System Description for T405 Team System

Juan Liu[1] , Ming Li[1,2] , Fei Su[1] , Cancan Li[1]

[1]Institute of Artificial Intelligence, School of Computer Science, Wuhan University

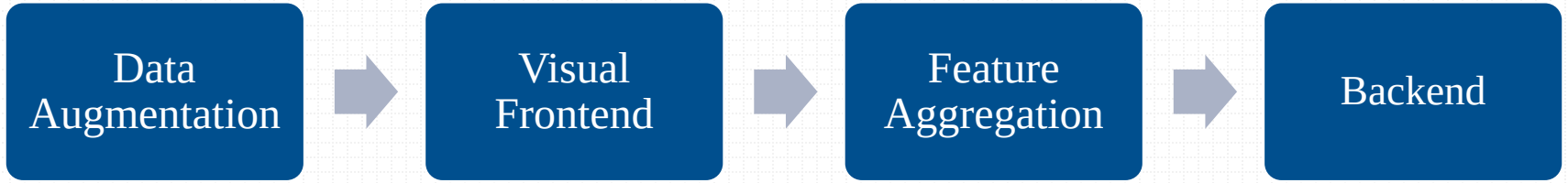
[2]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,

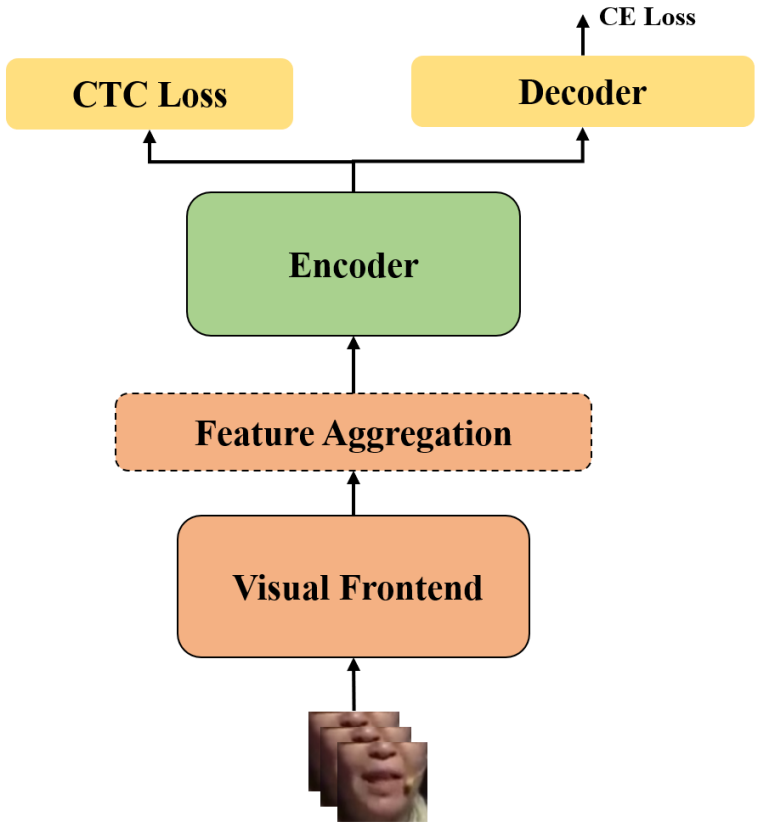
Data Science Research Center, Duke Kunshan University,

