Cross-Modal Robustness Transfer (CMRT): Training Robust Speech Translation Models Using Adversarial Text

Anonymous ACL submission

Abstract

End-to-End Speech Translation (E2E-ST), which involves translating directly from speech input to text in the target language, has achieved remarkable progress in recent years. However, previous work has been limited to clean, curated datasets (both for training and evaluation) and overlooks challenges such as robustness to inflectional morphology variation, which are common in non-native and dialectal speech. In this work, we extend an adversarial attack targeting inflectional morphology from text to speech, and show that E2E-ST models are vulnerable to it. While adversarial fine-tuning significantly improves robustness in the text domain, generating adversarial speech data is far more difficult. Therefore, we propose Cross-Modal Robustness Transfer (CMRT), a method for transferring robustness from the text to the speech modality, which alleviates the need for adversarial speech data. Experimental results[1] across four language directions show that our method improves adversarial robustness by an average of more than 3 BLEU points without using any adversarial speech data.

1 Introduction

In recent years, models that perform speech translation in an end-to-end fashion (E2E-ST)—mapping speech input directly to target-language text—have achieved remarkable results, (Dong et al., 2021; Tang et al., 2021; Ye et al., 2022; Fang et al., 2022; Ouyang et al., 2023; Zhou et al., 2023; Communication et al., 2023; Zhang et al., 2024, 2025), often surpassing cascaded solutions where Automatic Speech Recognition (ASR) and Machine Translation (MT) models are combined (Anastasopoulos et al., 2021; Ye et al., 2022; Fang et al., 2022; Ouyang et al., 2023). Unfortunately, the training and evaluation of E2E-ST models is mainly limited to datasets such as MuST-C (Di Gangi et al., 2019) and CoVoST2 (Wang et al., 2020a), which fall short from capturing the full complexity of non-native and dialectal speech especially during spontaneous conversations.

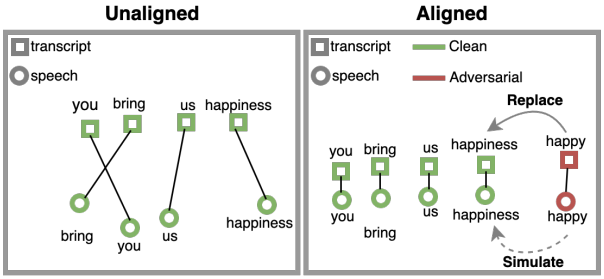

Figure 1: Our method aligns speech and text semantic spaces as shown in the right figure, then, it **replaces** the speech embeddings of the word "happiness" with the text embeddings of its adversarial inflection "happy", which **simulates** replacing with the speech embeddings of the inflection "happy"[2]. This can be used to fine-tune for robustness to inflectional morphology variation.

Non-native speech is characterized by a variety of phenomena such as accent, pronunciation, inflectional morphology, etc. (Moyer, 2011; Gut, 2009; Prévost and White, 2000; WHITE, 2003). Although CoVoST2 dataset contains speech recordings from non-native speakers, they are read from text segments, therefore, it is unlikely that they would capture inflectional morphology or fluency issues that are common in non-native speech (Prévost and White, 2000; WHITE, 2003; Akande, 2013; Adejare, 2019). This has been addressed in the context of MT often by fine-tuning on synthetic adversarial data (Anastasopoulos et al., 2019; Alam and Anastasopoulos, 2020; Tan et al., 2020). Unfortunately, these efforts benefit only cascaded speech translation models, while robustness to inflectional variation in E2E-ST is neglected. We attribute this to the relative difficulty of generating synthetic speech examples compared to the ease of

[1] Our code: https://anonymous.4open.science/r/CMRT-B691

[2] The example *"you bring us happy"* is taken from (Eaves, 2011).

generating synthetic text. Although it is possible to generate adversarial text examples and transform them into speech utterances using a Text-To-Speech (TTS) model, this builds on the assumption that a TTS model is available for the source language, and that its quality of generated speech will not hurt the original model during fine-tuning. Furthermore, running a TTS model on a large enough dataset requires extra resources and time.

In this work, we address these challenges by introducing **Cross-Modal Robustness Transfer (CMRT)**, a method for training robust E2E-ST models using only adversarial text data, therefore, transferring robustness from the text modality to the speech modality. Our hypothesis is that if the speech and text embeddings are aligned on the word level—meaning that the speech and text embeddings of the same word are closely similar—then it should be possible to simulate introducing an adversarial speech word using the text embeddings of the same word, which we can then leverage for robustness fine-tuning. Therefore, our method is composed of two stages: In the first stage, we train an E2E-ST model with an aligned semantic space between speech and text embeddings, building on previous work in E2E-ST (Fang et al., 2022; Ouyang et al., 2023). In the second stage, we fine-tune the model for robustness by introducing adversarial text embeddings into the speech manifold representations. As shown in figure 1, after building a shared semantic space, our method introduces adversarial text embeddings of inflections and trains the model to indirectly become robust to speech inflections.

We demonstrate the effectiveness of our method on handling adversarial inflectional morphology variation in E2E-ST. We introduce inflectional perturbations into text using MORPHEUS (Tan et al., 2020), however, for evaluation, we adapt MORPHEUS to E2E-ST and generate speech inputs with inflectional perturbations. We experiment with 4 language directions (i.e. En-De, En-Ca, En-Ar, Fr-En). According to Ethnologue (Eberhard et al., 2025), the number of non-native speakers (L2) of both English and French is more than 2 times the number of native speakers (L1), which makes them ideal for our experiments. Experimental results demonstrate the effectiveness of our method, where it leads to an average of more than 3 BLEU points improvement on adversarial speech. Compared to fine-tuning directly on adversarial speech data, our method achieves competitive results on the adversarial test set while maintaining better performance on the original CoVoST2 test set.

