# I2R's End-to-End Speech Translation System for IWSLT 2023 Offline Shared Task

**Muhammad Huzaifah, Kye Min Tan, Richeng Duan**
Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore

## Abstract

This paper describes I2R's submission to the offline speech translation track for IWSLT 2023. We focus on an end-to-end approach for translation from English audio to German text, one of the three available language directions in this year's edition. The I2R system leverages on pretrained models that have been exposed to large-scale audio and text data for our base model. We introduce several stages of additional pretraining followed by fine-tuning to adapt the system for the downstream speech translation task. The strategy is supplemented by other techniques such as data augmentation, domain tagging, knowledge distillation, and model ensemble, among others. We evaluate the system on several publicly available test sets for comparison.

## 1 Introduction

Historically, speech translation (ST) has involved combining automatic speech recognition (ASR) and machine translation (MT) systems in a cascade. The ASR system would transcribe speech signals into text in the source language, and the MT system would then translate this text into the target language. However, recent developments in deep learning have made it possible to use an end-to-end speech translation model (Bérard et al., 2016; Weiss et al., 2017), which directly translates speech in the source language into text in the target language, without relying on intermediate symbolic representations. This approach offers the advantages of lower latency and avoids error propagation. While cascaded models initially outperformed end-to-end models, recent results from IWSLT campaigns (Le et al., 2020; Bentivogli et al., 2021; Anastasopoulos et al., 2022) have shown that the performance of end-to-end models is now approaching that of cascaded solutions.

Large pretrained models (Lewis et al., 2020; Conneau et al., 2021; Raffel et al., 2020) have become a prevalent basis for speech and language processing work (Ma et al., 2021; Chen et al., 2022a). Through the utilization of pretrained models and subsequent finetuning using a small amount of labeled data, many tasks have exhibited significant improvements in performance (Baevski et al., 2020; Hsu et al., 2021; Guillaume et al., 2022; Navarro et al., 2022), some even reaching state-of-the-art results.

In this work, we describe our end-to-end system for the Offline Speech Translation Task at IWSLT 2023 (Agarwal et al., 2023) in the English-German (En-De) language direction. The current year's task not only includes the traditional TED talk evaluation set translated from English to German, but also introduces two additional test sets consisting of ACL presentations, press conferences and interviews (EMPAC), which are more complex and challenging. Furthermore, this year's constrained data track allows less data than previous years. Our team enhances the end-to-end ST system within the context of the pretrain-finetune paradigm. We introduce several pretraining stages before finetuning for the downstream ST task. Furthermore, we implemented dynamic audio augmentation methods to account for differences in audio recording quality. We boost the system's robustness by ensembling multiple individual models and use domain tagging to direct the model towards specific output styles. Here, we evaluate our system against various standard public test sets for both speech translation and text machine translation.

## 2 Methodology

In this section, we introduce the model architecture of our system, and describe some of the methods we incorporated into the design and training process.

*Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 202–210
July 13-14, 2023 ©2023 Association for Computational Linguistics

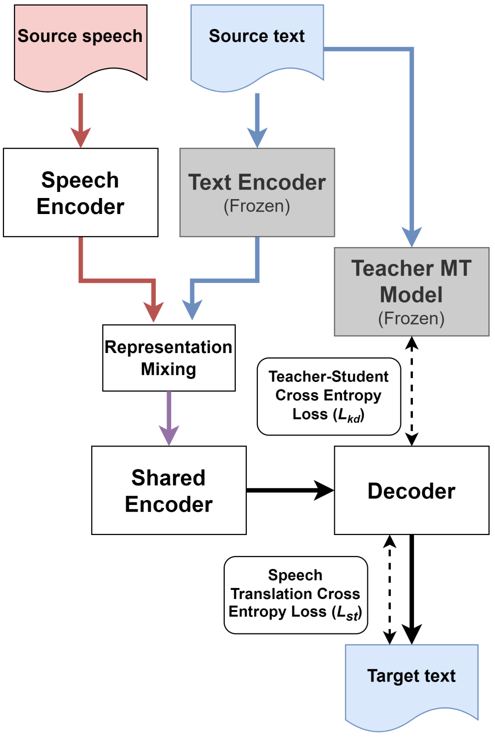
Source speech
Source text
Speech Encoder
Text Encoder (Frozen)
Teacher MT Model (Frozen)
Representation Mixing
Teacher-Student Cross Entropy Loss ($L_{kd}$)
Shared Encoder
Decoder
Speech Translation Cross Entropy Loss ($L_{st}$)
Target text

Figure 1: Our end-to-end ST model architecture

### 2.1 Model

As shown in Fig 1, our end-to-end ST model uses two separate encoders for speech and text, followed by a shared encoder and decoder. As the shared encoder is pretrained on text inputs while the final system has to work with speech inputs, we try to bring speech and text into a shared representation space by devising a training task using mixed speech and text inputs, described in Section 2.2.

