Enhancing End-to-End Speech-to-Speech Translation via Semantic Representation Learning with Cross-Attentive Regularization

Anonymous ACL submission

Abstract

Multitask learning has been widely explored to improve end-to-end speech-to-speech translation (S2ST) systems, typically by incorporating auxiliary speech-based tasks such as automatic speech recognition (ASR) and speech-to-text translation (S2T). However, these tasks provide only indirect semantic supervision and may introduce noise due to acoustic variability. In this work, we propose a semantically enhanced multitask framework that introduces a text-to-unit (T2U) auxiliary task to provide explicit text-level supervision. To further bridge the modality gap between speech and text, we employ Cross-Attentive Regularization (CAR), an attention-based loss that encourages alignment between speech and text encoder representations. We also adopt a teacher-student training strategy where a pretrained T2U model serves as a fixed semantic teacher to guide the speech encoder. Experiments on the CVSS-C corpus show that our method consistently improves over a basic S2UT baseline, achieving BLEU gains of +2.0 (Fr–En), +3.8 (Es–En), and +2.4 (De–Ee), along with substantial improvements in semantic similarity as measured by Sentence-BERT. Additional experiments under low-resource conditions and with alternative encoders (e.g., Branchformer) further validate the generalizability of our approach.[1]

1 Introduction

Speech-to-speech translation (S2ST) aims to directly translate spoken utterances from a source language into spoken utterances in a target language. Traditional S2ST systems are typically constructed through a cascade of automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) synthesis modules, or alternatively, through a combination of speech translation (ST) and TTS. Although these pipeline systems benefit from modularity, they suffer from error propagation between components and lack end-to-end optimization.

To address these challenges, recent research has shifted toward end-to-end S2ST models, which directly map source speech to target speech without relying on intermediate text representations. Among these, the speech-to-mel paradigm has been extensively studied. Translatotron introduced the first direct S2ST model using an attention-based encoder-decoder architecture to translate source mel spectrograms into target mel spectrograms, while jointly predicting phoneme sequences in both source and target languages to improve translation quality(Jia et al., 2019). Subsequent improvements, such as Translatotron 2 and Translatotron 3, further refined the modeling of speaker identity and prosody(Jia et al., 2022a; Nachmani et al., 2024).

In parallel, speech-to-unit translation has emerged as a promising alternative. These methods leverage self-supervised pre-trained models to extract speech representations from the target language, followed by clustering (e.g., via k-means) to obtain discrete acoustic units. The S2ST task is then reformulated as a two-stage pipeline: translating source speech into target discrete units, which are subsequently synthesized into waveforms using a unit vocoder(Lee et al., 2021a). Variants of this framework explore different enhancements, such as normalizing target units or employing stronger acoustic feature extractors(Lee et al., 2021b).

Despite these advances, both speech-to-mel and speech-to-unit systems rely heavily on multitask learning (MTL) with auxiliary tasks such as ASR and ST to improve performance. This is because directly learning the mapping from source speech to target speech is challenging, given the noisy and redundant nature of speech signals, which encode not only semantic information, but also speaker characteristics, acoustic variations, and environmental noise. However, existing MTL approaches primar-

[1] Audio samples are available at: https://cars2ut.github.io/cars2ut/

ily leverage auxiliary information from the speech mode. One notable exception is the UnitY model, which introduces text as an intermediate modality by decomposing S2ST into a speech-to-text stage followed by a text-to-unit translation stage, inserting a dedicated text-to-unit encoder(Inaguma et al., 2022). Although this approach successfully leverages rich semantic information, it introduces additional inference complexity, since the text-to-unit encoder must be executed during decoding.

In this work, we propose a novel multitask learning framework for S2ST that leverages text-to-unit translation as an auxiliary task, enabling explicit semantic-level supervision without introducing additional overhead during inference. Specifically, we integrate a separately pre-trained text-to-unit (T2U) encoder, whose parameters are frozen during joint training, to guide the model toward semantically richer acoustic representations. To further bridge the representational gap between speech and text modalities, we introduce a Cross-Attentive Regularization (CAR) mechanism, a bidirectional attention-based alignment loss that encourages the speech encoder to produce representations closely aligned with their text based counterparts.

