Scaling pseudo-labeling data for end-to-end low-resource speech translation (the case of Kurdish language)

Mohammad Mohammadamini, Aghilas Sini, Marie Tahon, Antoine Laurent

Scaling pseudo-labeling data for end-to-end low-resource speech translation (the case of Kurdish language)

Mohammad Mohammadamini[1], *Aghilas Sini*[1], *Marie Tahon*[1], *Antoine Laurent*[1]

[1]LIUM, Le Mans University, France

first.last@univ-lemans.fr

Abstract

In this paper we propose a pseudo-labeling pipeline to generate End-to-End Speech to Text Translation (E2E S2TT) data for low-resource languages. This pipeline allows us to achieve very promising results on S2TT task without having any parallel speech corpora. The proposed pipeline is composed of a speech segmentation, followed by a speech recognition system and a machine translation system. Our study is performed on Kurdish language which doesn't have resources for E2E S2TT. In our study, we firstly fine-tune and evaluate the ASR and MT systems to achieve a degree of reliability on the these components. The pipeline is used to pseudo-label 3,200 hours of Kurdish speech aligned with English translation. The pseudo-labeled data is extensively evaluated with different E2E S2TT systems such as Seq2Seq Transfomers and Whisper model. Achieving a 20.68 BLEU score on Fleurs benchmark shows the effectiveness of the our approach and its potential for other low-resource languages. The parallel pseudo-labeled data can be retrieved from: https://lium.univ-lemans.fr/en/ckbens2tt/

Index Terms: Speech Translation, Pseudo-labeling, S2TT, Kurdish language

1. Introduction

Speech translation is the task of translating audio from a source language into text or audio in a target language [1]. Developing a speech-to-text translation system requires a substantial amount of translated audio in the target language. Due to the novelty of this research area, the majority of languages suffer from a lack of sufficient data. In [2], a language with fewer than 1,000 hours of transcribed/translated publicly available audio is considered a low-resource language. Based on this definition, only a dozen languages out of approximately 7,000 can be considered high-resource languages. In the most recent multilingual speech translation system developed by the Seamless group at Meta, only 16 languages are categorized as high-resource languages [2].

Pseudo-labeling is a common approach to address the problem of data scarcity in speech translation. In the context of speech translation, pseudo-labeling refers to the automated process of translating transcribed speech or simultaneous transcription and translation of speech. Pseudo-labeling speech translation data is valuable not only for low-resource languages but also for high-resource languages [2], as the majority of available transcribed data, which is collected or designed for speech recognition, is insufficient to capture the linguistic content diversity required for speech translation. While the automatic generation of translated audio from high-resource to low-resource languages is more feasible, the lack of robust speech recognition systems makes the inverse direction more challenging. In this paper, we propose a comprehensive pipeline for pseudo-labeling speech-to-text translation data for low-resource languages. The proposed pseudo-labeling pipeline consists of three main components: speech segmentation, speech recognition, and text translation. Our goal is to optimize all of these components to generate reliable translated speech for end-to-end speech translation. We demonstrate how to use a limited amount of speech recognition and machine translation data to scale up speech translation data for low-resource languages.

Our experiments focus on Central Kurdish (CKB) to English speech-to-text translation. The Kurdish language, with more than 35 million speakers across different dialects, suffers severely from a lack of resources [3]. To the best of our knowledge, the only speech translation dataset that includes this language is the Fleurs benchmark [4], which is primarily used for the evaluation of ASR and S2TT systems. In this paper, we first collected more than 4,000 hours of clean Kurdish speech from publicly available audiobooks. The Seamless v2 large model was fine-tuned using available Kurdish ASR data. Additionally, we introduce a new machine translation dataset in this paper, comprising 222k pairs of CKB→EN sentences collected from various sources. This dataset was used to fine-tune NLLB 1.3B machine translation for the Kurdish language. Using a speech segmentation module along with the fine-tuned ASR and MT systems, we performed pseudo-labeling on the raw audio. For the output of three components in the pipeline (i.e. Segmentation, ASR, and MT) independent or joint filtering criteria are applied to exclude low-quality samples. Finally the pseudo-labeled data is used in training several E2E-S2TT models. The main contributions of this research are:.

