



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, Zhengzhou, China

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

Abstract

End-to-end speech translation (E2E ST) aims to directly convert speech in a source language into text in a target language, 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. Specifically, we remove the non-content features from speech through orthogonal projection and extract the purified 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 CoVoST-2 De-En benchmarks demonstrate that CMSP-ST effectively improves the speech translation performance of existing cross-modal mixup methods.

Index Terms: speech translation, cross-modal mixup, speech purification, orthogonal projection, adversarial training

1. Introduction

In recent years, end-to-end speech translation (E2E ST) models have attracted much attention. Compared with traditional cascade models, E2E ST models eliminate the intermediate transcription step, resulting in lower latency and avoiding error propagation [1, 2]. However, as a cross-modal and cross-lingual task, the data scarcity and modality gap make it challenging to train high-performance ST models. To address these problems, recent studies have explored methods of leveraging external machine translation (MT) data to assist ST training. This approach involves pre-training certain ST model modules with auxiliary tasks to enable cross-modal and cross-lingual translation, followed by fine-tuning them on ST data. The main techniques include pre-training [3, 4], multi-task learning [5, 6], and knowledge distillation [7, 8], all of which help transfer knowledge from MT task to ST task while alleviating the modality gap, thereby improving speech translation performance.

Although these transfer learning methods have demonstrated promising results in improving translation performance, their effectiveness often relies on the assumption that speech and text are well aligned in a common representation space. However, since speech features are derived from temporal variations in the signal, they contain non-content information such as timbre, pitch and rhythm in addition to content [9], which is the main reason why speech sequences are significantly longer than the textual one. Moreover, each text token conveys ex-

PLICIT meaning, whereas continuous speech frames often require mapping multiple frames to a single token for alignment, which complicates cross-modal alignment. Recent studies have adopted contrastive learning methods to enhance the alignment of speech and text representations [10], while others have introduced token-level mixup strategies to facilitate knowledge transfer across modalities [11, 12]. However, such methods often rely on strict Hard Alignment (H-Align) of individual speech and text segments, which is not easy to implement, and its effectiveness may be affected when combined with other methods [13]. Furthermore, these methods overlook the impact of redundant factors in speech on ST models. Previous studies have demonstrated that non-content information in speech can mislead the model toward context-irrelevant features, degrading ST performance [14].

To address these issues, we propose **Cross-modal Mixup with Speech Purification for Speech Translation (CMSP-ST)**¹. Specifically, 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 [15]. In addition, we use Soft Alignment (S-Align) [13] 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. 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.

2. Proposed Approach

2.1. Model architecture

Our model is shown in Figure 1. It 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). The A-Enc is implemented with a pre-trained model, processes speech input and converts it into token-level representations. The T-Emb maps tokenized text to the corresponding embedding vector. The T-Enc processes either tokenized text or the output from the A-Enc to extract semantic information. The SP removes non-content components from speech to obtain purified representations. The CMM implements cross-modal mixup of the outputs of the A-Enc and T-Enc. The T-Dec processes ei-

*Corresponding author.

¹Our code is available at <https://github.com/Akito-Go/CMSP-ST>

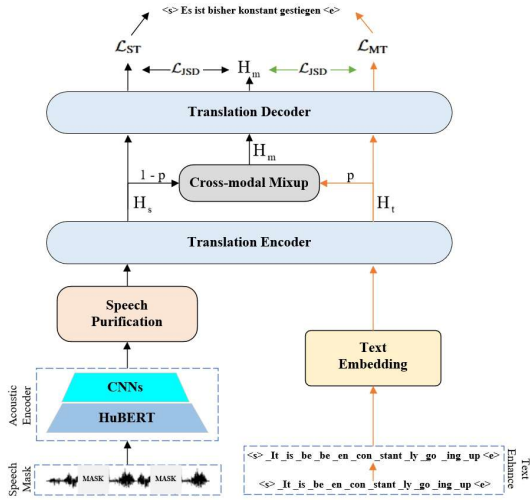


Figure 1: Overview of our proposed model, where text embedding and the MT forward path are removed during inference.

ther the single-modal representations or the cross-modal mixed representations of speech and text to generate translations.