Our work addresses the following central research question: **Can we transfer robustness from one modality to another by leveraging the semantic alignment between the two modalities?**

To answer this research question we make the following contributions:

- We build on bridging the modality gap works in E2E-ST (Fang et al., 2022; Ouyang et al., 2023) to train a model with a strongly aligned semantic space between speech and text.
- We adapt MORPHEUS adversarial attack to speech and introduce Speech-MORPHEUS where we show that even state-of-the-art techniques and models trained on clean data are vulnerable to inflectional variation.
- We demonstrate that we can significantly improve robustness to one modality using adversarial examples from another in the context of E2E-ST, where we use adversarial text to train robust E2E-ST models.

2 Related Works

Bridging the Modality Gap in E2E-ST aims at alleviating the discrepancy between speech and text modalities in E2E-ST, which has been shown to be a crucial ingredient for improving the performance of E2E-ST models and closing the gap with cascaded models (Liu et al., 2020; Wang et al., 2020b). Various techniques were shown to be effective for this purpose (Liu et al., 2019; Inaguma et al., 2021; Ye et al., 2022; Fang et al., 2022; Ouyang et al., 2023; Zhou et al., 2023; Zhang et al., 2025). Among these techniques, (Ye et al., 2022; Ouyang et al., 2023) directly optimize for this property by using contrastive learning to maximize the similarity between speech and text representations. Fang et al. (2022) point to the fact that the translation model is still taking either speech or text as input, which limits cross-model transfer, therefore, they introduce mixup training (Zhang et al., 2018), where they train on a mixup of speech and text representations. In this work, we show that combining both contrastive learning and mixup training leads to the best alignment between speech and text representation spaces, and we leverage this for robustness fine-tuning.

Robustness in ST previous work on robustness in speech translation has been limited to cascaded sys-

tems, where robustness has been studied either on ASR (Fukuda et al., 2018; Rosenberg et al., 2019; Ravanelli et al., 2020; Chen et al., 2022; Prabhu et al., 2023) or NMT (Li et al., 2018; Martucci et al., 2021; Padfield and Cherry, 2021; Wang et al., 2022; Indurthi et al., 2023) side, while E2E-ST models robustness has been neglected especially robustness to inflectional variation. Fortunately, in the context of NMT, robustness to inflectional variation has been explored in multiple works that target robustness to non-native speech (Anastasopoulos et al., 2019; Tan et al., 2020; Jayanthi and Pratapa, 2021). Our work is inspired by MORPHEUS (Tan et al., 2020), which introduces inflectional perturbations into clean sentences. We extend MORPHEUS to speech, and show that it significantly affects models trained on datasets such as CoVoST2. Furthermore, we introduce a method for improving robustness to inflectional variation in E2E-ST using only adversarial text.

3 Methodology

We begin this section by outlining the model architecture (§3.1). Then, in §3.2 we describe our method, which is composed of two stages: a training stage described in §3.2.1 and §3.2.2, and a fine-tuning stage described in §3.2.3. Finally, in §3.3 we describe our method for extending MORPHEUS to speech. Figure 2 illustrates the overview of our method.

Throughout this work, we denote the speech translation corpus as $D = (s, x, y)$, where s is the raw audio input, x is the transcription, and y is the translation in the target language. The triplets will be used during training, but at inference time the model will take the speech s as input and generate the target translation y.

3.1 Architecture

Similarly to previous work in E2E-ST (Ye et al., 2022; Fang et al., 2022), our model (Figure 2) is composed of a **speech encoder** that takes raw speech frames and transforms them into contextual embeddings, a text **embedding layer** that maps tokens into embeddings, and a **translation encoder-decoder** that takes either the text or speech embeddings and generates the target translation.

3.2 Cross-Modal Robustness Transfer (CMRT)

CMRT consists of two main steps: first, it maximizes the alignment between speech and text semantic spaces (§3.2.1, §3.2.2), and second, it leverages the aligned speech and text embeddings for adversarial fine-tuning using only adversarial text embeddings (§3.2.3).

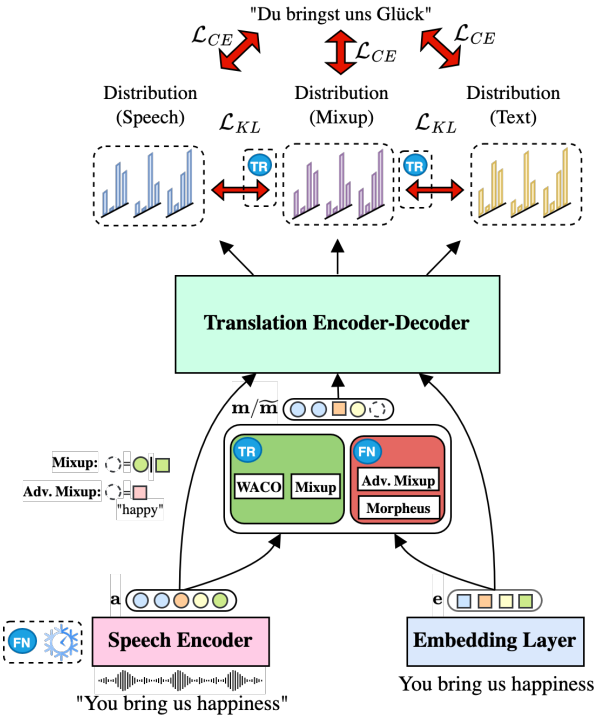

Figure 2: An overview illustration of our proposed method. Since our method is composed of two steps, CMRT-TR (§3.2.1 and §3.2.2) followed by CMRT-FN (§3.2.3), we refer to them as TR and FN respectively in the figure.

