GMU Systems for the IWSLT 2023 Dialect and Low-resource Speech Translation Tasks

Jonathan Kabala Mbuya
George Mason University
jmbuya@gmu.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Abstract

This paper describes the GMU Systems for the IWSLT 2023 Dialect and Low-resource Speech Translation Tasks. We submitted systems for five low-resource tasks and the dialectal task. In this work, we explored self-supervised pre-trained speech models and finetuned them on speech translation downstream tasks. We use the Wav2vec 2.0, XLSR-53, and Hubert as self-supervised models. Unlike Hubert, Wav2vec 2.0 and XLSR-53 achieve the best results when we remove the top three layers. Our results show that Wav2vec 2.0 and Hubert perform similarly with their relative best configuration. In addition, we found that Wav2vec 2.0 pre-trained on audio data of the same language as the source language of a speech translation model achieves better results. For the low-resource setting, the best results are achieved using either the Wav2vec 2.0 or Hubert models, while XLSR-53 achieves the best results for the dialectal transfer task. We find that XLSR-53 does not perform well for low-resource tasks. Using Wav2vec 2.0, we report close to 2 BLEU point improvements on the test set for the Tamasheq-French compared to the baseline system at the IWSLT 2022.

1 Introduction

Recently, speech-to-text translation (S2T) has received a lot of focus in the community where neural, end-to-end approaches outperform traditional statistical approaches (Weiss et al., 2017). Recent neural approaches to S2T have shown superior performance on this task (Fang et al., 2022; Tang et al., 2022). Despite the success of neural approaches to S2T, data scarcity is one of the significant challenges, given that neural networks require hundreds to thousands of hours of labeled data to train a good speech translation model (Sperber and Paulik, 2020). This makes developing such S2T models challenging, especially for low-resource languages.

The IWSLT 2023 Low-resource and dialectal shared tasks (Agarwal et al., 2023) give the possibilities for researchers to find innovative ways to develop speech translation systems for languages with limited data. Unlike previous years, this year noticed an addition of more low-resource languages language pairs (up to 6) in addition to a dialect language pair.

This paper describes the GMU submissions to the low-resource and dialectal tasks. Our systems use self-supervised pre-trained speech models to improve speech translation models' performance in general, particularly for low-resource languages. Self-supervised pre-training is possible because unlabeled data (i.e., audio or text) can be obtained easier compared to labeled data. Previous research has addressed using self-supervised speech models for speech translation (Wu et al., 2020; Nguyen et al., 2020; Popuri et al., 2022). However, these prior work did not consider exploring the impact of different layers of these self-supervised models to maximize the performance of S2T models.

In this paper, we consider three self-supervised speech models: Wav2vec 2.0 (Baevski et al., 2020), XLSR (Conneau et al., 2020) and Hubert (Hsu et al., 2021). Following the discussion by Pasad et al. (2022), we experimented to study the impact of removing the top n layers of these models for the speech translation task. By removing the last three layers of the Wav2vec 2.0 model, we achieve more than 2 BLEU improvement (8.03) on the blind test set for the Tamasheq-French pair compared to the best system submitted to the IWSLT 2022 low-resource shared task (Anastasopoulos et al., 2022; Zanon Boito et al., 2022). Similarly, using a pre-trained XLSR-53, we achieved a BLEU score of 16.3 on the Tunisian Arabic-to-English language pair without using the transcripts.

2 Task Descriptions

We are concerned with developing speech translation models in low-resource and dialectal tracks. Each track poses distinct challenges. The low-

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 269–276
July 13-14, 2023 ©2023 Association for Computational Linguistics

Language Pairs	Language Code	Train Set Hours	Shared Task
Irish to English (Agarwal et al., 2023)	ga-eng	11	Low-resource
Marathi to Hindi (Agarwal et al., 2023)	mr-hi	15.3	Low-resource
Pashto to French (ELRA)	pus-fra	61	Low-resource
Tamasheq to French (Boito et al., 2022)	tmh-fra	17	Low-resource
Quechua to Spanish	que-spa	1.6	Low-resource
Tunisian Arabic to English	aeb-eng	160	Dialectal

Table 1: Language pair details used in our experiments.

resource setting has limited training data, while the dialectal one lacks standard orthography and formal grammar. Both shared tasks allowed the submission of models trained under constrained and unconstrained conditions. In the constrained condition, models are only trained on data provided by the organizers. In contrast, models in the unconstrained condition can be trained on any available resources, including pre-trained models.

