

# FPCRL: Feature Projection and Contrastive Representation Learning for End-to-End Speech Translation

Jiale Ou, Hongying Zan

School of Computer and Artificial Intelligence

Zhengzhou University

1791088334@qq.com, iehyzan@zzu.edu.cn



INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS  
**IJCNN2025**  
30 JUNE - 5 JULY 2025 | ROME, ITALY  
INTERNATIONAL NEURAL NETWORK SOCIETY

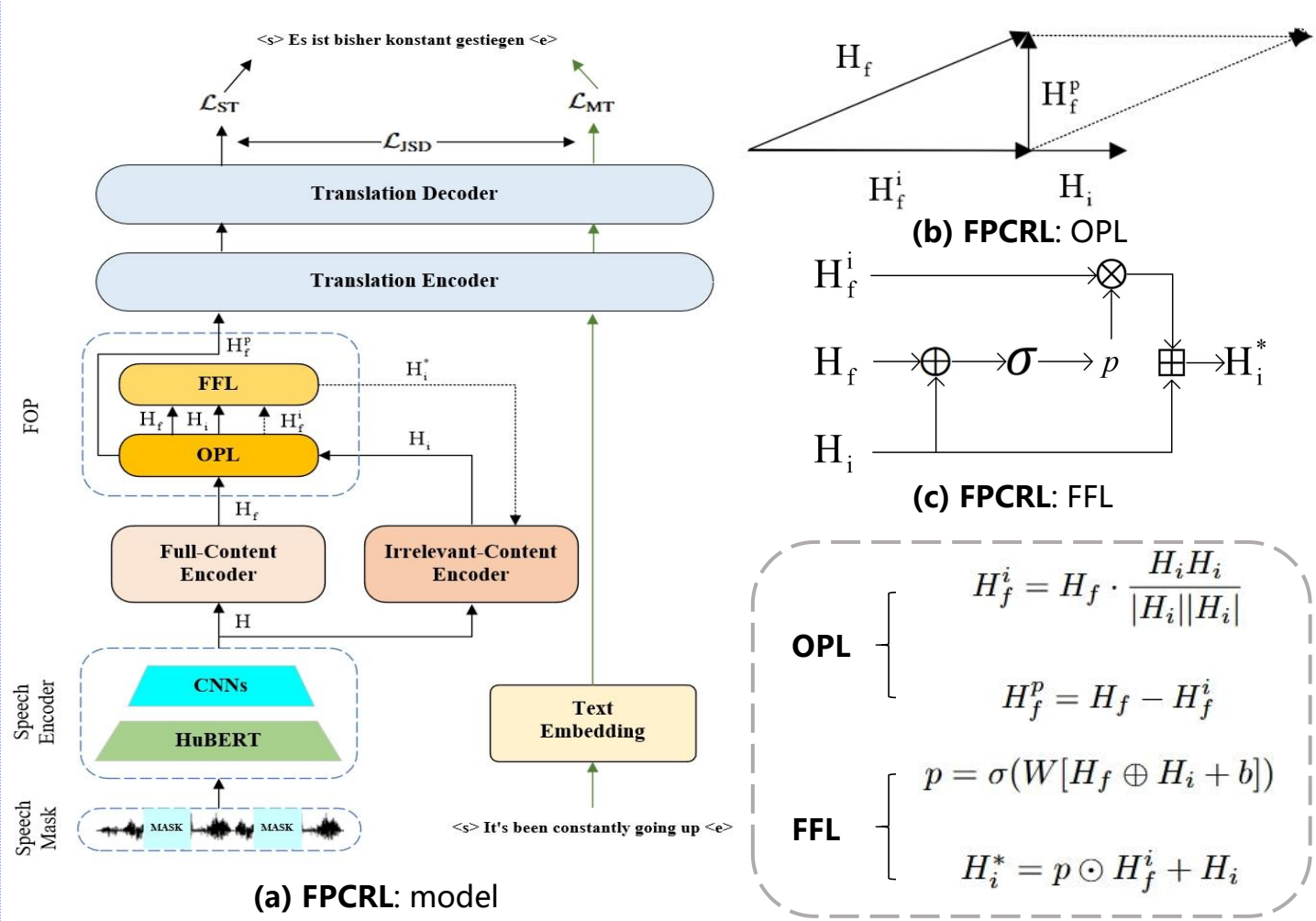


All Neural Network  
roads lead to  
Rome

## Introduction

Speech-to-text translation is a cross-modal and multilingual translation task. To alleviate the modality gap and data scarcity of this task, recent research has primarily focused on aligning speech and text representations to unify cross-modal features and incorporating external knowledge via multi-task learning. Although significant progress has been achieved, there remains potential for improvement, particularly in enhancing translation performance by purifying speech representations to extract more content-relevant information. In this paper, we propose a framework based on feature projection and contrastive representation learning for speech translation, FPCRL, which is adaptable to various training settings. FPCRL introduces additional full-content and irrelevant-content encoders, which separately extract full and irrelevant information from speech. Through a feature projection module, irrelevant components are removed from the full content representations, yielding purified speech representations. Furthermore, the extracted content-irrelevant information is utilized to guide the training of the irrelevant-content encoder via contrastive representation learning. Experiments on the MuST-C En-De, CoVoST-2 De-En, and CoVoST-2 Fr-En benchmarks demonstrate that FPCRL achieves significant improvements across all datasets. We open-source the model for future research at <https://github.com/Akito-Go/FPCRL>.

## Framework of FPCRL



## Main contributions:

- **E2E ST Training Framework.** We propose a **training framework FPCRL for E2E ST**, which can be applied in various settings. This framework purifies speech representations by introducing additional full-content and irrelevant-content encoders and employing a feature orthogonal projection method.
- **Speech Purification and Feature Fusion.** By leveraging the redundant content-irrelevant information extracted during **the purification process**, we introduce a **feature fusion method** to combine it with the output from the irrelevant-content encoder. This fusion further guides the encoder to effectively capture and learn the content-irrelevant information.