- Collecting a large-scale clean audio dataset from audiobooks for the Central Kurdish language.
- Introducing a fairly-large scale MT dataset fro Central Kurdish.
- Improvement the SOTA ASR and MT for Central Kurdish.
- Pseudo-labeling and evaluation of 3200 hours of Kurdish raw audio leveraging the trained ASR and MT for E2E S2TT.

2. Previous works

Recent initiatives to develop speech translation datasets for high-resource languages have been substantial. AugLibriSpeech, a French-translated version of the LibriSpeech corpus, comprising 236-hours EN→FR S2TT [5]. The largest publicly available speech translation dataset, CoVoST 2, provides bidirectional S2TT translations, covering English to 15 languages and 21 languages to English [6, 7]. VoxPopuli is a multilingual speech translation corpus, primarily sourced from European Parliament event recordings, featuring 15 European languages [8]. So far, three speech translation datasets have been derived from TED Talks. Among them, the MUST-C dataset offers English to 14 languages speech translation [9, 10], while TEDx contains translations from English to 7 languages [11]. Indic-TEDST, another TED-derived dataset, includes translations from English into 9 Indian languages [12]. The case study of the current research is Kurdish which is a low-resource language. The only dataset that incorporates Kurdish is Fleurs, a multilingual S2ST and S2TT corpus covering 101 languages, with 13 hours of X→CKB and CKB→X data [4].

The reviewed datasets show a limited number of languages, and in most cases, the parallel data belongs to high-resource languages. Therefore, pseudo-labeling data is a rapid and efficient way to expedite the development of speech translation systems. In the Seamless project [2], the most comprehensive speech translation project in terms of the number of languages, pseudo-labeling is done from English to 100 languages, while the inverse direction, from other languages to English, is limited to a very small number of high-resource languages. In [13], pseudo-labeling is investigated for joint speech transcription and speech translation. The included translation languages are English to German and Chinese. In [14], a pseudo-labeling pipeline is used to generate parallel data for S2TT translation. By adding speech synthesis, the pipeline is extended to S2ST. In a similar study, [15] used a cascade pseudo-labeling approach to generate data. In the major part of reviewed works the pseudo-labeling is done on high-resource languages, while in the current paper we focus on a low-resource case.

3. Pseudo-labeling pipeline

The proposed low-resource pseudo-labeling data generation pipeline (see Figure 1) is composed of three main components: a speech segmentation, a speech recognition and a machine translation system. In this pipeline we have three inputs: raw audio, small set human-annotated ASR dataset and a parallel text corpus. The output of this pipeline is a large-scale speech to text translation data compose of a triplet $\langle S_s, T_s, T_t \rangle$, where S_s, T_s, and T_t represent the source language speech, source language text, and target language text, that will be used in E2E S2TT. In out case S_s, and T_s are in Central Kurdish and the T_t is English translation.

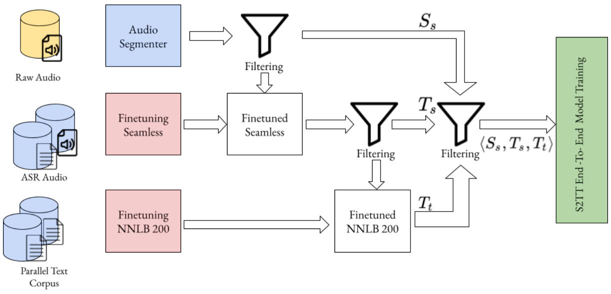

Figure 1: *The pseudo-labeling pipeline for low-resource speech translation*

To have a consistent text normalization across all parts of the pipeline, the same normalization is applied on the Kurdish text which includes applying Unicode correction, punctuation standardization, and number unification [16].

3.1. Speech segmentation

The first component is speech segmentation. Existing speech segmentation systems can be categorized into VAD-based and DNN-based models. Pre-trained speech segmentation DNNs are highly language-dependent [17]. For clean raw audio, we observed that VAD-based segmenters perform better. We implemented an energy-based VAD, where a speech signal is segmented if more than 30 consecutive frames are silent. Each frame is 10 ms long. To ensure consistent silence at the beginning and end of each segment, we maintain 15 frames of silence at both points.

