CMSP-ST: Cross-modal Mixup with Speech Purification 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

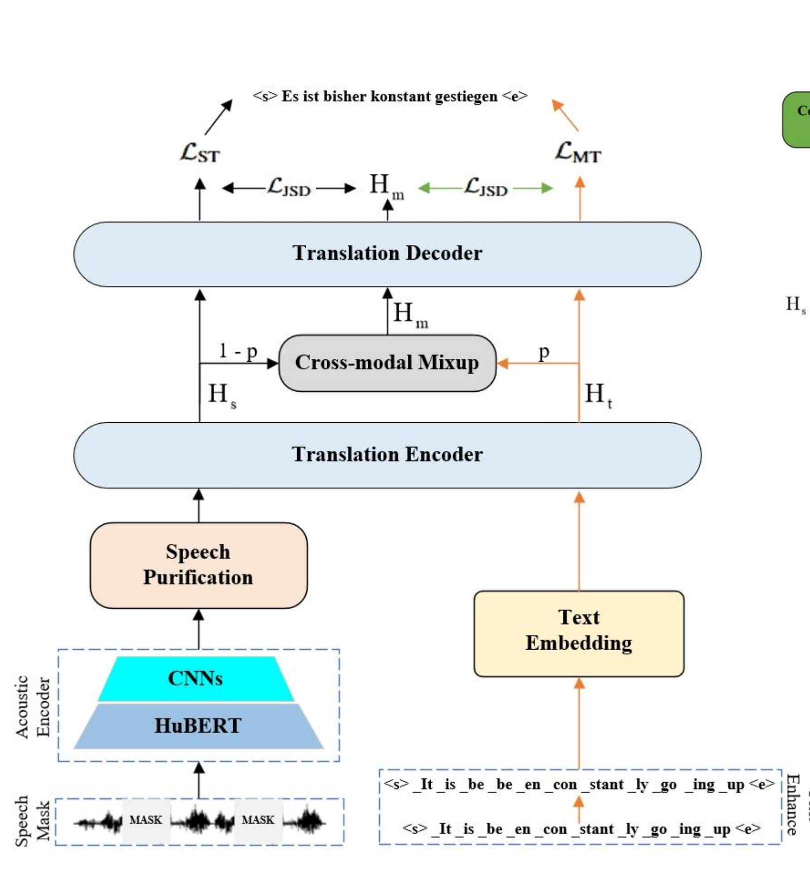
Fair and Inclusive Speech Science and Technology

Introduction

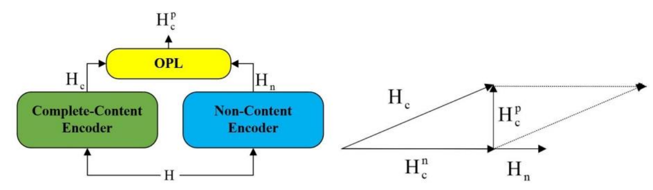
End-to-end speech translation (E2E ST) aims to directly convert speech in a source language into text in a target lan-guage, and its performance is constrained by the inherent modality gap. Existing methods attempt to align speech and text representations to perform cross-modal mixup at the token level, which overlooks the impact of redundant speech information. In this paper, we propose cross-modal mixup with speech purification for speech translation (CMSP-ST) to address this issue. Specifcally, we remove the non-content features from speech through orthogonal projection and extract the purifed speech features for cross-modal mixup. Additionally, we employ adversarial training under the Soft Alignment (S-Align) to relax the alignment granularity and improve robustness. Experimental results on the MuST-C En-De, CoVoST-2 Fr-En, and CoVST-2 De-En benchmarks demonstrate that CMSP-ST effectively improves the speech translation performance of existing cross-modal mixup methods.

We open-source the model for future research at ***https://github.com/Akito-Go/CMSP-ST***.

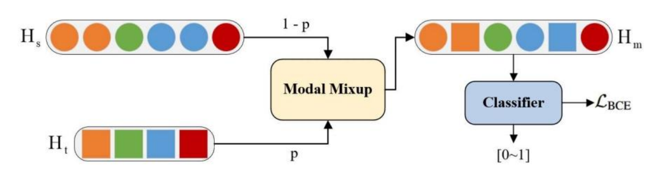
Framework of CMSP-ST



(a) **CMSP-ST**: model



(b) **CMSP-ST**: SP



(c) **CMSP-ST**: CM

SP:
$$H_f^i = H_f \cdot \frac{H_i H_i}{|H_i||H_i|}$$
$$H_f^p = H_f - H_f^i$$

CM:
$$\begin{cases} h_i^s, p > p^* \\ h_{a_i}^t, p \le p^* \end{cases}$$

Methods

(a) Model architecture:
Our model adopts an encoder-decoder architecture, comprising six main modules: the acoustic encoder (A-Enc), text embedding (T-Emb) module, translation encoder (T-Enc), speech purification (SP) module, cross-modal mixup (CMM) module, and translation decoder (T-Dec).

(b) Masking strategy:
We implement a masking strategy for the input of the A-Enc to enhance speech purification, following the configuration of CCSRD. Furthermore, we randomly insert repeated elements or padding into the input of the T-Emb with a predefined probability to simulate the characteristics of speech content information.

(c) Classifier network:
In the CMM module, we introduce a classifier network consisting of three feed-forward layers and an output layer followed by sigmoid activation for modality classification.