The key contributions of this paper are as follows:

- We propose a novel multitask framework for end-to-end S2ST that explicitly introduces semantic-level supervision via text-to-unit translation, complementing traditional speech-based auxiliary tasks.
- We utilize CAR, an attention-based regularization mechanism designed to explicitly align hidden representations from the speech and text encoders, significantly reducing their representational discrepancy.
- We demonstrate effective knowledge transfer from a pretrained text-to-unit model, serving as a frozen semantic teacher, to the speech encoder, improving semantic fidelity without increasing inference-time computational costs.

2 Related Work

Early end to end S2ST systems typically adopted attention-based encoder decoder architectures, where the model directly mapped source speech to target speech. To improve the encoder's ability to capture linguistic content, auxiliary tasks such as source phoneme recognition and source to target phoneme prediction were incorporated through multitask learning(Jia et al., 2019).

Lee et al. proposed S2UT, a transformer based encoder decoder model that outputs discrete acoustic units instead of mel spectrograms. This design improves attention learning by providing more stable and symbolic targets(Lee et al., 2021a). S2UT also incorporates auxiliary tasks such as source speech to source character, source speech to target character, and source speech to target text prediction within a multitask learning framework. To reduce variability in unit outputs across different speakers, unit normalization techniques were introduced(Lee et al., 2021b). Separately, Huang et al. proposed TransSpeech, which addresses input-side variability by applying bilateral perturbation to extract more speaker-invariant features(Huang et al., 2022). In addition, non-autoregressive (NAR) decoding methods have been explored to accelerate inference in S2UT-style models.

Inaguma et al. introduced UnitY, a two-pass framework combining speech to text and text to unit translation. By injecting text as an intermediate representation, it leveraged rich semantic information and improved translation accuracy. However, this also increased inference cost, since the intermediate text must be generated even when not part of the final output(Inaguma et al., 2022). Other approaches replace the speech encoder with self-supervised models like wav2vec 2.0, yielding stronger acoustic representations and better overall performance(Popuri et al., 2022).

Despite these advances, most existing multitask learning methods rely solely on speech inputs, overlooking the potential benefits of directly incorporating textual information. Although text representations inherently provide cleaner and more explicit semantic guidance, naively integrating them as auxiliary inputs may confuse the model due to modality differences. UnitY leveraged textual information, but only as an intermediate representation rather than a direct semantic learning signal. To more effectively integrate textual supervision, Tang et al. proposed Cross-Attentive Regularization, a method that explicitly aligns hidden representations from speech and text encoders to bridge the modality gap and enhance semantic representation learning and get the state-of-the-art result in speech translation(Tang et al., 2021).

Inspired by the success of CAR, we propose incorporating text-to-unit translation as an explicit

auxiliary task during training, directly injecting textual semantic knowledge into multitask S2ST. Unlike UnitY, our method employs text only during the training phase, thereby incurring no additional computational overhead at inference. Furthermore, we utilize CAR to explicitly align speech and text encoder representations, complemented by a teacher-student knowledge transfer scheme. In this scheme, a pre-trained text-to-unit encoder serves as a semantic teacher, providing effective guidance for learning speech representations.

3 Methods

3.1 Overview

Figure 1 presents an overview of our proposed framework for end-to-end speech-to-speech translation (S2ST), which follows the S2UT framework proposed in (Lee et al., 2021a), where a speech encoder and a unit decoder are trained to map source speech into discrete target units. On top of this base structure, we introduce several auxiliary tasks to support acoustic and semantic learning. Specifically, we add two auxiliary decoders that take intermediate outputs from the speech encoder to perform source-to-source and source-to-target character prediction. In addition, a CTC-based speech-to-text objective is applied to an intermediate layer of the unit decoder, these components form the basic S2UT backbone.

To further incorporate semantic information, we introduce a T2U translation branch and a CAR mechanism. The T2U branch supervises the unit decoder with representations from a pre-trained text encoder, while CAR explicitly aligns hidden representations from the speech and text encoders through attention-based reconstruction. These components are designed to encourage the speech encoder to learn semantically meaningful representations. All modules are jointly trained during the training phase. At inference time, however, only the speech encoder and unit decoder are used, ensuring that the proposed enhancements do not introduce any additional computational overhead.

