ON-TRAC consortium systems for the IWSLT 2023 dialectal and low-resource speech translation tasks

Antoine Laurent[1], **Souhir Gahbiche**[5], **Ha Nguyen**[2], **Haroun Elleuch**[4], **Fethi Bougares**[4], **Antoine Thiol**[5], **Hugo Riguidel**[1,3], **Salima Mdhaffar**[2], **Gaëlle Laperrière**[2], **Lucas Maison**[2], **Sameer Khurana**[6], **Yannick Estève**[2]
[1]LIUM - Le Mans University, France, [2]LIA - Avignon University, France, [3]Systran - France, [4]ELYADATA - Tunis, Tunisia, [5]Airbus - France,
[6]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA

Abstract

This paper describes the ON-TRAC consortium speech translation systems developed for IWSLT 2023 evaluation campaign. Overall, we participated in three speech translation tracks featured in the low-resource and dialect speech translation shared tasks, namely; i) spoken Tamasheq to written French, ii) spoken Pashto to written French, and iii) spoken Tunisian to written English. All our primary submissions are based on the end-to-end speech-to-text neural architecture using a pre-trained SAMU-XLSR model as a speech encoder and an mbart model as a decoder. The SAMU-XLSR model is built from the XLS-R 128 in order to generate language agnostic sentence-level embeddings. This building is driven by the LaBSE model trained on a multilingual text dataset. This architecture allows us to improve the input speech representations and achieve significant improvements compared to conventional end-to-end speech translation systems.

1 Introduction

IWSLT is a unique opportunity that allows each year the assessment of progress made in the area of Spoken Language Translation (SLT). This assessment is made possible throughout the organisation of an evaluation campaign including various shared tasks that address specific scientific challenges of the SLT domain. In addition to the well-established shared tasks, IWSLT organisers introduce new tasks to address the many challenges settings related to SLT area like data scarcity, multilingualism, time and computation constraints, etc.

In this context, the IWSLT 2023 proposes two interesting shared tasks: low-resource and dialect speech translation (ST). The former aims to assess the exploitability of current translation systems in data scarcity settings. The latter focuses on the assessment of the systems' capabilities in *noisy* settings: different dialects are mixed in a single dataset of spontaneous speech. For the low-resource task, several language pairs were proposed this year. In this paper, we focus on Tamasheq-French, Tunisian Arabic-English and Pashto-French.

This paper reports the ON-TRAC consortium submissions for the aforementioned tasks. The ON-TRAC Consortium is composed of researchers from three academic laboratories, LIUM (Le Mans University - France), LIA (Avignon University - France), MIT (Cambridge - USA) together with three industrial partners: Airbus France, ELYADATA and Systran. Our systems for the dialect task focus on both cascaded and end-to-end approaches for ST. For the low-resource task, we focus on the leveraging of models based on self-supervised learning (SSL), and on the training of ST models with joint automatic speech recognition (ASR), machine translation (MT) and ST losses.

This paper is organized as follows. Section 2 presents the related work. Section 3 is dedicated to detail our primary systems encoder-decoder approach. The experiments with the Tunisian Arabic-English dataset for low-resource and dialect ST tasks are presented in Section 4. Results for the Tamasheq-French and Pashto-French tracks are presented in Section 5 and 6 respectively. Section 7 concludes the paper and discusses future work.

2 Related work

Before the introduction of *direct* or *end-to-end* ST models (Berard et al., 2016; Weiss et al., 2017), the ST task was approached as a *cascaded* problem: the speech is transcribed using an ASR model, and the transcriptions are used to train a classic MT model. The limitations of this approach include the need for extensive transcriptions of the speech signal, and the error propagation between ASR and MT modules. In comparison to that, end-to-end ST models offer a simpler encoder-decoder architecture, removing the need for intermediate representations of the speech signal. Although at first, cas-

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 219–226

caded models were superior in performance compared to end-to-end models, results from recent IWSLT campaigns illustrate how end-to-end models have been closing this gap (Ansari et al., 2020; Bentivogli et al., 2021; Anastasopoulos et al., 2021, 2022). Moreover, the joint optimization of ASR, MT and ST losses in end-to-end ST models was shown to increase overall performance (Le et al., 2020; Sperber et al., 2020).