Due to limited computational resources, we make use of the allowed pretrained models in the constrained track. The speech encoder is initialized from the WavLM (Chen et al., 2022a) large checkpoint which was pretrained on Libri-Light, GigaSpeech and VoxPopuli data in a self-supervised fashion. WavLM was selected as it includes more data relevant to this year's test set, and showed better performance in our preliminary experiments compared to similar models like HuBERT. DeltaLM base (Ma et al., 2021) was used to initialize the text encoder, shared encoder and decoder sections. Prior to the final ST training, the DeltaLM model was first finetuned on text-to-text MT (described in Section 3.2). The text encoder includes the text and positional embedding layers of DeltaLM and is frozen in the final finetuning stage. The shared encoder encompasses the transformer layers of the DeltaLM encoder.

Given that ST data is commonly provided as a triplet of source speech, source text transcription and target text translations, we leverage both text and speech sources in our proposed architecture. Aside from the audio waveforms processed through the speech encoder, we take as input upsampled tokenized source text by repeating subword tokens according to a pre-calculated ratio given by an alignment system. For data with paired speech and text inputs, we mix representations from the two input encoders through random swapping. Otherwise, unimodal data is processed by their respective encoders and the mixing step is skipped, such as the case during speech-only ST inference. We also recognise that the flexible nature of the architecture allows the use of ASR and MT data as unimodal inputs to further expand the training data and train a multilingual model. However, due to time and computational constraints, this was not explored in this submission and is left as future work.

### 2.2 Representation Mixing

Recent work in unified representation learning of speech and text (Liu et al., 2020; Zhang et al., 2022; Chen et al., 2022b; Fang et al., 2022; Sainath et al., 2023) try to leverage abundant text data to supplement speech-based models. We similarly encourage our model to learn a joint multimodal representation by bringing speech and text inputs into a shared representation space.

To handle the large difference in sequence lengths of audio and text, systems from the literature often upsample text using a trained duration model or a resampling scheme. Here, we utilize offline forced alignment and upsampling to align the speech and text data. Specifically, a pretrained ASR model is used to first force align text transcripts to audio, returning an upsampling ratio between a particular subword and its corresponding speech segment. Each subword token is then repeated up to this ratio before being fed to the text encoder such that the final encoded subword is of the same length as its speech counterpart. The alignment and resampling procedure is described in detail in Section 3.1.

As the shared encoder was pretrained only on text, we hypothesize that the model may better adapt to the downstream speech task by using a mixed speech-text representation compared to training on pure speech inputs. When finetuning the ST model on data with both source speech and text, we feed both the audio and upsampled text tokens into the respective speech and text encoders, then mix the resultant embeddings at the individual subword token level using a fixed probability. In practice, a swapping mask is created before upsampling, with text embeddings being replaced with speech embeddings according to a swapping ratio $\alpha$, where $0 < \alpha < 1$. The tokens and swap mask are upsampled together and passed into the model so that sequences of identical upsampled tokens can be replaced with speech embeddings during the representation mixing step.

### 2.3 Knowledge Distillation

To fully utilize the larger amounts of text-only MT data allowed in the challenge, we train a separate MT model using DeltaLM large. This larger model is then frozen and used as a teacher during fine-tuning of the ST model via negative log-likelihood minimization between the hypotheses generated by both the models, similar to the knowledge distillation method proposed in Tang et al. (2021).

Our overall loss function therefore consists of cross entropy loss between the ground truth and hypothesis produced by the ST system ($L_{st}$) and negative log-likelihood loss between the teacher and student model hypotheses ($L_{kd}$), weighted by $\gamma$ and $\beta$ respectively: $L = \gamma L_{st} + \beta L_{kd}$

## 3 Experimental Setup

### 3.1 Data Preparation

Training data was compiled in accordance to constrained conditions. They can be divided into text and audio-based categories which were used to train the initial MT model and final ST model respectively.