CONTENTS

1. Introduction

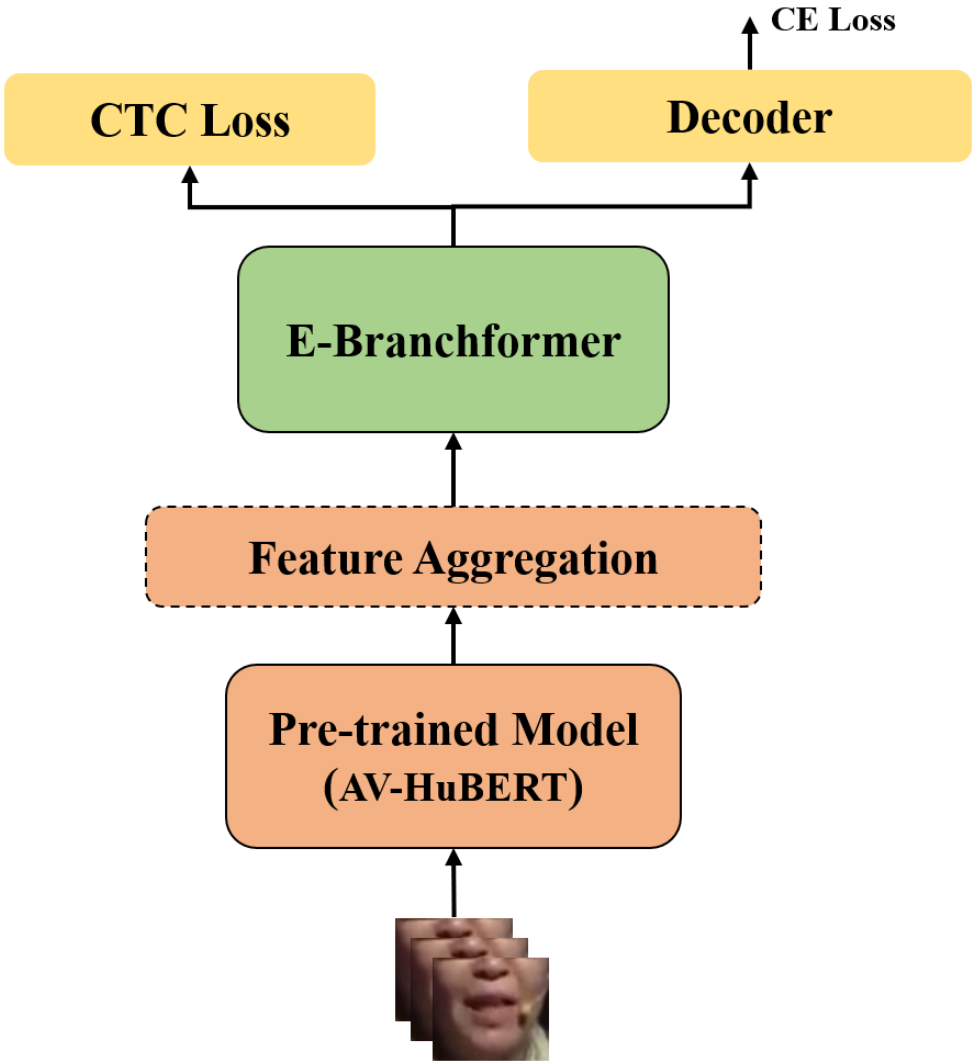
Overview T405 Team System





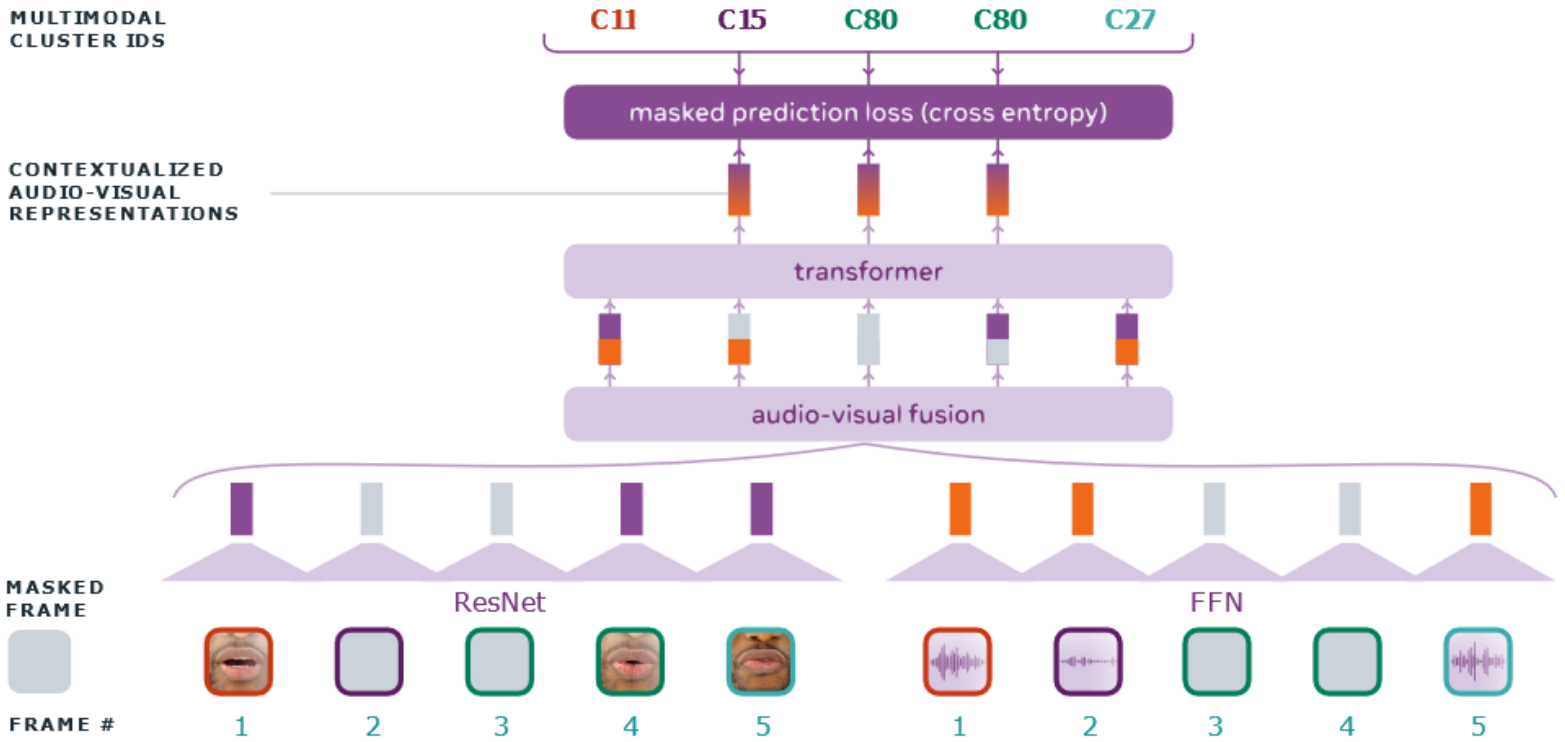
2. Methods

- **Proposed System:**
  - Pretrained Frontend
    - AV-Hubert
  - Multi-layer Feature Aggregation
  - Backend
    - E-Branchformer encoder
    - Bi-transformer decoder



2. Methods

- **Proposed System:**
  - Pretrained Frontend
    - AV-Hubert
  - Multi-layer Feature Aggregation
  - Backend
    - E-Branchformer encoder
    - Bi-transformer decoder



(c) Audio-visual HuBERT (proposed)

2. Methods

- **Proposed System:**
  - Pretrained Frontend
    - AV-Hubert
      - a: Self large vox 433h
      - b: Clean large vox iter5
      - c: Noise large vox iter5

The information of different pretrained frontend checkpoints. T represents Type. P represents the parameters used in our frontend. Pse represents the Pseudo-label data. Unl represents the Unlabeled data. Lab represents the Labeled data. NA represents the Noise-Augmented. WER represents the result of the checkpoint on LRS3.

Pretrain Frontend	P[M]	TS	Pse (h)	Unl (h)	Lab (h)	WER[%]
AV-Hubert	325.03	-	1326	1759	433	26.9
	325.03	-	-	1759	-	-
	325.03	NA	-	1759	-	-

2. Methods

- **Proposed System:**
  - Pretrained Frontend
    - AV-Hubert
  - Multi-layer Feature Aggregation
  - Backend
    - E-Branchformer encoder
    - Bi-transformer decoder

