NAVER LABS Europe's Multilingual Speech Translation Systems for the IWSLT 2023 Low-Resource Track

Edward Gow-Smith[1*] **Alexandre Bérard**[2†] **Marcely Zanon Boito**[2†] **Ioan Calapodescu**[2]

[1] University of Sheffield [2] NAVER LABS Europe

egow-smith1@sheffield.ac.uk first.last@naverlabs.com

Abstract

This paper presents NAVER LABS Europe's systems for Tamasheq-French and Quechua-Spanish speech translation in the IWSLT 2023 Low-Resource track. Our work attempts to maximize translation quality in low-resource settings using multilingual parameter-efficient solutions that leverage strong pre-trained models. Our primary submission for Tamasheq outperforms the previous state of the art by 7.5 BLEU points on the IWSLT 2022 test set, and achieves 23.6 BLEU on this year's test set, outperforming the second best participant by 7.7 points. For Quechua, we also rank first and achieve 17.7 BLEU, despite having only two hours of translation data. Finally, we show that our proposed multilingual architecture is also competitive for high-resource languages, outperforming the best unconstrained submission to the IWSLT 2021 Multilingual track, despite using much less training data and compute.

1 Introduction

The vast majority of speech pipelines are developed for *high-resource* languages, a small percentage of languages that have ample amounts of annotated data available (Joshi et al., 2020). However, the assessment of systems' performance based only on high-resource settings can be problematic, since it fails to reflect the real-world performance these approaches will have in diverse and smaller datasets. Moreover, as around half of the world's languages are considered to be not only *low-resource*, but also from oral tradition (i.e., without a written form), there is an urgent need for speech technology that can operate robustly in such *low-resource settings* (Bird, 2011). In this context, the *IWSLT conference*[1] proposes low-resource speech translation (ST) challenges that allow the speech community to realistically benchmark ST approaches using diverse and representative datasets. This paper describes NAVER LABS Europe's (NLE) submission to two of the language pairs from the IWSLT 2023 (Agarwal et al., 2023) Low-Resource Track: Tamasheq-French (*Taq-Fr*) and Quechua-Spanish (*Que-Es*).

Most successful approaches for tackling scenarios where ST data is scarce perform transfer learning across languages and modalities, leveraging multilingual pre-trained models for both speech and text (Anastasopoulos et al., 2022). However, due to the large number of parameters of current Transformer-based (Vaswani et al., 2017) approaches, training such systems is computationally expensive and not accessible to everyone. **NLE's submission focuses on a multilingual parameter-efficient training solution that allows us to leverage strong pre-trained speech and text models to maximize performance in low-resource languages.**

We present new SOTA results for the *Taq-Fr* pair (17 hours of training data) that represent a 57% BLEU increase compared to the results achieved by Khurana et al. (IWSLT 2022 post-evaluation).[2] This same system achieves 23.6 BLEU on the IWSLT 2023 test set, an improvement of 7.71 BLEU compared to the second best result submitted this year. We also present SOTA results in the unconstrained setting for the *Que-Es* pair (2 hours of training data), while maintaining most of the performance in the *Taq-Fr* pair. In addition, to showcase the usefulness of our parameter-efficient multilingual solution we evaluate it on the high-resource setting of the IWSLT 2021 Multilingual Task (Anastasopoulos et al., 2021). We find that our approach outperforms the best IWSLT 2021 submission (FAIR, Tang et al., 2021), despite training considerably fewer parameters (-64%), and using substantially

[*] Work done during an internship at NAVER LABS Europe.

