WACO: Word-Aligned Contrastive Learning for Speech Translation

Anonymous ACL submission

Abstract

End-to-end Speech Translation (E2E ST) aims to translate source speech into target translation without generating the intermediate transcript. However, existing approaches for E2E ST degrade considerably when only limited ST data are available. We observe that an ST model's performance strongly correlates with its embedding similarity from speech and transcript. In this paper, we propose **W**ord-**A**ligned **CO**ntrastive learning (**WACO**), a novel method for few-shot speech-to-text translation. Our key idea is bridging word-level representations for both modalities via contrastive learning. We evaluate WACO and other methods on the MuST-C dataset, a widely used ST benchmark. Our experiments demonstrate that WACO outperforms the best baseline methods by 0.7-8.5 BLEU points with only 1-hour parallel data.

1 Introduction

End-to-end speech translation (E2E ST) directly translates speech in a source language to text in a target language, without outputting the transcript text. E2E ST has experienced significant progress in translation performance (Inaguma et al., 2020; Wang et al., 2020a; Zhao et al., 2021; Zheng et al., 2021; Tang et al., 2021a; Dong et al., 2021; Han et al., 2021; Ye et al., 2021, 2022; Fang et al., 2022a; Zhang et al., 2022; Ao et al., 2022; Tang et al., 2022; Bapna et al., 2021). However, existing E2E ST methods degrade considerably when only a limited amount of parallel ST data are available (Wang et al., 2021). How can we build a well-performed ST model with no more than 10 hours of parallel data?

On the contrary, there are orders-of-magnitude more machine translation (MT) and automatic speech recognition (ASR) data than direct ST data for many languages. Plenty of recent works (Liu et al., 2020; Han et al., 2021; Xu et al., 2021; Bapna et al., 2021; Ye et al., 2022; Ao et al., 2022; Tang et al., 2022) leverage external MT and ASR data to improve the performance of E2E ST systems through model pre-training. However, we observe that the performance of the E2E ST model still degrades dramatically even though the model is pre-trained on a large-scale speech dataset and text translation dataset (Figure 1 blue line).

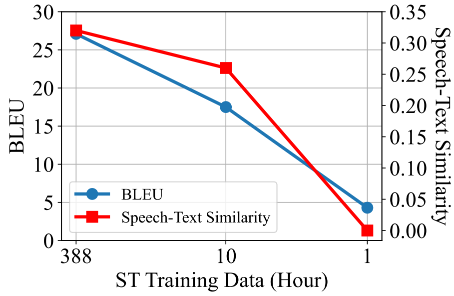
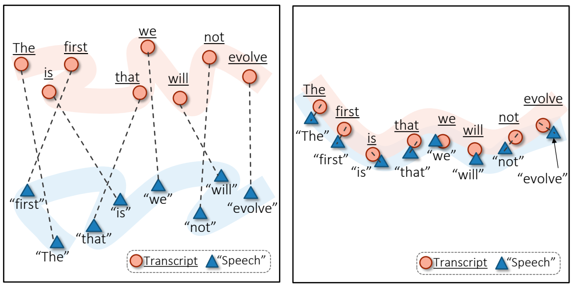


Figure 1: BLEU score of Transformer ST models trained on varying amount of ST data and their cosine similarity scores between speech and transcript word embeddings. Performance degrades significantly with fewer ST data. The ST performance highly correlates with speech-text representation similarity.

To figure out the cause of this phenomenon, we analyze speech and text representations from the directly trained ST model's encoder. We find that the translation performance highly correlates with the modality gap between speech and text representation. Specifically, we compute word-level aligned cosine similarity of speech and text embeddings (Figure 1 red line). The cross-modal similarity drops simultaneously with the BLEU score and almost reaches 0 given 1-hour ST training data. This means the model can map both modalities into a (partially) shared semantic space given enough ST data but fails when ST data is limited.

Based on the above analysis, we argue that reducing the modality gap is a key to a better E2E ST model in a few-shot ST setting. In this work, we propose WACO, a word-level contrastive learning method for few-shot speech-to-text translation. In-



(a) Baseline (b) Ideal

Figure 2: Schematic illustration of representations for speech and transcript text (projected to 2D). (a): representations learned by baseline model. (b): ideal representations — not only the sentence representations should be similar, but also the representations of each word should be close to each other.

tuitively, as shown in Figure 2, we extract speech and text representations for each word and apply contrastive learning on them to reduce the representational gap between corresponding speech and transcript text segments.

Our experiments on MuST-C dataset show that WACO outperforms all baseline methods by 0.7-8.5 BLEU points. Moreover, WACO achieves BLEU scores of 16.2 and 21.4 with only 1 and 10 hours of parallel ST data. Also, we demonstrate that WACO leads to more accurate translation than baseline methods by better speech-text alignment and fewer tokenization issues. We will make the model and code publicly available.

2 Related Work

End-to-end ST Due to error propagation and high latency in cascaded ST systems, Bérard et al. (2016); Duong et al. (2016) first proposed to translate source speech into target text directly without generating the intermediate transcript. The major difficulty in training end-to-end ST systems is the lack of direct ST data. Though many ST datasets (Wang et al., 2021; Cattoni et al., 2021) were proposed in recent years, the amount of ST data is still much less than that of MT and ASR. To overcome the data scarcity problem, methods including data augmentation (Park et al., 2019), self-training (Pino et al., 2020), multi-tasking (Le et al., 2020; Tang et al., 2021b,a; Ye et al., 2021; Zhang et al., 2022) and pre-training (Berard et al., 2018; Bansal et al., 2019; Wu et al., 2020; Wang et al., 2020b; Alinejad and Sarkar, 2020; Dong et al., 2021; Zheng et al., 2021; Bapna et al., 2021; Ao et al., 2022; Tang et al., 2022) have been proposed. WACO is a novel approach that can be applied in existing multi-tasking and pre-training frameworks to improve ST performance.

