The MINETRANS Systems for IWSLT 2023 Offline Speech Translation and Speech-to-Speech Translation Tasks

Yichao Du[♭‡], Zhengsheng Guo[♮], Jinchuan Tian[♮], Zhirui Zhang[♮], Xing Wang[♮], Jianwei Yu[♮], Zhaopeng Tu[♮], Tong Xu[♭‡] and Enhong Chen[♭‡]

[♭]University of Science and Technology of China [♮]Tencent AI Lab

[‡]State Key Laboratory of Cognitive Intelligence

[♭]duyichao@mail.ustc.edu.cn [♭]{tongxu, cheneh}@ustc.edu.cn [♮]zrustc11@gmail.com

[♮]{zhengshguo, tyriontian, tomasyu, brightxwang, zptu}@tencent.com

Abstract

This paper presents the MINETRANS English-to-Chinese speech translation systems developed for two challenge tracks of IWSLT 2023: Offline Speech Translation (S2T) and Speech-to-Speech Translation (S2ST). For the offline S2T track, MINETRANS employs a practical cascaded system consisting of automatic speech recognition (ASR) and machine translation (MT) modules to explore translation performance limits in both constrained and unconstrained settings. To this end, we investigate the effectiveness of multiple ASR architectures and two MT strategies, i.e., supervised in-domain fine-tuning and prompt-driven translation using ChatGPT. For the S2ST track, we propose a novel speech-to-unit translation (S2UT) framework to build an end-to-end system, which encodes the target speech as discrete units via our trained HuBERT and leverages the standard sequence-to-sequence model to learn the mapping between source speech and discrete units directly. We demonstrate that with a large-scale dataset, such as 10,000 hours of training data, this approach can well handle the mapping without any auxiliary recognition tasks (i.e., ASR and MT tasks). To the best of our knowledge, we are the first and only one to successfully train and submit the end-to-end S2ST model on this challenging track.

1 Introduction

In this paper, we describe the MINETRANS English-to-Chinese speech translation systems which participate in two challenge tracks of the IWSLT 2023 (Agarwal et al., 2023) evaluation campaign: Offline Speech Translation (S2T) and Speech-to-Speech Translation (S2ST).

The annual IWSLT evaluation campaign compares the models produced by different institutions on the task of automatically translating speech from one language to another. Traditional S2T/S2ST systems typically use a *cascade* approach (Ney, 1999; Sperber et al., 2017; Zhang et al., 2019; Wang et al., 2021b; Hrinchuk et al., 2022), which combines automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS, for S2ST) components. Recent advances in *end-to-end* models (Liu et al., 2019; Jia et al., 2019; Lee et al., 2022; Du et al., 2021, 2022; Zhang et al., 2022b,a) that directly translate one language speech to another without intermediate symbolic representations, have shown great potential in overcoming the problems inherent in cascaded systems, such as error propagation and slow inference. Despite this, there is still a gap between the two approaches, as *end-to-end* models have much less supervised training data than sub-tasks, i.e., ASR, MT, and TTS. Last year's IWSLT offline S2T track (Anastasopoulos et al., 2022) confirmed this, with the best *end-to-end* model submission scoring 1.7 BLEU points lower than the top-ranked cascade system. This year's competition aims to answer the question of *whether cascade solutions remain dominant*, particularly in the S2ST track, where there has large-scale data for training.

In the offline S2T track, MINETRANS employs a practical cascaded system to explore the limits of translation performance in both constrained and unconstrained settings, in which the entire system consists of automatic speech recognition (ASR), and machine translation (MT) modules. We also investigate the effectiveness of multiple ASR architectures and explore two MT strategies: supervised in-domain fine-tuning (Wang et al., 2022) and prompt-driven translation using ChatGPT[1] (Jiao et al., 2023; He et al., 2023).

In the S2ST track, MINETRANS utilizes a speech-to-unit translation (S2UT) framework to construct an *end-to-end* system, which is similar to Lee et al. (2021a) but removes all auxiliary recognition tasks (i.e., ASR and MT tasks). This framework converts target speech into discrete units via our pre-trained HuBERT and then

[1] https://chat.openai.com

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 79–88

leverages the standard sequence-to-sequence model to learn the mapping between source speech and discrete units directly. We found that with a large-scale dataset, such as 10,000 hours of training data, the previous multi-task learning technique (Jia; Lee et al., 2021a,b; Popuri et al., 2022; Dong et al., 2022) is not necessary for model convergence, and this approach can successfully handle the mapping between source speech and discrete units. We also explore various initialization strategies and several techniques to improve model performance, including (1) different self-supervised pre-trained speech encoders and pre-trained text-to-unit models, (2) data filtering and augmentation, consistency training, and model ensembles. To the best of our knowledge, we are the first and only one to successfully train and submit the end-to-end S2ST model on this challenging track. Our code is open-sourced at: https://github.com/duyichao/MINETrans-IWSLT23.