[†] Equal contribution

[1] https://iwslt.org/

[2] https://www.clsp.jhu.edu/jsalt-2022-closing-presentations/

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 144–158
July 13-14, 2023

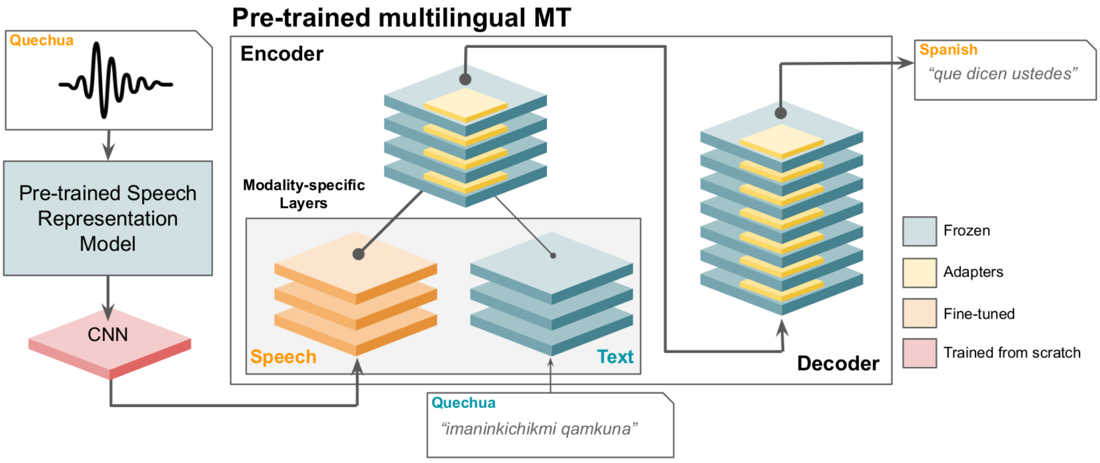

Figure 1: An illustration of our multilingual ST architecture as described in Section 2. The bold arrow path corresponds to the speech-to-text training path. At decoding time, we can choose between producing speech-to-text or text-to-text translations. Figure best seen in color.

less training data and compute.

This paper is organized as follows. We first describe the architecture and training settings of our multilingual ST systems in Section 2. We next list the resources we use in Section 3. Section 4 presents our results in both low and high-resource settings. Lastly, we highlight the zero-shot potential of our approach in Section 5 and present our concluding remarks in Section 6.

2 System Description

In this work we focus on a parameter-efficient training solution that allows us to input the features from a pre-trained speech representation model into a pre-trained multilingual MT model, producing translations from both speech and text in multilingual settings. This setting also allows us to leverage automatic speech recognition (ASR; i.e. speech-to-transcript) data. The general architecture is presented in Figure 1. The architecture is considered *parameter-efficient* because a small portion of its parameters are trained (bottom encoder layers and small adapters layers).

Architecture. We initialize our models with a pre-trained multilingual MT model, which we adapt to the ST task by inputting features extracted with a frozen pre-trained speech representation model. The MT model is also frozen, except for the bottom 2 or 3 encoder layers and small adapter modules (those introduced by Bapna and Firat (2019), with bottleneck dimension 64) added after each encoder and decoder layer. As we show in our results, the fine-tuned encoder layers are able to map the speech features into the representation space of the pre-trained MT model and the adapters can help with domain adaptation (and possibly help alleviate the length mismatch). At inference, this model can be used for MT with very little memory overhead: the convolutional layers and adapters are disabled, and the bottom encoder layers are swapped with those of the initial pre-trained model.

Training settings. We train on 4 V100 GPUs (80GB) for up to 200 000 updates, with a maximum batch size of 4 000 source features (or 80 seconds of audio) and accumulated gradients over two batches.[3] We sample language pairs with a temperature of 3.[4] We validate every 5 000 updates and perform early stopping on valid BLEU for the language pair(s) of interest, with a patience of 5, averaging model weights across the last 3 checkpoints.[5] We find best results using a single convolutional layer with stride 2, which downsamples the sequence of speech features by a factor of 2. The other hyperparameters are listed in Appendix Section A.1.

[3] This corresponds to a total of 32 000 features per update, or 640 seconds of audio. In practice, with padding, each update corresponds to approximately 80 utterances or 530 seconds of audio.

[4] $p_k = u_k^{1/3} / \sum u_i^{1/3}$ where u_k is the utterance count for language pair k.

[5] While all the configurations presented in this paper use checkpoint averaging, we later re-trained our contrastive submission for *Taq-Fr* and found virtually the same results without it.

Model	# params	Transformer layers	Feature dimension
Tamasheq (Boito et al., 2022b)	95M	12	768
Niger-Mali (Boito et al., 2022b)	95M	12	768
mHuBERT-Tamasheq	95M	12	768
XLSR-53 (Conneau et al., 2021)	317M	24	1024
XLS-R (Babu et al., 2022)	317M	24	1024

Table 1: Speech representation models. The top portion presents *Tamasheq-dedicated* models, while the bottom lists large *general purpose* multilingual models.