Cross-modal representation learning Researchers realized recently that the modality gap between speech and text representations hinders the knowledge transfer from external ASR and MT data (Liu et al., 2020; Xu et al., 2021; Han et al., 2021; Ye et al., 2022). Liu et al. (2020) shrank the speech representation to match the length of text representation and also closed the representational gap by minimizing their L2 distance. Xu et al. (2021) mapped speech representation to text representation through both the Connectionist Temporal Classification (CTC) (Graves et al., 2006) distribution and a mapping layer. Han et al. (2021) developed a novel architecture enabling fixed-length shared semantic space for both modalities. Ye et al. (2022) employed sentence-level contrastive loss to reduce the modality gap and achieved state-of-the-art results on MuST-C. Our method, however, works on word-level instead of sentence-level and empirically provides both better performance and higher data efficiency. Fang et al. (2022b) also proposes to close the word-level representational gap between speech and text, but their method heavily relies on target translation while our method only requires ASR data for modality reduction. Also, we note that Tang et al. (2022) explores the possibility of pre-training MT models with phoneme tokenizations, but it is unclear if the phoneme-based MT model has an advantage over the traditional BPE-based MT model and we leave the comparison to future works.

3 Proposed Method: WACO

In this section, we describe problem formulation (Section 3.1), our model architecture (Section 3.2), word-aligned contrastive method (Section 3.3) and training strategy (Section 3.4).

3.1 Problem Formulation

A typical ST corpus $\mathcal{D}^{\text{ST}}$ contains speech s and its transcript x in a source language and translation y in another language. Equivalently, $\mathcal{D}^{\text{ST}} = \{(s, x, y)\}$ and ASR corpus can be similarly defined as $\mathcal{D}^{\text{ASR}} = \{(s, x)\}$.

Given $\mathcal{D}^{\text{ST}}$ and $\mathcal{D}^{\text{ASR}}$ as training sets, the E2E ST model needs to translate speech s into translation y accurately without generating transcript x in the intermediate steps. Specifically, we consider

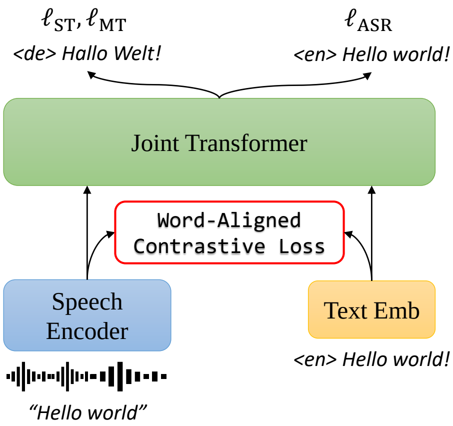


Figure 3: Model architecture of WACO. It accepts both speech and text input and outputs text sequence. In particular, we apply word-aligned contrastive loss to reduce modality gap between speech and text embeddings.

two settings in this work:

- **Few-shot ST**: we have very limited ST data but plenty of ASR data, i.e., $|\mathcal{D}^{\text{ST}}| \ll |\mathcal{D}^{\text{ASR}}|$. For example, we have ASR and ST data from 100-hour and 1-hour subsets of the MuST-C training set respectively.
- **Regular ST**: we have full ST triplet data. For example, $\mathcal{D}^{\text{ST}}$ contains the entire MuST-C 400-hour training set.

3.2 Model Architecture

Figure 3 illustrates our model architecture. WACO consists of 3 modules: a speech encoder, a text embedding layer and a joint Transformer. This architecture enables multi-tasking of both speech and text-related tasks (details of training in Section 3.3 and 3.4).

Speech Encoder extracts contextualized acoustic embeddings from the raw waveform. It consists of wav2vec 2.0 (Baevski et al., 2020) and 2 downsampling layers. Wav2vec 2.0 is one of the state-of-the-art self-supervised models pre-trained on unlabeled English speech corpus to produce contextualized speech embeddings. It has a hybrid architecture with 7 convolutional layers as the feature extractor and a Transformer as the contextualized encoder. After wav2vec 2.0, we further downsample the embedding sequence with 2 convolutional layers by a factor of 4 to alleviate the length discrepancy between speech and text embeddings.

Text Embedding embeds text tokens into a sequence of token embeddings. This is the text counterpart of the speech encoder.

Joint Transformer accepts outputs from both the speech encoder and the text embedding layer. We are using the same configuration as the vanilla Transformer (Vaswani et al., 2017). Specifically, the encoder further extracts contextualized high-level semantic features from both modalities and the decoder generates a token sequence for different tasks. Besides, since we are using general Transformer architecture, both the text embedding layer and the joint Transformer can be pre-trained on additional MT data.

3.3 Word-Aligned Contrastive Learning (WACO)

To reduce the modality gap between speech and text, we propose word-aligned contrastive learning to bring speech and text embeddings closer in a fine-grained level (Figure 4).

Suppose we have a speech-transcript pair (s, x). The transcript is tokenized by a Byte-Pair-Encoding (BPE) tokenizer into a sequence of BPE tokens $x = (x_1, x_2, \cdots, x_n)$. We group n BPE tokens back into m whole words where $w_i = x[l_i^t : r_i^t]$ for $i = 1, 2, \cdots, m$.

Then we align whole words $w_1, w_2, \cdots, w_m$ with speech $s = (s_1, s_2, \cdots, s_{|s|})$ by a forced aligner. This provides us time interval $1 \leq l_i^s \leq r_i^s \leq |s|$ for each of the word w_i.