3.2. Speech Recognition

The second component of the pipeline is speech recognition for which we use Seamless v2 large model. Seamless is a set of complex models for T2TT, S2TT, S2SS, and ASR. We use the ASR component, which consists of a Wav2Vec-BERT speech encoder and an NLLB-200 decoder. The model is jointly optimized for ASR and S2TT tasks [1, 2]. We fine-tune the Seamless model using Mel-filter bank (bins = 80) features over 10 epochs, with a batch size of 16 and a learning rate of 1e-4.

3.3. Machine Translation

In order to achieve a reliable MT system in our pipeline, we fine-tune NLLB 1.3B with a new collection of parallel data. The NLLB 1.3B contains 200 languages including Central Kurdish [18]. Our goal is to improve the performance of the MT translation system using available parallel text data. The NLLB is fine-tuned with a Adam optimizer, a learning rate equal to 1-e4 and weight decay equal to 1-e3 in 500k steps.

3.4. E2E S2TT system

3.4.1. S2TT Transfomer

The first E2E S2TT used to evaluate the pseudo-labeled S2TT data is speech translation transformer which includes 12 encoder and 6 decoder layers [19, 20]. We are using three Seq2Seq transformers: Small Transformer, Medium Transformer and Large Transformer. The size of Fully connected (FC) layers and the key (K), value (V), and query (Q) projection dimensions within the attention mechanism are as follows. In the Small Transformer the FC size is 2048, while the K, V, and Q projection sizes are set to 256. In Medium Transformer FC layer retains a hidden dimension of 2048, but the K, V, and Q projections are increased to 512. In the Large Transformer a FC of 4096, with K, V, and Q projections set to 1024 are used. The models are trained from scratch with Adam optimizer, lr 2e-3, in 25 epochs. .

3.4.2. Whisper Large V3

Whisper is a transformer-based Seq2Seq model trained on 680,000 hours of labeled speech data, encompassing tasks such as ASR, S2TT, VAD, and Speaker Recognition (SR) [21]. Whisper supports more than 80 languages for ASR and S2TT; however, the Kurdish language is not currently supported. Our motivation for using Whisper is twofold: First, given its promising results with low-resource languages [22], we plan to fine-tune it using a portion of pseudo-labeled data. Second, we aim to leverage the potential of this multilingual system to enhance generalizability to out-of-domain data, such as proper names. We are fine-tuning the Whisper v3 large model using the AdamW optimizer with a learning rate of 1e-5, a batch size of 16, in 10 epochs.

4. Data and resources

4.1. Raw audio

The main source of our raw data is freely available audiobooks on the web. We collected 1026 audiobooks covering several topics. The size of collected raw audio is 4,300h. After filtering the poetry, dialectical speech and local folk tales, 884 audiobooks in 11 main categories remained which are described in Table 1.

Table 1: *Topics of audiobooks used as raw audio*

Topic	**Number**
Short stories	90
Novel	204
Language and critical theory	45
Politics	75
Children	20
Religion	81
History	85
Auto/biography	89
Miscellaneous	56
Feminism	25
Philosophy, Psychology, Sociology	114
Total	884

4.2. Transcribed audio

The human-annotated ASR data, comes from two sources. The primary source is Common Voice 18 [1], which includes 117,000 validated samples. Additionally, we utilized the training portion of Asosoft v1 [23], which comprises 42,500 samples recorded from 700 sentences designed to reflect the distribution of Kurdish di-phones (i.e., pairs of consecutive phonemes). While the dataset contains a substantial number of recordings, its linguistic diversity is limited, with only 19,100 unique short sentences (Table 2).

Table 2: *Human annotated speech recognition data*

Part	**Samples**	**Unique Sents**	**Hours**
Common Voice 18	117k	18.4k	134
Asosoft corpus v1 [23]	42.5k	700	43
Total	163.5k	19.1k	177

4.3. Machine translation data

The third dataset used in the proposed pipeline is parallel En→CKB dataset. The list of different sources of MT data is shown in Table 3. A major part of this data was collected during this research. The "Certified Translators" category includes translations provided by certified translators covering various domains. About 16k sentence pairs were collected from books provided by the writers or publishers. "Stanford Plato" refers to the Kurdish translations of the Stanford Encyclopedia of Philosophy [2]. KUTED is the Kurdish version Ted corpus [3].