3.2.1 Aligning Speech and Text Representations

Our hypothesis is that a strong alignment between speech and text representations is required for robustness to transfer from text to the speech modality, this is motivated by previous work showing that modality alignment allows for cross-modal transfer (Castrejón et al., 2016). Furthermore, we believe that our method can be understood through modal complementarity (Wang et al., 2023), where previous work has shown that speech models are weaker than text models at capturing syntactic features (Shen et al., 2023), and by aligning speech and text representations our method complements the speech modality with textual syntactic features, which previous work has shown to be a crucial component of the robustness of MT models (Issam et al., 2025). Therefore, inspired by previous work on bridging the modality gap between speech and text modalities, we train the speech encoder model to output representations that are aligned

with the text embedding representations. For this purpose, we use Word-Aligned Contrastive Learning (WACO) (Ouyang et al., 2023), where we train the model with contrastive learning to maximize the similarity between the speech and text embeddings of words in the sentence. However, this is not enough as the transformer encoder-decoder still takes either speech or text embeddings as input, which firstly, limits the cross-modal transfer between the modalities, and secondly for our purpose, the translation encoder-decoder has to be accustomed to receiving a mix of speech and text representations. Therefore, we rely on mixup training (Fang et al., 2022), where we sample either speech or text embeddings of words based on a mixup probability.

Both methods require word-level forced alignment between speech and text to identify words. For a sequence of words $w = [w_1, w_2, ..., w_{|w|}]$, a sequence of speech embeddings $a = [a_1, a_2, ..., a_{|a|}]$, and a sequence of text embeddings $e = [e_1, e_2, ..., e_{|e|}]$, an alignment of the word w_i consists of the start position l_i^s and end position r_i^s in the speech sequence, and the start position l_i^t and end position r_i^t in the text sequence. Therefore, the speech embeddings of the word w_i are $a[l_i^s : r_i^s]$ while its text embeddings are $e[l_i^t : r_i^t]$.

For WACO, we derive word level representations by mean pooling the speech and text embeddings of each word as follows:

$$f_i^s = \text{MeanPool}(a[l_i^s : r_i^s]) \tag{1}$$

$$f_i^t = \text{MeanPool}(e[l_i^t : r_i^t]) \tag{2}$$

f_i^s and f_i^t are considered as positive pairs, while other words in the batch are considered negatives. Therefore, the contrastive loss is computed as follows:

$$\ell_{\text{CTR}}(\mathcal{B}) = - \mathop{\mathbb{E}}_{f_i^s, f_i^t \in \mathcal{B}} \left[\log \frac{\exp(\text{sim}(f_i^s, f_i^t)/\tau)}{\sum_{f_{j \neq i}^t \in \mathcal{B}} \exp(\text{sim}(f_i^s, f_j^t)/\tau)} \right] \tag{3}$$

where $\mathcal{B}$ is the current batch, τ is the temperature, and sim is the cosine similarity function.

During mixup training, we similarly leverage alignments. For each word w_i in the sequence, we choose either its speech embeddings or text embeddings based on a mixup probability p^*:

$$m_i = \begin{cases} a[l_i^s : r_i^s] & p < p^* \\ e[l_i^t : r_i^t] & p \geq p^* \end{cases} \tag{4}$$

Where p is sampled from a uniform distribution $\mathcal{U}(0, 1)$.

The final mixup sequence is the result of concatenating all m_i segments:

$$m = Concat(m1, m2, ..., m_{|w|}) \tag{5}$$

3.2.2 Training

Similar to previous work in E2E-ST, we start by pre-training the text embedding layer and the translation encoder-decoder on transcription-translation pairs $D = (x, y)$ for the task of machine translation using cross entropy as follows:

$$\mathcal{L}_{MT} = -\mathbb{E}_{x,y} \log P(y|x) \tag{6}$$

The second involves multi-task learning with the addition of WACO and mixup training as introduced earlier. Below we briefly introduce each training objective.

ST Loss: We train with an ST objective where the speech encoder takes raw speech input s and generates the target translation y. The ST loss is formalized as follows:

$$\mathcal{L}_{ST} = -\mathbb{E}_{s,x,y} \log P(y|s) \tag{7}$$

MT Loss: To maximize cross modal transfer, we train with an MT objective as well. The MT loss is formalized as follows:

$$\mathcal{L}_{MT} = -\mathbb{E}_{s,x,y} \log P(y|x) \tag{8}$$

CTR Loss: We use contrastive learning as formalized in Equation 3. The contrastive loss over the full dataset is formalized as:

$$\mathcal{L}_{CTR} = -\mathbb{E}_{\mathcal{B} \subseteq D_{s,x,y}} [\ell_{\text{CTR}}(\mathcal{B})] \tag{9}$$

Mixup: We train with cross entropy where the translation encoder-decoder is supposed to take the mixup representations m and translate to the target language:

$$\mathcal{L}_M = -\mathbb{E}_{m,y} \log P(y|m) \tag{10}$$

This is an important step for familiarizing the model with translating from mixed representations. Another crucial step for bridging the modality gap is training with symmetric Kullback-Leibler (KL) divergence between the output probability distribution of the mixup m and both the output probability distributions of the ST and MT tasks as follows:

$$\mathcal{L}_{KL_{m \leftrightarrow s}} = \mathbb{D}_{KL}(P(y|s)||P(y|m)) + \mathbb{D}_{KL}(P(y|m)||P(y|s)) \tag{11}$$

$$\mathcal{L}_{KL_{m \leftrightarrow x}} = \mathbb{D}_{KL}(P(y|x)||P(y|m)) + \mathbb{D}_{KL}(P(y|m)||P(y|x)) \tag{12}$$

Therefore, the final loss combines the previous losses:

$$\mathcal{L} = \mathcal{L}_{ST} + \mathcal{L}_{MT} + \lambda_{ctr}.\mathcal{L}_{CTR} + \mathcal{L}_{M} + \lambda_{kl}.(\mathcal{L}_{KL_{s\leftrightarrow m}} + \mathcal{L}_{KL_{x\leftrightarrow m}})/2 \quad (13)$$

where λ_{ctr} is the weight hyper-parameter for the contrastive loss, and λ_{kl} is the weight for KL divergence loss that is part of the mixup training.

We term this stage of training **CMRT-TR**, where we train the model for speech translation with the priority of maximizing speech-text semantic alignment, which is achieved by combining both WACO and mixup training.

