Direct Speech to Text Translation: Bridging the Modality Gap Using SimSiam

Balaram Sarkar
IIT Indore
Madhya Pradesh, India
ms2204101006@iiti.ac.in

Chandresh Kumar Maurya
IIT Indore
Madhya Pradesh, India
chandresh@iiti.ac.in

Anshuman Agrahri
IIT Indore
Madhya Pradesh, India
phd2201101008@iiti.ac.in

Abstract

Learning similar representations for spoken utterances and their written text involves understanding both forms in a shared manner. This process of developing similar representations for semantically related speech and text is essential, particularly for tasks like speech-to-text (S2T) translation. To that end, we propose a SimSiam-based S2T (**S3T**) model that leverages the SimSiam network, a state-of-the-art unsupervised learning architecture, to bridge the modality gap between speech and text. The proposed model does not require negative sample mining. The comparative study using four directions of the standard MuST-C (Di Gangi et al., 2019) dataset demonstrates that the proposed S3T translation model beats all the existing methods, and achieves an average metric of 30.02 BLEU score. Our analysis affirms that S3T effectively bridges the representation gap between the two modalities.

1 Introduction

Speech-to-text (S2T) translation is to map speech input in a given language to text output in another language. It has applications in video subtitling, facilitating communication across different demographics, education, etc. Traditional approaches for solving S2T tasks cascade two models: machine translation (MT) and automatic speech recognition (ASR). Cascade models suffer from high latency, error propagation, and memory cost. Therefore, recent works addressing S2T use end-to-end (E2E) models based on pre-trained models such as (Inaguma et al., 2020; Bérard et al., 2018; Wang et al., 2020b; Bansal et al., 2019; Le et al., 2021) or multi-task learning (joint-training) approaches (Chuang et al., 2020; Anastasopoulos and Chiang, 2018; Wang et al., 2019; Ye et al., 2022; Sperber et al., 2019; Le et al., 2020; Tang et al., 2021b). A very recent work (Ye et al., 2022) hypothesizes that the low performance of E2E models is due to the modality gap between speech and text representations. Building on the same hypothesis, we present a novel methodology based on the SimSiam (Chen and He, 2021) network, leveraging the cosine similarity (CS) loss, to mitigate the modality gap between speech and textual representations. Unlike (Ye et al., 2022), the proposed model learns joint representations in an unsupervised way and does not need negative sample mining. Our major contributions are given as follows: (a) We utilize SimSiam architecture to reduce the modality gap between textual and speech representation for the first time. As per our knowledge, such a study has not been done before, and (b) Empirical results on benchmark MuST-C data show the superiority of our approach where it outperforms the baseline by 0.17 BLEU score. Analysis indicates that the proposed approach is able to fill the modality gap.

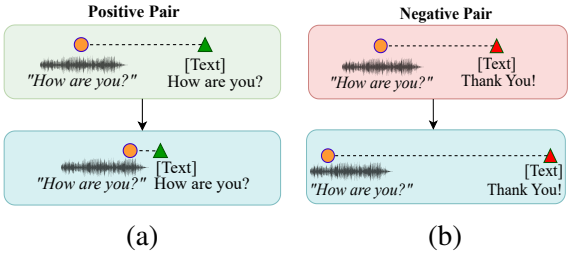


Figure 1: Depiction of representations for speech and textual transcripts. An ideal representation is where two different modalities with the same meaning (positive pair) should be close to each other as shown in (a) and it's the opposite for negative pairs in (b).

2 Related Work

Our work jointly studies end-to-end (E2E) S2T tasks and methods to counter the modality gap between speech and text.

2.1 Speech-to-Text

The traditional approach to solve S2T problem consists of cascaded systems using an ASR followed

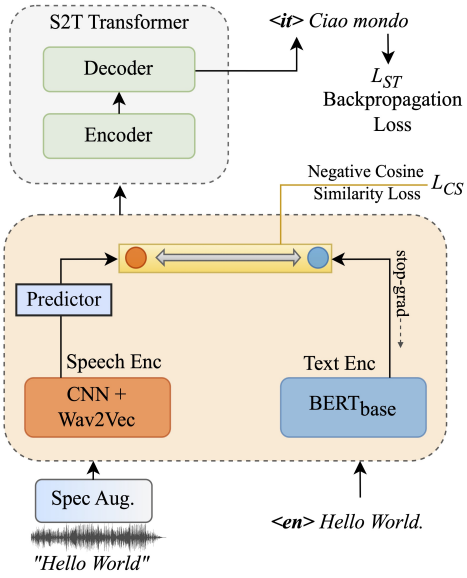


Figure 2: Proposed model architecture.

