The BIGAI Offline Speech Translation Systems for IWSLT 2023 Evaluation

Zhihang Xie
Beijing Institute of General Artificial Intelligence
zhihangxie@gmail.com

Abstract

This paper describes the BIGAI's submission to IWSLT 2023 Offline Speech Translation task on three language tracks from English to Chinese, German and Japanese. The end-to-end systems are built upon a Wav2Vec2 model for speech recognition and mBART50 models for machine translation. An adapter module is applied to bridge the speech module and the translation module. The CTC loss between speech features and source token sequence is incorporated during training. Experiments show that the systems can generate reasonable translations on three languages. The proposed models achieve BLEU scores of 22.3, 10.7 and 33.0 on tst2023 en→de, en→ja and en→zh TED datasets. It is found that the performance is decreased by a significant margin on complex scenarios like presentations and interviews.

1 Introduction

Speech translation aims to solve the problem of translating speech waveform in source language into written text in target language. Cascade systems decompose the problem into automatic speech recognition (ASR) to transcribe source speech into source text and machine translation (MT) to translate source text into target text (Wang et al., 2021b; Zhang et al., 2022a). It is clear that such architecture has the advantage of ensembling results from state-of-the-art (SOTA) ASR models and MT models and the disadvantages of accumulating subsystem errors and discarding paralinguistic features. Recent end-to-end speech translation (E2E ST) systems have shown the potential to outperform cascade systems (Hrinchuk et al., 2022; Shanbhogue et al., 2022). However, due to the lack of high-quality parallel training data, it is difficult to quantify the gap between the two categories.

Inspired by Zhang et al.'s (2022b) work, this submission explores various techniques to address problems in speech translation. 1) Perform fine-grained data filtering by calculating WERs for speech data and alignment scores for translation data. 2) Apply a straightforward split-and-merge method to split long audio clips into short segments. 3) Employ a three-stage training strategy to concatenate the finetuned speech module and the translation module. 4) Incorporate connectionist temporal classification (CTC) loss to leverage the divergence between speech features and source token sequences (Graves et al., 2006). Experiments are carried out to perform speech translation at sentence level and corpus level. The performance of the three PT36 models is finally evaluated on the tst2023 datasets with automatic metrics.

The rest of this paper is organized as follows. Section 2 describes how speech data and translation data are processed in the experiments. Section 3 explains how finetuned models are assembled to perform speech translation on all three languages. Section 4 illustrates experiment setups, results and analysis. Section 5 concludes the submission.

2 Data Processing

2.1 Speech Corpora

Under the constrained condition, there are five speech datasets used to train ASR models, namely LibriSpeech (Panayotov et al., 2015), Mozilla Common Voice v11.0 (Ardila et al., 2019), MuSTC (Cattoni et al., 2021), TEDLIUM v3 (Hernandez et al., 2018) and VoxPopuli (Wang et al., 2021a). Statistics on each dataset are shown as Table 1. Note that only the MuSTC datasets are used to train speech translation systems on the three language tracks, English-to-German (en→de), English-to-Japanese (en→ja) and English-to-Chinese (en→zh).

In general, all speech files are unified to single channel 16kHz format. During training, utterances shorter than 0.2s or longer than 20s are removed. An extra W2V model with 24 Transformer layers is finetuned on the LibriSpeech dataset and calculates WER scores by performing CTC greedy decoding

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 123–129
July 13-14, 2023

Table 1: Statistics on speech datasets

Dataset	Utterances	Hours
CommonVoice	948,736	1,503.28
LibriSpeech	281,241	961.05
MuSTC en→de v3	269,851	440.18
MuSTC en→ja v2	328,637	541.04
MuSTC en→zh v2	358,852	596.20
TEDLIUM	268,263	453.81
VoxPopuli	182,466	522.60
Total, loaded	2,638,046	5,018.17
Total, filtered	2,528,043	4,713.35

Table 2: Statistics on translation datasets

Dataset	en→de	en→ja	en→zh
MuSTC	0.269m	0.328m	0.358m
OpenSubtitles	22.512m	2.083m	11.203m
Commentaries	0.398m	0.002m	0.322m
Total	23.181m	2.414m	11.884m

at character level on the other speech datasets, so utterances with WER scores over 75% are discarded as well. As a result, the speech corpora contains nearly 2.53 million valid utterances with the total duration of 4,713.35 hours.