We use the Transformer [16] as the backbone network and apply data augmentation to both speech and text sequences. Specifically, we implement a masking strategy for the input of the A-Enc to enhance speech purification, following the configuration of CCSRD [9]. Furthermore, considering the length differences between the speech and text sequences, 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 [10, 17]. 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 [13].

2.2. Speech purification

As shown in Figure 2 (a), 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 .

As shown in Figure 2 (b), 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 . These processes can be described as follows:

$$H_c^n = H_c \cdot \frac{H_n H_n^T}{|H_n| |H_n|} \quad (1)$$

$$H_c^p = H_c - H_c^n \quad (2)$$

Through these processes, we eliminate redundancy in the complete speech features and generate purified speech representations.

2.3. Cross-modal mixup

For the speech representations H_s and the corresponding text representations H_t output by the T-Enc, we first align them to obtain the alignment $A = (a_1, \dots, a_n)$ using the optimal transport (OT) [18] strategy. Then we perform token-level mixup

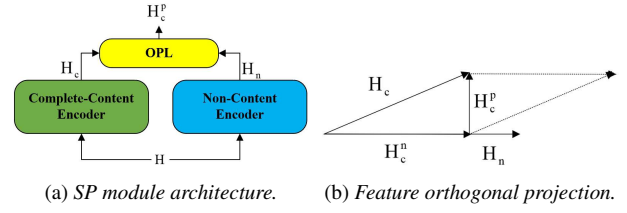


Figure 2: Overview of the architecture and principles of the SP module. (a) is the SP module architecture, and (b) is the display of the feature orthogonal projection method.

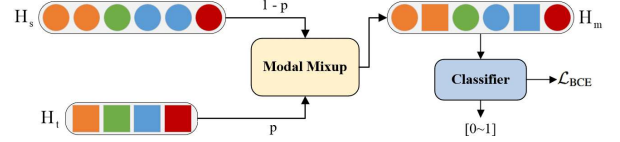


Figure 3: Overview of the CMM module. The classifier predicts the mixing probability of different modalities under S-Align.

based on A with the mixup probability p , which is sampled from the uniform distribution $U(0, 1)$, to obtain the mixed representations H_m [12]. The process is illustrated in Figure 3.

Based on the predefined mixup probability p^* , we generate the $H_m = (h_1, \dots, h_n)$ as follows:

$$h_i = \begin{cases} h_i^s, & p > p^* \\ h_{a_i}^t, & p \leq p^* \end{cases} \quad (3)$$

where $h_i^s \in H_s$, $h_{a_i}^t \in H_t$, and a_i is obtained from the alignment A. This alignment represents the most probable position of a fragment of the text representations corresponding to certain fragments in the speech representations.

Considering that achieving the ideal H-Align is difficult and may conflict with cross-modal mixup, we introduce S-Align [13] 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:

$$\mathcal{L}_{ADV} = -\log P(p_s | h_s) - \log P(p_t | h_t) - \log P(p_f | h_s) - \log P(p_f | h_t) \quad (4)$$

where p_s and p_t denote the mixup probability of the speech and text representations, respectively, and p_f denotes the fake mixup probability.

2.4. Training objective

We adopt a progressive training strategy [6] to train the model. First, we pre-train the T-Emb, T-Enc, and T-Dec on the MT task, and then fine-tune them on the ST task with multi-task learning. The ST and MT training sets are denoted as (s, x, y) and (x, y) respectively, where s , x , and y represent the speech, transcription, and translation sequences, respectively. The training objectives can be described as follows:

$$\mathcal{L}_{ST} = - \sum_{(s,y) \in D} \log P(y|s) \quad (5)$$

$$\mathcal{L}_{\text{MT}} = - \sum_{(x,y) \in D} \log P(y|x) \quad (6)$$

In addition, we minimize the average Jensen-Shannon divergence (JSD) between the probability distributions of the two modality representations and the mixed representations, which helps transfer knowledge from the MT task to the ST task:

$$\begin{aligned} \mathcal{L}_{\text{MIX}} = \frac{1}{2} [& \sum_{(s,y) \in D} \text{JSD}[P(y|s)||P(y|m)] \\ & + \sum_{(x,y) \in D} \text{JSD}[P(y|x)||P(y|m)]] \end{aligned} \quad (7)$$

The overall training objectives for both the multi-task and external data settings are as follows:

$$\mathcal{L} = \mathcal{L}_{\text{ST}} + \mathcal{L}_{\text{MT}} + \mathcal{L}_{\text{MIX}} + \mathcal{L}_{\text{ADV}} \quad (8)$$