3.2.3 Robustness Fine-tuning

This section builds on CMRT-TR to fine-tune the model for robustness to adversarial speech. In a typical adversarial fine-tuning setup, we would need adversarial speech examples $\widetilde{s}$ and their target translations y. However, generating synthetic text examples is much easier than generating speech examples, therefore, our method, which we term **CMRT-FN**, leverages the adversarial text examples $\widetilde{x}$ instead, along with clean speech s. We denote $\widetilde{D} = (s, \widetilde{x}, y)$ the adversarial dataset for training an adversarially robust model using CMRT-FN.

Text adversarial attacks generally do not alter all words in the sentence but only a subset, otherwise, the meaning of the sentence is difficult to preserve. We denote the word indices of the words that were affected by the adversarial attack as $\widetilde{\mathcal{I}}$. For example, for a given sequence $w = [w_1, w_2, w_3, ..., w_{|w|}]$, a replacement adversarial attack that affects $\widetilde{\mathcal{I}} = [2, 3]$ would replace the words w_2 and w_3 with adversarial copies $\widetilde{w_2}$ and $\widetilde{w_3}$ respectively.

Our strategy for fine-tuning CMRT-TR model is to introduce the adversarial text embeddings directly into the aligned speech manifold coming from the speech encoder, therefore, we freeze the speech encoder to maintain its alignment, and we train the rest of the model. Furthermore, we modify the mixup training strategy to use the adversarial embeddings of the inflection $\widetilde{w_i}$ if $i \in \widetilde{\mathcal{I}}$, otherwise, we apply standard mixup between clean speech and text embeddings. We denote the adversarial text embeddings of the word $\widetilde{w_i}$ as $\tilde{e}[\tilde{l}_i^t : \tilde{r}_i^t]$, where $\tilde{l}$ and $\tilde{r}$ are the start and end positions in the sequence of adversarial embeddings $\tilde{e}$. We term this step adversarial mixup:

$$\widetilde{m_i} = \begin{cases} \tilde{e}[\tilde{l}_i^t : \tilde{r}_i^t] & i \in \widetilde{\mathcal{I}} \\ a[l_i^s : r_i^s] & p < p^*, i \notin \widetilde{\mathcal{I}} \\ e[l_i^t : r_i^t] & p \geq p^*, i \notin \widetilde{\mathcal{I}} \end{cases} \quad (14)$$

We use cross entropy to train on adversarial mixup embeddings:

$$\mathcal{L}_{\widetilde{M}} = -\mathbb{E}_{\widetilde{m},y} \log P(y|\widetilde{m}) \quad (15)$$

Finally, we use asymmetric KL divergence to match output probability distribution of adversarial mixup with both the output probability distributions of ST and MT:

$$\mathcal{L}_{KL_{\widetilde{m}\rightarrow s}} = \mathbb{D}_{KL}(P(y|s)||P(y|\widetilde{m})) \quad (16)$$
$$\mathcal{L}_{KL_{\widetilde{m}\rightarrow x}} = \mathbb{D}_{KL}(P(y|x)||(P(y|\widetilde{m})) \quad (17)$$

Note that unlike CMRT-TR where we train with symmetric KL, we use asymmetric KL because the mixup representations are adversarial, therefore, we seek to train the model to handle adversarial noise by making the distribution of adversarial mixup similar to clean input. This part resembles virtual adversarial training (VAT) (Miyato et al., 2017). We term this stage of training CMRT-FN, where we fine-tune the model to be robust to adversarial attacks. The final loss is formalized as:

$$\mathcal{L} = \mathcal{L}_{ST} + \mathcal{L}_{MT} + \mathcal{L}_{\widetilde{M}} + \lambda_{kl}.(\mathcal{L}_{KL_{\widetilde{m}\rightarrow s}} + \mathcal{L}_{KL_{\widetilde{m}\rightarrow x}})/2 \quad (18)$$

3.3 Speech-MORPHEUS

MORPHEUS (Tan et al., 2020) was introduced to simulate inflectional errors that are common in non-native or dialectal speech. It was confirmed through a human evaluation study that it generates sentences that are meaning-preserving, and that it is effective at simulating non-native speakers writing. Although this has been demonstrated in written language, inflectional errors can be more common in spoken language (Weissberg, 2000; Gardner et al., 2021). This motivated our effort to introduce Speech-MORPHEUS, which simply extends MORPHEUS to speech. Similarly to MORPHEUS, we focus on verbs, nouns and adjectives in the sentence, and we find all candidate inflections with a similar Part-Of-Speech (POS) tag. We greedily search over the candidate inflections to find the ones that lead to significantly altering the model output. Given a sequence of words $w = [w_1, w_2, ..., w_{|w|}]$, MORPHEUS replaces each candidate word w_i with each of its possible inflections until finding the set of candidate inflections $\widetilde{w}$.

However, this process is infeasible in the context of ST where the input is continuous, therefore, we target the translation encoder-decoder instead by inflecting the transcriptions, then, we transform them into speech using a TTS model. However, even though this would work on languages such as English, it does not guarantee similar effect on languages with a high number of homophones such as French, where an inflection that might hurt the model in written form would not affect it in spoken form because the word is pronounced similarly to the original word. To address this, we add another step of filtering out candidate inflections that have the same phonemes as the original word.

4 Experiments

4.1 Dataset

We conduct our experiments on CoVoST-2 dataset (Wang et al., 2020a), a large multilingual ST dataset that is based on Common Voice project (Ardila et al., 2020). CoVoST-2 covers translation from 21 source languages to English and from English to 15 target languages. However, we only experiment with 4 language directions, namely En-De, En-Ca, En-Ar, Fr-En. We report the results on the test set.

For adversarial data experiments, we use Speech-MORPHEUS (described in §3.3 and Appendix A) to generate an adversarial copy of 50k examples sampled from the train set, and a copy of the dev and test sets. We report the results on the test set.