The remainder of this paper is organized as follows: Section 2 describes data preparation, including data statistics, data preprocessing, and data filtering. Section 3 describes our solution for the offline speech translation track. Section 4 describes our solution to the speech-to-speech track. In Section 5, we conclude this paper.

2 Data Preparation

2.1 Data Statistics

Table 1 lists statistics of the speech corpus we used for MINETRANS training, which can be divided into four categories: unlabeled speech, ASR, TTS and S2ST Corpus.

Unlabeled Speech. As shown in Table 1, we integrate source side speech from VoxPopuli (Wang et al., 2021a) and GigaSS[2] to build a large-scale unlabeled English speech corpus for self-supervised training of speech encoders Wav2vec2.0 (Baevski et al., 2020) and HuBert (Hsu et al., 2021), which are used for initializing the S2UT model in the S2ST track. Similarly, we also integrate target speech from GigaSS and AISHELL-3 (Shi et al., 2020) to train the Chinese HuBert, which is used for discretizing Chinese speech.

ASR Corpus. To train data-constrained English ASR models, we merge MuST-C (Gangi et al., 2019), Common Voice v11 (Ardila et al., 2019), Librispeech (Panayotov et al., 2015), and Europarl-ST (Iranzo-Sánchez et al., 2019), resulting in approximately 4500 hours of labeled ASR corpus, as shown in Table 1. For MuST-C and Europarl-ST, we collect source speech for all translation directions and de-duplicated them based on audio identifiers. In addition, GigaSpeech (Chen et al., 2021) is used to construct data-unconstrained ASR model, which includes 10k hours data covering various sources (audiobooks, podcasts, and stream media), speaking styles (reading and spontaneous), and topics (arts, science, sports, etc.). Of these corpus, we use MuST-C as the in-domain for the Offline track and the rest as the out-of-domain.

MT Corpus. To train data-constrained English-to-Chinese MT models, MuST-C v1&v2 are considered in-domain corpora, while OpenSubtitles2018 (Lison et al., 2018) and NewsCommentary[3] corpora are considered out-of-domain. Additionally, we utilize in-house corpora to train data-unconstrained MT models, although we cannot provide further details about it.

TTS Corpus. To ensure target speech timbre matching with the S2ST track, we consider the single-speaker GigaSS-S, a small subset of GigaSS, as in-domain and the multi-speaker AISHELL-3 (Shi et al., 2020) as out-of-domain. These corpora are used to train the TTS model and its corresponding vocoder.

S2ST Corpus. The full version of GigaSS is used to train our end-to-end S2UT model, which is an large-scale S2ST corpora derived from GigaSpeech (Chen et al., 2021) via MT and TTS. We also construct S2ST pseudo-data, the details of which will be presented in Section 4.1.2.

2.2 Data Pre-processing and Filtering

In general, a simple way to improve model performance is to provide them with better data. However, through a careful review of the data, we identified issues with the quality of the original data. To address this, we performed the following preprocessing and filtering:

- We convert all audio data to mono-channel 16kHz wav format. Since the sentences of spoken translation are generally short, we discarded sentences with text longer than 100 and speech frames longer than 3000. Then 80-dimensional

[2] https://github.com/SpeechTranslation/GigaS2S

[3] https://opus.nlpl.eu/News-Commentary.php

	Corpus	Utterances (k)	Duration (h)	S2T CST.	S2ST CST.
Unlabeled	VoxPopuli	22,905	28,708	✓	✓
ASR	MuST-C ASR v1&v2	342	617	✓	–
	Common Voice v11.0	1680	3,098	✓	–
	Librispeech	281	960	✓	–
	Europarl-ST	34	81	✓	–
	GigaSpeech	8,030	10,000	×	–
MT	NewsCommentary	32	–	✓	–
	OpenSubtitles	9,969	–	✓	–
	MuST-C v1&v2	543	–	✓	–
	In-house	–	–	×	–
TTS	AISHELL 3	88	85	–	✓
	GigaSS-S	210	244	–	✓
S2ST	GigaSS	7,635	9,000	–	✓
	CoVoST synthetic	288	288	–	✓
	MuST-C synthetic	358	587	–	✓

Table 1: Statistics of the training data. The "CST." indicates that a corpus is in the task constrained corpus list of corresponding S2T or S2ST. The "-" indicates this corpus is not available in that column.

log-mel filter banks acoustic features are extracted with a stepsize of 10ms and a window size of 25ms. The acoustic features are normalized by global channel mean and variance.