3.2 Text-to-Unit Semantic Supervision

The text encoder and unit decoder are initialized from a separately pre-trained T2U model, and the encoder remains trainable to adapt to the multitask setting. This branch reinforces semantic grounding in the unit decoder by complementing speech-based inputs with textual representations.

In parallel, we adopt a teacher–student strategy in which the T2U branch serves as a semantic teacher for the speech encoder. During Cross-Attentive Regularization, gradients are blocked from the text encoder to prevent interference, allowing it to act as a fixed reference. This regularization encourages the speech encoder to align with the semantics encoded by the text encoder, without altering the teacher's parameters.

3.3 Cross-Attentive Regularization

To promote semantic consistency between modalities, we introduce a CAR loss that encourages the speech encoder to produce representations aligned with those of the text encoder. CAR serves to reduce the representation gap between speech and text encoders, enabling more effective cross-modal supervision. The alignment is achieved by reconstructing speech representations conditioned on text features and enforcing similarity to a self-attentive transformation of the text representations.

Let $H_s = (h_1^s, h_2^s, \ldots, h_n^s) \in \mathbb{R}^{n\times d}$, and $H_t = (h_1^t, h_2^t, \ldots, h_m^t) \in \mathbb{R}^{m\times d}$ denote the sequences of hidden vectors from the speech and text encoders, respectively, where n and m are the sequence lengths.

We compute a similarity matrix $S \in \mathbb{R}^{m\times n}$, where each entry measures the cosine similarity between text and speech hidden vectors:

$$S_{i,j} = \frac{(h_i^t)^\top h_j^s}{\|h_i^t\|_2 \, \|h_j^s\|_2}, \quad \forall\, i \in [1,m],\ j \in [1,n] \tag{1}$$

Row-wise softmax is applied to obtain attention weights from the text to the speech modality:

$$\tilde{S}_{i,j} = \frac{\exp(S_{i,j})}{\sum_{j'=1}^{n} \exp(S_{i,j'})} \tag{2}$$

These weights are used to reconstruct speech-side representations:

$$\hat{H}_s = \tilde{S} \cdot H_s \in \mathbb{R}^{m\times d} \tag{3}$$

To further enhance semantic guidance, we apply a self-attention operation over the text representations. Specifically, we compute:

$$A = \frac{H_t^\top H_t}{\|H_t\|_2} \in \mathbb{R}^{d\times d} \tag{4}$$

This transformation is applied to the text encoder output to obtain refined representations:

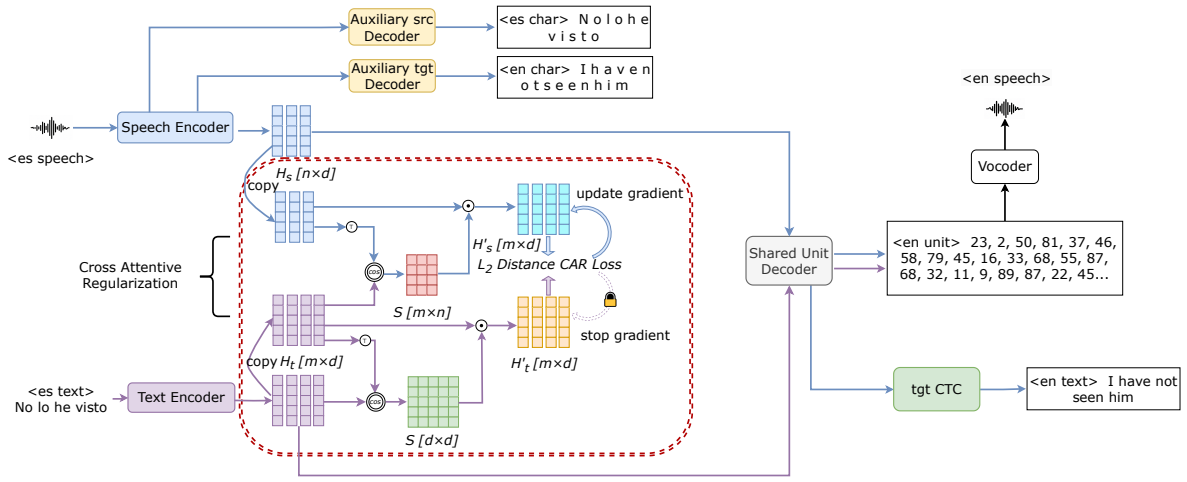

Figure 1: Overall architecture of the proposed S2ST framework with text-to-unit supervision and cross-attentive regularization