3. Experiment and Results

3.1. Dataset setup

We conducted our experiments on the MuST-C English-German (En-De), CoVoST-2 French-English (Fr-En), and CoVoST-2 German-English (De-En) datasets. MuST-C [19] is a multilingual speech translation corpus comprising speeches, transcripts, and translations from TED Talks. CoVoST-2 [20] is a large-scale speech translation corpus based on the Common Voice project, covering translations from 21 languages to English and from English to 15 languages. We used the official train/dev/test split. Furthermore, we utilized the WMT16 En-De dataset as external data for the MuST-C En-De task for MT pre-training.

3.2. Experimental setup

Pre-processing For speech input, we used 16-bit 16kHz mono-channel raw audio and filtered out samples with frames greater than 480k or fewer than 1k for efficiency. For transcripts and translations, we tokenized them using a unigram SentencePiece [21] model with a vocabulary size of 10k to build a shared vocabulary.

Model Configurations All models were built using the Fairseq² [22, 23] toolkit with a hidden size of 512. The A-Enc was initialized with the HuBERT³ [24] model, which was pre-trained on LibriSpeech-960h [25] without fine-tuning. The kernel size, stride, and hidden size of the two CNN layers on HuBERT were set to 5, 2, and 512, respectively. The SP module set 1 layer for both encoders, while the T-Enc and T-Dec were each set to 6 layers. Euclidean distance was used in the OT to measure the modality gap between speech and text representations.

Training and Inference We set the learning rate to 1e-4, the warm-up steps to 25k. For data augmentation, we set the probability of speech masking to 0.75, and text enhancement to 0.2. For cross-modal mixup, we referred to [12] and set p^* to 0.2. Training was stopped early if the BLEU score on the dev set did not improve for 10 epochs. We used beam search with a beam size of 8 and a length penalty of 1.4, using the average of the last 10 epochs. We evaluated the case-sensitive detokenized BLEU scores on the test set using sacreBLEU [26]. All experiments were performed on two Nvidia GeForce RTX 4090 GPUs.

Baselines We compared the proposed model with three types of baseline models: (1) Strong multi-task baselines, including XSTNet [6], STEMM [11], ConST [10], and CMOT [12],

which significantly improve ST performance in a multi-task learning mode; (2) JT-S-MT [5] and S-Align-ST [13], which discuss the interaction between ST tasks and auxiliary tasks in the multi-task learning mode; (3) CCSRD [9] and SRPSE [14], which focus on processing non-content information in speech and enhance ST performance through decoupling and purification techniques. To ensure a fair comparison with the above models, we adopt the training mode and Transformer-based framework, consistent with these baselines.

3.3. Main results

Comparison with Baselines The main results on the MuST-C and CoVoST-2 datasets are shown in Table 1 and 2, respectively. HuBERT-Transformer serves as the multi-task baseline model in the CMOT research. In the multi-task setting, CMSP-ST outperforms HuBERT-Transformer by 2.0 BLEU⁴ and surpasses CMOT, which also uses OT and cross-modal mixup, by 0.4 BLEU. With the introduction of external MT data, our model also slightly outperforms CMOT and achieves performance comparable to SRPSE. These results demonstrate the effectiveness of our method. Notably, the OT we use is the relaxed OT without the window strategy in CMOT. SRPSE has achieved significant improvements in speech purification performance by enhancing the supervision of non-content information and minimizing mutual information between purified speech representations and non-content information representations. We believe that the simple orthogonal projection method alone can obtain encouraging performance gains, so we do not incorporate these additional enhancement strategies.

Multilingual Experiments To verify the applicability of the model in a multilingual and diverse dataset environment, we selected the two most common language pairs from the CoVoST-2 dataset, including Fr-En and De-En, and employed multi-task learning to train the model. Additionally, we selected several baseline models that have been evaluated on the CoVoST-2 dataset for comparison. 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.