**Text data** Parallel En-De lines were gathered from both MT and ST datasets, seen in Table 1. These were split into in-domain and out-of-domain based on whether the text was derived from TED-like sources. The in-domain sources include a combination of MuST-C v1, v2 and v3 (Cattoni et al., 2021), ST TED (Niehues et al., 2018), and TED 2020 (Reimers and Gurevych, 2020), whereas the out-of-domain sources mostly comprised of OpenSubtitles (Lison and Tiedemann, 2016) and Europarl (Koehn, 2005), but also include CoVoST v2 (Wang et al., 2021b), ELRC-CORDIS News, Europarl-ST (Iranzo-Sánchez et al., 2020), NewsCommentary (Tiedemann, 2012) and Tatoeba. A common pre-processing pipeline was applied to the text data, namely removing any tags and control codes, normalizing bullet points, simplifying punctuation by removing repeats (with the exception of '...') and normalizing whitespace characters. Sentence pairs where source and target differed by more than three times in length were then removed given that they were likely to be misaligned. Finally, the remaining sentences were deduplicated. The out-of-domain data was further filtered using Language-agnostic BERT Sentence Embedding (LaBSE) (Feng et al., 2022). Specifically, we removed sentence pairs with sentence representations lower than 0.5 cosine similarity. We opted not to use any backtranslation data for training since the provided monolingual dataset was found to largely overlap with OpenSubtitles. The final dataset contained 850,003 in-domain and 13,083,335 out-of-domain sentence pairs.

| Dataset | Lines |
|---|---|
| ***in-domain*** | |
| MuST-C v1/v2/v3 | 391K |
| ST TED corpus | 170K |
| TED2020 v1 | 288K |
| ***out-of-domain*** | |
| CoVoST v2 | 300K |
| ELRC-CORDIS News v1 | 111K |
| Europarl v10 | 1.7M |
| Europarl-ST v1.1 | 69K |
| NewsCommentary v16 | 380K |
| OpenSubtitles v2018 apptek | 10.1M |
| Tatoeba v1 | 288K |
| **Total** | 13.9M |

Table 1: Breakdown of text training data. For ST datasets only transcription and translation pairs were used.

**Audio data** Audio data sources include both ASR and ST corpora, listed in Table 2. ASR data consist of Commonvoice (Ardila et al., 2020), Librispeech (Panayotov et al., 2015), TED LIUM (Rousseau et al., 2012), and Vox Populi (Wang et al., 2021a), whereas the ST data include CoVoST (Wang et al., 2021b), Europarl-ST (Iranzo-Sánchez et al., 2020), MuSTC v3 (Cattoni et al., 2021) and ST TED (Niehues et al., 2018). Speech was first converted to mono channel and resampled to 16kHz if required before being saved in FLAC format. Only utterances between 800 to 480,000 samples (i.e. 0.05-30s) were kept and utilized for

| **Dataset** | **Utterances** | **Hours** |
|---|---|---|
| ***ASR data*** | | |
| Commonvoice v11.0 | 949K | 2320 |
| Librispeech v12 | 281K | 960 |
| TED LIUM v3 | 268K | 453 |
| Vox Populi | 177K | 543 |
| ***ST data*** | | |
| CoVoST v2 | 289K | 364 |
| Europarl-ST v1.1 | 68K | 89 |
| MuST-C v3 | 265K | 273 |
| ST TED corpus | 169K | 252 |
| **Total** | 2.47M | 5254 |

Table 2: Breakdown of available audio training data

training. The provided segmentation was used for all speech training data.

To increase the amount of available ST data, we generated additional translations from ASR transcription data using our trained MT model. These synthetic speech-text pairs were used as part of the ST dataset during the finetuning phase.

**Forced alignment and upsampling** To prepare text inputs for mixing with speech inputs, we preprocessed the text by upsampling and aligning it to its corresponding speech features using a pretrained HuBERT ASR model. First, we normalized the transcripts from ASR and ST datasets by deleting non-verbal fillers and converting numbers into their corresponding words. Characters not found among the HuBERT labels were then removed after tokenizing the text. Next, we obtained an alignment between the subword tokens and parallel speech using a pretrained HuBERT large model (Hsu et al., 2021) and, following those alignments, duplicated the input tokens to match the lengths of the speech representation produced by the speech encoder. The frequency of the upsampled text tokens is 50 Hz (equivalent to 16 kHz input audio downsampled 320 times by the WavLM feature extractor).