3 Resources

3.1 Pre-trained Speech Representation Models

We experiment with different versions of two speech representation models: HuBERT (Hsu et al., 2021) and wav2vec 2.0 (Baevski et al., 2020). We do not fine-tune these models in any of our configurations, but instead use them as feature extractors (see Figure 1). Because of this, our models are sensitive to the layer we extract features from. Pasad et al. (2021) argue that, for wav2vec 2.0 models that are not fine-tuned on ASR, speech features from *middle* layers tend to have a higher abstraction from the speech signal, which is beneficial to downstream tasks. The results from Boito et al. (2022b) seem to confirm this observation holds for low-resource ST. To the best of our knowledge, there is no similar investigation for HuBERT models.[6]

Table 1 presents the speech representation models we experiment with. The *Tamasheq* model is a monolingual wav2vec 2.0 Base model trained on 243 h of Tamasheq speech. The *Niger-Mali* is a wav2vec 2.0 Base model trained on the same Tamasheq speech data plus 111 h of French, 109 h of Fulfulde, 100 h of Hausa, and 95 h of Zarma. This gives 658 h in total. The data for both models is sourced from the Niger-Mali audio collection (Boito et al., 2022a). The unreleased *mHuBERT-Tamasheq* model uses this same audio collection for training, while also including Common Voice (Ardila et al., 2020) data in four other languages (English, French, Arabic and Kabyle), resulting in 5 069 h of speech. *XLSR-53* (56k hours) and *XLS-R* (500k hours) are massively multilingual wav2vec 2.0 Large models covering 53 and 128 languages, respectively. Neither of these two multilingual models have seen Tamasheq or Quechua speech during training.[7]

[6]We hypothesize that layer selection is less important for HuBERT architectures due to the multi-iteration approach that increases signal abstraction at each iteration.

Task	Source	Target	hours:minutes	# utterances
ASR	Quechua	Quechua	51:39	8,301
ST	Quechua	Spanish	2:42	698
ST	Tamasheq	French	15:43	5,025

Table 2: Speech Translation (ST) and Speech Recognition (ASR) data provided by the organizers (train+valid). The ASR data is outside of the constrained setting.

3.2 Pre-trained Multilingual MT Models

To initialize our ST models, we first experimented with mBART for many-to-many translation (mBART50NN; Tang et al., 2020), but found the NLLB-200 models (Costa-jussà et al., 2022) to give better results. We experiment with the dense NLLB models of various sizes: the distilled 600M-parameter and 1.3B-parameter versions, and the 3.3B-parameter version. We end up using the larger versions in our submissions (1.3B and 3.3B). Note that NLLB covers 202 languages, including Tamsheq and Quechua, which is not the case for mBART. At the same model size, despite covering more languages, NLLB is also a stronger machine translation model overall than mBART. Also, unlike mBART, it is not English-centric.

Contrary to Tang et al. (2021), we keep the original mBART or NLLB vocabularies of size 250k and do not train any embeddings. Instead, like Berard et al. (2021), we find that it is possible to filter the vocabulary at test time to only cover the languages of interest, significantly reducing the memory footprint of the model with a minor reduction in performance.[8] We can also filter the vocabulary and embeddings before ST fine-tuning and achieve the same performance as with the full vocabulary without needing to train any embeddings. See Table 14 in Appendix for a comparison of these approaches. In order to study the zero-shot translation capabilities of our models (i.e., translating to languages and language pairs unseen at training), we do not apply vocabulary filtering to the configurations presented in the main paper.

[7]Appendix Table 16 lists all models with links for downloading checkpoints, when available.

[8]With NLLB, 44k tokens are enough for a 100% coverage of the training data (mTEDx, TED-LIUM, Quechua, Tamasheq), or 35k when restricting to our *Taq-Fr* setting. This represents a reduction of more than 200M parameters.