Now we have identified m corresponding pairs of speech segments $s[l_i^s : r_i^s]$ and words $x[l_i^t : r_i^t]$. The representations of them are obtained as follows,

$$f_i^s = \text{MeanPool}(\text{S-Enc}(s)[\tilde{l}_i^s : \tilde{r}_i^s]) \quad (1)$$

$$f_i^t = \text{MeanPool}(\text{T-Emb}(x)[l_i^t : r_i^t]) \quad (2)$$

where S-Enc is speech encoder, T-Emb is text embedding layer, $\tilde{l}_i^s = \frac{l_i^s}{|s|}|\text{S-Enc}(s)|$ and $\tilde{r}_i^s = \frac{r_i^s}{|s|}|\text{S-Enc}(s)|$ refer to the relative indices given the audio representation length shrinkage after Speech Encoder.

We treat f_i^s and f_i^t as a positive pair and treat f_i^s and other words in the same batch as negative pairs and we apply multi-class N-pair contrastive loss (Sohn, 2016) on them:

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

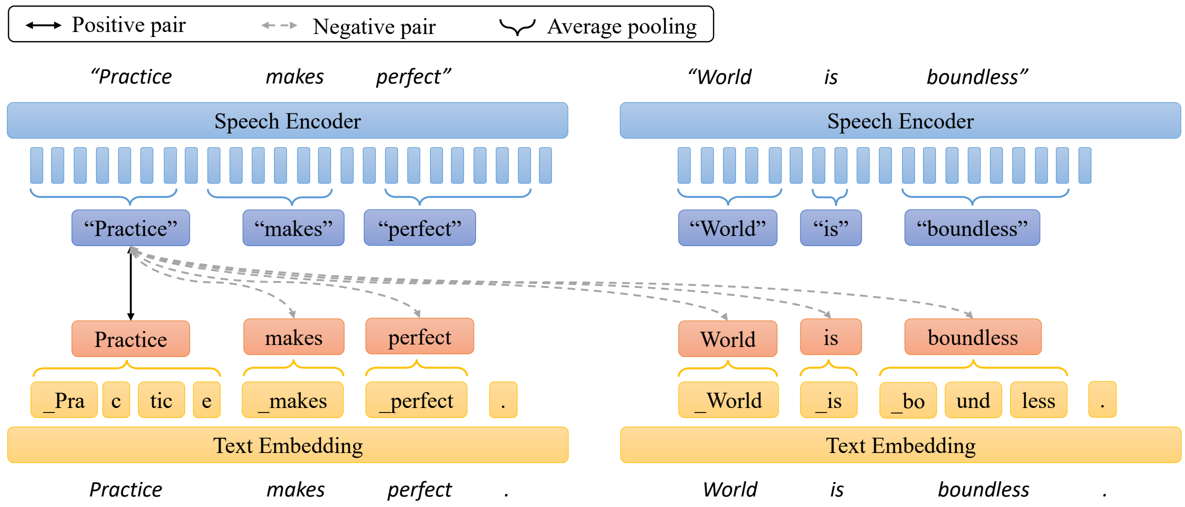


Figure 4: An illustration of word-aligned contrastive learning for a batch of two data points. Speech and text are passed through speech encoder and text embedding respectively to obtain embeddings. Then we group embeddings by word-level average pooling for both modalities. Average speech and text embeddings for the same word are treated as the positive pair and average embeddings for different words are treated as the negative pairs.

where $\mathcal{B}$ is the current batch, τ is the temperature hyper-parameter, $sim()$ is used to measure the distance between two representations, and we use cosine similarity $sim(a, b) = a^\top b/\|a\|\|b\|$.

3.4 Training Strategy

Cross-Modal Pre-training We first train a forced aligner on $\mathcal{D}^{\text{ASR}}$, then we pre-train our model using word-aligned contrastive loss

$$\mathcal{L}^{\text{PT}} = \underset{\mathcal{B} \subseteq \mathcal{D}^{\text{ASR}}}{\mathbb{E}} \left[\ell_{\text{CTR}}(\mathcal{B}) \right]. \tag{4}$$

Pre-training stage aims to map speech and text embeddings into a shared semantic space using ASR data. If the model is already pre-trained on MT corpus, this stage can also be regarded as using ASR data to distill MT knowledge.

Multi-task Fine-tuning We fine-tune our model using the multi-task cross-entropy losses and (optionally) contrastive loss.

$$\mathcal{L}^{\text{FT}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{CTR}} \tag{5}$$

where

$$\mathcal{L}_{\text{CE}} = \underset{(s,x,y) \in \mathcal{D}^{\text{ST}}}{\mathbb{E}} \left[\ell_{\text{ST}} + \ell_{\text{MT}} + \ell_{\text{ASR}} \right] \tag{6}$$

$$\mathcal{L}_{\text{CTR}} = \underset{\mathcal{B} \subseteq \mathcal{D}^{\text{ST}}}{\mathbb{E}} \left[\ell_{\text{CTR}}(\mathcal{B}) \right]. \tag{7}$$

Cross entropy losses are derived directly from the triplet dataset $\mathcal{D}^{\text{ST}}$,

$$\ell_{\text{ST}}(s, y) = -\log P(y|s) \tag{8}$$

$$\ell_{\text{MT}}(x, y) = -\log P(y|x) \tag{9}$$

$$\ell_{\text{ASR}}(s, x) = -\log P(x|s). \tag{10}$$

λ is the hyper-parameter controlling the weight of contrastive loss. When $\lambda = 0$, we are only optimizing the multi-task cross-entropy losses.