$$\tilde{H}_t = H_t \cdot A \in \mathbb{R}^{m \times d} \tag{5}$$

The CAR loss is then defined as the average L2 distance between the reconstructed speech representation and the transformed text representation:

$$\mathcal{L}_{\text{CAR}} = \frac{1}{m} \left\| \hat{H}_s - \text{sg}(\tilde{H}_t) \right\|_2^2 \tag{6}$$

Here, $\text{sg}(\cdot)$ denotes the stop-gradient operation, which prevents gradients from flowing into the text encoder. This allows the text side to act as a fixed semantic teacher, guiding the speech encoder without being updated.

3.4 Training Objectives

The model is trained using a combination of the main speech-to-unit translation loss and several auxiliary objectives. The overall loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{S2UT}} + \lambda_2 \mathcal{L}_{\text{SRC-CHAR}} + \lambda_3 \mathcal{L}_{\text{TGT-CHAR}} \\ + \lambda_4 \mathcal{L}_{\text{ST}} + \lambda_5 \mathcal{L}_{\text{T2U}} + \lambda_6 \mathcal{L}_{\text{CAR}} \end{aligned} \tag{7}$$

Here, $\mathcal{L}_{\text{S2UT}}$ denotes the cross-entropy loss for the main speech-to-unit translation task. $\mathcal{L}_{\text{SRC-CHAR}}$ and $\mathcal{L}_{\text{TGT-CHAR}}$ are computed using two auxiliary decoders, each predicting source and target language characters from the intermediate speech encoder outputs via a standard cross-entropy objective. $\mathcal{L}_{\text{ST}}$ is a CTC loss applied to the intermediate layer of the shared unit decoder to predict target text. $\mathcal{L}_{\text{T2U}}$ supervises the T2U loss, and $\mathcal{L}_{\text{CAR}}$ is the cross attentive regularization loss.

The weights λ_1 through λ_6 control the relative contributions of each objective and are predefined hyper-parameters.

4 Experimental Setup

4.1 Datasets

All experiments are conducted on the CVSS-C corpus (Jia et al., 2022b), which provides multilingual speech-to-speech translation data. We use French-English (Fr-En), Spanish-English (Es-En), and German-English (De-En) as our main evaluation language pairs. To evaluate the model's generalization in limited resource settings, we also conduct experiments on Fr-En using 30% and 10% of the training data, as well as on two additional language pairs with relatively low-resource conditions: Italian-English (It-En) and Russian-English (Ru-En). For the T2U pretraining, we only use the source side text from the corresponding CVSS-C corpus of each language. No additional data or multilingual corpus is used. Details of data preprocessing and splits are provided in Appendix A.

4.2 Model Configuration and Training

Our base model follows the S2UT framework (Lee et al., 2021a), consisting of a Transformer encoder and a Transformer-based unit decoder. We also add auxilary speech to character and text translation tasks. Following Lee et al. (2021a), we use Pretrained HuBERT-base (6-layer) representations followed by KMeans clustering (with $K = 100$) to extract discrete units from target speech. During inference, the predicted unit sequences are converted back to waveform audio using a unit-based

Table 1: BLEU and Sentence-BERT scores on Fr-En, Es-En, and De-En. The ground truth BLEU is computed between the ASR transcription of the reference audio and the reference text. Cascade system uses separate ASR, MT, and TTS modules. All other systems are end-to-end S2UT-based models.