- We use a pre-trained ASR model on Librispeech to filter the audio with very poor quality, i.e., word error rate (WER) more than 75.
- Since the annotation format is not uniform across multiple datasets, we remove non-printing characters, speaker names, laughter, applause and other events. In addition, we also regularize punctuation marks.
- For the English-to-Chinese direction of MuST-C, we first merge the v1 and v2 versions and then remove duplicates based on audio identifiers.

3 Offline Speech Translation

3.1 Cascaded MINETRANS S2T System

3.1.1 Speech Recognition

A standard RNN-Transducer (Graves, 2012) model is used for speech recognition. It consists of an acoustic encoder, a prediction network and a joint network. The acoustic encoder contains 18 Conformer (Gulati et al., 2020) layers with the following dimensions: attention size is 512, feed-forward size is 2048, number of attention heads is 4, and convolutional kernels is 31. The prediction network is a standard 1-layer LSTM with a hidden size of 1024. The joint network is linear with a size of 512. The input acoustic features are 80-dim Fbank plus 3-dim pitch, which are down-sampled by a 2-layer CNN with a factor of 6 in the time-axis before being fed into the acoustic encoder. The overall parameter budget is 126M. During training, SpecAugment (Park et al., 2019) is consistently adopted for data augmentation. The training on both GigaSpeech and MuST-C datasets lasts for 50 epochs each, which consumes 32 Nvidia V100 GPUs. The Adam optimizer is adopted, with peak learning rate of 5e-3, warmup steps of 25k and inverse square root decay schedule(Vaswani et al., 2017a). Model weights from the last 10 epochs are averaged before decoding. The default decoding method described in Graves (2012) is adopted with a beam size of 10. External language models in any form are not adopted.

ASR Output Adaptation. In the realm of automatic speech recognition (ASR) and machine translation (MT), it is common for ASR output to lack punctuation, whereas MT models are sensitive to punctuation. To address this issue, we propose an ASR output adaptation method by incorporating a punctuation model between ASR and MT. Specifically, we adopt a BERT-based punctuation model that can automatically recover the original punctu-

ation. The objective of this approach is to bridge the disparity between ASR and MT, leading to improved overall performance in speech translation tasks.

Speech Segmentation. Speech translation is a multi-faceted task that requires overcoming the challenges of bridging the gap between automatic speech recognition (ASR) and machine translation (MT) systems. To address these challenges, we employ several text augmentation techniques to improve the quality and accuracy of our training data. Specifically, we have utilized speech-based audio segmentation (SHAS (Tsiamas et al., 2022)) to identify and segment meaningful units of speech that can be accurately translated by the MT system.

3.1.2 Machine Translation

In our systems, we adopt four different types of translation strategies:

- **TRANSFORMER** is a system trained on the constrained data. We train the Transformer-base (Vaswani et al., 2017b) model on the constrained general data and finetune the model on the in-domain MuST-C data.
- **M2M-100**[4] (Fan et al., 2021) is a multilingual model trained for many-to-many multilingual translation. We employ the supervised in-domain fine-tuning strategy to finetune the M2M-100 1.2B-parameter model on the downstream MuST-C data.
- **CHATGPT** is a large language model product developed by OpenAI. Previous studies (Jiao et al., 2023; Wang et al., 2023) have demonstrated that ChatGPT is a good translator on high-resource languages. Therefore we utilize the proper translation prompts with ChatGPT to carry out the translation task.
- **IN-HOUSE MODEL** We fine-tune our in-house translation model (Huang et al., 2021) using the MuST-C data. Our in-house model is a Transformer-big (Vaswani et al., 2017b) model with a deep encoder (Dou et al., 2018).