Furthermore, SSL models for speech processing are now a popular foundation blocks in speech pipelines (Schneider et al., 2019; Hsu et al., 2021; Baevski et al., 2019, 2020). These models are large trainable networks with millions, or even billions (Babu et al., 2021b), of parameters that are trained on unlabeled audio data only. The goal of training these models is providing a powerful and reusable abstraction block, which is able to process raw audio in a given language or in multilingual settings (Conneau et al., 2020; Babu et al., 2021b), producing a richer audio representation for the downstream tasks to train with, compared to surface features such as MFCCs or filterbanks. Recent work found considerable performance gains and/or state-of-the-art performance by including these blocks in their target tasks, and more importantly, the final models can be trained with a smaller amount of labeled data, increasing the *accessibility* of current approaches for speech processing (Kawakami et al., 2020; Schneider et al., 2019; Hsu et al., 2021; Baevski et al., 2019, 2020).[1] Recent work found considerable performance gains and/or state-of-the-art performance by including these blocks in downstream tasks. Most of them focused on ASR (Kawakami et al., 2020; Schneider et al., 2019; Hsu et al., 2021; Baevski et al., 2019, 2020), but recent speech benchmarks (Evain et al., 2021b,a; Yang et al., 2021) cover tasks such as ST, spoken language understanding, emotion recognition from speech and more.

3 Primary systems encoder-decoder architecture

3.1 SAMU-XLS-R (SAMU−XLS−R)

SAMU−XLS−R is a multilingual multimodal semantic speech representation learning framework where the speech transformer encoder XLS−R (Babu et al., 2021a) is fine-tuned using semantic supervision from the pre-trained multilingual semantic text encoder LaBSE (Feng et al., 2022). The training and modeling details can be found in the original paper (Khurana et al., 2022). In this work, we use the same training framework but train the model using transcribed speech collected from approximately 100 spoken languages from several datasets such as CommonVoice-v10 (Ardila et al., 2020a), Multilingual Speech (MLS) (Pratap et al., 2020), Babel, IndicSuperb (Javed et al., 2022), Shrutilipi (Bhogale et al., 2023), Voxpopuli (Wang et al., 2021), MGB-2 Arabic (Ali et al., 2019) and Wenetspeech (Zhang et al., 2022).

3.2 Translation model

We use the standard encoder-decoder architecture for our translation model. We initialize the encoder using the pre-trained SAMU−XLS−R. Following (Li et al., 2020), the decoder is initialized with the decoder of a pre-trained text-to-text translation model, namely MBART[2]. The encoder-decoder model is trained using corpora that consist of tuples $(\mathbf{a}_{1:S}, \mathbf{y}_{1:L})$, where $\mathbf{y}_{1:L}$ is the text translation sequence of the speech sequence $\mathbf{a}_{1:S}$.

To maintain the pre-trained SAMU−XLS−R values of the speech encoder, we leave its parameters unchanged. However, we introduce task-specific parameters in the form of adapters (Houlsby et al., 2019), consisting of a bottleneck Feed-Forward layer, which are added after the Multi-Headed Self-Attention and fully-connected blocks in each transformer layer. While most parameters of the decoder remain fixed from pre-training, we fine-tune the Layer Normalization and Encoder-Decoder Cross-Attention blocks based on (Li et al., 2020).

4 Tunisian Arabic-English track

In this section, we present our experiments for translating Tunisian Arabic to English in the context of the dialect and low-resource tasks from IWSLT 2023. Section 4.1 describes the data used in our experiments. Results on the ST task are presented in Section 4.3.