**Audio segmentation** As segmentation information was not provided in this year's evaluation data, we used the pretrained Supervised Hybrid Audio Segmentation (SHAS) model (Tsiamas et al., 2022) to perform voice activity detection and segmentation on the input audio from the IWSLT test sets. SHAS has been evaluated on MuST-C and mTEDx and shows results approaching manual segmentation.

### 3.2 Training configuration

**On-the-fly audio augmentation** To make our model more robust against the bigger variances in recording quality of the evaluation data introduced this year, we implemented an on-the-fly augmentation pipeline for input audio via the Audiomentations library. In addition to initial utterence cepstral mean and variance normalization (CMVN), we apply gain, seven-band parametric equalization, gaussian noise, time stretch, pitch shift and a lowpass filter, where each augmentation independently has a 20% chance of being utilized. During inference only CMVN is used.

**Machine translation** We finetuned several configurations of DeltaLM base and large for En-De machine translation. DeltaLM base has 12 encoder and six decoder layers, with an embedding dimension of 768 and 12 attention heads per transformer layer. In contrast, DeltaLM large contains 24 encoder and 12 decoder layers, an embedding dimension of 1024 and 16 attention heads per layer.

We used a two phase approach to finetuning. In the first phase, we directly initialized the MT model with DeltaLM pretrained weights and trained on all available MT data. We then continued finetuning only on in-domain data after checkpoint averaging the best five checkpoints from the first phase in terms of BLEU on the validation set that comprised of IWSLT test sets from 2015, 2018, 2019 and 2020, plus MuST-C v3 tst-COMMON split. We also tried progressive finetuning (Li et al., 2020) during the second phase for the DeltaLM base configuration where the depth of the encoder was increased to 16 with four extra randomly initialized layers.

All models were implemented with the Fairseq library. Models were trained with Adam optimization, an inverse square root learning rate (LR) schedule and a peak LR of 1e-4 for the first phase and 1e-5 for the second phase. Label smoothing of 0.1 was also used. Training was carried out on four NVIDIA V100 GPUs. We employ subword tokenization for all text inputs using a Sentencepiece model inherited from the original DeltaLM, with a vocabulary size of 250,000.

**Speech translation finetuning** As described in section 2.1, the end-to-end speech translation model consists of separate speech encoder and text embedding input layers, followed by a shared encoder and decoder. The speech encoder is initialized with a pretrained WavLM large model that contains a seven layer convolutional feature extractor followed by 24 transformer layers. We initialize the text embeddings, shared encoder and decoder layers with the DeltaLM base model previously finetuned for MT. The input text embeddings are frozen throughout the ST finetuning. Meanwhile, the teacher text model was instead initialized with the finetuned DeltaLM large configuration.

Domain tagging has been shown in previous MT (Britz et al., 2017) and ST (Li et al., 2022) work to be effective for domain discrimination and to condition the model towards certain output styles. Given the distinct TED-style outputs of the evaluation data, we introduce '<*indomain*>' and '<*outdomain*>' tags as prefix tokens during decoding to help the model better distinguish the data distribution and style of the in-domain data from the other parts of the dataset.

Similar to the approach employed during MT training, we initially trained the end-to-end ST model on all available ST data, including those synthesized from ASR data. Adam optimization with inverse square root LR schedule and peak LR of 1e-5 was used. A swapping ratio of 0.8 was used during training but 1.0 (i.e. pure speech representation) was used for inference and testing. In the second phase we continued finetuning two separate models with different data splits, while swapping ratio was kept at 1.0. To target the usual TED evaluation data, we trained one with only MuST-C and ST-TED data, while the other also included CoVoST and Europarl to help deal with the more diverse speech patterns found in the ACL and EMPAC parts of the evaluation data (given that no direct development data was provided). We weight the ST loss and knowledge distillation loss with $\gamma = 1$ and $\beta = 0.1$ respectively. Training was carried out on four NVIDIA V100 GPUs for both phases.

## 4 Results and Analysis

We present our experimental results and analyses in this section.

### 4.1 Effect of audio augmentations and pretrained speech encoder

As a preliminary experiment, we tested whether the input audio augmentations have a tangible impact on downstream applications. We finetuned a pretrained WavLM large model together with a six layer transformer decoder for ASR using MuST-C v2 data, with and without input augmentations (Table 3). Furthermore, we trained a HuBERT large model in the same setup to contrast between different pretrained speech encoders.

| Model | WER |
|---|---|
| HuBERT large without augmentation | 7.59 |
| WavLM large without augmentation | 5.86 |
| WavLM large with augmentation | 5.56 |

Table 3: ASR results on MuST-C v2 tst-COMMON.