2.1 Data

Six low-resource languages were made available, and one dialectal. However, due to data quality issues (see Section 5) we do not report results on the Maltese to English task. Table 1 shows the data details for each language pair. The organizers shared additional data for specific languages, including data for automatic speech recognition (ASR) and machine translation (MT). However, our approach used the data described in table 1. The exception is for Tamasheq-French, where we used the provided 234 hours of unlabeled Tamasheq audio to pre-train a self-supervised speech model.

For the unconstrained condition, we used data from MUST-C[1] (Di Gangi et al., 2019) to train an ASR model for which we used its encoder to initialize the speech translation training. We used publicly available pre-trained self-supersized models (Wav2vec 2.0 (Baevski et al., 2020), XLSR-53 (Conneau et al., 2020), and Hubert (Hsu et al., 2021)). The Wav2vec 2.0 and Hubert checkpoints we used were trained on the Librispeech 960hr English-only data (Panayotov et al., 2015), while XLSR-53 was trained on 53 different languages (Conneau et al., 2020). No source language of all language pairs appears in any self-supervised models except Tamasheq-French, where we pre-trained the Wav2vec 2.0 model we used for Tamasheq-French was pre-trained on Tamasheq audio-only data. Though Tunisian Arabic is not part of the XLSR-53, the XLSR-53 contains Arabic data that may be related to Tunisian Arabic.

3 Proposed Methods

Our methods consist of three different architectures. The first is an end-to-end based transformer-based architecture (E2E) trained on only provided data. The second architecture, which we name E2E-ASR, is the same as the first, except that we initialize the encoder with an ASR encoder. The third architecture uses self-supervised speech models as an encoder and a transformer-based decoder. We used three different self-supervised models, Wav2vec 2.0, XLSR-53, and Hubert, and refer to these architectures as W2V-E2E, XLSR-E2E, and Hubert-E2E respectively.

We used the Fairseq ST (Wang et al., 2020) framework for all our experiments and modified this framework to accommodate our new custom model architectures.

3.1 End-to-end and End-to-end with ASR

For End-to-end (E2E) architecture, we used a transformer-based encoder-decoder architecture (Vaswani et al., 2017) (`st_tranformer_s`) as implemented in the Fairseq S2T framework (Wang et al., 2020). The E2E architecture consists of a 6-block transformer encoder and a 6-block transformer decoder and is optimized using the cross-entropy loss with label smoothing. We used this model architecture to train the model for the primary constrained category (**primary-constrained**).

The End-to-end with ASR (E2E-ASR) architecture, similar to (Stoian et al., 2019) and (Bansal et al., 2019), uses the same architecture as the E2E. The difference is that we use a pre-trained ASR model to initialize its encoder. We used a transformer-based architecture identical to the one

[1] English to French only

for E2E to train the ASR on the English data of the English-French Must-C dataset (Di Gangi et al., 2019). We chose this architecture for the ASR model to facilitate the transfer of the ASR encoder weights to initialize the E2E-ASR encoder. The decoder of the E2E-ASR was randomly initialized and did not use the ASR decoder because it was trained on a different language with a different vocabulary. We used this model architecture to train the model for the second contrastive unconstrained category (**contrastive2-unconstrained**).

3.2 Self-Supervised Approaches

The self-supervised approach uses self-supervised speech models as acoustic encoders with a transformer-based decoder. The use of these self-supervised models is motivated by the scarcity of data in the low-resource setting. However, we found these models useful even for the dialectal task. The self-supervised architecture is illustrated in figure 1.

We used three different self-supervised models, Wav2vec 2.0, XLSR-53, and Hubert, which correspond to the respective architectures W2V-E2E, XLSR-E2E, and Hubert-E2E. These models consist of a feature encoder and a context network. The feature encoder has seven temporal convolution blocks, and the context network consists of several transformer blocks. The Wav2vec 2.0 and Hubert models used in our experiments have 12 transformer blocks, whereas the XLSR-53 has 24.[2]

We use these self-supervised models as encoders following the traditional encoder-decoder model architecture. The decoder consists of a transformer network with six layers preceded by a linear layer.