4 Experiments

4.1 Datasets

MuST-C We conduct experiments on the MuST-C dataset (Di Gangi et al., 2019), one of the largest ST benchmark datasets[1] containing translations from English to 8 languages[2] collected from TED Talks. Each language direction involves around 400 hours of audio recordings. Limited by computing resources, we examine our method on three language directions: En-De, En-Fr and En-Es.

MuST-C Few-Shot To examine few-shot ST performance, we manually create ASR and ST subsets from the MuST-C En-De training set. Specifically, we build 10-hour, 100-hour and 370-hour ASR subsets and 1-hour and 10-hour ST subsets respectively through random sampling.

External ASR We also introduce LibriSpeech (Panayotov et al., 2015) as the external ASR dataset.

[1]Released under CC BY NC ND 4.0 International
[2]Here we refer to MuST-C v1.0.

LibriSpeech is the *de facto* public English ASR benchmark[3] containing 960 hours of speech data. We build a 1330-hour ASR dataset by combining MuST-C and LibriSpeech. We use LibriSpeech mainly to evaluate how out-of-domain ASR corpus can help in-domain ST performance through cross-modal methods.

External MT Additionally, we introduce external WMT En-De/Fr/Es datasets (Bojar et al., 2016) for each language direction to pre-train text embedding and joint Transformer. As shown in previous works (Xu et al., 2021; Ye et al., 2021), MT pre-training greatly improves ST performance.

The statistics of datasets above are listed in Appendix A.1.

4.2 Experimental Setups

Model Configurations In MuST-C experiments, we use wav2vec 2.0 base model[4] in our S-Enc which is solely pre-trained on 960-hour English audio. It consists of a 7-layer convolutional feature extractor and 12 Transformer encoder blocks with 768 hidden units. Two down-sampling convolutional layers have kernel size 5, stride size 2 and hidden size 512. Joint Transformer has 6 encoder and decoder layers with hidden size 512, 2048 FFN hidden units and 8 attention heads. Joint Transformer and text embedding are pre-trained on the external WMT dataset (MT training details can be found in Appendix A.4).

Preprocess We filter speech that is either too long (>480k frames) or too short (<1k frames) out. This results in 388/471/480 hours of speech being retained as ST training data for En-De/Fr/Es directions. We jointly tokenize the transcripts and translations for each language direction using SentencePiece (Kudo and Richardson, 2018) with a vocabulary size set to 10k. Before forced alignment, we remove punctuations and group whole words by identifying special space token in the vocabulary. We use Montreal Forced Aligner (MFA) [5] to train forced aligners on $\mathcal{D}^{\text{ASR}}$ to align English speech and words. Due to vocabulary mismatch between MFA and our SentencePiece model, a small number of speeches and transcripts (e.g., 18h for En-De) cannot be aligned and we simply ignore them when doing contrastive learning.

Training The input is the raw 16-bit 16kHz mono-channel waveform. For both cross-modal pre-training and multi-task fine-tuning, we set contrastive temperature $\tau = 0.05$ and optimize our model by Adam optimizer (Kingma and Ba, 2015) ($\beta_1 = 0.9, \beta_2 = 0.98$) with learning rate 1e-4 and 25k warm-up steps. After the warm-up, the learning rate is decayed following the inverse square root schedule. The effective batch size is 16 million frames. We set dropout rate to 0.1. For pre-training, we save the checkpoints with the best contrastive loss on the validation set. For fine-tuning, we save the checkpoints with the best BLEU on the validation set and average the last 10 saved checkpoints. Also, we set label smoothing to 0.1 for the cross-entropy losses, $\lambda = 0$ in few-shot ST and $\lambda = 1$ in ST with full data. All models are trained on Nvidia A6000 GPUs.

Inference and Evaluation During inference, we run beam search with beam size 10 and length penalty 0.6/1.0/0.1 for En-De/Fr/Es directions respectively. For evaluation, we report case-sensitive detokenized BLEU scores on MuST-C tst-COMMON using sacreBLEU (Post, 2018)[6].

Baselines In few-shot ST settings, we compare our method with three baselines:

- **Base**: This baseline ignores $\mathcal{D}^{\text{ASR}}$ and only optimizes cross entropy loss in Equation 6 on $\mathcal{D}^{\text{ST}}$.
- **Base+CTC**: This baseline, on top of **Base**, applies CTC loss on $\mathcal{D}^{\text{ASR}}$ to align speech and text representations. In particular, we add a linear layer after the speech encoder to predict the text BPE token at each frame and fix its weight with text embedding. We only include CTC with BPE tokenization here since it performs consistently better than its phoneme counterpart (details in Section 5.2).
- **ConST**: This baseline adds a coarse-grained contrastive loss on $\mathcal{D}^{\text{ASR}}$ on top of **Base** to reduce modality gap as in Ye et al. (2022), one of the state-of-the-art ST methods. Instead of word-level alignment, **ConST** conducts contrastive learning on sentence-level

[3]Released under CC BY 4.0

[4]https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt

[5]https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

[6]BLEU signature: nrefs:1|bs:1000|seed:12345|case:mixed|eff:no|tok:13a| smooth:exp|version:2.0.0

Method	Few-Shot						
ASR Data	10h	100h	370h	1330h	100h	370h	1330h
ST Data	1h				10h		
Base	4.3	4.3	4.3	4.3	17.5	17.5	17.5
Base+CTC	0.2	12.6	14.6	13.6	18.3	20.4	19.4
ConST	3.0	7.3	11.7	13.7	16.9	18.6	19.6
WACO	**12.8**	**14.7**	**15.3**	**16.2**	**20.1**	**20.8**	**21.4**

Table 1: Case-sensitive detokenized BLEU scores on MuST-C En-De `tst-COMMON` set of models pre-trained on ASR data using different cross-modal methods and fine-tuned on ST data. All models share the same W2V2-Transformer architecture. **Base** ignores ASR data and only conducts multi-task fine-tuning on ST data, while other three baselines pre-train on ASR data using **CTC**, sentence-level contrastive (**ConST**) and word-aligned contrastive (**WACO**) losses.

average speech and text embeddings. Hyper-parameters are directly borrowed from Ye et al. (2022).