Table 3: *Parallel text corpus used to fine-tune the MT system*

Part	Pairs	EN tokens	CKB tokens
Certified translators	107.5k	1.52m	1.50m
KUTED	91.08k	1.65m	1.40m
Scentific books	4.34k	97.80k	88.65k
Biography books	5.44 k	95.70k	86.81k
Politics books	6.30k	112.38k	105.41k
Stanford Plato	2.53k	58.84k	54.69k
COVID Initiative [24]	3.07k	70.34k	66.00k
Miscellaneous	2.90k	82,21k	78,54k
All	222,16k	3,69m	3,38m

5. Pseudo-labeled data generation

Several factors can contribute to low-quality pseudo-labeled samples, such as noisy audio, background music, or distorted speech form ASR, and translation errors from MT module. For a given triplet $\langle S_s, T_s, T_t \rangle$, where S_s, T_s, and T_t represent the source speech, source text, and target text, respectively, a series of filtering steps are applied to ensure data quality. To remove unwanted samples, we apply the following filtering criteria:

- **Partial Transcription:** Some utterances may be partially transcribed. To identify such cases, we remove samples where the Words Per Minute (WPM) metric of T_s falls outside the range of (90, 200). This threshold is chosen based on the fact that the average number of clearly pronounced WPM typically ranges from (117, 239) [25]. A flexible margin is applied to retain a larger number of valid utterances.
- **Short Utterances:** If S_s is shorter than 1 second or T_s contains fewer than 3 space separated tokens, the sample is discarded.
- **Long Utterances:** If S_s exceeds 30 seconds in duration or T_s contains more than 50 tokens, the sample is removed.
- **ASR Confidence:** The quality of T_s is assessed using ASR confidence. The confidence score is computed as: $\frac{1}{N}\sum_{i=1}^{N} \max_j \text{Softmax}(\text{logits}_{i,j})$ where N is the number of tokens in T_s, and $\text{Softmax}(\text{logits}_{i,j})$ represents the probability assigned to token j at position i. Samples with a confidence score below 0.9 are discarded.
- **ASR and MT Prediction Loops:** In some cases the ASR or MT produce a repetitive meaningless output. If T_s or T_t contains n-grams ($n = 1, 2, 3$) with more than two consecutive repetitions, the sample is discarded. This helps filter out low-quality audio, particularly cases of language switching or dialectical speech. Also it filters some prediction errors of MT system.
- **Length Ratio:** The ratio between the lengths of T_s and T_t must satisfy: $0.5 < \text{Length}(T_s)/\text{Length}(T_t) < 1.5$. Samples outside this range are discarded.
- **Proper Names:** If the proportion of proper names in T_t exceeds 50%, the sample is removed using NLTK toolkit.

[1] https://commonvoice.mozilla.org/en/datasets

[2] https://zanistname.com/

[3] https://lium.univ-lemans.fr/en/kuted/

The statistics of the pseudo-labeled data is presented in Table 4. The volume of the filtered pseudo-labels data is 1,71m files, comprising 3,231 hours of audio, with 22,66 million aligned English tokens.

Table 4: *Pseudo-labled data for E2E S2TT from Central Kurdish to English*

Part	Samples	Length	CKB tokens	EN tokens
E2E S2TT	1.71m	3,231h	21.58m	22.66m

6. Results

6.1. ASR results

We evaluated the ASR model on two test sets: Asosoft and Fleurs. The Asosoft test set consists of 800 utterances from eight speakers, recorded in a controlled environment [23]. Fleurs is an extended version of the Flores benchmark for ASR and S2TT translation. The Kurdish test set of Fleurs includes 921 sentences totaling three hours of audio. Table 5 presents the ASR model results. The first row displays the Seamless baseline (before fine-tuning), which includes Kurdish. The baseline yields a Word Error Rate (WER) of 24.04 on the Asosoft test set and 38.33 on the Fleurs test set. After fine-tuning the model with our human-annotated dataset, we achieved a WER of 8.18 on the Asosoft test set and 20.31 on the Fleurs test set. Notably, compared to previous studies [23, 26], our model achieves a significant SOTA performance improvement for Kurdish ASR.