4.2 Experimental Setup

Preprocessing We use raw 16 bit 16kHz mono-channel speech input. We filter out examples with a number of speech frames that exceeds 480k or less than 1k. For text, we remove punctuation from the input transcriptions. We use uni-gram SentencePiece model with a vocabulary of 10k that is shared between the source and target.

Since our method requires word level alignments between speech and transcrption, we use NeMo Forced Aligner (NFA) [3] which was shown to achieve state-of-the-art results in terms of alignment accuracy (Rastorgueva et al., 2023).

Model Our model consists of two main components, a speech encoder which is composed of a pre-trained HuBERT model (Hsu et al., 2021) in the case of En-X direction, and mHuBERT (Zanon Boito et al., 2024) in the case of Fr-En, plus two additional 1-dimensional convolution layers of kernel size 5, stride 2, padding 2, and hidden dimension 1024, to help shrink down the output representation over the time axis. The second component is a translation encoder-decoder model (pre-trained on CoVoST2 train set), which has 6 encoder and 6 decoder layers. Each layer is comprised of 512 hidden units, 8 attention heads, and 2048 feed-forward hidden units.

Baselines

MT-Transformer: The translation encoder-decoder with the embedding layer trained with the MT objective (Equation 6).

HuBERT-Transformer: Similar architecture to our model trained with ST objective (Equation 7).

HuBERT-CMOT: Trained using state-of-the-art Cross-modal Mixup via Optimal Transport (CMOT) method (Zhou et al., 2023).

CMRT-TR: Trained using WACO and Mixup as described in §3.2.2.

TTS-Morpheus-FN: HuBERT-Transformer baseline fine-tuned on Speech-MORPHEUS synthetic train set of 50k examples with an ST objective.

CMRT-FN: CMRT-TR baseline fine-tuned with adversarial mixup training as described in §3.2.3.

We provide further details on our experimental setup in Appendix A.

5 Results and Analysis

5.1 Robustness Results

Table 1 shows that CMRT-FN is effective in transferring robustness from text to speech modality. Our method improves robustness by an average of 3.4 BLEU points on the 4 directions compared to HuBERT-Transformer without using any adversarial speech data (and 3.1 compared to CMRT-TR). CMRT-FN still falls short compared to fine-tuning on adversarial speech (an average of 1.4 BLEU points), as shown in the results of TTS-Morpheus-FN, which was fine-tuned on 50k examples from the train set that were adversarially attacked using Speech-MORPHEUS [4]. However, this also leads to a significant drop in performance on the Original test set (i.e. 3.6 from HuBERT-Transformer), while our method witnesses a smaller drop in performance (i.e. 0.6 from CMRT-TR which is the model we fine-tuned in this case). Surprisingly, our method leads to a slightly better robustness on Fr-En, although this is partly due to CMRT-TR's

[3] https://github.com/NVIDIA/NeMo/tree/main/tools/nemo_forced_aligner

[4] We do not fine-tune on the full dataset because generating examples using the TTS model is time and resource consuming. We also consider 50k examples to be a reasonable amount of data for fine-tuning.

	En-De			En-Ca			En-Ar			Fr-En		
	Original	Morpheus	Clean	Original	Morpheus	Clean	Original	Morpheus	Clean	Original	Morpheus	Clean
MT-Transformer	-	13.2	26.7	-	20.2	34.6	-	11.3	18.3	-	21.3	32.3
HuBERT-Transformer	21.4	14.4	25.3	27.4	19.7	32.7	15.7	12.8	19.2	28.4	20.5	27.9
HuBERT-CMOT	**21.8**	14.6	**25.9**	**28.2**	18.4	**33.5**	**16.2**	14.5	**19.5**	**30.9**	22.0	**30.9**
CMRT-TR	20.8	14.4	25.1	26.5	19.8	31.6	14.8	12.7	18.3	29.5	21.9	29.9
TTS-Morpheus-FN (50K)	18.2	**19.9**	22.5	22.9	**26.1**	28.7	13.0	**15.4**	17.0	24.7	25.1	25.5
CMRT-FN (50K)	19.9	17.4	24.0	25.4	23.6	31.0	13.9	14.5	17.3	28.3	24.6	28.6
CMRT-FN	20.3	17.6	24.5	25.8	23.6	31.7	14.4	14.7	17.9	28.8	**25.2**	29.2

Table 1: Results of CMRT-FN against baselines on 4 languages from CoVoST2 dataset. Original represents the original CoVoST2 test set, Morpheus shows the results on Speech-MORPHEUS adversarial copy of the same test set, and Clean shows the results on TTS of clean transcriptions. We underline the model scores that achieved the best result on Morpheus without using adversarial speech. On average, CMRT-FN improves robustness to MORPHEUS adversarial attack by more than 3 BLEU points compared to HuBERT-Transformer without using any adversarial speech data.

performance on the Original data, it might suggest that our method can close the gap with adversarial fine-tuning on other source languages.

To demonstrate that the drop in performance is due to Speech-MORPHEUS attack and not due to the quality of TTS data, Table 1 also shows the results of applying the same TTS model on clean transcriptions, where the performance is even higher than the Original test set because the TTS speech is normalized. The table also shows the performance of the MT-Transformer on clean and adversarial MORPHEUS transcriptions, which we find to be correlated with the effect of Speech-MORPHEUS on HuBERT-Transformer.

CMOT achieves the best performance on Original and Clean test sets, however, when looking at adversarial robustness, this method still witnesses a significant drop in performance (7.4 on average). This demonstrates the fact that even state-of-the-art methods are vulnerable to adversarial attacks when trained on clean data, and our method can be advantageous in terms of robustness. In Appendix B we take this further by fine-tuning a pre-trained state-of-the-art MT model, the results show that although this significantly improves the results, the model still witnesses a drop in performance on Morpheus (an average of 5.2), and CMRT leads to a significant improvement in robustness over the baseline (an average of 4.4) without using any adversarial speech data.