In regular ST with full MuST-C data, we compare our method with other existing works.

Models	**En-De**	**En-Fr**	**En-Es**
(Zhang et al., 2022)	23.0	33.5	28.0
W-Transf. (Ye et al., 2021)	23.6	34.6	28.4
SpeechT5 (Ao et al., 2022)	25.2	35.3	-
FAT-ST (Zheng et al., 2021)	25.5	-	30.8
JT-S-MT (Tang et al., 2021a)	26.8	37.4	31.0
Chimera (Han et al., 2021)	27.1	35.6	-
XSTNet (Ye et al., 2021)	27.8	38.0	30.8
SATE (Xu et al., 2021)	28.1	-	-
STEMM (Fang et al., 2022b)	**28.7**	37.4	31.0
ConST (Ye et al., 2022)	28.3	**38.3**	**32.0**
WACO	28.1	38.1	**32.0**
STPT (Tang et al., 2022)*	29.2	39.7	33.1

Table 2: Case-sensitive detokenized BLEU scores on MuST-C En-De `tst-COMMON` set of models trained on full MuST-C training set. *Note that STPT is trained on 60k hours speech data instead of 960 hours in WACO and contains more parameters (169M) than WACO (151M).

4.3 Main Results

Few-Shot ST Results are shown in Table 1. The ASR data for cross-modal pre-training varies from 10 hours to 1330 hours, and the ST data for multi-task fine-tuning varies from 1 hour to 10 hours. **WACO** consistently outperforms baseline methods in all data configurations. In particular, our model achieves a BLEU score of 12.8 with only 1h ST and 10h ASR data and 20.1 with only 10h ST and 100h ASR data. With 1330h ASR data, WACO even pushes the BLEU score to 16.2 and 21.4. More surprisingly, we find that **WACO** has a further advantage when using less ASR data. When reducing ASR data from 388 hours to 100 hours, the BLEU score increases (**WACO** vs **Base+CTC**, **ConST**) are enlarged from +0.7,+3.6 to +2.1,+7.4 in 1h ST setting and from +0.4,+2.2 to +1.8,+3.2 in 10h ST setting respectively. This demonstrates that WACO is more data-efficient than the baseline methods.

Regular ST Results are shown in Table 2. Here we are using the entire MuST-C training set as in previous works to enable fair comparison, which means $\mathcal{D}^{\text{ST}}$ has full MuST-C training data. WACO is competitive with previous state-of-the-art models such as STEMM and ConST in all three language directions. Note that STPT achieves that highest BLEU scores in all directions, but STPT trains on 60k hours of speech data instead of 960 hours in WACO (wav2vec 2.0 base) and employs a different model architecture with more parameters (169M) than WACO (151M).

5 Analysis

In this section, we analyze why word-level alignment (WACO) is better than sentence-level one (ConST) and why CTC learning is sub-optimal than WACO.

5.1 Why Word-level Contrastive Loss is Better than Sentence-level Contrastive Loss?

Intuitively, only reducing the representational gap between speech and text at the sentence level cannot assure that model captures the accurate word correspondence between these two modalities. Here we substantiate it both quantitatively and qualitatively.

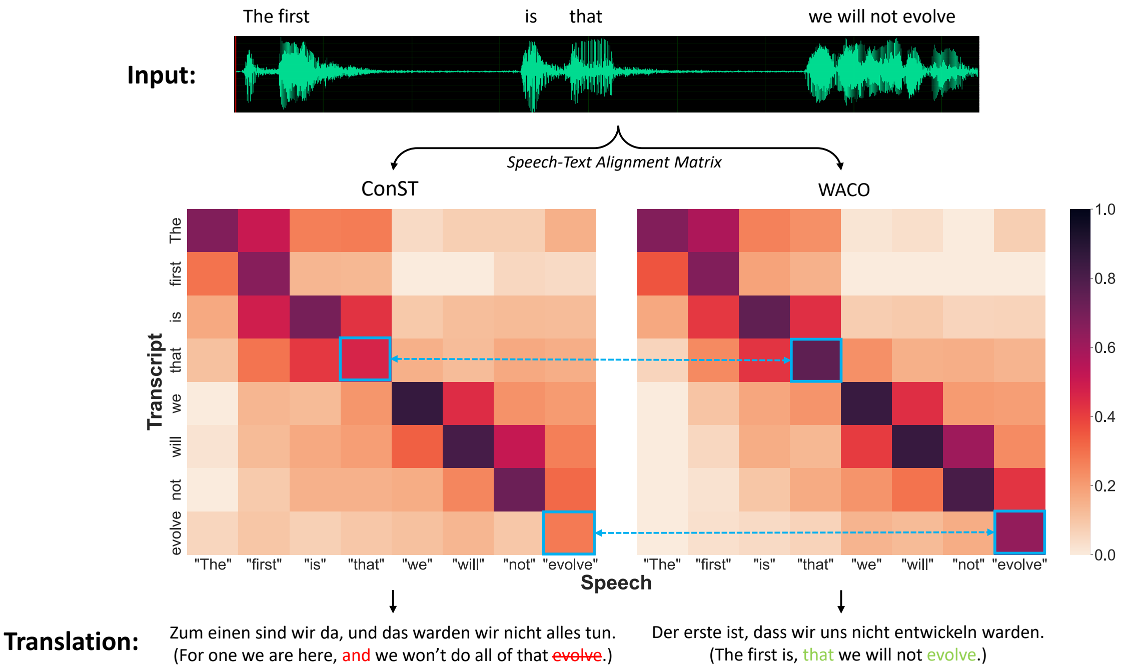


Figure 5: An example showing that WACO can capture the word-level details better than ConST. The matrix illustrates pairwise cosine similarity between word-level average embeddings of speech and transcript. WACO aligns two modalities well while ConST fails to align word "that" and "evolve". Though ConST still provides higher sentence-level similarity than WACO (0.60 for ConST and 0.58 for WACO), its translation is not as accurate as our method due to misaligned words.