by an MT module. This method still has some limitations such as being susceptible to error propagation and having high latency (Anastasopoulos and Chiang, 2018). Recently, various authors have explored the end-to-end S2T models (Le et al., 2020; Weiss et al., 2017; Tang et al., 2022a; Di Gangi et al., 2019; Inaguma et al., 2020). Earlier, major work in this domain only produced modest results for S2T data (Tang et al., 2022a; Weiss et al., 2017), whereas the current work approached the results of the cascaded S2T models closely (Ye et al., 2022, 2021; Bentivogli et al., 2021; Xu et al., 2021; Tang et al., 2021a).

2.2 Speech and Text Alignment

The previous S2T models have worked on aligning text and speech embeddings, e.g., using an adversarial loss (Alinejad and Sarkar, 2020) in supervised pre-training, in self-supervised pre-training (Ao et al., 2022; Chen et al., 2022; Bapna et al., 2021), and using Euclidean distance (Dong et al., 2021; Liu et al., 2020; Tang et al., 2021a), cosine distance (Chuang et al., 2020), Kullback–Leibler divergence (Tang et al., 2022b), and contrastive loss (Han et al., 2021; Ye et al., 2022; Ouyang et al., 2022) in multi-task learning. All these methods require negative samples from the corpus to train the model, whereas our approach works without the need for any negative sample.

3 Problem Definition

The problem of S2T is defined as follows. Given the sequence of input audio features $x = (x_1, \ldots, x_{|x|})$ and its transcript $t = (t_1, \ldots, t_{|t|})$, the goal is to learn a representation as shown in Figure 1. More formally, the cosine similarity (CS) between positive pairs of speech and text representations (x_p, t_p) is less than the CS between negative pairs (x_n, t_n) in the embedding space.

$$\text{CS}(f(x_p), f(t_p)) < \text{CS}(f(x_n), f(t_n)) \quad (1)$$

Where f is the representation learning function. The new representations are used for the downstream S2T task, where the S2T model seeks to optimize the following objective function:

$$\theta^* = \arg\max_{\theta} \mathcal{L}(f(x, t), y) \quad (2)$$

Where $\mathcal{L}(\cdot)$ denotes the loss function of the S2T model and $y = (y_1, \ldots, y_{|y|})$ is the sequence of target text translations.

4 Method

The S2T baseline used to optimize the objective function (2) is a transformer-based encoder-decoder model. The core idea behind our approach is to use CS to align the source speech and transcript pairs and use it for downstream S2T tasks. The hypothesis is that source speech and corresponding transcript representations should be closer in the embedding space since they represent the same semantics. To that end, we seek to employ the approach originally proposed for visual recognition task handling similarity learning using SimSiam (Chen and He, 2021). Motivated by its recent application, we ask the following research question: Will the same approach be able to learn similar representations in an S2T setting? We confirm that using Siamese-like encoders for speech and transcript in an earlier stage can yield better results for the S2T task and help bridge the modality gap without negative sample mining.

4.1 SimSiam Network

Our main goal is to reduce the modality gap in S2T, which arises due to the distance between speech and textual representations. To propose a solution for this issue, we introduce an architecture influenced by (Chen and He, 2021) comprising two encoders as shown in Figure 2: One for speech and

the other for text input.

$$H \triangleq (h_1, \ldots, h_{|x|}) \triangleq \texttt{ENCODE}(x; \theta_m)$$
$$K \triangleq (k_1, \ldots, k_{|t|}) \triangleq \texttt{ENCODE}(t; \theta_n)$$

where H and K are the hidden feature vectors of audio speech sequences and their transcripts, and θ_m and θ_n are the parameters of the text and speech encoders respectively. We use Wav2Vec (Baevski et al., 2020) followed by CNN as speech encoder and as the text encoder we use a BERT base uncased (Devlin et al., 2019) model. The input pair of speech x and its parallel text t are fed to the corresponding encoders as shown in Figure 2. The SimSiam network is trained by minimizing negative cosine similarity in an unsupervised manner to generate features that are close to each other in the embedding space. The gradients from the text encoder's contribution to the loss are not used to update the speech encoder's parameters in (3) and vice versa, and this is achieved by applying the stop-gradient (SG) operation. We utilize SG with symmetric CS loss defined as follows:

$$\mathcal{L}_{CS} = \frac{1}{2}\mathcal{D}(H, \texttt{SG}(K)) + \frac{1}{2}\mathcal{D}(K, \texttt{SG}(H)) \quad (3)$$

This allows the model to learn more meaningful features from the input data.