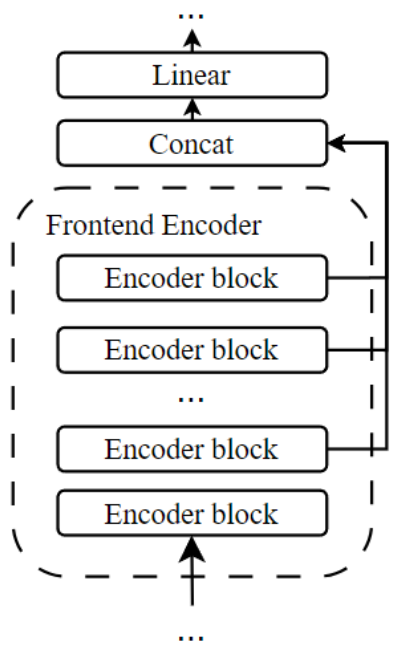


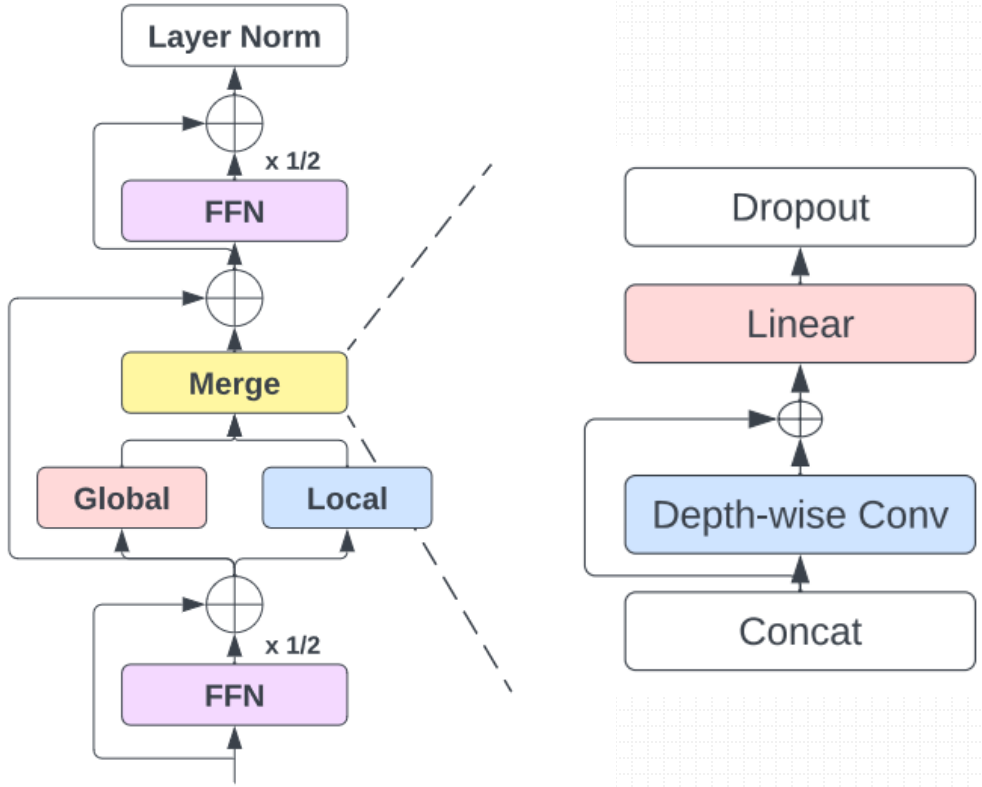
Illustration of Multi-layer Feature Aggregation on the results of the penultimate layers of the encoder.

$$H' = Concat(h_L, h_{L-1}, ..., h_{L-N+1})$$
$$H = Linear(H') \quad (1)$$

where L represents the number of encoder layers in frontend, $h_{L-n}, n \in [0, N]$ represents the feature from the last n encoder layer. If the dimension of the input of the backend is D, then the $H \in \mathbb{R}^{T \times D}$ represents the output feature of the MFA module.