3.4. Ablation studies

To evaluate the contribution of each method, we gradually eliminated them, and the results are shown in Table 3. First, we removed the adversarial training process under S-Align, and the BLEU score drops by 0.4, indicating that this training strategy contributes to improving ST performance under cross-modal mixup method. When the cross-modal mixup method is removed, the BLEU score drops by 0.4, indicating that the token-level mixup strategy using the OT method positively impacts model performance, which is consistent with the findings of the CMOT study. After further removal of the data augmentation operations on the speech and text sequences, the BLEU score drops slightly by 0.1, indicating that data augmentation plays a certain but not obvious role in improving the robustness of the model. Finally, when the orthogonal projection strategy is removed, the BLEU score drops significantly by 1.1, proving the significant effectiveness of the speech purification method in improving translation performance.

²<https://github.com/facebookresearch/fairseq>

³<https://github.com/facebookresearch/fairseq/tree/main/examples/hubert>

⁴the inference time per sample also decreased (10.54ms→10.38 ms)

Table 1: Main results on MuST-C En-De tst-COMMON set. "Speech Pretraining" denotes using pretrained speech models.

Models	Speech Pretraining	BLEU	
		Multi-tasks	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

Table 2: Main results on CoVoST-2 Fr-En and De-En test set.

Models	Speech Pretraining	BLEU	
		Fr-En	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

Table 3: Ablation results on MuST-C En-De tst-COMMON set.

Models	BLEU
CMSP-ST _{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

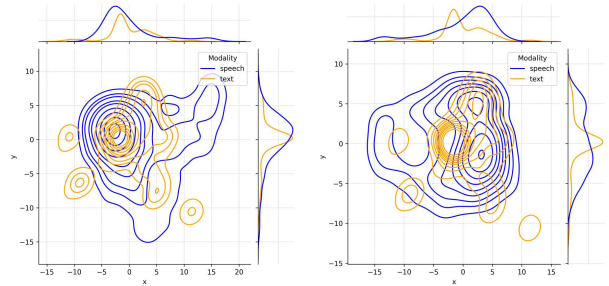
4. Analysis

4.1. Effect of speech purification

We extracted representations of speech and text modalities from the encoder of the MuST-C En-De dev set and used principal component analysis (PCA) to reduce the dimensionality of these features to two dimensions. The experimental results are shown in Figure 4. Compared with cross-modal mixup without speech purification, the representations of speech and text modalities show a larger overlap area after applying the method, indicating that the speech and text modalities are more aligned in their representation. This result demonstrates that speech purification can effectively generate content-based purified speech representations, thereby improving the modality alignment.

4.2. Effect of hard and soft alignment on mixup

To evaluate the performance improvement of S-Align compared with H-Align on cross-modal mixup, we also evaluated the performance of H-Align on the MuST-C En-De tst-COMMON set. The results, as shown in Table 4, indicate that although ST performance improves with the application of H-Align, the improvement is modest and comparable to the performance gains



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

Figure 4: Effects of speech purification on representations.

observed with non-adversarial training under S-Align. This also suggests that H-Align may limit the effectiveness of adversarial training.

Table 4: Two alignment results.

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

5. Conclusion

In this paper, we propose CMSP-ST, a cross-modal mixup with speech purification for speech translation. The model obtains purified speech representations by introducing an additional speech purification module and performs token-level mixup of different modal representations accordingly. In addition, we further enhance the effectiveness of modal mixup by introducing adversarial training under S-Align. Our experimental results on the MuST-C En-De, CovoST-2 Fr-En, and CovoST-2 De-En benchmarks verify the effectiveness of our methods. The combination of speech purification and cross-modal mixup also offers new insights for alleviating the modality gap, though currently the purification is limited to speech. We will explore text purification as well as more fine-grained speech purification, and combine them with cross-modal fusion in the future.

6. Acknowledgements

We thank anonymous reviewers for their insightful comments. This work is supported by the Key Program of Natural Science Foundation of China (Grant No. U23A20316).