As observed, the audio augmentations were found to be beneficial, leading to a reduction of WER by 0.3. We found WavLM large together with augmentations to perform the best overall and so was adopted for the rest of the experiments.

### 4.2 Machine translation results

The results of the MT systems for En-De are shown in Table 4, separated into the full-domain training phase and the in-domain training phase. Performance was evaluated using cased BLEU with default `SacreBLEU` options (13a tokenization).

It was evident that the continuous finetuning with in-domain data improves performance on similar datasets such as past year IWSLT evaluation data or MuST-C. While the DeltaLM large models achieved the best results, the base variants were not far behind and generally performed within 1 BLEU score of the former. However, we found no added benefit to the progressively finetuned models. It may be the case that the extra representative power of the expanded encoder layers were not beneficial at the relatively small scale of the in-domain data, which was less than 1 million sentence pairs. Some training runs produced better scores by checkpoint averaging the best five checkpoints. Nevertheless, the improvement was not consistent throughout all test sets.

An ensemble of model variants 6 and 9 further improved the BLEU scores on the test sets. We utilize the ensemble model to generate translations from ASR transcriptions to supplement the available ST data. The best checkpoint for DeltaLM base (model 5) and DeltaLM large (model 9) were subsequently used to initialize the end-to-end ST model and teacher text model respectively for the final finetuning.

|  | Model | BLEU |  |  |  |
|---|---|---|---|---|---|
|  |  | tst2020 | tst2019 | MuST-C v3 | MuST-C v2 |
|  | ***full-domain*** |  |  |  |  |
| 1 | base (best) | 31.76 | 28.81 | 33.11 | 33.77 |
| 2 | base (avg 5) | 32.86 | 29.43 | 34.05 | 34.67 |
| 3 | large (best) | 31.82 | 29.01 | 33.20 | 34.21 |
| 4 | large (avg 5) | 32.52 | 29.54 | 33.65 | 34.68 |
|  | ***in-domain*** |  |  |  |  |
| 5 | base (best) | 33.64 | 30.67 | 35.29 | 35.99 |
| 6 | base (avg 5) | 33.73 | 30.64 | 35.26 | 36.11 |
| 7 | base-progressive (best) | 33.40 | 30.51 | 34.25 | 34.83 |
| 8 | base-progressive (avg 5) | 33.26 | 30.48 | 34.37 | 35.09 |
| 9 | large (best) | **34.44** | **31.47** | 35.60 | 36.26 |
| 10 | large (avg 5) | 34.32 | 31.42 | **35.89** | **36.48** |
|  | Ensemble (6 + 9) | **34.91** | **31.77** | **36.14** | **36.93** |

Table 4: MT results on various test sets.

|  | Model | BLEU |  |  |  |  |
|---|---|---|---|---|---|---|
|  |  | tst2020 | tst2019 | MuST-C v3 | MuST-C v2 | CoVoST v2 |
|  | ***in-domain*** |  |  |  |  |  |
| 1 | base (best) | **25.70** | **22.68** | **30.29** | **30.56** | 27.92 |
| 2 | base (avg 5) | 24.81 | 22.25 | 29.98 | 30.29 | 28.11 |
|  | ***extended-domain*** |  |  |  |  |  |
| 3 | base (best) | 22.80 | 21.17 | 29.33 | 29.50 | 28.63 |
| 4 | base (avg 3) | 23.21 | 21.20 | 29.61 | 29.95 | **29.30** |
|  | Ensemble (1 + 2 + 4) | 24.99 | 22.64 | 29.99 | 30.35 | 29.13 |

Table 5: ST results on various test sets.

### 4.3 Speech translation results

Results from our end-to-end ST systems for English speech to German text are provided in Table 5. As mentioned in section 3.2, we trained two models during the second ST finetuning phase, which are labelled here as 'in-domain', targeting more TED-like inputs, and 'extended-domain' for other input domains. As reference segmentation information was not provided for IWSLT-tst2019 and IWSLT-tst2020 test sets, we used SHAS to segment the audio. The translation hypotheses were then compared to the references provided by using the `SLT.KIT` evaluation script listed on the challenge website, that uses the mwerSegmenter resegmentation tool and the BLEU calculation script from the `Moses` toolkit. The provided segmentation and `SacreBLEU` were utilized for the other test sets.

