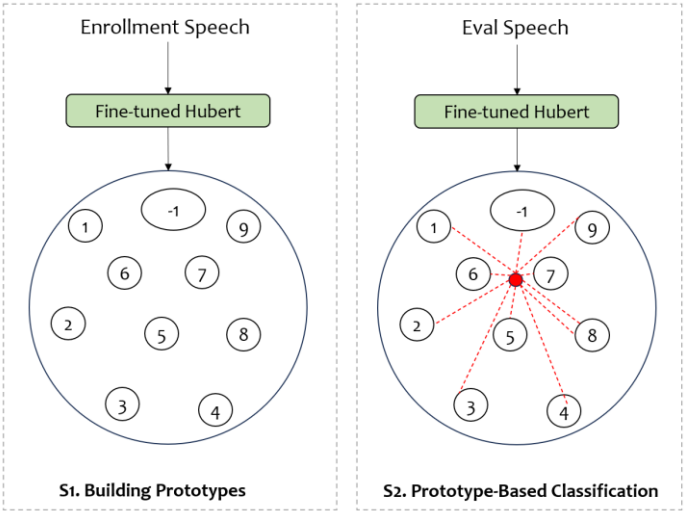
System Description

1. Methods:
   a) Dataset: We solely utilized the dataset provided by the competition organizers, without incorporating any additional datasets.
   b) Pre-trained models: chinese-hubert-base.pt
      i. The pre-training dataset is a subset of WenetSpeech[1] training L, about 10,000 hours.
      ii. Download link: https://huggingface.co/TencentGameMate/chinese-hubert-base .
   c) Model architecture:
      i. Overview:




**Training Phase**




**Inference Phase**

1. Training phase:
   a) Stage 1: Building Speaker-Independent Control KWS model (SIC):
      i. Incorporate a full connection layer following the pre-trained Hubert[2] model, employ cross-entropy loss for fine-tuning, and ensure the Hubert parameters remain unfrozen.
      ii. Perform fine-tuning utilizing the LRDWWS_train/Control dataset.
      iii. The training, validation, and testing splits remain consistent with those of the baseline.
   b) Stage 2: Building Speaker-Independent Dysarthria KWS model (SID)
      i. Perform fine-tuning on the SIC model built in stage 1.
      ii. Perform fine-tuning utilizing the LRDWWS_train/Uncontrol dataset.
      iii. The training, validation, and testing splits remain consistent with

those of the baseline.

c) Stage 3: Building Speaker-Dependent Dysarthria KWS systems (SDD)
   i. Perform fine-tuning on the SID model built in stage 2.
   ii. Perform fine-tuning utilizing the enrollment speech data of the target speaker.
   iii. Both the training set and the validation set consist of the enrollment speech of the target speaker.

2. Inference phase:
   a) Stage 1 Building Prototypes: utilizing the fine-tuned Hubert of the SDD model to extract features from the enrollment speech of the target speaker. Only the first frame features are retained, and the features of all samples within each class are averaged to build a prototype representing the class. In practice, as there is only one sample per keyword in the enrollment speech, the first frame of the sample is used as the prototype for the class. However, for non-keyword samples, their features are averaged to build the prototype.
   b) Stage 2 Prototype-Based Classification:
      i. Prepare the test feature by extracting the features of the target speaker's test speech using the fine-tuned Hubert of the SDD model. Retain only the first frame features.
      ii. Utilize the cosine similarity API provided by the Faiss[3] library to compute the cosine similarity between the test feature and each prototype. Assign the class ID associated with the nearest prototype as the classification result.

ii. Hubert Model architecture:

| Transformer layers | Embedding dim. | FFN dim. | Attention heads | Projection dim. |
|---|---|---|---|---|
| 12 | 768 | 3072 | 12 | 256 |

2. Experimental results:

| TeamID | Score(%) | FAR(%) | FRR(%) |
|---|---|---|---|
| T021 | 0.009801 | 0.004801 | 0.005000 |

3. References


[1] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu,X. Chen, C. Zeng et al., "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 6182–6186.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451–3460, 2021.

[3] J. Johnson, M. Douze, and H. J´egou, "Billion-scale similarity search with gpus," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2019.