4.1 Data

The training and development data conditions are identical to IWSLT 2022 edition. It consisted of two types of datasets: (1) 383h of manually transcribed conversational speech and (2) 160h, subpart of it, augmented with their English translations to form a three-way parallel corpus (audio, transcript,

[1] Recent benchmarks for SSL models can be found in Evain et al. (2021b,a); Yang et al. (2021); Conneau et al. (2022).

[2] Text-to-text translation model: MBART

translation). This dataset is made available by LDC under reference LDC2022E01. The goal of this track is to train speech translation systems under two training conditions: constrained, in which only the provided dataset resources are allowed, and unconstrained where participants may use any public or private resources.

4.2 End-to-end ST

We used the end-to-end translation model presented in section 3.2. The model was trained directly on the Tunisian to English task (no pre-training of the encoder-decoder model), using SAMU−XLS−R trained on 100 languages. We used adapters (Houlsby et al., 2019) inside the encoder to keep the semantic information while fine-tuning.

4.3 Results

Table 1 presents our ST results for the Tunisian to English Dialectal and Low-resource track. Our primary system obtained a BLEU of 20.7 on our validation set. As shown in the tables, the official evaluation scores appear to be low compared to the good result obtained on the validation set. We suspect that our test submission was not conform to the evaluation specifications. We speculate that this difference between validation and test scores is due to the fact we did not remove the punctuation nor the disfluencies tokens from the case-sensitive translation we submitted, while the evaluation is made lowercase and no punctuation. We mistakenly expected this normalization step to be applied by the organizers instead of the participant. We were able to ask the organizers to evaluate our normalized output after the evaluation period. The results are reported in Table 1. Test2 refers to the IWSLT 2022 evaluation campaign test, and test3 refers to the one of IWSLT 2023. This normalization before the training of our translation model is expected to further improve our results because we believe that the post-deadline fix more accurately reflects our system's true performance.

System	Description	valid	test2	test3
primary	SAMU−XLS−R 100	20.7	9.6	8.8
post-deadline fix	SAMU−XLS−R 100	20.7	18.2	16.3

Table 1: Results for Tunisian Arabic to English translation systems in terms of %BLEU for low-resource (LR)track.

5 Tamasheq-French Experiments

In this section we present our experiments for the Tamasheq-French dataset in the context of the low-resource ST track.

5.1 Data

This dataset, recently introduced in Boito et al. (2022), contains 14 h of speech in the Tamasheq language for the training split which corresponds to 4,444 utterances translated to French. The development set contains 581 utterances (a little bit less than 2 h of speech), the 2022 test set contains 804 utterances (approximatively 2 h of speech). The 2023 test set contains 374 utterances (approximatively 1 h of speech). Additional audio data was also made available through the *Niger-Mali audio collection*: 224 h in Tamasheq and 417 h in geographically close languages (French from Niger, Fulfulde, Hausa, and Zarma).[3] For all this data, the speech style is radio broadcasting, and the dataset presents no transcription.

5.2 Models

For the Tamasheq to French task, we performed several experiments. First of all, we did the same experiment that was done for Pashto-French and Tunisian-English tasks. We used the end-to-end translation model presented in section 3.2, directly trained on the Tamasheq→French task. Directly means that we used SAMU−XLS−R-xx (xx corresponds to the number of languages in the training set, equals to 53, 60 and 100) to initialise the encoder and performed the training of the encoder-decoder model using the Tamasheq→French training set.

We used the CoVoST-2 (Wang et al., 2020) $\mathcal{X}$→EN speech-translation dataset in which we translated the EN text into French (using Mbart Many-to-Many). Additionally, we exploited the Europarl benchmark, which comprises 72 translation tasks (denoted as $\mathcal{X}$→$\mathcal{Y}$), with the source language set ($\mathcal{X}$) consisting of nine languages: FR, DE, ES, IT, PL, PT, RO, NL, and EN. The target language set ($\mathcal{Y}$) is equivalent to the source language set. For the specific training data distribution of each of the 72 translation tasks, refer to (Iranzo-Sánchez et al., 2019).