We note that although our method requires two stages of training, the second stage CMRT-FN is only a fine-tuning stage for 1/8th of the training stage [5], therefore, training with CMRT does not double the training time. Moreover, given the benefits of using CMRT, which alleviates the need for a TTS model as well the associated computational costs on a large dataset, the fine-tuning stage is a beneficial and necessary trade-off.

[5] A CMRT-TR job on one H100 took 15:56:03 while CMRT-FN took only 1:46:19. This is significantly better than running the TTS model on GeForce RTX 2080 Ti which took 17:53:01 on En-De train sample of 50k examples.

5.2 Effect of Shared Semantic Space

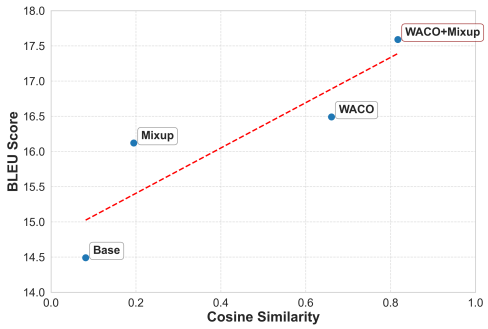

Figure 3: The figure shows the correlation between the cosine similarity of speech and text representations on En-De dev set, and the BLEU score on the adversarial test set. We see a correlation between speech and text alignment and CMRT-FN's effectiveness.

Figure 3 shows that the cosine similarity between speech and text representations strongly correlates with performance on adversarial data after fine-tuning with CMRT. Furthermore, although mixup training alone falls significantly behind WACO in terms of semantic similarity, the difference in BLEU score between the two is not as significant. This suggests that mixup offers a clear advantage: it familiarizes the translation encoder-decoder to mixup (speech-text) inputs, which is crucial for CMRT during the fine-tuning stage. Critically, the figure also shows that our method of combining WACO and Mixup training yields the strongest semantic alignment and the highest robustness to

adversarial inputs. This confirms our hypothesis that a stronger alignment between speech and text representations should allow robustness to transfer from text to speech, and that mixup training is necessary to familiarize the model on mixup inputs.

5.3 Effect of λ_{kl} on CMRT-FN

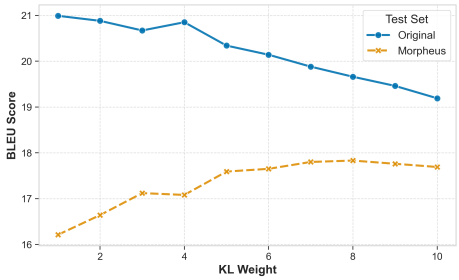

Figure 4: Effect of KL divergence weight λ_{kl} on adversarial Morpheus and original CoVoST2 test sets.

When fine-tuning the model with CMRT-FN, KL divergence serves as a regularization mechanism that penalizes the model for being sensitive to the adversarial input. As formulated in §3.2.3, this is done by minimizing asymmetric KL between the output probability distribution of adversarial mixup with both speech and text output probability distributions. The weight λ_{kl} is used to control the strength of KL as shown in Equation 18. Figure 4 shows the effect of increasing the value of λ_{kl} between 1 and 10 on the robustness to adversarial examples, and its effect on original CoVoST2 test set performance. The figure shows that there is a trade-off between improved performance on Morpheus adversarial subset and performance on the original test set as we increase λ_{kl}. This is a documented trade-off between adversarial robustness and accuracy (Tsipras et al., 2019). Furthermore, increasing λ_{kl} indirectly down-weights the ST loss which means the model is learning less from the original training samples. The figure also shows that adversarial robustness plateaus after $\lambda_{kl} = 8$, suggesting that the method exhausted the amount of adversarial speech robustness it can learn from text. The fact that CMRT-FN plateaus below TTS-Morpheus-FN suggests that the model retains modality specific information that limits cross-modal transfer. We attribute this to the limitations of contrastive learning (Liang et al., 2022), and will research other techniques for minimizing modality specific information in future work.

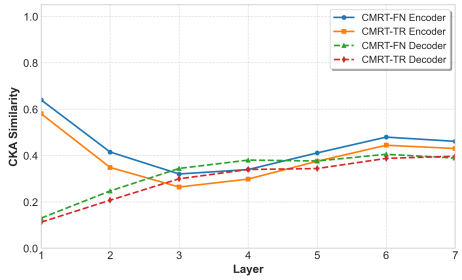

Figure 5: We compare the CKA similarity of representations of adversarial sentences of CMRT-TR and CMRT-FN models against TTS-Morpheus-FN. CMRT-FN model is more aligned with TTS-Morpheus-FN because both are trained to handle Morpheus errors.

5.4 Layer-wise Similarity Analysis

CMRT-FN behaves similarly to fine-tuning which is robustness to inflectional morphology variation, but are there similarities on the representation level as well? To answer this, we measure how the model trained with CMRT represents adversarial speech examples compared to the model that is fine-tuned on adversarial speech. We take the mean of the sequence representations of each sentence and use Centered Kernel Alignment (CKA) (Kornblith et al., 2019) to compute the similarity of CMRT-FN and CMRT-TR with TTS-Morpheus-FN on En-De dev set. Although representations are averaged over sentences that contain both clean and adversarial words, the similarity of CMRT-FN is still higher than CMRT-TR. This suggests that CMRT-FN not only matches the behavior of adversarially fine-tuned models but also converges towards similar internal representations, which confirms a main claim of CMRT: robustness can be transferred from text to speech modality, even in the absence of adversarial speech data.

6 Conclusion

In this paper, we introduce Speech-Morpheus, a method for generating adversarial speech inputs targeting inflectional variation and propose Cross-Modal Robustness Transfer (CMRT) to improve E2E-ST robustness using only adversarial text. CMRT improves speech adversarial robustness by over 3 BLEU points across 4 translation directions (without requiring adversarial speech data), which demonstrates that robustness can be effectively transferred across modalities. In future work, we will experiment with the application of Speech-Morpheus and CMRT to other speech-to-text tasks and architectures.