2.2 Translation Corpora

In addition to the MuSTC datasets, the OpenSubtitles v2018 (Lison et al., 2018) and the News Commentaries v16 (Farhad et al., 2021) datasets are added up to train MT models. Statistics on these translation datasets are described as Table 2. Since translation pairs do not perfectly match all the time, the translation quality is measured by the *fast-align*[1] toolkit in terms of the percentage of aligned words. Word sequences are obtained by splitting English texts and German texts using whitespaces and converting Chinese texts and Japanese texts into character sequences. Parallel training examples are filtered out if: 1) the source sentence contains more than 150 words; 2) the alignment score in either forward translation or backward translation is lower than a certain threshold.

3 Method

3.1 Pretrained Models

Two state-of-the-art models pretrained with self-supervised objectives are employed as base models for downstream tasks with labeled data, namely the *wav2vec2-large-960h-lv60-self*[2] model for speech recognition and the *mbart-large-50-one-to-many-mmt*[3] model for machine translation.

The W2V models (Baevski et al., 2020) are trained with contrastive learning to distinguish whether two transformations of convolution features result in similar latent representations. The first transformation is to learn high-level contextual speech representations through a sequence of Transformer layers (Vaswani et al., 2017). The second transformation is to create discrete targets for self-training by the quantization module. The best partial representations chosen from multiple codebooks with the Gumbel softmax (Jang et al., 2016) are concatenated and transformed to a quantized representation with a linear layer.

The mBART25 models (Liu et al., 2020) are Transformer-based encoder-decoder models that are pretrained on monolingual sentences from many languages and finetuned with parallel translation data on 25 languages. The pretraining objective is a denoising loss so that the model learns to reconstruct corrupted sentences to their original forms. The noise function randomly masks 35% of input sentences in consecutive spans and permutes sentence orders for document-level MT if multiple sentences are given. The mBART50 models (Tang et al., 2020) extend embedding layers with an extra set of 25 languages and are finetuned on translation task from English to the other 49 languages.

3.2 Finetuned Models

The two base models result in one ASR model, three MT models and three E2E ST models. Written texts in the four languages are tokenized into subword tokens in byte-pair encoding (BPE) using the SentencePiece toolkit (Kudo and Richardson, 2018). The tokenizer is inherited from the mBART50 model with a multilingual configuration by prepending language symbols and the total number of BPE tokens in the vocabulary is 250k.

For speech recognition, the finetuned model (ASR12) takes the first 12 Transformer layers from the base model. An adapter module (Li et al., 2020; Shanbhogue et al., 2022) compresses the feature vectors by a factor of eight, which consists of three one-dimensional convolution layers with a stride of two. A linear layer transforms the compressed representations into output probabilities.

[1] https://github.com/clab/fast_align

[2] facebook/wav2vec2-large-960h-lv60-self

[3] facebook/mbart-large-50-one-to-many-mmt

For end-to-end speech translation, the models have similar architecture as the PT36 models in Zhang et al.'s (2022b) work instead of the PT48 models to reduce computational complexity. Within a PT36 model, the speech module and the translation module are initialized with the ASR12 model and the MT24 model respectively. The adapter module that connects the two modules is not trained from random initialization, because it has been trained with the ASR12 model on the first stage. The training loss combines the cross entropy loss for machine translation and the CTC loss for speech recognition with a hyperparameter to balance the weights between the two losses.