Method	BLEU-FR	BLEU-ES	BLEU-DE	SBERT-FR	SBERT-ES	SBERT-DE
S2UT Baseline(Lee et al., 2021a)	23.09	14.43	13.11	0.659	0.566	0.516
+ T2U	22.16	13.99	12.55	0.654	0.540	0.512
+ Pretrained T2U	23.86	16.04	13.80	0.672	0.613	0.548
+ T2U + CAR	23.51	14.07	12.67	0.644	0.548	0.510
+ Pretrained T2U + CAR	**25.11**	**18.22**	**15.51**	**0.699**	**0.634**	**0.575**
Cascade (ASR+MT+TTS)	33.36	31.21	32.45	0.768	0.744	0.775
Ground Truth	84.60	88.60	88.40	–	–	–

HiFi-GAN vocoder(Polyak et al., 2021).

We conduct three groups of experiments to evaluate the effectiveness and generality of our proposed method. First, we investigate the contribution of each component: T2U multitask learning, CAR, and T2U pretraining through ablation comparisons. Second, we evaluate the model under limited resource settings, using only 30% and 10% of the Fr-En training data, as well as two additional language pairs (It-En and Ru-En) with relatively scarce paired speech. Lastly, we test the generalization of our method across architectures by replacing the Transformer encoder with a Branchformer (Peng et al., 2022) and observing performance changes with and without CAR. Model hyperparameters, layer configurations, and loss coefficients are summarized in the Appendix A. All our experiments were conducted using the SpeechBrain framework(Ravanelli et al., 2024).

4.3 Evaluation Metrics

We evaluate translation quality by following the procedure introduced in Lee et al. (2021a), where an automatic speech recognition (ASR) model is applied to the generated speech output. BLEU scores are then computed between the ASR-transcribed text and the corresponding reference translations. We report corpus-level BLEU using the `sacreBLEU` toolkit with default settings to ensure reproducibility (Post, 2018). For ASR, we adopt a publicly available English Transformer-based model that achieves a word error rate of 2.27% on the LibriSpeech test-clean set.

To assess semantic preservation beyond surface-level lexical similarity, we also report Sentence-BERT similarity (Reimers and Gurevych, 2019) between the transcribed hypothesis and the reference. This embedding-based metric provides a more robust measure of semantic fidelity in generation.

We compare our approach against basic direct S2UT baseline without text-based supervision proposed by (Lee et al., 2021a) and a ASR (Conneau et al., 2021) + MT (Devlin et al., 2019) + TTS (Pratap et al., 2024) cascade system.

5 Results

5.1 Effectiveness of Text-to-Unit and CAR Supervision

Table 1 presents the BLEU and Sentence-BERT scores for three language pairs (Fr-En, Es-En, and De-En). We can find that our best configuration combining pre-trained T2U translation with CAR achieves consistent improvements across all languages. Compared to the S2UT baseline, this setup yields gains of up to +4.1 BLEU (on Es-En) and improves semantic similarity by +0.06 to +0.08 on Sentence-BERT, demonstrating the effectiveness of incorporating stable semantic supervision.

However, these gains are not observed uniformly across configurations. Simply adding the T2U task without pretraining slightly degrades BLEU while offering only marginal improvement in semantic similarity. This is likely due to decoder over-reliance on the text encoder, which shares parameters with the speech-to-unit decoder. Since the text encoder provides easier input during early training, the decoder may prioritize this path to minimize loss, while under-utilizing the speech encoder. At inference time, when the text encoder is not present, this discrepancy results in degraded generation quality.

Pretraining the T2U branch mitigates this issue by providing a more stable decoder initialization and better text–unit alignment. In this case, the decoder is less incentivized to overfit to the text path, and begins to learn more effectively from speech inputs. The result is a small BLEU improvement

Table 2: Qualitative examples comparing model outputs. Errors such as repetition and malformed expressions are shown in bold.

Source:	Es la máxima autoridad administrativa en la materia electoral en la República de Chile
Target:	In the Republic of Chile it is the highest administrative authority in electoral matters
S2UT:	**She** is the **adominization material** and the electoral material at the Chile
+ T2U:	It is the **dynast rate of administrative** in the material of Chile
+ Pretrained T2U:	It is the **dominated material** and the electoral material at the Chile
+ CAR:	It is the **dominant territory** in the electoral matter of Chile
+ PreT2U + CAR:	It is the **largest** administrative authority in the electoral material at the Chile
Source:	Durante este tiempo su principal objetivo era informar de las actividades eclesiásticas del reino
Target:	During this time the main goal was to inform all the ecclesiastical activities of the kingdom
S2UT:	During this time their main **ejective** was inform all **the reign attivities** of the kingdoms
+ T2U:	During this time his main **ejective mane** was an **ecclesiasticities** of the kingdom
+ Pretrained T2U:	During this time his main goal was informed of the **equipment of the kingdom** of the kingdom
+ CAR:	During this time their main objective was the **informal activities of the kingdom** of the kingdom
+ PreT2U + CAR:	During this time his main objective was informed the ecclesiastical activities of the kingdom