7 Limitations

Our method has two main limitations:

First, it is composed of two stages (i.e. CMRT-TR and CMRT-FN) which can increase the complexity of E2E-ST training, therefore, in future work we will aim at training for both ST and robustness in one step.

Secondly, our first stage of training where we focus on maximizing the alignment between speech and text representations leads to worse results on CoVoST2 test set than HuBERT-Transformer baseline on En-X directions. We will investigate this negative effect of the alignment and explore ways to eliminate it.

References

Roseline Abonego Adejare. 2019. The acquisition and use of the -s plural morpheme by third-year students of english in a nigerian university. *Open Journal of Modern Linguistics*, 9(6):145–160.

Akinmade T. Akande. 2013. Non-standard syntactic features in nigerian university graduates' english. *Awka Journal of English Language and Literary Studies*, 4(1).

Md Mahfuz Ibn Alam and Antonios Anastasopoulos. 2020. Fine-tuning MT systems for robustness to second-language speaker variations. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 149–158, Online. Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. 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 3070–3080, Minneapolis, Minnesota. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. pages 4978–4982.

Lluís Castrejón, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Learning aligned cross-modal representations from weakly aligned data. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2940–2949.

Chen Chen, Nana Hou, Yuchen Hu, Shashank Shirol, and Eng Siong Chng. 2022. Noise-robust speech recognition with 10 minutes unparalleled in-domain data. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4298–4302.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. 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 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021. Consecutive decoding for speech-to-text translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12738–12748.

Megan Eaves. 2011. English, chinglish or china english?: Analysing chinglish, chinese english and china english. *English Today*, 27(4):64–70.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, twenty-eighth edition. SIL International, Dallas, Texas.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. STEMM: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7050–7062, Dublin, Ireland. Association for Computational Linguistics.

Takashi Fukuda, Raul Fernandez, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, Alexander Sorin, and Gakuto Kurata. 2018. Data augmentation improves recognition of foreign accented speech. In *Proc. Interspeech 2018*, pages 2409–2413.

Qingyuan Gardner, Holly P. Branigan, and Vasiliki Chondrogianni. 2021. Spoken and written production of inflectional morphology among l1 mandarin speakers of english. *Journal of Memory and Language*, 120:104250.

U. Gut. 2009. *Non-native Speech: A Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. English corpus linguistics. Peter Lang.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe. 2021. Source and target bidirectional knowledge distillation for end-to-end speech translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1872–1881, Online. Association for Computational Linguistics.

Sathish Indurthi, Shamil Chollampatt, Ravi Agrawal, and Marco Turchi. 2023. CLAD-ST: Contrastive learning with adversarial data for robust speech translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9049–9056, Singapore. Association for Computational Linguistics.

Abderrahmane Issam, Yusuf Can Semerci, Jan Scholtes, and Gerasimos Spanakis. 2025. A representation level analysis of NMT model robustness to grammatical errors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8579–8601, Vienna, Austria. Association for Computational Linguistics.

Sai Muralidhar Jayanthi and Adithya Pratapa. 2021. A study of morphological robustness of neural machine translation. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 49–59, Online. Association for Computational Linguistics.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Xiang Li, Haiyang Xue, Wei Chen, Yang Liu, Yang Feng, and Qun Liu. 2018. Improving the robustness of speech translation. *Preprint*, arXiv:1811.00728.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. pages 1128–1132.

Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2020. Bridging the modality gap for speech-to-text translation. *ArXiv*, abs/2010.14920.

Giuseppe Martucci, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021. Lexical modeling of asr errors for robust speech translation. pages 2282–2286.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*.

Alene Moyer. 2011. Foreign accent: The phenomenon of non-native speech. *Foreign Accent: The Phenomenon of Non-Native Speech*, pages 1–218.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Siqi Ouyang, Rong Ye, and Lei Li. 2023. WACO: Word-aligned contrastive learning for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3891–3907, Toronto, Canada. Association for Computational Linguistics.

Dirk Padfield and Colin Cherry. 2021. Inverted projection for robust speech translation. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 236–244, Bangkok, Thailand (online). Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. ACL '02, page 311–318, USA. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Darshan Prabhu, Preethi Jyothi, Sriram Ganapathy, and Vinit Unni. 2023. Accented speech recognition with accent-specific codebooks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7175–7188, Singapore. Association for Computational Linguistics.

Philippe Prévost and Lydia White. 2000. Missing surface inflection or impairment in second language acquisition? evidence from tense and agreement. *Second Language Research*, 16(2):103–133.

E. Rastorgueva, V. Lavrukhin, and B. Ginsburg. 2023. Nemo forced aligner and its application to word alignment for subtitle generation. In *Proc. Interspeech 2023*, pages 5257–5258.

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. 2020. Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993.

Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. 2019. Speech recognition with augmented synthesized speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 996–1002.

Gaofei Shen, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. 2023. Wave to syntax: Probing spoken language models for syntax. pages 1259–1263.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

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. In *Proceedings 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)*, pages 4252–4261, Online. Association for Computational Linguistics.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*.

Changhan Wang, Anne Wu, and Juan Pino. 2020a. Covost 2: A massively multilingual speech-to-text translation corpus. *Preprint*, arXiv:2007.10310.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020b. Bridging the gap between pretraining and fine-tuning for end-to-end speech translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9161–9168.

Daheng Wang, Tong Zhao, Wenhao Yu, Nitesh V. Chawla, and Meng Jiang. 2023. Deep multimodal complementarity learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10213–10224.

Shuaibo Wang, Yufeng Chen, Songming Zhang, Deyi Xiong, and Jinan Xu. 2022. Adversarially improving NMT robustness to ASR errors with confusion sets. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 221–227, Online only. Association for Computational Linguistics.

Bob Weissberg. 2000. Developmental relationships in the acquisition of english syntax: writing vs. speech. *Learning and Instruction*, 10(1):37–53.