(d) Speech purification:
The SP module consists of a complete-content encoder (CC-Enc), a non-content encoder (NC-Enc), and an orthogonal projection layer (OPL). The output of the T-Enc is first processed by the CC-Enc to obtain the complete feature representations H_c, while the NC-Enc extracts the non-content feature representations H_n. we project H_c onto H_n using the OPL to extract the redundant non-content information H_c^n from H_c. Then we project H_c onto the orthogonal hyperplane to H_c^n to obtain the purified speech representations H_c^p.

(e) Cross-modal mixup:
Considering that achieving the ideal H-Align is difficult and may conflict with cross-modal mixup, we introduce S-Align to relax the alignment granularity. The classifier adjusts the classification target from the modality ID (0 or 1) to p, achieving a shift from S-Align to H-Align, aiming to learn a unified representation space by identifying the modality spaces of the input representations. To further enhance its effectiveness, we use a pseudo-label with a fixed mixup probability of 0.5 for adversarial training and employ binary cross-entropy (BCE) loss for modality classification. The overall adversarial training objective can be described as follows.

$$\mathcal{L}_{\text{ADV}} = -\log P(p_s|h_s) - \log P(p_t|h_t) - \log P(p_f|h_s) - \log P(p_f|h_t)$$

(f) Training objective:
The overall training objectives for both the multi-task and external data settings are as follows.

$$\mathcal{L} = \mathcal{L}_{\text{ST}} + \mathcal{L}_{\text{MT}} + \mathcal{L}_{\text{MIX}} + \mathcal{L}_{\text{ADV}}$$

Main contributions:

- ***E2E ST Training Framework.*** We propose **Cross-modal Mixup** with **Speech Purification** for End-to-end Speech Translation (CMSP-ST).
- ***Speech Purification.*** We introduce two additional encoders, one for extracting non-content information from speech and the other for extracting complete speech features, and obtain **content-focused purified speech features** by removing non-content information from complete speech features through **an orthogonal projection strategy**.
- ***Adversarial Training with Soft Alignment.*** We use **Soft Alignment (S-Align)** to relax alignment granularity by aligning **the representation spaces** of speech and text, and further improve the robustness of the model through **adversarial training**. Based on this, we implement token-level mixup of text and purified speech.
- ***Significant improvements.*** Experimental results on the MuST-C En-De, CoVoST-2 Fr-En, and CoVoST-2 De-En datasets show that the CMSP-ST method can enhance the knowledge transfer of existing cross-modal mixup methods and effectively **alleviate the modality gap** in ST tasks.

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.

Models	Speech Pretraining	BLEU Multi-tasks	BLEU Exter.data
JT-S-MT [5]	×	24.1	26.8
XSTNet [6]	✓	25.5	27.8
STEMM [11]	✓	25.6	28.7
ConST [10]	✓	25.7	28.3
CCSRD [9]	✓	26.1	28.1
S-Align-ST [13]	✓	26.5	28.6
SRPSE [14]	✓	26.9	**29.2**
CMOT [12]	✓	27.0	29.0
HuBERT-Transformer [12]	✓	25.4	27.5
CMSP-ST	✓	**27.4**	29.1

In the multi-task setting, CMSP-ST outperforms HuBERT-Transformer by **2.0 BLEU** and surpasses CMOT, which also uses OT and cross-modal mixup, by **0.4 BLEU**. With the introduction of external MT data, CMSP-ST also slightly outperforms CMOT and achieves performance comparable to SRPSE.

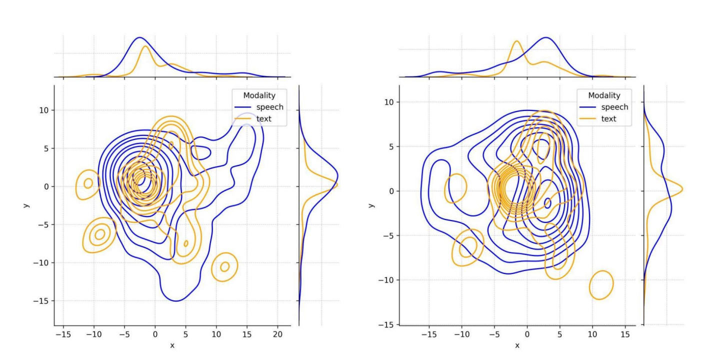
The experimental results demonstrate that, despite some baseline models leveraging large-scale external ASR and MT data in the pre-training stage to train encoder/decoder modules, or employing back-translation techniques, **the CMSP-ST model still achieves performance that is comparable**.

Models	Speech Pretraining	BLEU Fr-En	BLEU De-En
Transformer-ST [20]	✓	26.3	17.1
Revisit ST [27]	×	26.9	14.1
Siamese-PT [28]	✓	28.4	20.4
DUB [29]	✓	29.5	19.5
SRPSE [14]	✓	29.3	21.4
CMSP-ST	✓	**31.3**	**22.4**

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

Ablation Study

Models	BLEU
CMSP-ST$_{\text{MTL}}$	**27.4**
w/o Adv Training (S-Align)	27.0
w/o Cross-modal Mixup	26.6
w/o Data Augmentation	26.5
w/o Speech Purification	25.4



(a) *w/o speech purification.* (b) *w/ speech purification.*

Models	Adv Training	BLEU
CMSP-ST	×	27.0
CMSP-ST w/ S-Align	×	27.2
CMSP-ST w/ S-Align	✓	**27.4**
CMSP-ST w/ H-Align	✓	27.2

Ext. Methods Evaluation: ablation studies && visualization