3.3 Speech Resegmentation

Past years' systems (Anastasopoulos et al., 2021; Antonios et al., 2022) have proved that speech resegmentation has a great impact on the translation performance at corpus level. During evaluation, audio clips are splitted into segments with a simple two-stage strategy using the *WebRTCVAD*[4] toolkit. On the split stage, long audios are processed with three-level settings of aggressiveness modes increasing from 1 to 3 and frame sizes decreasing from 30ms to 10ms. In this way, most segments are no longer than a maximum duration dur_{max} and the outliers are further segmented into $\lfloor \frac{duration}{0.75\times\theta} \rfloor$ chunks brutally. On the merge stage, consecutive segments are merged into final segments no shorter than a minimum duration dur_{min}.

4 Experiments

4.1 Settings

All the models are implemented with the SpeechBrain toolkit (Ravanelli et al., 2021). The total number of parameters in a PT36 model is about 794.0M, 183.2M in the speech module and 610.9M in the translation module. The feature extractor processes speech waveform with seven 512-channel convolution layers, in which kernel sizes and strides are [10,3,3,3,3,2,2] and [5,2,2,2,2,2,2]. There are 12 Transformer layers with 16 attention heads, model dimension of 1024 and inner dimension of 4096 in speech encoder, text encoder and decoder. The adapter module has three Conv1D layers with kernel sizes and strides being [3,3,3] and [2,2,2].

On the first stage, the ASR12 model is finetuned on the speech corpora using 16 NVIDIA A100 GPUs for 21 epochs with the batch size of 3 and the update frequency of 8. The parameters in the Wav2Vec2 module and the linear layer are separately optimized by the Adam optimizer (Kingma and Ba, 2014). The learning rates are initialized with $1e^{-4}$ and $4e^{-4}$ with the annealing factors set to 0.9 and 0.8. The learning rates are updated based on the improvement of the training losses between the previous epoch and the current epoch. During training, speech waveform is perturbed with a random speed rate between 0.9 and 1.1 and speech features are augmented with the SpecAugment technique (Park et al., 2019).

[4]https://github.com/wiseman/py-webrtcvad

Table 3: WER scores on test speech datasets

LibriSpeech	TEDLIUM	MuSTC
27.23	32.17	34.73

Table 4: BLEU scores on tst-COMMON datasets

Model	**en→de**	**en→ja**	**en→zh**
MT24	31.04	14.74	22.80
+ finetune	33.00	17.11	23.44
PT36	26.45	14.28	19.65

On the second stage, three MT24 models are finetuned on the translation corpora with the batch size of 12 and the update frequency of 4. The en→de MT24 model is trained using 8 A100 GPUs for 2 epochs and the other two models are trained using 4 A100 GPUs for 6 epochs and 3 epochs. The model parameters are optimized with the Adam optimizer and the initial learning rates are set to $5e^{-5}$ with the annealing factor set to 0.9.

On the third stage, three PT36 models are finetuned on the corresponding MuSTC datasets, each of which is trained using 4 A100 GPUs for 10 epochs with the batch size of 12 and the update frequency of 4. The learning rates are initialized to $3e^{-5}$ for the W2V module and $5e^{-5}$ for the mBART module with the annealing factors set to 0.9. The loss weights are set to 0.1 for the ASR module and 0.9 for the MT module since the performance of the ASR module is not good enough.

4.2 Speech Recognition

Table 3 lists WER scores on test speech datasets, where 34.73% is the average WER score of the three MuSTC datasets. Obviously, the performance of the ASR12 model is much worse than that of other systems (Zhang et al., 2022b; Wang et al., 2021b) with WERs around 10%. Due to extremely large vocabulary size, the model requires a long

Table 5: Statistics on short segments in the tst2020 dataset with different dur_{min} and dur_{max} settings.

id	dur_{min}	dur_{max}	**level1**	**level2**	**level3**	**brutal**	**split**	**merge**
1	5	20	3,473	342	449	185	4,449	2,621
2	10	30	3,568	146	258	69	4,041	1,699
3	15	60	3,624	35	115	0	3,774	1,237
4	20	90	3,635	9	73	0	3,717	970