Quantitatively, we compute the average cosine similarity between speech embedding and text embedding using models (ConST and WACO) pre-trained on 370h ASR dataset and fine-tuned on 1h ST dataset. Specifically, we produce embeddings following Equation 1 and 2. The result is shown in Table 3. WACO achieves more accurate word-level alignment, which indicates WACO can handle word-level details inside a sentence better.

We show an example in Figure 5 to further demonstrate the importance of such details. From the similarity matrix, we can see that WACO aligns both modalities quite well for all words but ConST struggles on words "that" and "evolve" as highlighted in blue boxes. This directly results in two translation errors of ConST. First, it fails to recover the clause structure implied by "that". Second, it omits "evolve" entirely in the translation. Though ConST still provides higher sentence similarity than WACO, it fails to understand the subtlety inside the sentence. More examples are in Figure 8.

Methods	Similarity
ConST	0.44
WACO	**0.51**

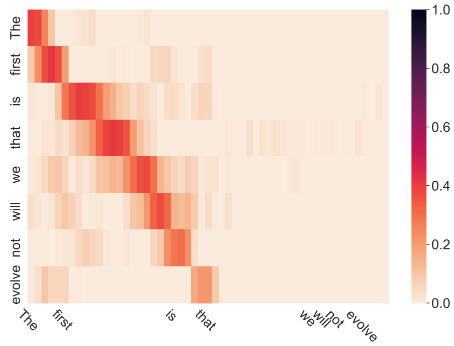
Table 3: Average cosine similarity between words from speech and transcript.

5.2 Why WACO is better than CTC?

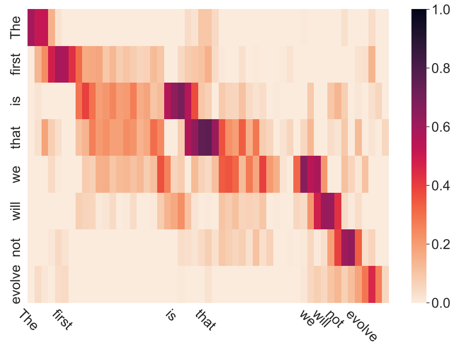
WACO treats the word as the base unit which preserves acoustic boundaries and also enables the model to leverage knowledge from the pre-trained MT model. CTC cannot benefit from word tokenization due to its extremely large vocabulary. Instead, CTC usually employs BPE, phoneme or character tokenization to learn speech-text alignment. Among these, BPE does not guarantee acoustic boundaries of each token and may lead to inconsistent tokenization (Table 4). Phoneme and character tokenization, however, make it hard to exploit the existing MT model pre-trained on large corpus since most MT methods are based on BPE tokenization.

Word	Sustainable	sustainable
BPE Tokens	_Su st ain able	_sustainable

Table 4: BPE leads to inconsistent tokenization even for the same word with different capitalization.



(a) CTC



(b) WACO

Figure 6: Token-to-Frame embedding alignment matrix produced by models trained with CTC and WACO respectively. Each row corresponds to a word and each column stands for a frame. Words in X-axis are placed according to their timestamps in speech to show how well the alignments are.

To support our claim above, we first empirically verify the disadvantage of BPE tokenization. Except direct BLEU scores reported in Table 1, Figure 7a illustrates CTC losses on training and dev set during pre-training. CTC using BPE cannot generalize well to unseen speech in the dev set (cannot even reach <2). In Figure 6, we can see that CTC indeed learns inaccurate alignment compared to WACO.

As for other tokenizations, we evaluate phoneme tokenization as an example. Specifically, we use the same phoneme vocabulary and grapheme-to-phoneme package as in (Tang et al., 2022). Different from Base+CTC introduced in Section 4.2, we randomly initialize the linear layer on top of the speech encoder since text embedding is still pre-trained using BPE tokenization. In this way, the pre-trained MT model is only used in multi-task fine-tuning. The results are shown in Table 5. CTC with phoneme tokenization is consistently outperformed by its BPE counterpart, not to mention our method. This demonstrates the importance of leveraging pre-trained MT embedding in cross-modal training.

Tokenization	100h ASR	370h ASR
BPE	18.3	20.4
Phoneme	14.3	19.0

Table 5: Case sensitive detokenized BLEU score on MuST-C En-De tst-COMMON of CTC models with BPE and phoneme tokenizations. Fine-tuning ST data is fixed at 10h.

In conclusion, CTC learning suffers from either broken acoustic boundaries (BPE) or inefficient knowledge transfer (phoneme), while WACO outperforms CTC by keeping acoustic boundaries intact and enabling efficient knowledge transfer in cross-modal training.

6 Conclusion

In this work, we propose WACO to align word-level speech and text embeddings. Experiments demonstrate the effectiveness of our method in both few-shot and regular ST settings. Analysis shows that our method can achieve better speech-text alignment and avoid tokenization issues compared to baseline methods.

Limitations

There are two main limitations in this work.