Comparing CoVoST against the rest of the test sets reveals that the in-domain and extended-domain models show better results in their respective domain specializations, as was intended. We unexpectedly get poor results on IWSLT-tst2019 and IWSLT-tst2020 relative to last year's best performing entries, which may point to a weakness in the current training procedure, a domain mismatch since training was more aligned to MuST-C, or compounded errors due to resegmentation. We plan to investigate the reasons more precisely in future papers. The ensemble model of variants 1, 2 and 4 shows balanced performance across both domains, and we submit this as our primary submission, with variants 1 and 4 as our contrastive systems.

## 5 Conclusion

In this paper we outline our proposed end-to-end system that incorporates pretrained models trained on large-scale audio and text data to enhance the ST performance. The system underwent several stages of additional pretraining followed by finetuning for the downstream speech translation task. We explored several techniques including audio augmentation, domain tagging, knowledge distillation and model ensemble to improve the system's performance. We utilize both speech and text inputs, and propose a mixing procedure to unify representations from both modalities to not only increase the amount of available training data but also better adapt the model to downstream speech tasks. We plan to carry out more experiments to further explore the effect of modality mixing and improve the performance of such models for speech-to-text tasks.

## Acknowledgements

This work was supported by the A*STAR Computational Resource Centre through the use of its high performance computing facilities.

## 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 *International Workshop on Spoken Language Translation (IWSLT 2022)*, pages 98–157. Association for Computational Linguistics.

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

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

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. Association for Computational Linguistics.

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning and speech and audio precessing*, Barcelona, Spain.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126, Copenhagen, Denmark. Association for Computational Linguistics.

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 & Language*, 66:101155.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022a. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro J. Moreno, Ankur Bapna, and Heiga Zen. 2022b. MAESTRO: Matched speech text representations through modality matching. In *Interspeech*, pages 4093–4097.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Interspeech*, pages 2426–2430.

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

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for automatic speech recognition, experiments on a fieldwork corpus of japhug (trans-himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland. Association for Computational Linguistics.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: How much can a bad teacher benefit ASR pre-training? In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537.

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

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 *28th International Conference on Computational Linguistics*, pages 3520–3533, Barcelona, Spain.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bei Li, Ziyang Wang, Hui Liu, Yufan Jiang, Quan Du, Tong Xiao, Huizhen Wang, and Jingbo Zhu. 2020. Shallow-to-deep training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 995–1005, Online. Association for Computational Linguistics.

Yinglu Li, Minghan Wang, Jiaxin Guo, Xiaosong Qiao, Yuxia Wang, Daimeng Wei, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang, and Ying Qin. 2022. The HW-TSC's offline speech translation system for IWSLT 2022 evaluation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 239–246, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. DeltaLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders.

David Fraile Navarro, Mark Dras, and Shlomo Berkovsky. 2022. Few-shot fine-tuning SOTA summarization models for medical dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 254–266, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Jan Niehues, Rolando Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The IWSLT 2018 evaluation campaign. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 2–6, Brussels. International Conference on Spoken Language Translation.

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)*, pages 5206–5210.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi

Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2012. TED-LIUM: an automatic speech recognition dedicated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 125–129, Istanbul, Turkey. European Language Resources Association (ELRA).

Tara N Sainath, Rohit Prabhavalkar, Ankur Bapna, Yu Zhang, Zhouyuan Huo, Zhehuai Chen, Bo Li, Weiran Wang, and Trevor Strohman. 2023. JOIST: A joint speech and text streaming model for asr. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 52–59. IEEE.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4252–4261, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ioannis Tsiamas, Gerard I. Gállego, Carlos Escolano, José Fonollosa, and Marta R. Costa-jussà. 2022. Pretrained speech encoders and efficient fine-tuning methods for speech translation: UPC at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 265–276, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. 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 993–1003, Online. Association for Computational Linguistics.

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

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In *Interspeech*, pages 2625–2629.

Ziqiang Zhang, Sanyuan Chen, Long Zhou, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun Gong, Lirong Dai, Jinyu Li, and Furu Wei. 2022. SpeechLM: Enhanced speech pre-training with unpaired textual data. *arXiv preprint arXiv:2209.15329*.