Table 6: BLEU scores on calculated on past years' IWSLT en→de test sets with hypotheses automatically resegmented by the *mwerSegmenter* toolkit (Ansari et al., 2021) based on source transcriptions and target translations.

id	dur_{min}	dur_{max}	**2010**	**2013**	**2014**	**2018**	**2019**	**2020**	**Δ**
1	5	20	21.44	27.37	25.87	12.41	18.95	20.14	21.03
2	10	30	23.79	30.33	28.53	16.29	21.22	22.60	+2.76
3	15	60	24.17	31.16	29.23	**18.38**	22.04	23.46	+3.71
4	20	90	**24.31**	**31.73**	**30.05**	17.98	**22.16**	**23.55**	**+3.93**

time to train. As a result, the model is still far from converge at the time of this submission.

4.3 Sentence-level Translation

The *tst-COMMON* datasets are used to evaluate the translation performance at sentence level and the BLEU scores are calculated by the *SacreBLEU*[5] toolkit, where Japanese texts are tokenized by the *Mecab*[6] morphological analyzer and Chinese texts are tokenized into characters. The BLEU scores on the three datasets are listed in Table 4.

For machine translation, compared with the base MT24 models, the performance of the fine-tuned MT24 models is improved by 1.96 (~6.3%), 2.37 (~16.1%) and 0.64 (~2.8%) BLEU scores on en→de, en→ja and en→zh translations. It indicates that adding out-of-domain corpora like OpenSubtitles and NewsCommentaries is able to boost the machine translation quality.

For speech translation, compared with the fine-tuned MT24 models, the performance of PT36 models is degraded by a large margin with 6.55 (~19.8%), 2.83 (~16.5%) and 3.79 (~16.2%) BLEU scores on en→de, en→ja and en→zh translations. Compared with the base MT24 models, the gaps are still relatively large with 4.59 (~14.8%), 0.46 (~3.1%) and 3.15 (~13.8%) BLEU scores.

4.4 Corpus-level Translation

The translation performance of en→de PT36 model is further evaluated on past years' test datasets with challenging scenarios. To keep consistency, all test audios are resegmented using the method described in Section 3.3. Statistics on short segments in the tst2020 dataset are shown as Table 5. It is noticed that the number of brutal segments is decreased to zero when dur_{min} is set to more than 15s.

Table 6 lists BLEU scores on past years' test datasets with different dur_{min} and dur_{max} settings. It is found that the performance is boosted as the segment duration gets longer, which means that more contextual information is provided to the model. When dur_{min} and dur_{max} are set to 20s and 90s, the best BLEU scores are achieved on most test datasets with an increment of 3.93 (~18.7%) mean BLEU score. Further investigation on long audio segments finds that avoiding brutal segmentation is another factor of such improvement. Comparing experiment 2 and experiment 3, the mean BLEU score is increased by 0.95 (~3.9%) points, when the number of brutal segments is decreased from 69 to 0. Comparing experiment 3 and experiment 4, the mean BLEU score is merely increased by 0.22 (~0.8%) points.

4.5 Submissions

The three PT36 models are finally evaluated on tst2023 datasets (Agarwal et al., 2023) with more challenging scenarios like presentations and interviews. Test audios are resegmented with dur_{min} and dur_{max} set to 20s and 90s. Official metrics are presented as Table 7 for en→de datasets, Table 8 for en→ja datasets and Table 9 for en→zh datasets.

Comparing the performance between in-domain TED datasets and out-of-domain ACL datasets, the BLEU scores are decreased by 2.7 (~12.1%), 0.3 (~2.8%) and 5.6 (~16.9%) points on en→de, en→ja and en→zh translations. Noticeably, the perfor-

[5]https://github.com/mjpost/sacrebleu
[6]https://github.com/taku910/mecab

Table 7: Official metrics on the tst2023 en→de subsets with hypotheses automatically resegmented by the *mwerSegmenter* toolkit (Ansari et al., 2021) based on source transcriptions and target translations.