4.2 S2T Transformer

The S2T Transformer model is a variant of the Transformer architecture adapted for processing the aligned speech-text representation as input. These features are passed through the S2T encoder containing multiple layers of self-attention mechanisms that allow the model to process different parts of the input sequence and effectively capture long-range dependencies. A self-attention mechanism computes attention weights to emphasize important features while decoding the output. During training, the model is typically tuned to a ground truth target transcript of the spoken audio by optimizing the following loss function:

$$\mathcal{L} = \mathcal{L}_{CS} + \mathcal{L}_{ST} \quad (4)$$

where

$$\mathcal{L}_{ST} = -\sum_{n} \log P(x_n|y_n)$$

$\mathcal{L}_{ST}$ is the label-smoothed-cross-entropy loss on <*speech, target text*> pairs. The output of the S2T transformer is a sequence of predicted tokens representing the translated text.

Methods	BLEU				
	De	Fr	Nl	It	Avg
NeurST	22.8	33.3	27.2	22.9	26.55
ESPnet-ST	22.9	32.7	27.4	23.8	26.7
Dual-decoder	23.6	33.5	27.6	24.2	27.22
FAIRSEQ S2T	24.5	34.9	28.6	24.6	28.15
XSTNet	25.5	36	30	25.5	29.25
ConST	25.7	36.8	**30.6**	**26.3**	29.85
S3T	**26.8**	**37**	30.2	26.1	**30.02**

Table 1: Performance of baselines and proposed model on MuST-C test split.

5 Experiment

In this section, we explain the (a) datasets, (b) baselines, (c) training and testbed followed by (d) metrics used during the evaluation.

5.1 Dataset

We conduct experiments on four pairs of translation directions available in **MuST-C**[1] (Di Gangi et al., 2019) dataset: English (En) to German (De), French (Fr), Dutch (Nl) and Italian (It). It contains audio, transcript and translation from TED talks for each direction.

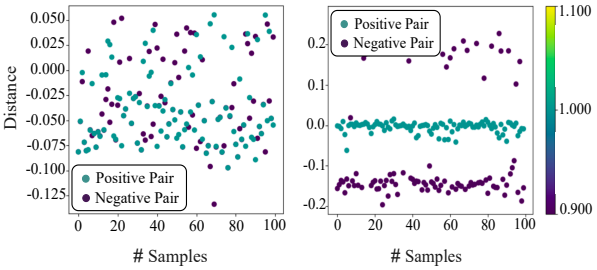
5.2 Baselines

We compare our model with two kinds of baseline: (1) standard E2E S2T models, and (2) E2E S2T models with modality bridging techniques. In the first category, we compare performance with NeurST (Zhao et al., 2021), ESPNet-ST, S2T with Dual Decoder, FAIRSEQ-S2T, and XSTNet (Ye et al., 2021). For the second category, we compare with ConST which uses contrastive loss to attract positive pairs and repel negative pairs. Note that such a scheme requires negative sample mining which is costly.

5.3 Training and Testbed

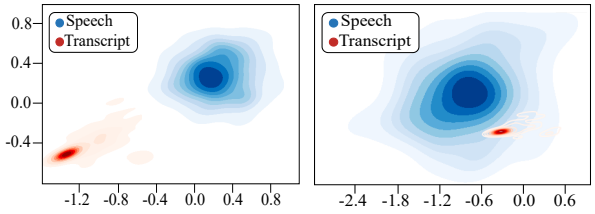
The method in this work is implemented using FAIRSEQ S2T toolkit (Wang et al., 2020a). The backbone framework consists of an S2T Transformer encoder-decoder model as shown in Figure 2. The number of self-attention layers for both the encoder and decoder is set to 6, with 8 attention

[1]We use v1.0. https://ict.fbk.eu/must-c/



(a) (b)

Figure 3: Scatter plot showing distances between positive and negative speech-text pairs (a) before, and (b) after training. The positive and negative pairs form separate clusters.