(approximately +1 point on average) and a clearer gain in semantic similarity. These findings suggest that even without structural changes, semantic pretraining can guide the decoder toward more semantically meaningful representations.

When applying CAR without pretraining, performance declines further. Although CAR is designed to align speech and text encoder representations, its effectiveness depends on the quality of the text encoder as a semantic reference. If the text encoder is not well-trained, it may serve as a poor teacher and introduce noise into the alignment objective. In contrast, combining CAR with a pretrained text encoder significantly improves both BLEU and semantic similarity. The speech encoder benefits from the additional regularization, learning to produce representations more consistent with semantically grounded text features.

Qualitative Examples. To further validate the semantic improvements of our method, we present several translation examples from the Es-En test set in Table 2. The first example presents a case where the output yields a low BLEU score yet preserves the full semantic content of the reference. This highlights our method's capacity to recover meaning even when surface forms diverge. The second example involves lexical errors in baseline outputs, such as repeat or misinterpreted terms. In contrast, our approach generates fluent and semantically faithful translations that more accurately reflect the source.

5.2 Low-Resource Generalization

To assess the effectiveness of our method in data-scarce scenarios, we conduct experiments on limited-resource conditions using (i) 30% and 10% of the Fr-En training data, and (ii) two additional low-resource language pairs: It-En and Ru-En. Notably, the T2U pretraining for all settings is performed using the full available text data in CVSS-C. Table 3 reports BLEU and Sentence-BERT scores for the S2UT baseline and our proposed PreT2U+CAR configuration. In the Fr-En setting, where the T2U model is pretrained on the full corpus, we observe consistent improvements under both the 30% and 10% speech training conditions. BLEU increases by +0.93 and +2.52 respectively, while semantic similarity also improves.

Table 3: BLEU and Sentence-BERT scores under low-resource settings. PreT2U is pretrained using the full text available in CVSS-C for each language.

Language Pair	Method	BLEU / SBERT
Fr-En (30%)	S2UT	12.42 / 0.482
	+ PreT2U + CAR	**13.35 / 0.505**
Fr-En (10%)	S2UT	3.14 / 0.243
	+ PreT2U + CAR	**5.66 / 0.266**
Ru-En	S2UT	2.61 / 0.211
	+ PreT2U + CAR	**2.81 / 0.215**
It-En	S2UT	7.28 / 0.377
	+ PreT2U + CAR	**8.43 / 0.411**

For It-En, our method yields modest gains in both BLEU and semantic similarity, indicating that even limited in-domain text can provide useful supervision. In contrast, Ru-En shows little improvement, which we attribute to the extremely limited availability of both paired and monolingual data, leading to an undertrained T2U module that offers insufficient guidance for CAR.

Beyond aggregated scores, we observe that CAR contributes positively to semantic alignment at the sentence level, particularly for shorter or moder-

Table 4: Performance on Es-En using a Branchformer encoder and Transformer decoder.

Encoder	BLEU-ES	SBERT-ES
Branchformer	14.73	0.565
+ PreT2U + CAR	**17.67**	**0.620**

ately long inputs. However, its effectiveness diminishes for longer sentences: while the initial segments are often translated correctly, semantic content tends to degrade toward the middle or end. This suggests that CAR helps the model capture local semantic structure but may struggle with long-range dependencies due to limited contextual modeling. Representative cases illustrating this pattern are available on our demo page.

5.3 Encoder Robustness

To assess the generalizability of our method across different encoder architectures, we replace the Transformer encoder with a Branchformer encoder (Peng et al., 2022), while keeping the Transformer-based unit decoder unchanged. This configuration allows us to isolate the impact of CAR on the encoder side, and verify whether the observed gains are architecture-agnostic.