3.2.1 Using Wav2vec 2.0 and XLSR-53

Instead of using all the layers of the context network for the Wav2vec 2.0 and XLSR-53 models, we explored the impact of removing the top *n* most layers. The exploration of removing the top layers was inspired by Pasad et al. (2022), who analyzed self-supervised speech models and measures the acoustic, phonetic, and word-level properties encoded in individual layers of the context network. For Wav2vec 2.0 and XLSR, the analyses show that the initial and the final layers are more similar to the inputs than the intermediate layers. Instead of re-initializing the top *n* layers and then fine-tuning these models on a downstream task as done in Pasad et al. (2022), we explored the idea of removing these layers and then fine-tuning the modified model on a downstream task. Through a series of experiments, we found that removing the last three layers for the Wav2vec 2.0 and XLSR-53 models yields the highest BLEU score.

We found the Wav2vec 2.0 helpful for the low-resource languages, while the XLSR-53 was more beneficial for the dialectal language. Therefore, we used the Wav2vec 2.0 for the primary unconstrained category (**primary unconstrained**) for the low-resource task. The XLSR-53 was used as the primary unconstrained category (**primary unconstrained**) for the dialectal transfer task.

The Wav2vec 2.0 we used for all the low-resource languages (except Tamasheq-French) was trained on the English raw audio of the Librispeech 960hr data (Panayotov et al., 2015). However, due to the availability of Tamasheq raw audio, we also trained a Wav2vec 2.0 model on Tamasheq raw audio that used this model on the Tamasheq to French language pair. The XLSR-53 model we used was trained on 53 raw audio data from 53 different languages.

3.2.2 Using Hubert

Unlike Wav2vec 2.0 and XSLR-53, we did not remove any layers for the Hubert model. We rather fine-tuned the out-of-the-box pre-trained Hubert model on the English raw audio data of Librispeech 960hr. As discussed by (Pasad et al., 2022), Hubert does not follow the autoencoder pattern, given that the higher layers appear to encode more phonetic and word information. The choice of not removing top layers for the Hubert model was also corroborated through our empirical experiments, where we achieved the highest BLEU score for the Hubert model when we did not remove any top layers.

We used the Hubert model for the first contrastive constrained category (**contrastive1 unconstrained**) for the low-resource and dialectal tasks.

3.3 Data

The input to architectures E2E and E2E-ASR consist of 80-channel log-mel filterbank features computed on a 25 ms window with a 10 ms shift. We used raw audio as input for all the architectures using self-supervised models. For the translation text, we use the byte pair encoding (BPE) (Sennrich et al., 2016) algorithm with the sentencepiece toolkit from the Fairseq ST framework (Wang et al.,

[2] We refer the reader to the following papers (Baevski et al., 2020), (Conneau et al., 2020) and (Hsu et al., 2021) for more details on these models.

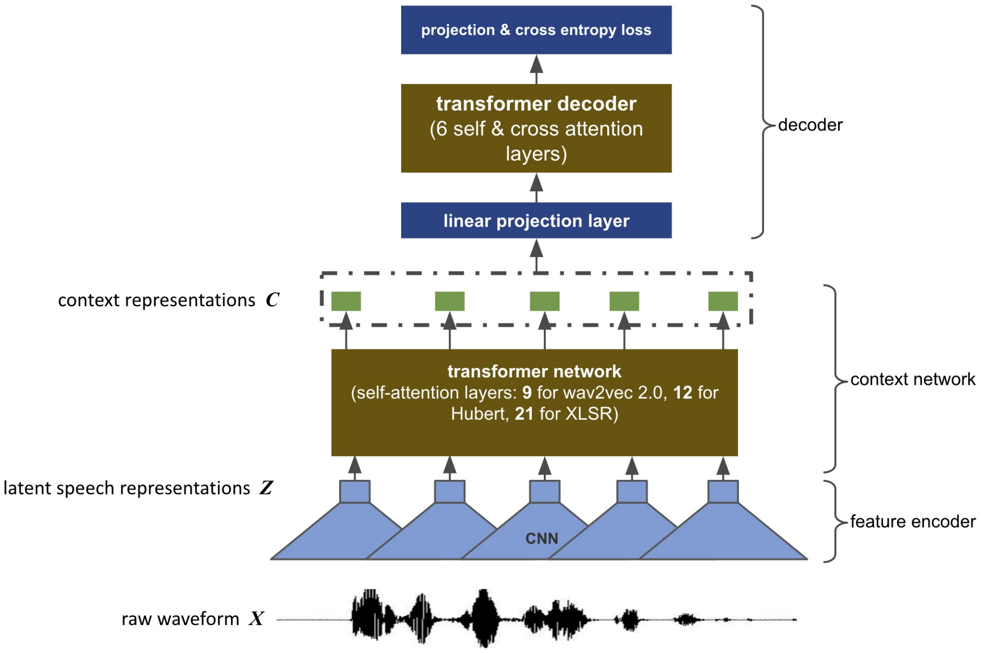
projection & cross entropy loss
transformer decoder (6 self & cross attention layers)
linear projection layer
decoder
context representations C
transformer network (self-attention layers: 9 for wav2vec 2.0, 12 for Hubert, 21 for XLSR)
context network
latent speech representations Z
CNN
feature encoder
raw waveform X