We trained a translation model using CoVost-2 X→FR,EN and Europarl X→FR, namely models

[3] https://demo-lia.univ-avignon.fr/studios-tamani-kalangou/

System	Description	valid	test 2023
primary	samu100l[cv2_xx→(en,fr)+europarl_xx→fr] + test22	21.39	16.00
contrastive1	samu100l[cv2_xx→(en,fr)+europarl_xx→fr]	**21.41**	**16.52**
contrastive2	samu60l[cv2_xx→(en,fr)+europarl_xx→fr] + test22	20.80	15.84
contrastive3	samu60l[cv2_xx→(en,fr)+europarl_xx→fr]	20.66	15.35
contrastive4	samu100l continue training + test22	21.39	16.30
contrastive5	samu100l continue training	20.78	15.60
baseline	best system from IWSLT2022	8.34	5.70

Table 2: Results of the Tamasheq-French ST systems in terms of BLEU score.

samu60l[cv2_xx→(en,fr)+europarl_xx→fr] and samu100l[cv2_xx→(en,fr)+europarl_xx→fr]). We also translated the French translation of the Tamasheq speech into Spanish, Portuguese and English (still using MBart Many to Many).

Using the pre-trained models, we trained a translation model from Tamasheq to French, Spanish, English and Portugese. We added the 2022 test set inside the training corpus for the Primary model.

Moreover, we used the last checkpoint of the `SAMU−XLS−R` training (100 languages) and pushed further the training using the `LaBSE` embeddings of the translations of the Tamasheq into French, Spanish, English and Portuguese. Then using the specialized Tamasheq `SAMU−XLS−R`, we trained a Tamasheq to French, Spanish, English, Portuguese model.

5.3 Results

Table 2 presents our ST results for the Tamasheq to French task. Our first contrastive model performed better than the Primary model (16.52 for the contrastive model compare to 16.00 for the primary model). This was unexpected because the 2022 test set was added inside the training corpus for the Primary model and not in the contrastive one. The constrative4 and contrastive5 performances (in which we push the training of the `SAMU−XLS−R`-100 model further) are very close to the primary and contrastive1 (16.30 BLEU vs 16.52 BLEU).

We did not use the 224 hours of unlabelled data. We could probably get better results by using pseudo-labelling using our best model and then using the translation for the training of the translation model. Another direction could be the use of another decoder like the recently proposed NLLB model (Costa-jussà et al., 2022).

6 Pashto-French Experiments

In this section, we present our experiments for the first edition of translating Pashto speech to French in the context of the low-resource ST track for IWSLT 2023.

6.1 Data

The Pashto-French dataset used in our experiments was provided by ELDA. This dataset is available in the ELRA catalog, *TRAD Pashto Broadcast News Speech Corpus* (ELRA catalogue, 2016b) concern audio files and *TRAD Pashto-French Parallel corpus of transcribed Broadcast News Speech - Training data* (ELRA catalogue, 2016a) are their transcriptions.

This dataset is a collection of about 108 hours of Broadcast News with transcriptions in Pashto and translations in French text. Dataset is build from collected recordings from 5 sources: Ashna TV, Azadi Radio, Deewa Radio, Mashaal Radio and Shamshad TV. Training data contains 99h of speech in Pashto, which corresponds to 29,447 utterances translated into French.

We participated for Pashto to French task for both types of submissions: constrained and unconstrained conditions. For constrained conditions, systems are trained only on the dataset provided by the organizers, while for unconstrained conditions, systems can be trained with any resource, including pre-trained models.

We investigate two types of ST architectures: end-to-end architectures 6.2, and pipeline 6.3 models.