- **Significant improvements.** Experiments on the MuST-C En-De, CoVoST-2 De-En, and CoVoST-2 Fr-En benchmarks show that the methods we proposed can lead to significant improvements over strong E2E ST baselines across three settings: **transcript-free**, **multi-task**, and **expanded data**.

## Main Results

### Evaluation Metrics:

- ✓ **BLEU:** An automatic metric used to evaluate the quality of machine-generated text, especially in translation tasks. It measures how closely a candidate translation matches one or more reference translations based on overlapping n-grams. Scores range from 0 to 1 (often shown as 0–100), with higher scores indicating better translation quality.

## Methods

### (a) Speech Mask:

We modify the speech input  $s$  to mask continuous segments and obtain the masked waveform  $s'$ , which is utilized as the input of the model. Then we mask with a probability of 0.75 for each speech input. The selected speech input is then masked for at least 2 spans, each containing at least 3600 consecutive frames.

### (b) Orthogonal Projection:

We use the OPL to remove content-irrelevant information from the full content representations and firstly project  $H_f$  onto  $H_i$  to extract the content-irrelevant information  $H_f^i$ . Then we obtain the expected purified speech representations  $H_f^p$  by removing  $H_f^i$  from  $H_f$ .

### (c) Feature Fusion:

We fuse  $H_f^i$  and  $H_i$  into  $H_i^*$ , and the future fusion process automatically controls the selection of  $H_f^i$  based on  $H_f$  and  $H_i$ , and then adds the selected portion to  $H_i$  to obtain the fused features  $H_i^*$ .

### (d) Contrastive Learning:

To further enhance the encoding ability of irrelevant-content encoder for content-irrelevant information, we leverage  $H_i$  and  $H_i^*$  for contrastive learning.

$$\mathcal{L}_{CRL} = -\left[\sum_{m=1}^N \log \frac{\text{sim}(H_i, H_{im}^*)}{\tau} + \sum_{m=1}^N \sum_{n \neq m} \log \left(1 - \frac{\text{sim}(H_i, H_{in}^*)}{\tau}\right)\right]$$

### (e) Consistency Constraints:

For the speech input  $s_i$ , we apply Gaussian noise with a perturbation level defined randomly by the signal-to-noise ratio  $\text{snr} \in [5, 50]$  to obtain new noisy data:  $\tilde{s}_i = s_i + \text{Guss}(s_i, \text{snr})$  to verify the effectiveness of the FOP in removing content-irrelevant information.

$$\mathcal{L}_{CST} = \sum_{i=1}^{|D|} \|\text{Mean}(H_f^p) - \text{Mean}(\tilde{H}_f^p)\|_2$$

### (f) Loss Warm-up:

We propose a loss warm-up method, where the loss coefficient is initialized to a small value of  $1e-8$  at the beginning of training. As the number of training steps increases, the coefficient gradually increases to 1, allowing the loss to fully contribute to optimization.