Figure 1: Self-supervised model architecture. This is an end-to-end architecture that uses self-supervised speech models as the encoder. The encoder is one of the Wav2vec 2.0, XLSR, or Hubert models. We removed the top 3 layers of the Wav2vec 2.0 and XLSR models.

Language Pairs	Vocab. Size
Irish-English	1000
Marathi-Hindi	1000
Pashto-French	3000
Tamasheq-French	1000
Quechua Spanish	400
Tunisian Arabic-English	8000

Table 2: BPE vocabulary for each language.

2020) to create vocabularies for all the target languages. We chose the vocabulary size based on the amount of text data we had for each language. Table 2 shows the BPE vocabulary size we used for each language pair. Though we used the training data size as a heuristic for choosing these BPE vocabulary sizes, we empirically tested a few configurations. We kept the sizes that gave the best BLEU score.

4 Results and Analyses

Table 3 shows results for all the systems we submitted. Our primary system reports the best results for the unconstrained setting where we used the *W2V-E2E* and *XLSR-E2E* architectures for the low-resource and dialectal tasks, respectively.

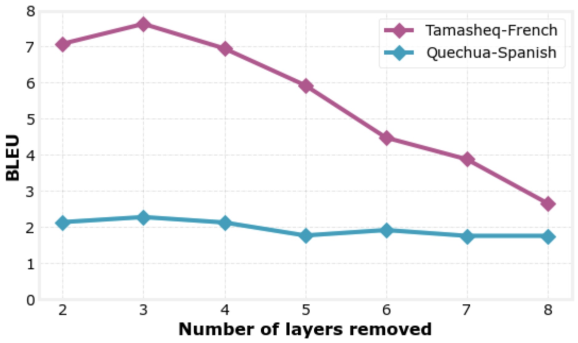

Figure 2: BLEU score on the test set for Tamasheq-French (tmh-fra) and Quechua-Spanish [3](que-spa) after removing top n number of layers of the Wav2vec 2.0. These results are run using the W2V-E2E architecture. For both Tamasheq-French and Quechua-Spanish, the best BLEU is achieved after removing the top 3 layers.

We explored the impact of removing the top n layers for the Wav2vec 2.0 model used in the *W2V-E2E* architecture. As illustrated in figure 2, the highest BLEU was achieved by removing the top three layers of the Wav2vec 2.0 model. We, therefore, used the same heuristic for the XLSR-53 model, given that it has the same architecture as the Wav2vec 2.0 model.

[3] The results for Quechua to Spanish are different from those in Table 3 because they were run after the evaluation period.