6.2 Pipeline models

For the cascaded approach, i.e. the task of using an ASR model followed by a MT model, we focused on Wav2Vec2.0 (Baevski et al., 2020) as a Speech-to-Text system. The architecture used is Wav2Vec2-

XLSR-53 (Conneau et al., 2020), a large version of Wav2Vec2 pre-trained on the multilingual dataset Common-Voice (Ardila et al., 2020b). Once adding a language modeling head on top of the model for fine-tuning on the Pashto dataset, we observed a score of less than 20% of WER and a good modeling of the reference language since the difference of the scores for translating written Pashto to written French when using either the reference or the generated Pashto text, was always less than 0.5 of BLEU. For the MT system, we tested multiple approaches using auto regressive Sequence-2-Sequence models.

We mainly focused on transformers encoder-decoder systems from small basic transformers (contrastive3 in Table 3) to large pre-trained multilingual text-to-text transformers such as T5 (Raffel et al., 2020) and mT5 (multilingual T5). For primary cascaded system, models are based on a convolutional model (fconv) (Gehring et al., 2017) upgraded (fconv-up). We reduced the depth and the width of both the encoder and decoder to adapt the size of our fconv model to our dataset. Our fconv-up model achieves 14.52 of BLEU on valid set and 15.56 on the test set, while fconv would give 13 of BLEU. Compared to the cascaded baseline system, based on small basic transformers (contrastive3), fconv-up cascaded system outperforms by 6 BLEU points.

Experiments have been carried out in order to extract the encoder of the fine-tuned W2V and use the latent representation of the audio to train an auto-regressive decoder and thus to skip the Speech-to-Text part, but without any success.

6.3 End-to-end models

We used the end-to-end translation model presented in section 3.2. The model was trained directly on the Pashto to French task (no pre-training of the encoder-decoder model), using SAMU−XLS−R trained on 53 and 100 languages. We used adapters (Houlsby et al., 2019) inside the encoder to keep the semantic information while fine-tuning.

Two constrained contrastive end-to-end systems were submitted for this task. Both share the same encoder-decoder architecture using transformers (Vaswani et al., 2017). The system encoder is the encoder from a Whisper small (768) (Radford et al., 2022) pre-trained model. The decoder has a dimension of 512 using 8 heads and 6 layers. It is not pre-trained. A feed forward network projection layer is used between the encoder and decoder to connect both modules. The difference between both systems lies in the use of a transformer language model trained from scratch on the provided dataset.

Both of these systems were also trained on additional Pashto data and submitted as contrastive unconstrained systems 2 and 3. The language model was not trained on the additional data.

6.4 Results

Results for constrained and unconstrained conditions are presented in Table 3 and Table 4 respectively.

System	Description	Constrained valid	Constrained test
primary	Pipeline, fconv-up	14.52	15.56
contrastive1	E2E, without LM	11.06	15.29
contrastive2	E2E, with LM	11.11	15.06
contrastive3	Pipeline	10.5	9.2

Table 3: Results for Constrained Pashto-to-French ST systems in terms of %BLEU score.

As for constrained setting, we noted that a pipeline of two E2E ASR and NMT system gives better results compared to using one speech translation E2E system. Although the usage of a LM improves the E2E ST further, we were not able to exceed the pipeline of the two E2E systems (ASR+NMT).

System	Description	Unconstrained valid	Unconstrained test
primary	SAMU_XLSR 100l	24.82	24.87
contrastive1	SAMU_XLSR 53l	23.38	23.87
contrastive2	E2E, without LM	12.26	15.18
contrastive3	E2E, with LM	12.16	15.07

Table 4: Results for Unconstrained Pashto-to-French ST systems in terms of %BLEU score.

When we switch to the unconstrained setting, we see a significant improvement demonstrated by a dramatic increases of the BLEU score with the SAMU−XLS−R system. SAMU−XLS−R obtained a BLEU of 24.87 on the test set when trained starting from a pretrained encoder with 100 languages (SAMU−XLS−R-100) and a full BLEU point less (23.87) when we start from a 53 languages encoder (SAMU−XLS−R-53).