First, the source language is always English, which has more than a thousand hours of public speech data to pre-train our speech encoder, while other languages like Manx have no access to even ten hours of that. As shown in previous works (Baevski et al., 2020; Babu et al., 2021), the self-supervised model (speech encoder in WACO) heavily relies on the amount of speech data especially when downstream tasks only have limited labeled data. Thus, it remains a question to which extent other languages can benefit from WACO.

Second, instead of best ST performance given full data, our cross-modal pre-training only aims to demonstrate the effectiveness of our method in the few-shot ST setting. We realize that unified pre-training for both speech and text gradually becomes a dominant paradigm for ST and our future work is to fuse WACO into a joint pre-training framework.

Ethics Statement

WACO has the potential to benefit speakers of low-resource languages. For example, their published video or speech can be better translated into other languages, so more viewers in the world can understand them, enabling deeper communication between different cultures. Though WACO may be beneficial to cross-language communication, we do not encourage users to treat the translation generated by the E2E ST model as fully correct since they are far from perfect in practice.

References

Ashkan Alinejad and Anoop Sarkar. 2020. Effectively pretraining a speech translation decoder with machine translation data. In *Proc. of EMNLP*, pages 8014–8020.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proc. of NAACL-HLT*, pages 58–68.

Ankur Bapna, Yu-an Chung, Nan Wu, Anmol Gulati, Ye Jia, Jonathan H. Clark, Melvin Johnson, Jason Riesa, Alexis Conneau, and Yu Zhang. 2021. Slam: A unified encoder for speech and language modeling via speech-text joint pre-training.

Alexandre Berard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 6224–6228.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS workshop on End-to-end Learning for Speech and Audio Processing*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech and Language*, 66:101155.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proc. of NAACL-HLT*, pages 2012–2017.

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021. Consecutive decoding for speech-to-text translation. In *Proc. of AAAI*.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proc. of NAACL-HLT*, pages 949–959.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022a. STEMM: Self-learning with speech-text manifold mixup for speech translation. In *the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022b. 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.

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. of ICML*, volume 148 of *ACM International Conference Proceeding Series*, pages 369–376.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2214–2225.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one speech translation toolkit. In *Proc. of ACL*, pages 302–311.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. of EMNLP*, pages 66–71.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3520–3533.

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

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Proc. of INTERSPEECH*, pages 2613–2617.

Juan Miguel Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-training for end-to-end speech translation. In *Proc. of INTERSPEECH*, pages 1476–1480.

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.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. 2022. Unified speech-text pre-training for speech translation and recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1488–1499, Dublin, Ireland. Association for Computational Linguistics.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021a. 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.

Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. 2021b. A general multi-task learning framework to leverage text data for speech to text tasks. In *Proc. of ICASSP*, pages 6209–6213. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings 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*, pages 33–39.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. CoVoST 2 and Massively Multilingual Speech Translation. In *Proc. Interspeech 2021*, pages 2247–2251.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020b. Curriculum pre-training for end-to-end speech translation. In *Proc. of ACL*, pages 3728–3738.

Anne Wu, Changhan Wang, Juan Miguel Pino, and Jiatao Gu. 2020. Self-supervised representations improve end-to-end speech translation. In *Proc. of INTERSPEECH*, pages 1491–1495.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2021. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. In *Proc. of ACL*, pages 2619–2630.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. In *Proc. of INTERSPEECH*, pages 2267–2271.

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.

Biao Zhang, Barry Haddow, and Rico Sennrich. 2022. Revisiting end-to-end speech-to-text translation from scratch. In *Proceedings of the 39th International*

Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26193–26205. PMLR.

Chengqi Zhao, Mingxuan Wang, Qianqian Dong, Rong Ye, and Lei Li. 2021. NeurST: Neural speech translation toolkit. In *Proc. of ACL*, pages 55–62.

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *Proc. of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 12736–12746.

A Appendix

A.1 Statistics of Datasets

We show statistics of MuST-C, LibriSpeech and WMT datasets in Table 6,7,8 and 9.

Direction	Hours	# Sentence
En-De	408	234K
En-Fr	492	280K
En-Es	504	270K

Table 6: Statistics of MuST-C.

Type	Hours	# Sentence
ST	1	0.6K
	10	5.8K
ASR	10	5.8K
	100	58K
	370	216K
	1330	497K

Table 7: Statistics of ST and ASR subsets in MuST-C En-De Few Shot.

Language	Hours	# Sentence	# Speaker
En	960	281K	2338

Table 8: Statistics of LibriSpeech.

Direction	Name	# Sentence
En-De	WMT16	4.6M
En-Fr	WMT14	40.8M
En-Es	WMT13	15.2M

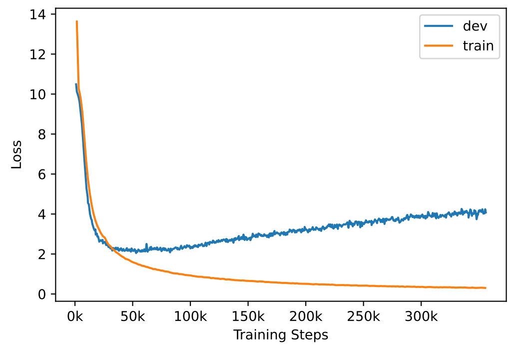
Table 9: Statistics of WMT.

A.2 More Examples of WACO versus ConST

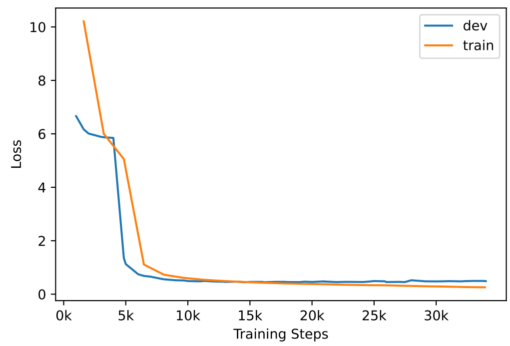
We show two more examples that WACO achieves more accurate translation than ConST by better speech-text alignment in Figure 8.