TED							**ACL**			**Sub**		
Comet		BLEU			chrF		Comet	BLEU	chrF	Comet	BLEU	chrf
ref2	ref1	ref2	ref1	both	ref1	ref2						
0.7128	0.7055	22.3	19.3	27.4	0.49	0.50	0.6295	19.6	0.46	0.3555	11.5	0.45

Table 8: Official metrics on the tst2023 en→ja subsets.

TED					**ACL**	
Comet		BLEU			Comet	BLEU
ref2	ref1	ref2	ref1	both		
0.7201	0.7228	10.7	13.2	16.8	0.6769	10.4

mance is almost halved (~48.4%) with only 11.5 BLEU scores on the en→de Sub dataset. The results indicate that the proposed PT36 models have inadequate abilities of handling non-native speakers, different accents, spontaneous speech and controlled interaction with a second speaker.

5 Conclusion

In conclusion, this paper describes the end-to-end speech translation systems for IWSLT 2023 offline tasks. Built upon pretrained models, the systems are further trained on large amount of parallel data using the three-stage finetuning strategy. The PT36 model consists of an ASR12 module with an adapter module for ASR and an MT24 module for MT. The training loss sums up the CTC loss for ASR and the cross entropy loss for MT. Experiments demonstrate that the proposed methods have the potential to achieve a reasonable performance. However, due to limited resources, some modules has not well trained, which has a negative impact on subsequent tasks. Therefore, the end-to-end models still underperform SOTA systems.

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

Ebrahim Ansari, Ondřej Bojar, Barry Haddow, and Mohammad Mahmoudi. 2021. Sltev: Comprehensive evaluation of spoken language translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 71–79.

Anastasopoulos Antonios, Barrault Loc, Luisa Bentivogli, Marcely Zanon Boito, Bojar Ondřej, Roldano Cattoni, Currey Anna, Dinu Georgiana, Duh Kevin, Elbayad Maha, et al. 2022. Findings of the iwslt 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1912.06670*.

Table 9: Official metrics on the tst2023 en→zh subsets.

TED					**ACL**	
Comet		BLEU			Comet	BLEU
ref2	ref1	ref2	ref1	both		
0.7428	0.7014	33.0	23.3	38.6	0.6534	27.4

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.

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.

Akhbardeh Farhad, Arkhangorodsky Arkady, Biesialska Magdalena, Bojar Ondřej, Chatterjee Rajen, Chaudhary Vishrav, Marta R Costa-jussa, España-Bonet Cristina, Fan Angela, Federmann Christian, et al. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. Tedlium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pages 198–208. Springer.

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 *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 225–231.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2020. Multilingual speech translation with efficient finetuning of pretrained models. *arXiv preprint arXiv:2010.12829*.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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. IEEE.

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. *arXiv preprint arXiv:1904.08779*.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Akshaya Shanbhogue, Ran Xue, Ching Yun Chang, and Sarah Campbell. 2022. Amazon alexa ai's system for iwslt 2022 offline speech translation shared task. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 169–176.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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. *arXiv preprint arXiv:2101.00390*.

Minghan Wang, Yuxia Wang, Chang Su, Jiaxin Guo, Yingtao Zhang, Yujia Liu, Min Zhang, Shimin Tao, Xingshan Zeng, Liangyou Li, et al. 2021b. The hwtsc's offline speech translation systems for iwslt 2021 evaluation. *arXiv preprint arXiv:2108.03845*.

Weitai Zhang, Zhongyi Ye, Haitao Tang, Xiaoxi Li, Xinyuan Zhou, Jing Yang, Jianwei Cui, Pan Deng, Mohan Shi, Yifan Song, et al. 2022a. The ustcnelslip offline speech translation systems for iwslt 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 198–207.

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 preprint arXiv:2206.05777*.