We conduct this experiment on the Es-En language pair. As shown in Table 4, the baseline Branchformer model achieves 14.73 BLEU and a Sentence-BERT similarity of 0.565. When CAR is applied in conjunction with a pretrained T2U module, performance improves significantly to 17.67 BLEU and 0.620 in semantic similarity.

These results confirm that CAR provides consistent benefits even when the encoder architecture changes, and does not rely on the inductive biases of a specific model.

6 Conclusion

In this work, we proposed a semantically guided multitask framework for speech-to-speech translation by introducing a text-to-unit auxiliary task and a Cross-Attentive Regularization mechanism. Our approach leverages pretrained text encoders to provide stable semantic supervision and encourages alignment between speech and text representations during training. Experiments on the CVSS-C corpus demonstrate consistent improvements over strong multitask baselines in both lexical accuracy and semantic similarity. Further evaluations under low-resource conditions and with alternative encoders confirm the robustness and generalizability of our method. These findings highlight the importance of semantic-level supervision in improving the quality and stability of direct speech-to-speech translation systems.

Limitations

While our proposed framework achieves consistent improvements in both lexical and semantic metrics, several limitations remain. First, the effectiveness of our method in extremely low-resource settings is limited. When the available text data is insufficient, the pretrained T2U model fails to provide reliable semantic supervision, thereby diminishing the benefit of CAR. Second, the success of cross-attentive regularization depends heavily on the quality of the pretrained text encoder. In scenarios where the text encoder is not well trained, CAR may misguide the speech encoder and harm performance. Finally, the multitask training setup introduces additional model complexity and requires careful balancing of multiple loss components. This increases the cost of hyperparameter tuning and may limit scalability in practical deployments.

References

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Rongjie Huang, Jinglin Liu, Huadai Liu, Yi Ren, Lichao Zhang, Jinzheng He, and Zhou Zhao. 2022. Transpeech: Speech-to-speech translation with bilateral perturbation. *arXiv preprint arXiv:2205.12523*.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. 2022. Unity: Two-pass direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2212.08055*.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022a. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International Conference on Machine Learning*, pages 10120–10134. PMLR.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022b. Cvss corpus and massively multilingual speech-to-speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6691–6703.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, and 1 others. 2021a. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and 1 others. 2021b. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*.

Eliya Nachmani, Alon Levkovitch, Yifan Ding, Chulayuth Asawaroengchai, Heiga Zen, and Michelle Tadmor Ramanovich. 2024. Translatotron 3: Speech to speech translation with monolingual data. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10686–10690. IEEE.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. In *International Conference on Machine Learning*, pages 17627–17643. PMLR.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. In *INTERSPEECH 2021-Annual Conference of the International Speech Communication Association*.

Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. 2022. Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. *arXiv preprint arXiv:2204.02967*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, and 1 others. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, and 14 others. 2024. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021. Improving speech translation by understanding and learning from the auxiliary text translation task. *arXiv preprint arXiv:2107.05782*.

A Appendix

Table 5: Training data statistics for each language pair

Language Pair	Source (h)	Target (h)	Utterances
Fr–En	264.3	174.0	207,364
Es–En	113.1	69.5	79,012
De–En	184.3	112.4	127,822
Fr–En (30%)	79.9	52.8	62,209
Fr–En (10%)	26.1	17.2	20,736
It–En	44.2	29.4	31,698
Ru–En	18.2	13.3	12,122

Table 6: Model architecture, training hyperparameters, and loss weighting coefficients.

Module	Setting
Speech Encoder layers	12
Aux. Src Char Encoder layers	6
Aux. Tgt Char Encoder layers	8
Aux. Char Decoder layers	2
Text Encoder layers	6
Unit Decoder layers	12
Aux. ST Decoder layers	3
Attention Heads	8
Hidden Size (d_{model})	512
Dropout	0.1
Learning Rate	0.0005
Optimizer	AdamW
Loss Coefficients	$\lambda_1 = 1.0, \lambda_2 = 0.2$ $\lambda_3 = 0.2, \lambda_4 = 0.2$ $\lambda_5 = 0.1, \lambda_6 = 0.1$