Language	System	Task	Architecture	dev/valid	test1	test2	test3
ga-eng	primary constr.	LR	E2E	-	-	15.1	-
	primary unconstr.		W2V-E2E	-	-	66.5	-
	contrastive1 unconstr.		Hubert-E2E	-	-	**77.4**	-
	contrastive2 unconstr.		E2E-ASR	-	-	15.1	-
mr-hi	primary constr.	LR	E2E	0.77	-	3.3	-
	primary unconstr.		W2V-E2E	4.76	-	7.7	-
	contrastive1 unconstr.		Hubert-E2E	**5.78**	-	**8.6**	-
	contrastive2 unconstr.		E2E-ASR	4.07	-	5.9	-
pus-fra	primary constr.	LR	E2E	2.66	-	5.92	-
	primary unconstr.		W2V-E2E	**11.99**	-	**16.87**	-
	contrastive1 unconstr.		Hubert-E2E	11.27	-	15.24	-
	contrastive2 unconstr.		E2E-ASR	9.72	-	13.32	-
tmh-fra	primary constr.	LR	E2E	1.24	1.0	0.48	-
	primary unconstr.		W2V-E2E	**12.07**	**7.63**	**8.03**	-
	contrastive1 unconstr.		Hubert-E2E	4.79	2.77	1.3	-
	contrastive2 unconstr.		E2E-ASR	5.24	3.77	2.1	-
que-spa	primary constr.	LR	E2E	1.46	-	1.46	-
	primary unconstr.		W2V-E2E	1.2	-	1.78	-
	contrastive1 unconstr.		Hubert-E2E	**1.84**	-	**1.86**	-
	contrastive2 unconstr.		E2E-ASR	1.63	-	1.63	-
aeb-eng	primary constr.	DT	E2E	11.49	8.94	5.0	4.5
	primary unconstr.		XLSR-E2E	**19.35**	**16.31**	**16.6**	**14.6**
	contrastive1 unconstr.		Hubert-E2E	17.69	14.52	15.0	13.4
	contrastive2 unconstr.		W2V-E2E	16.7	14.4	14.1	12.9

Table 3: BLEU score for all the submitted systems. LR and DT indicate low-resource and dialectal transfer, respectively. dev/valid refers to the validation or development sets we used during training. test1 refers to the test set we used during training (some language pairs did not have this set). test2 refers to the blind test set. Some language pairs (i.e., aeb-eng) had an additional blind test set called test3. The "-" character indicates that we do not have BLEU results for that category. We did not report the dev/valid results for the Irish to English (ga-eng) task due to the data quality issue discussed in section 5.

4.1 Low-Resource Task

For the low-resource shared task, the highest BLEU is obtained on average by the architecture that uses the Wav2vec 2.0 model (W2V-E2E). However, the Hubert (Hubert-E2E) architecture yields competitive BLEU compared to the W2V-E2E architecture. In fact, for Marathi-Hindi and Quechua-Spanish language pairs, the highest BLEU is achieved by using the Hubert model. Based on our experiments, we think both the Hubert and the Wav2vec 2.0 models may have similar performance though each model may require different configurations. In the future, we hope to have a detailed analysis of the conditions under which one model performs better than the other. Table 3 shows the BLEU results for the low-resource task.

The *W2V-E2E* architecture achieves a relatively high BLEU score compared to *Hubert-E2E* for Tamasheq-French. This behavior is explained by the fact that the Wav2vec 2.0 models used for Tamasheq-French were pre-trained on 234 hours of Tamasheq audio, while the Hubert was pre-trained on 960 hours of English data from the Librispeech dataset. Therefore, pre-training a self-supervised model on audio data from the same source language helps improve the model's performance on a downstream task.

Interestingly, pre-training on audio data from a different language than the source language for the speech translation task still yields improvement compared to starting with random weights. While Bansal et al. (2019) reported this behavior for ASR pre-training, we still see the same pattern for self-supervised pre-training.

Particularly for Tamasheq-French, which had a baseline BLEU score of 5.7 for the best IWSLT 2022 system (Anastasopoulos et al., 2022), we nevertheless improved upon the baseline by more than 2 BLEU on the blind test set.

4.2 Dialectal Task

Unlike the low-resource task, the highest BLEU for the dialectal task was achieved by using the XLSR-53 model (*XLSR-E2E*). Therefore, we used this architecture for our primary unconstrained setting. Table 3 shows the results for Tunisian Arabic-English.

For this task, Wav2vec 2.0 and Hubert had comparable BLEU scores. However, surprisingly, they did not perform as well as XLSR-53. This finding was counterintuitive given that the XLSR-53 model did not perform as well as the Wav2vec 2.0 or Hubert on all the low-resource languages. The XLSR-53 model was also reported to have poor performance by Zanon Boito et al. (2022) on a low-resource language. Based on our experiments, we think that the poor performance of the XLSR-53 model for the low-resource task was related to its size. We speculate that the XLSR-53 model size may fail to adapt while fine-tuning it on little data. However, fine-tuning it on a lot of data, like the case of Tunisian-Arabic-English, may yield overall improvement.

It is also possible that the best performance of the XLSR-53 model on the Tunisian Arabic-English data is because it was trained on more languages. It will be interesting to investigate the impact of the model size and multilinguality for self-supervised pre-trained speech models to improve the performance of speech translation downstream tasks. In addition, we think there may be room to study further the speech representation of the XLSR-53 model across layers so that they can be better adapted in low-resource settings.