7 Conclusion

This paper presents results obtained on three tasks from the IWSLT 2023 Dialectal and Low-resource ST track, namely Tunisian to English, Tamasheq to French and Pashto to French. Given an unconstrained condition, our submission relies heavily on the semantic speech representation learning framework SAMU-XLS-R that greatly improves results compared to the other submitted end-to-end ST models by leveraging multilingual data from other languages. These data can thus come from high resource languages and help to alleviate the low-resource setting difficulty. We indeed observe slightly improved results when using a SAMU-XLS-R model trained on more languages (Tamasheq to French : 15.35 BLEU when using 60 languages, 16.52 BLEU when using 100 languages). We believe results could be further improved by using the unlabelled data available for the Tunisian to English and the Tamasheq to French tasks, and by investigating other decoders in our encoder-decoder framework.

Acknowledgements

This work was partially funded by the following projects: French Research Agency (ANR) ON-TRAC project under contract number ANR-18-CE23-0021, European Commission SELMA project under grant number 957017, European Union's Horizon 2020 ESPERANTO research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101007666 and the DGA RAPID COMMUTE project. This work was partially performed using HPC resources from GENCI–IDRIS, grants AD011012527. The Pashto-French data was totally provided by ELRA. We acknowledge 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.

References

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2019. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Yannick Estéve, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel M. 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, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alex 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)*, Dublin, Ireland. 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, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, et al. 2020. Findings of the iwslt 2020 evaluation campaign. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020a. Common Voice: A massively-multilingual speech corpus. *arXiv:1912.06670*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020b. Common voice: A massively-multilingual speech corpus.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021a. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv:2111.09296*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021b. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2019. Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912*.

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? *CoRR*, abs/2106.01045.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *CoRR*, abs/1612.01744.

Kaushal Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2023. Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier, and Yannick Estéve. 2022. Speech resources in the tamasheq language. *Language Resources and Evaluation Conference (LREC)*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, et al. 2022. Xtreme-s: Evaluating cross-lingual speech representations. *arXiv preprint arXiv:2203.10752*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

ELRA catalogue. 2016a. Trad pashto broadcast news speech corpus. https://catalogue.elra.info/en-us/repository/browse/ELRA-S0381/. ISLRN: 918-508-885-913-7, ELRA ID: ELRA-S0381.

ELRA catalogue. 2016b. Trad pashto-french parallel corpus of transcribed broadcast news speech - training data. http://catalog.elda.org/en-us/repository/browse/ELRA-W0093/. ISLRN: 802-643-297-429-4, ELRA ID: ELRA-W0093.

Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, et al. 2021a. Task agnostic and task specific self-supervised learning from speech with *LeBenchmark*. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2021b. *LeBenchmark*: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *Interspeech*, pages 1439–1443.

F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th ACL*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proc. ICML*.

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.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2019. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. *arXiv:1911.03167*.

Tahir Javed, Kaushal Santosh Bhogale, Abhigyan Raman, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2022. Indicsuperb: A speech processing universal performance benchmark for indian languages. *arXiv preprint arXiv:2208.11761*.

Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, and Aaron van den Oord. 2020. Learning robust and multilingual speech representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1182–1192, Online. Association for Computational Linguistics.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–13.

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. *arXiv preprint arXiv:2011.00747*.

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:2010.12829*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *arXiv:2012.03411*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

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.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Matthias Sperber, Hendra Setiawan, Christian Gollan, Udhyakumar Nallasamy, and Matthias Paulik. 2020. Consistent transcription and translation of speech. *Transactions of the Association for Computational Linguistics*, 8:695–709.

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, Juan Pino, Anne Wu, and Jiatao Gu. 2020. Covost: A diverse multilingual speech-to-text translation corpus. *arXiv:2002.01320*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. 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.

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

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Interspeech*, pages 1194–1198.

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