Models	External Data	Speech	ASR	MT	BLEU
Training in <i>transcript-free</i> setting					
Fairseq ST [5]	-	-	-	-	22.7
Revisit ST [41]	-	-	-	-	23.0
Self-training [17]	✓	✓	-	-	25.2
CCSRD [11]	✓	-	-	-	25.4
DUB [40]	✓	-	-	✓	26.2
SRPSE [15]	✓	-	-	-	<b>26.2</b>
W2V2-Transformer	✓	-	-	-	24.3
HuBERT-Transformer	✓	-	-	-	24.4
FPCRL	✓	-	-	-	<b>25.7</b>

Models	External Data	Speech	ASR	MT	CoVoST-2
De-En					
Transformer-ST [35]	-	✓	-	-	17.1
Revisit ST [41]	-	-	-	-	14.1
Siamese-PT [43]	-	✓	✓	-	19.7
DUB [40]	✓	-	-	✓	19.5
FPCRL	✓	-	-	-	16.4
FPCRL-MTL	✓	-	-	-	<b>22.8</b>

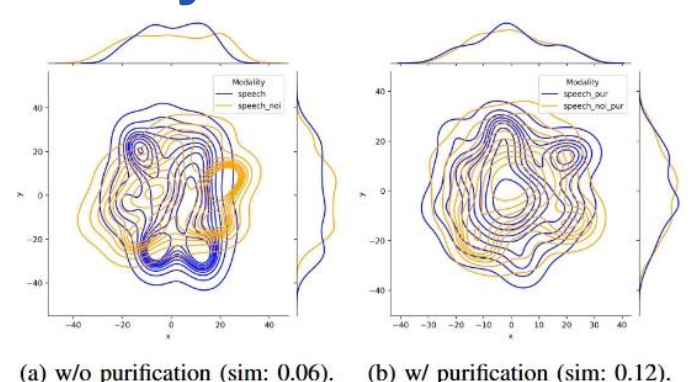
Models	External Data	Speech	ASR	MT	BLEU
Training in <i>multi-task</i> setting					
XSTNet [6]	✓	-	-	-	25.5
STEMM [7]	✓	-	-	-	25.6
ConST [8]	✓	-	-	-	25.7
CCSRD [11]	✓	-	-	-	26.1
M <sup>3</sup> ST [42]	✓	-	-	-	26.4
SRPSE [15]	✓	-	-	-	26.9
CMOT [9]	✓	-	-	-	<b>27.0</b>
FPCRL	✓	-	-	-	26.8

Models	External Data	Speech	ASR	MT	CoVoST-2
Fr-En					
XSTNet [6]	✓	-	-	✓	27.8
CCSRD [11]	✓	-	-	✓	28.1
ConST [8]	✓	-	-	✓	28.2
STEMM [7]	✓	-	-	✓	28.7
CMOT [9]	✓	-	-	✓	29.0
SRPSE [15]	✓	-	-	✓	<b>29.2</b>
FPCRL	✓	-	-	✓	28.8

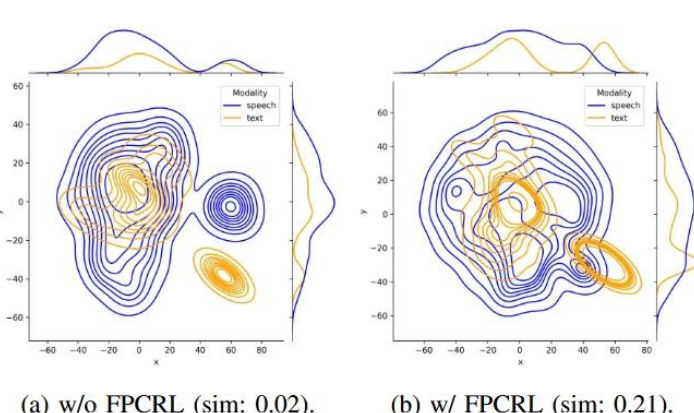
**Ext. Main Results:** comparison with baselines & multilingual verification

## Ablation Study

Models	BLEU
Our FPCRL	<b>25.7</b>
w/o $\mathcal{L}_{CST}$	25.5
w/o Speech Mask	25.3
w/o $\mathcal{L}_{CRL}$	24.5
w/o OPL	24.3



Models	LW <sub>CRL</sub>	LW <sub>CST</sub>	BLEU
FPCRL <sub>CRL</sub>	0	-	24.8
	5k	-	25.1
	10k	-	25.3
	15k	-	<b>25.5</b>
FPCRL <sub>CRL-CST</sub>	15k	0	24.9
	15k	5k	25.0
	15k	10k	25.5
	15k	15k	<b>25.7</b>



**Ext. Methods Evaluation:** ablation studies & visualization