5 Data Quality Issues

The low-resource shared tasks of the IWSLT 2023 consists of six tasks, each task corresponding to one language pair. As we worked on these shared tasks, we noticed issues with the data of two tasks: Maltese to English and Irish to English.

The Maltese to English data had a number of issues that made it hard to work with. For instance, the metadata of about 1001 out of 1698 samples mentioned zero or less than zero duration for audio samples (`start_time >= end_time`) while the aligned utterances had several words in most cases. Therefore, we were not able to align most audio data with their utterances.

The Irish to English data had an issue with the development set. Initially, the samples in the development were also present in the training set. However, the organizer later fixed this issue by updating the development set data. However, no matter how we trained our models, we never achieved more than 1 BLEU score on the updated development set. After troubleshooting our model on the training data, we were confident that we should have gotten a BLEU score that was well above 1. We proceeded with submitting our system for this task. However, we are very suspicious of the high BLEU score reported on the blind test, as shown in Table 3, as it suggests that there's an overlap between training and test sets.

6 Conclusion

In this paper, we presented the GMU Systems for the IWSLT 2023 Dialect and Low-resource Speech Translation Tasks. Our approach mainly focused on using self-supervised pre-trained speech models to improve the performance of speech translation on downstream tasks. The self-supervised pre-trained speech models used in this paper are the Wav2vec 2.0, XLSR-53, and Hubert. We showed that the Wav2vec 2.0 and the Hubert model have comparable results in low resource and dialectal transfer tasks. However, the Wav2vec 2.0 performs well when we remove the top three layers, while the Hubert model has no such requirements.

Our experiments showed that the XLSR-53 model performs poorly in the low-resource setting compared to the Wav2vec 2.0 and Hubert models. However, in the dialectal task, the XLSR-53 model outperforms the Wav2vec 2.0 and Hubert models.

In the future, we plan to conduct an in-depth analysis to understand the advantages and limitations of these self-supervised pre-trained speech models while fine-tuning them on downstream speech translation tasks.

Acknolwedgements

We are thankful to the organizers of the IWSLT 2023 low resource and dialectal shared tasks. This work was generously supported by NSF grant

IIS-2125466 and by a Meta Sponsored Research Award. We are also thankful to the Office of Research Computing at George Mason University (https://orc.gmu.edu), funded in part by grants from the National Science Foundation (Awards Number 1625039 and 2018631), for the computing resources we used to train our models.

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, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

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

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier, and Y. Estève. 2022. Speech resources in the tamasheq language. In *International Conference on Language Resources and Evaluation*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdel rahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech*.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

ELRA. Elra catalogue (http://catalog.elra.info), trad pashto-french parallel corpus of transcribed broadcast news speech - training data, islrn: 802-643-297-429-4, elra id: Elra-w0093, trad pashto broadcast news speech corpus, islrn: 918-508-885-913-7, elra id: Elra-s0381.

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.

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.

Ha Nguyen, Fethi Bougares, Natalia Tomashenko, Yannick Estève, and laurent besacier. 2020. Investigating self-supervised pre-training for end-to-end speech translation. In *ICML 2020 Workshop on Self-supervision in Audio and Speech*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev 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.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2022. Comparative layer-wise analysis of self-supervised speech models. *ArXiv*, abs/2211.03929.

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*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Annual Meeting of the Association for Computational Linguistics*.

Mihaela Stoian, Sameer Bansal, and Sharon Goldwater. 2019. Analyzing asr pretraining for low-resource speech-to-text translation. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7909–7913.

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

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Miguel Pino. 2020. Fairseq s2t: Fast speech-to-text modeling with fairseq. In *AACL*.

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

Anne Wu, Changhan Wang, Juan Miguel Pino, and Jiatao Gu. 2020. Self-supervised representations improve end-to-end speech translation. *ArXiv*, abs/2006.12124.

Marcely Zanon Boito, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani, Ha Nguyen, Florentin Barbier, Souhir Gahbiche, and Yannick Estève. 2022. ON-TRAC consortium systems for the IWSLT 2022 dialect and low-resource speech translation tasks. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 308–318, Dublin, Ireland (in-person and online). Association for Computational Linguistics.