Data Re-Annotation. We have identified two issues with the annotation of the English-to-Chinese translation direction in the MuST-C v2.0 test set[5]. Firstly, we have observed samples of incorrect literal translations. For example, for the parallel sentence pair, "I remember my first fire. 我记得我第一场火", we usually translate the English word "fire" into Chinese word "火灾 (huo zhai)" not "火 (huo)". Secondly, we have noticed inconsistencies in the punctuation annotation, as most Chinese translations lack proper full stop marks. To address these challenges, we have employed the services of a professional translator to accurately translate the English sentences. We will release the data, aiming to facilitate future research in the field.

Domain Augmentation. The MuST-C v2.0 training data contains considerable bilingual sentence pairs that are partially aligned. In the specific pair "Thank you so much Chris. 非常谢谢，克里斯。的确非常荣幸", we are unable to locate the corresponding translation for the Chinese phrase "的确非常荣幸" in the English sentence. As Koehn and Knowles (2017); Wang et al. (2018) pointed out, data noise (partially aligned data) has been demonstrated to impact the performance of Neural Machine Translation (NMT). To address this issue, we employ a data rejuvenation strategy (Jiao et al., 2020). Specifically, we first finetune the model using the raw parallel data and then rejuvenate the low-quality bilingual samples to enhance the training data.

3.2 Experiment

The Cascaded MINETRANS S2T System we propose comprises an Automatic Speech Recognition (ASR) model and a machine translation (MT) model. In our evaluation, we assess the performance of each component separately. For the ASR system evaluation, we employ the Word Error Rate (WER) metric, while the BLEU score is utilized to evaluate the performance of our machine translation model.

The evaluation results obtained on the MuST-C dataset, with and without fine-tuning, are presented in Table 2. When the GigaSpeech ASR system is used without fine-tuning, we observe a WER of 10.0 on the MuST-C test set. However, when the system is fine-tuned using the MuST-C dataset, a significant improvement in performance is observed, resulting in a noticeable decrease in the error rate from WER of 10.0 to 5.8. This highlights the effectiveness of fine-tuning on the MuST-C dataset in enhancing the overall performance of our system.

[4] https://github.com/facebookresearch/fairseq/tree/main/exa\mples/m2m_100

[5] https://ict.fbk.eu/MuST-C/

System	Dev	Test
Gigaspeech	9.3	10.0
+ MuST-C Finetune	4.8	5.8

Table 2: ASR performance measured in terms of word error rates.

We evaluate various translation strategies using the MuST-C test set. The experimental results are presented in Table 2. In the constrained scenario, TRANSFORMER achieved a test BLEU score of 25.04, whereas M2M-100 attained a marginally higher score of 25.40. In the unconstrained setting, CHATGPT demonstrated superior performance with a BLEU score of 28.25, while IN-HOUSE MODEL obtained the highest BLEU score of 30.91. These results emphasize the significance of utilizing in-domain data for achieving optimal performance in spoken language translation.

System	Dev	tst-COMMON
TRANSFORMER	13.93	25.04
M2M-100	16.53	25.40
CHATGPT	—	28.25
IN-HOUSE MODEL	21.52	30.91

Table 3: Offline speech translation performance measured in terms of the BLEU score.

4 Speech-to-Speech Translation

4.1 End-to-End MINETRANS S2ST System

As shown in Figure 1, we construct an end-to-end S2UT (Lee et al., 2021a) model comprising a speech encoder, length adapter, and unit decoder. Following (Lee et al., 2021a), we encode target speech as discrete units via our trained Chinese HuBert and remove consecutive repetitive units to generate a reduced unit sequence. Unlike (Lee et al., 2021a), our S2UT model directly learns the mapping between source speech and discrete units without any auxiliary recognition tasks (i.e., ASR and MT tasks), which hyper-parameters are difficult to tune. Then we leverage a unit-based HiFi-GAN Vocoder to achieve unit-to-waveform conversion (Polyak et al., 2021). Next, we detail the efforts making in pre-training for model initialization, data augmentation, consistency training and model ensemble, which are used to improve the translation quality of our system.

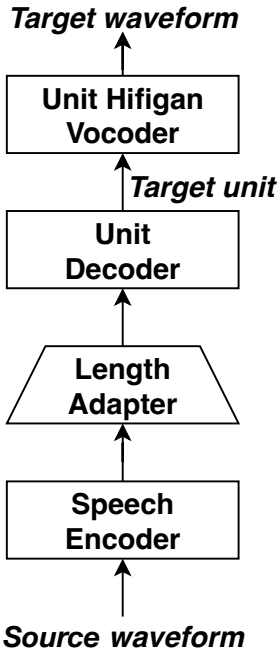

Figure 1: The overall architecture of the end-to-end S2ST system.