7. References

- [1] A. Bérard, O. Pietquin, L. Besacier, and C. Servan, “Listen and translate: A proof of concept for end-to-end speech-to-text translation,” in *Proc. of NIPS Workshop on end-to-end learning for speech and audio processing*, 2016.
- [2] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, “An attentional model for speech translation without transcription,” in *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 949–959.
- [3] J. Pino, Q. Xu, X. Ma, M. J. Dousti, and Y. Tang, “Self-training for end-to-end speech translation,” in *Proc. of Interspeech 2020*, 2020, pp. 1476–1480.
- [4] Y. Tang, H. Gong, N. Dong, C. Wang, W.-N. Hsu, J. Gu, A. Baevski, X. Li, A. Mohamed, M. Auli *et al.*, “Unified speech-text pre-training for speech translation and recognition,” in *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 1488–1499.
- [5] Y. Tang, J. Pino, X. Li, C. Wang, and D. Genzel, “Improving speech translation by understanding and learning from the auxiliary text translation task,” in *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 4252–4261.
- [6] R. Ye, M. Wang, and L. Li, “End-to-end speech translation via cross-modal progressive training,” *arXiv preprint arXiv:2104.10380*, 2021.
- [7] Y. Liu, H. Xiong, J. Zhang, Z. He, H. Wu, H. Wang, and C. Zong, “End-to-end speech translation with knowledge distillation,” *Proc. of Interspeech 2019*, 2019.
- [8] Y. Lei, Z. Xue, X. Zhao, H. Sun, S. Zhu, X. Lin, and D. Xiong, “Ckdst: Comprehensively and effectively distill knowledge from machine translation to end-to-end speech translation,” in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 3123–3137.
- [9] X. Zhao, H. Sun, Y. Lei, S. Zhu, and D. Xiong, “Ccsrd: Content-centric speech representation disentanglement learning for end-to-end speech translation,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 5920–5932.
- [10] R. Ye, M. Wang, and L. Li, “Cross-modal contrastive learning for speech translation,” in *Proc. of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 5099–5113.
- [11] Q. Fang, R. Ye, L. Li, Y. Feng, and M. Wang, “Stemm: Self-learning with speech-text manifold mixup for speech translation,” in *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 7050–7062.
- [12] Y. Zhou, Q. Fang, and Y. Feng, “Cmot: Cross-modal mixup via optimal transport for speech translation,” in *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 7873–7887.
- [13] Y. Zhang, K. Kou, B. Li, C. Xu, C. Zhang, T. Xiao, and J. Zhu, “Soft alignment of modality space for end-to-end speech translation,” in *Proc. of ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 041–11 045.
- [14] C. Zhang, Y. Zhou, R. Zhao, Y. Chen, and X. Shi, “Representation purification for end-to-end speech translation,” in *Proc. of the 31st International Conference on Computational Linguistics*, 2025, pp. 6255–6269.
- [15] Q. Qin, W. Hu, and B. Liu, “Feature projection for improved text classification,” in *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8161–8171.
- [16] A. Vaswani, “Attention is all you need,” *Advances in Neural Information Processing Systems*, 2017.
- [17] Y. Zhang, C. Xu, B. Li, H. Chen, T. Xiao, C. Zhang, and J. Zhu, “Rethinking and improving multi-task learning for end-to-end speech translation,” in *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 10753–10765.
- [18] G. Peyré, M. Cuturi *et al.*, “Computational optimal transport: With applications to data science,” *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.
- [19] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, “Must-c: a multilingual speech translation corpus,” in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 2012–2017.
- [20] C. Wang, A. Wu, and J. Pino, “Covost 2 and massively multilingual speech-to-text translation,” *arXiv preprint arXiv:2007.10310*, 2020.
- [21] T. Kudo, “Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” in *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 66–71.
- [22] M. Ott, “fairseq: A fast, extensible toolkit for sequence modeling,” in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 48–53.
- [23] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, “Fairseq s2t: Fast speech-to-text modeling with fairseq,” in *Proc. of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, 2020, pp. 33–39.
- [24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, 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.
- [25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *Proc. of 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [26] M. Post, “A call for clarity in reporting bleu scores,” in *Proc. of the Third Conference on Machine Translation: Research Papers*, 2018, pp. 186–191.
- [27] B. Zhang, B. Haddow, and R. Sennrich, “Revisiting end-to-end speech-to-text translation from scratch,” in *Proc. of International conference on machine learning*. PMLR, 2022, pp. 26 193–26 205.
- [28] P.-H. Le, H. Gong, C. Wang, J. Pino, B. Lecouteux, and D. Schwab, “Pre-training for speech translation: Ctc meets optimal transport,” in *Proc. of International Conference on Machine Learning*. PMLR, 2023, pp. 18 667–18 685.
- [29] D. Zhang, R. Ye, T. Ko, M. Wang, and Y. Zhou, “Dub: Discrete unit back-translation for speech translation,” in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 7147–7164.