A.3 Loss Curves for Cross-Modal Pre-training

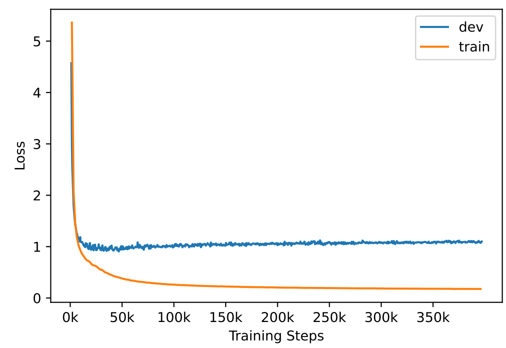
We present pre-training loss curves of CTC with both BPE and phoneme tokenizations, and WACO in Figure 7.



(a) CTC with BPE tokenization



(b) CTC with phoneme tokenization



(c) WACO

Figure 7: Loss curves of various cross-modal pre-training method. CTC with BPE tokenization cannot generalize well to unseen speech (cannot reach below 2 on dev set).

A.4 MT Pre-training

We use the same vocabulary and SentencePiece model described in Section 4.2 to tokenize the WMT datasets. The model is optimized with Adam. The learning rate starts at 1e-7, warmed up to 7e-4 by 4k steps and then decays following the inverse square root schedule with a minimum learning rate of 1e-9. The maximum number of tokens in a batch is 8192. We select the checkpoint with the highest BLEU (beam size 4, length penalty 0.6) on the WMT validation set.

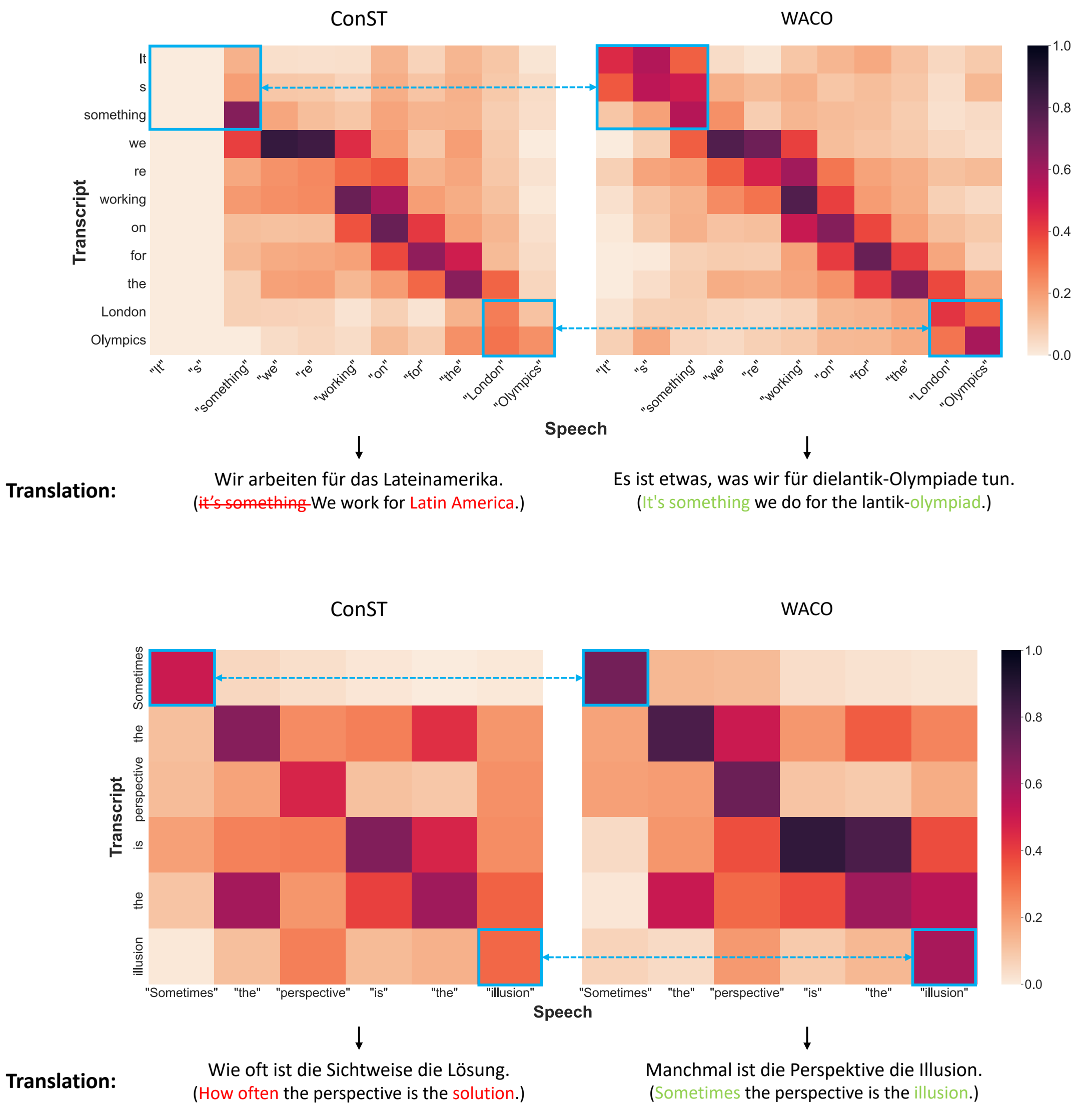


Figure 8: Two additional examples with speech-text alignment matrices and translations of WACO and ConST.