Table 5: *ASR Results (WER)*

-	Seamless baseline	Best reported	Ours
Asosoft test	24.04	11.80	**8.18**
Fleurs test	38.33	-	**20.31**

6.2. MT results

The results given by NLLB 1.3B baseline and the fine-tuned system on Flores benchmark are presented in Table 6. The NLLB base gives a BLEU equal to 10.4 for En→CKB direction and 35.34 BLEU fo the inverse direction. Our fine-tuned model using the dataset introduced in Table 3 shows significant improvement giving 17.12 BLEU for ENG→CKB direction while for the CKB→EN direction obtaining a 36.24 BLEU score, we achieved a marginal improvement. We compared the results obtained by our system to Google[4], however it pass the Google system for En→CKB direction significantly, the results for CKB→Eng are very close.

Table 6: *MT Results (BLEU/ChrF++) on Flores benchmark [27]*

-	Google	NLLB Base	Ours
Flores en-ckb	13.59	10.40/45.10	**17.12/49.62**
Flores ckb-en	**36.84**	35.34/58.21	36.24/58.35

6.3. E2E S2TT

The pseudo-labeled data generated using the described pipeline is used in E2E S2TT. We firstly train three E2E transformers(described in Section 3.4). In all cases, we used a SentencePiece unigram tokenizer with a vocabulary size of 10k. The tokenizer is trained using the pseudo-labeled data generated by the machine translation system. The Small model achieves a BLEU score of 18.45 on the Fleurs test set, while the Large Transformer model achieves a BLEU score of 20.68. The last column of Table 7, belong to models that are firstly trained on the pseudo-labeled data, then fine-tuned using the train part of Fleurs dataset. The fine-tuning set includes 3,040 samples [4]. In all cases, this adaptation improves the system performance, with the Large Transformer model achieving a BLEU score of 21.97.

In another experiment, the Whisper v3 Large model is fine-tuned. We used 200k randomly chosen samples from the pseudo-labeled data to fine-tune this model. With this portion of the training data, we obtained the same results as the Large Transformer model trained from scratch using the entire generated dataset. Then we performed a second fine-tuning step using the training portion of the Fleurs dataset on the Whisper model, which had already been fine-tuned with the pseudo-labeled data. This resulted in a BLEU score of 21.19. In the last experiment, we used all the pseudo-labeled data to fine-tune Whisper V3 model but we didn't observed more improvement in comparison to using 200K samples.

Table 7: *E2E S2TT results obtained by Transformers trained from scratch and fine-tuned Whisper (BLEU/ChrF++)*

Model	Pseudo-label	Fleurs-Finetune
Transformer Small	18.45/44.16	19.83/45.88
Transformer Medium	19.50/46.51	20.12/46.93
Transformer Large	20.42/46.85	21.97/48,07
Whisper V3 Large	20.68/46.54	21.19/48.09

7. Conclusion

In this paper we proposed a pseudo-labeling pipeline for speech translation in low-resource languages. The main idea was the development an E2E S2TT system using limited ASR and MT resources. Our research was focused on Central Kurdish that is a low-resource language with minimal speech translation resources. Our obtained results are competitive to the average results given by SOTA speech translation models [1] which shows that our proposed approach can be used to expand the speech translation technology for many low-resource languages that already have a degree of resources for ASR and MT tasks. Future improvements could include replacing the VAD-based segmenter with a sentence boundary segmentation module to enhance data quality. Integrating a language model that help to revive from the ASR errors can be another possible direction of research.

[4] https://translate.google.com/

8. Acknowledgments

This work has been funded by the DGA RAPID COMMUTE project. This work was performed using HPC resources from GENCI–IDRIS under allocation A0181012527.

9. References

[1] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, et. all., and S. C. Team, "Joint speech and text machine translation for up to 100 languages," *Nature*, vol. 637, no. 8046, pp. 587–593, January 2025. [Online]. Available: https://doi.org/10.1038/s41586-024-08359-z

[2] S. Communication, L. Barrault, Y.-A. Chung, M. C. Meglioli, and et. al., "Seamless: Multilingual expressive and streaming speech translation," 2023. [Online]. Available: https://arxiv.org/abs/2312.05187

[3] H. Veisi, M. MohammadAmini, and H. Hosseini, "Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus," *Digital Scholarship in the Humanities*, vol. 35, no. 1, pp. 176–193, 02 2019. [Online]. Available: https://doi.org/10.1093/llc/fqy074

[4] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 798–805.

[5] A. C. Kocabiyikoglu, L. Besacier, and O. Kraif, "Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation," in *LREC 2018*. Miyazaki, Japan: European Language Resources Association (ELRA), may 2018.