LYDIA WHITE. 2003. Fossilization in steady state l2 grammars: Persistent problems with inflectional morphology. *Bilingualism: Language and Cognition*, 6(2):129–141.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5099–5113, Seattle, United States. Association for Computational Linguistics.

Marcely Zanon Boito, Varun Iyer, Nathanaël Lagos, Laurent Besacier, and Ionut Calapodescu. 2024. mhubert-147: A compact multilingual hubert model. In *Proc. Interspeech 2024*, pages 3939–3943.

Chengwei Zhang, Yue Zhou, Rui Zhao, Yidong Chen, and Xiaodong Shi. 2025. Representation purification for end-to-end speech translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6255–6269, Abu Dhabi, UAE. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Yuhao Zhang, Kaiqi Kou, Bei Li, Chen Xu, Chunliang Zhang, Tong Xiao, and Jingbo Zhu. 2024. Soft alignment of modality space for end-to-end speech translation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11041–11045.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. CMOT: Cross-modal mixup via optimal transport for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7873–7887, Toronto, Canada. Association for Computational Linguistics.

A Experimental Setup Details

Speech-Morpheus Morpheus was introduced initially for English only, although it was extended to other languages in (Jayanthi and Pratapa, 2021), we found the French inflections to be limited to verbs. Therefore, we extend Morpheus to French by using inflecteur [6] to find inflections for French, while we use LemmInflect [7] for English similar to the original work. For tokenization and POS tagging, we use Spacy [8] and Spacy-lefff [9] for English and French respectively. Furthermore, Speech-Morpheus requires two extra tools for phonemization and TTS. For phonemization, we use eSpeak NG [10], while for TTS, we use XTTS-v2 [11] as a state-of-the-art multilingual TTS model (Casanova et al., 2024).

Training Our models are trained in two stages, first we train the translation encoder-decoder on transcription-translation pairs from the CoVoST2 train set. We train with a learning of 1e-4, a maximum of 33k tokens per batch, for a maximum of 100k steps, where we early stop the training if the loss doesn't decrease for 10 epochs on the dev set. In the second stage, we fine-tune the speech encoder and translation encoder-decoder for ST with a learning rate of 1e-4, a maximum of 16M audio frames, for 40k steps. For CMRT-TR, we train with a contrastive loss weight $\lambda_{ctr} = 1.0$, contrastive temperature $\tau = 0.2$, a KL weight $\lambda_{kl} = 2.0$, and a mixup probability $p^* = 0.8$ [12].

Fine-tuning We fine-tune models from the training stage with a learning rate of 1e-4, a maximum of 16M audio frames, for 5k steps. In this stage, the speech encoder is kept frozen while we train the rest of the model. For CMRT-FN, we increase the KL weight to 5.0, and use the same mixup probability $p^* = 0.8$.

The MT models are trained using one A100 GPU and ST models are trained using one H100 GPU. We use Fairseq [13] (Ott et al., 2019) for the implementation.

Inference For models from the training stage, we average the last 10 epoch checkpoints and evaluate, while for the fine-tuning stage, we take the average when the number of epochs exceeds 10 otherwise we evaluate the last checkpoint. We generate with a beam size of 5 and use SacreBLEU (Post, 2018) to compute detokenized case-sensitive BLEU score (Papineni et al., 2002).

B CMRT Improves Robustness of Massively Pre-trained Models

Table 2 shows the results of using No Language Left Behind (NLLB) (NLLB Team et al., 2022) as a massively pre-trained multilingual model for initializing the MT transformer. We specifically use the efficient distilled version [14]. The results demonstrate that although using a massively pre-trained model significantly improves the results, we still see a significant decrease in performance when evaluating on Morpheus (a drop of 5.2 from Original test set on average). This shows that the issue of robustness is inherent in training on clean data, and that fine-tuning on adversarial data can be crucial for improving robustness. More importantly, using a pre-trained model does not hinder the the robustness benefits of fine-tuning using CMRT but rather the opposite, where fine-tuning with CMRT leads to 4.4 and 3.3 BLEU points improvement on average over HuBERT-NLLB and CMRT-TR respectively. Furthermore, on average, CMRT leads to a minimal drop in performance on the Original test set (e.g. 0.2 and 0.3 compared to HuBERT-NLLB and CMRT-TR respectively).

[6] https://github.com/Achuttarsing/inflecteur
[7] https://github.com/bjascob/LemmInflect
[8] https://github.com/explosion/spaCy
[9] https://github.com/sammous/spacy-lefff
[10] https://github.com/espeak-ng/espeak-ng
[11] https://huggingface.co/coqui/XTTS-v2
[12] We found that decreasing the mixup probability hurts the performance on the original CoVoST2 test set without improving speech-text alignment
[13] https://github.com/facebookresearch/fairseq
[14] https://huggingface.co/facebook/nllb-200-distilled-600M

	En-De			**En-Ca**			**En-Ar**			**Fr-En**		
	Original	Morpheus	Clean	Original	Morpheus	Clean	Original	Morpheus	Clean	Original	Morpheus	Clean
HuBERT-NLLB	**28.1**	22.9	**32.8**	30.8	25.5	37.5	**20.6**	17.4	24.1	31.9	24.7	31.4
CMRT-TR	27.6	23.4	**32.8**	**31.0**	26.1	**38.1**	19.6	17.9	**24.4**	**33.7**	27.2	**34.1**
TTS-Morpheus-FN (50K)	24.9	**29.1**	30.3	27.9	**33.3**	35.0	18.4	**21.8**	22.5	30.4	**31.5**	31.6
CMRT-FN	26.6	<u>26.5</u>	32.6	30.9	<u>30.6</u>	**38.1**	19.5	<u>20.0</u>	24.1	33.6	<u>30.8</u>	33.8

Table 2: Results of using NLLB pre-trained model for initializing the MT transformer. We underline the model scores that achieved the best result on Morpheus without using adversarial speech. On overage, CMRT-FN improves robustness to Speech-Morpheus adversarial attack by more than 4 BLEU points compared to HuBERT-NLLB without using any adversarial speech data.