4.1.1 Pretrained Models

Previous experiences (Dong et al., 2022; Popuri et al., 2022) shown that better initialization can reduce learning difficulty, we explore pre-training of both the speech encoder and unit decoder.

Speech Encoder Pre-training. We use Wav2vec 2.0 (Baevski et al., 2020) and HuBert (Hsu et al., 2021), which are trained in a self-supervised manner, as speech encoders. Due to the data limitation of the S2ST track, we use the unlabeled speech described in Table 1 for training speech encoder:

- **Wav2vec 2.0** uses a multi layer convolution neural network to encode audio and then uses a transformer-based context encoder to construct a contextual representation. The model is trained by having a masked span of contrast loss on the input of the context encoder. In this paper, we modify Transformer as Conformer to obtain better performance.
- **HuBert** has the same model architecture as Wav2vec 2.0. However, its training process differs primarily in the use of cross-entropy and additionally in the construction of targets through a separate clustering process.

Unit Decoder Pre-training. We use the standard sequence-to-sequence model to model the Text-to-unit (T2U) task on GigaSS, and the decoder of

this model will be used for the initialization of the unit decoder of S2UT. The T2U model contains 12 transformer layers for the encoder and coder, respectively. More specifically, we set the size of the self-attention layer, the feed-forward network, and the head to 1024, 4096, and 8, respectively.

4.1.2 Model Finetuning

We combine the pre-trained speech encoder and unit decoder, and adding a randomly initialized length adapter between the pre-trained modules. The length adapter consists of a one-dimensional convolutional layer with a stride of 2, which mitigates the length difference between the source audio and the reduced target unit, as well as the mismatch between representations.

Consistency Training. To further improve the consistency of our model, we employ the R-Drop algorithm (Liang et al., 2021) with a weight α set to 5. The R-Drop algorithm reduces inconsistencies predicted by the model between training and inference through dropout, thereby improving generalization. Specifically, it randomly drops out parts of the model during training, forcing it to learn more robust representations that are less sensitive to small changes in the input. For a more detailed description of the R-Drop algorithm and its implementation, please refer to the paper by (Liang et al., 2021).

4.1.3 Unit-based Vocoder

We utilize the unit-based HiFi-GAN (Polyak et al., 2021) vocoder to convert discrete units into waveform for the speech-to-unit model. Following the (Lee et al., 2021a) setup, we augment the vocoder with a duration prediction module for the reduced unit output, which consists of two 1D convolutional layers, each with ReLU activation, followed by layer normalization and a linear layer.

4.1.4 Ensemble

Model ensemble can reduce the inconsistency of the system to some extent, and we consider the ensemble of four variants of S2UT models:

- **W2V2-CONF-LARGE**: The speech encoder is initialized using Conformer-based Wav2vec 2.0 LARGE model. The unit decoder is initialized randomly.
- **W2V2-CONF-LARGE+T2U**: The speech encoder is initialized using Conformer-based Wav2vec 2.0 LARGE model. The unit decoder is initialized from the T2U model.
- **W2V2-TRANS-LARGE+T2U**: The speech encoder is initialized using Transformer-based Wav2vec 2.0 LARGE model. The unit decoder is initialized from the T2U model.
- **HUBERT-TRANS-LARGE+T2U**: The speech encoder is initialized using Transformer-based HuBert LARGE model. The unit decoder is initialized from the T2U model.

4.1.5 Data Augmentation

We utilize well trained Fastspeech2 (Ren et al., 2020) TTS models (see Section 4.2 for details) to generate speech for MuST-C and CoVoST Chinese texts to construct pseudo-corpora. These pseudo-corpora are used as training data together with the original labeled S2ST corpus.