2. Methods

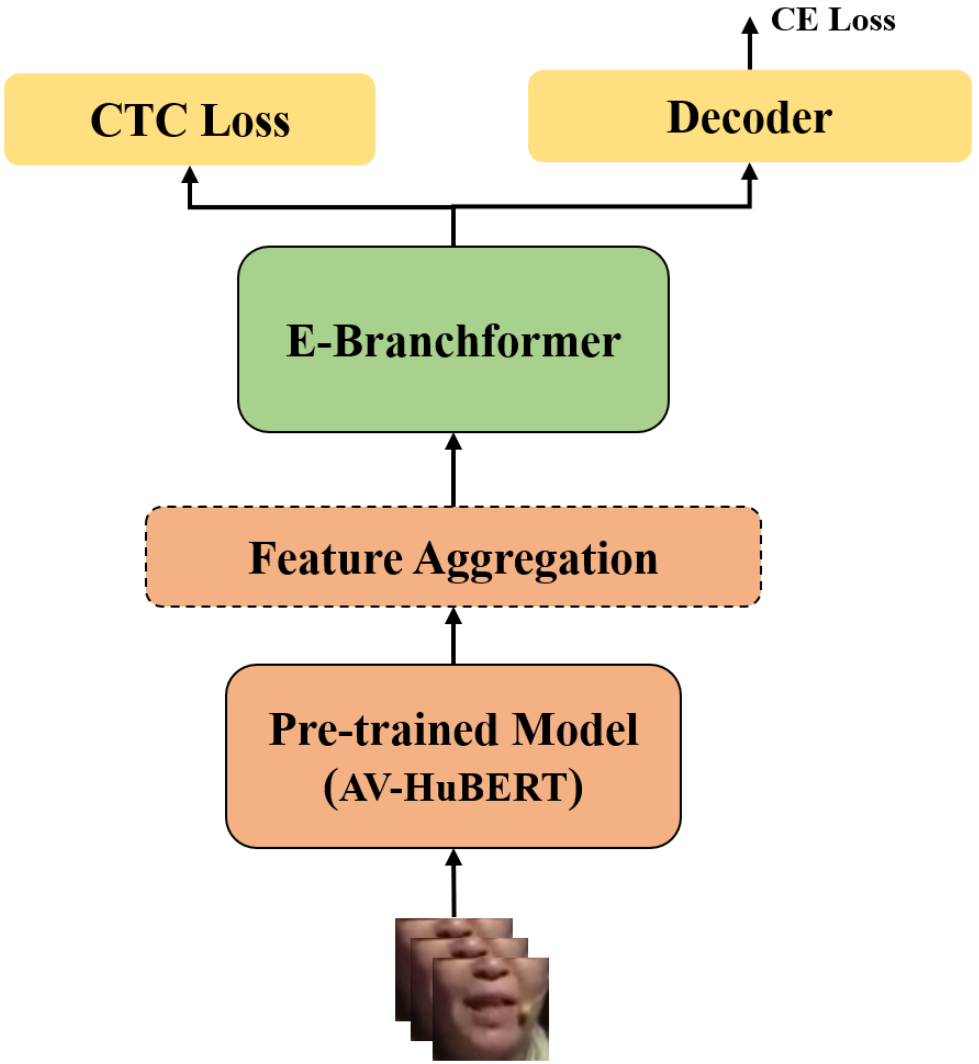
- **Proposed System:**
  - Pretrained Frontend
    - AV-Hubert
  - Multi-layer Feature Aggregation
  - Backend
    - E-Branchformer encoder
    - Bi-transformer decoder



(a)

2. Methods

- **Proposed System:**
  - Pretrained Frontend
    - AV-Hubert
  - Multi-layer Feature Aggregation
  - Backend
    - E-Branchformer encoder
    - Bi-transformer decoder



3. Results

- Frontend init by
  - a: Self large vox 433h
  - b: Clean large vox iter5
  - c: Noise large vox iter5
- Pretraining
  - CN-CVS
  - CN-CVS2-P1
- Fine-tuning
  - CNVSRC-Multi.Dev

System	Model	CER(%)
BS	Baseline	58.37
ID1	Proposed System	

3. Results

Conclusion

- Explored the application of a front-end for self-supervised representation extraction, coupled with a feature analysis back-end, for visual speech recognition;
- Proposed the Multi-layer Feature Aggregation by utilizing features from multiple layers of the pretrained frontend instead of solely relying on the output of the final layer;
- Due to our delayed involvement, we were constrained to conduct experiments on a limited of data and networks, yet these preliminary findings indicate that the unsupervised model exhibits enhanced robustness in the downstream task.

System	Model	CER(%)
BS	Baseline	58.37
ID1	Proposed System	

3. Results

Thanks