(a) (b)

Figure 4: Bivariate KDE contour plot for the embeddings of English speech and text(a) before and (b) after training. The red lines denote the text and the blue lines denote the speech representations.

heads in each layer. Due to training resource constraints, the encoder and decoder architecture is *medium* and consists of 512 hidden units. The training is halted when the performance is not improved for 15 consecutive epochs. The SpecAugment (Park et al., 2019) is used for data augmentation, and the GELU activation function is used to shift normalization and improve convergence and training stability. The S2T model is trained using label-smoothed-cross-entropy loss with a value of 0.1 as the label smoothing factor. Adam optimizer with a learning rate of 1e-4, and the learning rate schedule using an inverse square root scheduler was used.

5.4 Performance Metric

Case-sensitive detokenized BLEU using sacreBLEU is used to report the performance of the model. We average the ten best checkpoints and predict the output using a beam size of five. All experiments are repeated with three different random seeds, and we report the average BLEU on the MuST-C `tst-COMMON` set.

6 Results

This section presents the results of the comparative evaluation followed by an analysis of our proposed method.

6.1 Comparative Evaluation

Table 1 shows the main results. We compare our method with several S2T baselines. Many existing works utilize external data, such as ASR/MT data, to boost their model performance. We include models without external MT data for fair comparison and compare results with the model's *medium* architecture due to computational constraints. Comparison with standard E2E S2T models shows that our method consistently outperforms in all directions with an average BLEU of 30.02. Compared to ConST, the proposed method outperforms in two directions (De and Fr) and achieves a gain in average BLEU score of 0.17. Additionally, our approach does not need to mine any negative samples as ConST does.

6.2 Analysis

The effectiveness of our approach is shown in Figure 3. Although our method works without any positive or negative sample mining, we aim to determine its capacity to distinguish between positive and negative pairs without requiring explicit labeling. We plot the distance between pairs of speech and text samples (positive pairs with the same meaning and negative ones with different meanings) before and after the model is trained. It shows a reduction in the distance between the positive pair of samples and an increase in the distance between the negative pair of samples. To look more into it, the bivariate kernel density estimation (Parzen, 1962) (KDE) contour of the features are plotted as shown in Figure 4. If the speech and its parallel text embeddings are similar, their contour lines will overlap as much as possible. As shown in Figure 4(b), the proposed method is able to align the two representations and close the gap.

7 Conclusion

We propose S3T, a S2T framework bridging the speech-text modality gap in an unsupervised way. Results on MuST-C indicate the effectiveness of the proposed method compared to baselines. Future works may explore designing even better modality bridging techniques leveraging external data.

Limitations

Although our proposed method outperforms most baselines on the S2T benchmark, it still has some limitations: (1) the choice of hyperparameters such as learning rates, batch sizes, and the length of the projection network can significantly impact the training process and the quality of learned representations, so we need to make careful choices about it's settings; (2) with a smaller dataset, this approach might not work as effectively, because there is less variety and fewer examples for the model to learn from during training; (3) how to apply our method to other tasks also needs to be studied further.

References

Ashkan Alinejad and Anoop Sarkar. 2020. Effectively pretraining a speech translation decoder with machine translation data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8014–8020.

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91.

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.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

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 *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 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. *CoRR*, abs/2110.10329.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? 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 2873–2887.

Alexandre Bérard, 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)*, pages 6224–6228.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753.

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro J. Moreno, Ankur Bapna, and Heiga Zen. 2022. Maestro: Matched speech text representations through modality matching. In *Interspeech*.

Shun-Po Chuang, Tzu-Wei Sung, Alexander H. Liu, and Hung-yi Lee. 2020. Worse WER, but better BLEU? leveraging word embedding as intermediate in multitask end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5998–6003.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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.

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. In *AAAI Conference on Artificial Intelligence*.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. *CoRR*, abs/2105.03095.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight Adapter Tuning for Multilingual Speech Translation. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Hang Le, Juan Miguel 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 (COLING)*, pages 3520–3533.

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

Siqi Ouyang, Rong Ye, and Lei Li. 2022. WACO: Word-Aligned Contrastive Learning for Speech Translation. *arXiv e-prints*, page arXiv:2212.09359.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*.

Emanuel Parzen. 1962. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019. Attention-passing models for robust and data-efficient end-to-end speech translation. *Transactions of the Association for Computational Linguistics*, 7:313–325.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. 2022a. 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.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. 2022b. 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.

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.

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 *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6209–6213.

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.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and M. Zhou. 2019. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. In *AAAI Conference on Artificial Intelligence*.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020b. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3728–3738.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Z. Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Interspeech*.

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 *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 2619–2630.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. 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.

Chengqi Zhao, Mingxuan Wang, Qianqian Dong, Rong Ye, and Lei Li. 2021. NeurST: Neural speech translation toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 55–62.