4.2 Experiments

4.2.1 Implementation Details

All end-to-end S2UT models are implemented based on the FAIRSEQ[6] (Ott et al., 2019) toolkit. We use pre-trained Chinese HuBERT model and k-means model to encode Chinese target speech into a vocabulary of 250 units. The Chinese HuBERT and k-means models are learned from the TTS data in Table 1. The architectural details of the S2UT models are detailed in section 4.1.4. During training, we use the adam optimizer with a learning rate set to 5e-5 to update model parameters with 8K warm-up updates. The label smoothing and dropout ratios are set to 0.15 and 0.2, respectively. In practice, we train S2UT with 8 Nvidia Tesla A100 GPUs with 150K update steps. The batch size in each GPU is set to 1200K, and we accumulate the gradient for every 9 batches. For the first 5K steps of S2UT model training, we freeze the update of the speech encoder. The Unit HiFi-GAN Vocoder is trained using SPEECH-RESYNTHESISRES[7] toolkit for 500k steps. For FastSpeech2 and HiFi-GAN, we followed the paddlespeech AISHELL recipe[8] for training. During inference, we average the model parameters on the 30 best checkpoints based on the performance of the GigaSS dev set, and adopt beam search strategy with beam size of 10.

[6] https://github.com/facebookresearch/fairseq

[7] https://github.com/facebookresearch/speech-resynthesis

[8] https://github.com/PaddlePaddle/PaddleSpeech/tree/develop/examples/aishell3/tts3

ID	Model	BLEU	chrF
1	W2V2-CONF-LARGE	27.7	23.4
2	W2V2-CONF-LARGE+T2U	**27.8**	**23.7**
3	W2V2-TRANS-LARGE+T2U	25.2	22.3
4	HUBERT-TRANS-LARGE+T2U	26.2	23.2
5	HUBERT-TRANS-LARGE+T2U*	25.7	22.6
6	Ensemble(1, 2, 4)	**28.0**	**23.9**
7	Ensemble(2, 4, 5)	27.2	23.0

Table 4: ASR-BLEU and ASR-chrF on GigaSS validation set. '*' indicates adding the GigaST test set to the training data and fine-tuning it for one round.

4.2.2 Results

To evaluate the speech-to-speech translation system, we use a Chinese ASR system[9] trained on WenetSpeech (Zhang et al., 2021) to transcribe the speech output with the `ctc_greedy_serach` mode. Based on this, we report case-sensitive BLEU and chrF scores between the produced transcript and a textual human reference using sacreBLEU. The results on the GigaSS validation set is shown in Table 4. Comparing W2V2-CONF-LARGE+T2U and W2V2-TRANS-LARGE+T2U, using Conformer-based architecture pre-trained speech encoder for initialization has better performance. In addition, we find that adding the GigaST test set to training leads to a weak performance degradation on the validation set, possibly because the annotations of the test set are calibrated by humans and their style differs from that of the training data.

5 Conclusion

This paper presents the MINETRANS system for two challenge tracks of the IWSLT 2023: Offline Speech Translation (S2T) and Speech-to-Speech Translation (S2ST). For the S2T track, MINETRANS employs a cascaded system to investigate the limits of translation performance in both constrained and unconstrained settings. We explore two machine translation strategies: supervised in-domain fine-tuning and prompt-guided translation using a large language model. For the S2ST track, MINETRANS builds an end-to-end model based on the speech-to-unit (S2U) framework. To the best of our knowledge, we are the first and only team to successfully train and submit the end-to-end S2ST on this track. This model uses our trained HuBERT to encode the target speech as discrete units and leverages the standard sequence-to-sequence model to directly learn the mapping between source speech and discrete units without the need for auxiliary recognition tasks such as ASR and MT. We use several techniques to improve MINETRANS's performance, including speech encoder pre-training on large-scale data, data filtering, data augmentation, speech segmentation, consistency training, and model ensemble.

[9] https://github.com/wenet-e2e/wenet/blob/main/docs/pretrained_models.en.md

Acknowledgements

This work is supported by the grants from National Natural Science Foundation of China (No.62222213, U20A20229, 62072423), and the USTC Research Funds of the Double First-Class Initiative (No.YD2150002009). The authors would like to thank anonymous reviewers for their valuable comments. Zhirui Zhang and Tong Xu are the corresponding authors.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutail Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondrej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Y. Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, B. Hsu, Dávid Javorský, Věra Kloudová, Surafel Melaku Lakew, Xutai Ma, Prashant Mathur,

Paul McNamee, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John E. Ortega, Juan Miguel Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander H. Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the iwslt 2022 evaluation campaign. In *IWSLT*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. In *International Conference on Language Resources and Evaluation*.

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Weiqiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10, 000 hours of transcribed audio. *ArXiv*, abs/2106.06909.

Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, Qibing Bai, and Yu Zhang. 2022. Leveraging pseudo-labeled data to improve direct speech-to-speech translation. In *Interspeech*.

Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Shuming Shi, and Tong Zhang. 2018. Exploiting deep representations for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4253–4262.

Yichao Du, Weizhi Wang, Zhirui Zhang, Boxing Chen, Tong Xu, Jun Xie, and Enhong Chen. 2022. Non-parametric domain adaptation for end-to-end speech translation. In *Conference on Empirical Methods in Natural Language Processing*.

Yichao Du, Zhirui Zhang, Weizhi Wang, Boxing Chen, Jun Xie, and Tong Xu. 2021. Regularizing end-to-end speech translation with triangular decomposition agreement. In *AAAI Conference on Artificial Intelligence*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Mattia Antonino Di Gangi, R. Cattoni, L. Bentivogli, Matteo Negri, and M. Turchi. 2019. Must-c: a multilingual speech translation corpus. In *NAACL*.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *arXiv preprint arXiv:2305.04118*.

Oleksii Hrinchuk, Vahid Noroozi, Ashwinkumar Ganesan, Sarah Campbell, Sandeep Subramanian, Somshubra Majumdar, and Oleksii Kuchaiev. 2022. Nvidia nemo offline speech translation systems for iwslt 2022. In *IWSLT*.

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 Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. Transmart: A practical interactive machine translation system. *arXiv preprint arXiv:2105.13072*.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Alberto Sanchís, Jorge Civera Saiz, and Alfons Juan-Císcar. 2019. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

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

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Wenxiang Jiao, Xing Wang, Shilin He, Irwin King, Michael Lyu, and Zhaopeng Tu. 2020. Data rejuvenation: Exploiting inactive training examples for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2255–2266.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *First Workshop on Neural Machine Translation*, pages 28–39. Association for Computational Linguistics.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Miguel Pino, and Wei-Ning Hsu. 2021a. Direct speech-to-speech translation with discrete units. In *Annual Meeting of the Association for Computational Linguistics*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3327–3339, Dublin, Ireland. Association for Computational Linguistics.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Juan Miguel Pino, Jiatao Gu, and Wei-Ning Hsu. 2021b. Textless speech-to-speech translation on real data. *ArXiv*, abs/2112.08352.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, M. Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. *ArXiv*, abs/2106.14448.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *International Conference on Language Resources and Evaluation*.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. In *INTERSPEECH*.

H. Ney. 1999. Speech translation: coupling of recognition and translation. *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, 1:517–520 vol.1.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, S. Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and S. Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 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. *Interspeech 2019*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. *ArXiv*, abs/2104.00355.

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

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *ArXiv*, abs/2006.04558.

Yao Shi, Hui Bu, Xin Xu, Shaojing Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. In *Interspeech*.

Matthias Sperber, Graham Neubig, J. Niehues, and A. Waibel. 2017. Neural lattice-to-sequence models for uncertain inputs. In *EMNLP*.

Ioannis Tsiamas, Gerard I Gállego, José AR Fonollosa, and Marta R Costa-jussà. 2022. Shas: Approaching optimal segmentation for end-to-end speech translation. *arXiv preprint arXiv:2202.04774*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. *Advances in neural information processing systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. *Advances in neural information processing systems*, 30.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Miguel Pino, and Emmanuel Dupoux. 2021a. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Annual Meeting of the Association for Computational Linguistics*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Minghan Wang, Yuxia Wang, Chang Su, Jiaxin Guo, Yingtao Zhang, Yujiao Liu, M. Zhang, Shimin Tao, Xingshan Zeng, Liangyou Li, Hao Yang, and Ying Qin. 2021b. The hw-tsc's offline speech translation system for iwslt 2022 evaluation. In *IWSLT*.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 133–143.

Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael Lyu. 2022. Understanding and improving sequence-to-sequence pretraining for neural machine translation.

In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2591–2600.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2021. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186.

Peidong Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2019. Lattice transformer for speech translation. In *ACL*.

Weitai Zhang, Zhongyi Ye, Haitao Tang, Xiaoxi Li, Xinyuan Zhou, Jing Yang, Jianwei Cui, Dan Liu, Junhua Liu, and Lirong Dai. 2022a. The ustc-nelslip offline speech translation systems for iwslt 2022. In *IWSLT*.

Ziqiang Zhang, Junyi Ao, Shujie Liu, Furu Wei, and Jinyu Li. 2022b. The yitrans end-to-end speech translation system for iwslt 2022 offline shared task. *ArXiv*, abs/2206.05777.