[6] C. Wang, J. Pino, A. Wu, and J. Gu, "Covost: A diverse multilingual speech-to-text translation corpus." Marseille, France: European Language Resources Association, may 2020, pp. 4197–4203.

[7] C. Wang, A. Wu, J. Gu, and J. Pino, "Covost 2 and massively multilingual speech translation," in *Proc. Interspeech 2021*, 2021, pp. 2247–2251.

[8] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation." Online: Association for Computational Linguistics, aug 2021, pp. 993–1003. [Online]. Available: https://aclanthology.org/2021.acl-long.80

[9] M. A. D. Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, jun 2019, pp. 2012–2017. [Online]. Available: https://aclanthology.org/N19-1202

[10] R. Cattoni, M. Antonino, L. Bentivogli, M. Negri, and M. Turchi, "Must-c: A multilingual corpus for end-to-end speech translation," *Computer Speech and Language*, vol. 66, p. 101155, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230820300887

[11] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and M. Post, "Multilingual tedx corpus for speech recognition and translation," in *Proceedings of Interspeech*, 2021.

[12] N. Sethiya, S. Nair, and C. Maurya, "Indic-tedst: Datasets and baselines for low-resource speech to text translation," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, may 2024, pp. 9019–9024. [Online]. Available: https://aclanthology.org/2024.lrec-main.790

[13] M. Gheini, T. Likhomanenko, M. Sperber, and H. Setiawan, "Joint speech transcription and translation: Pseudo-labeling with out-of-distribution data," in *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 7637–7650. [Online]. Available: https://aclanthology.org/2023.findings-acl.483/

[14] C. Wang, H. Inaguma, P.-J. Chen, I. Kulikov, Y. Tang, W.-N. Hsu, M. Auli, and J. Pino, "Simple and effective unsupervised speech translation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 771–10 784. [Online]. Available: https://aclanthology.org/2023.acl-long.602/

[15] J. Pino, Q. Xu, X. Ma, M. J. Dousti, and Y. Tang, "Self-training for end-to-end speech translation," in *Interspeech 2020*, 2020, pp. 1476–1480.

[16] A. Mahmudi, H. Veisi, M. MohammadAmini, and H. Hosseini, "Automated kurdish text normalization," 2019.

[17] I. Tsiamas, G. I. Gállego, J. A. R. Fonollosa, and M. R. Costa-jussà, "Shas: Approaching optimal segmentation for end-to-end speech translation," in *Interspeech 2022*, 2022, pp. 106–110.

[18] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, et. all., and N. Team, "Scaling neural machine translation to 200 languages," *Nature*, vol. 630, no. 8018, pp. 841–846, June 2024. [Online]. Available: https://doi.org/10.1038/s41586-024-07335-x

[19] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, "Fairseq S2T: Fast speech-to-text modeling with fairseq," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 33–39. [Online]. Available: https://aclanthology.org/2020.aacl-demo.6/

[20] ——, "Fairseq S2T: Fast speech-to-text modeling with fairseq," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 33–39. [Online]. Available: https://aclanthology.org/2020.aacl-demo.6

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[22] Y. Liu, X. Yang, and D. Qu, "Exploration of whisper fine-tuning strategies for low-resource asr," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 29, 2024.

[23] H. Veisi, H. Hosseini, M. MohammadAmini, W. Fathy, and A. Mahmudi, "Jira: a central kurdish speech recognition system, designing and building speech corpus and pronunciation lexicon," *Language Resources and Evaluation*, vol. 56, no. 3, pp. 917–941, 2022.

[24] A. Anastasopoulos, A. Cattelan, Z.-Y. Dou *et al.*, "TICO-19: The translation initiative for COVID-19," in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online: Association for Computational Linguistics, Dec. 2020.

[25] M. Brysbaert, "How many words do we read per minute? a review and meta-analysis of reading rate," *Journal of Memory and Language*, vol. 109, p. 104047, 2019.

[26] A. A. Abdullah, H. Veisi, and T. Rashid, "Breaking walls: Pioneering automatic speech recognition for central kurdish: End-to-end transformer paradigm," 2024. [Online]. Available: https://arxiv.org/abs/2406.02561

[27] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation," 2021.