Task	Source	Target	hours:minutes	# utterances
ASR	English	English	208:00	91,003
ASR	French	French	218:59	117,081
ASR	Spanish	Spanish	214:15	103,076
ST	French	English	57:39	31,207
ST	French	Spanish	42:14	21,862
ST	Spanish	English	79:37	37,168
ST	Spanish	French	9:34	4,568

Table 3: ASR and ST data in English, French and Spanish sourced from TED talks (unconstrained setting).

3.3 Datasets

We tackle the low-resource setting by building multilingual systems that utilize both ASR and ST data in the languages of interest (Tamasheq and Quechua), and in high-resource directions whose target language is of interest (French and Spanish). Note that we also include X→English data, as we initially planned to participate in the Irish-English task. Including more data in high-resource languages has several advantages. Firstly, it has a regularization effect that prevents us from immediately overfitting the low-resource training data. Secondly, this enables knowledge transfer from common target languages and from similarly-sounding source languages.[9] Thirdly, as we build multilingual ST systems by mapping the speech representation vectors into the same space as the multilingual MT model, our goal is to produce a model that is *as multilingual as possible*, not specializing in one specific language. Our results show that training on multiple languages at once achieves this effect, while also producing good zero-shot ST results.

Table 2 presents statistics for the datasets provided by the IWSLT 2023 organizers. The *Que-Es* dataset[10] is an unreleased dataset prepared for this year's challenge. It corresponds to a translated subset of the Quechua ASR data ("Siminchik") from Cardenas et al. (2018). The *Taq-Fr* dataset was introduced by Boito et al. (2022a). Table 3 presents statistics for the datasets in high-resource languages. English ASR data comes from TED-LIUMv2 (Rousseau et al., 2014), and the other data comes from mTEDx (Salesky et al., 2021). Appendix Table 15 lists the datasets used in each of our submissions. In Section 4.3, we also run experiments in the setting of the IWSLT 2021 Multilingual Task to measure how good our approach is on high-resource languages. The datasets used for this setting are presented in Appendix Table 10.

[9] Manual inspection revealed that audio from both datasets presents some degree of target language borrowing (e.g., Spanish words present in the Quechua speech, French words present in the Tamasheq speech).

[10] We are aware the dataset reference is *Que-Spa*. We chose to use the ISO 639-1 two letters abbreviation for Spanish for consistency with the other datasets used in this work.

		Taq-Fr IWSLT 2022	Taq-Fr IWSLT 2023	Que-Es IWSLT 2023
Taq-Fr	primary	**20.75**	**23.59**	✗
	contrastive 1	19.06	21.31	✗
	contrastive 2	18.58	18.73	**17.74**
Que-Es	primary	18.58	18.73	**17.74**
	contrastive 1	16.84	✗	15.67
	contrastive 2	16.21	✗	15.25

Table 4: Results on the official test sets for the IWSLT 2023 Low-Resource Task. We also show results on the IWSLT 2022 *Taq-Fr* test set. Note that all Quechua models are trained on Tamasheq data, but the reverse is not true (see Appendix Table 15). Lines 3 and 4 correspond to the same model.

4 Experiments and Results

All our submissions to the low-resource ST task are in the *unconstrained* setting, due to the use of pre-trained models, and from training on data in other languages. The datasets used in each submission are listed in Appendix Table 15. This section is organized as follows. We present our *Taq-Fr* results (4.1) with a detailed ablation study justifying our architectural choices. We then present our *Que-Es* results (4.2). Lastly, we evaluate and analyze our approach in a high-resource setting (4.3).

4.1 Tamasheq-French Results

We submit two systems that have *Taq-Fr* as the only low-resource language pair (**primary** and **contrastive 1**). Additionally, we take our primary submission for *Que-Es*, which has also been trained on *Taq-Fr*, and submit this as **contrastive 2**. The top portion of Table 4 gives the test BLEU scores, and the top portion of Appendix Table 11 presents the valid BLEU scores. Table 12 shows statistics (average and standard deviation) over multiple runs when applicable.

System description. The **contrastive 1** model uses as a speech feature extractor the *Niger-Mali* wav2vec 2.0 model (8^{th} layer). It was initialized with NLLB 1.3B, whose bottom 3 encoder layers were finetuned. We took three runs of this setting with different random seeds and picked the best performing one on the validation set (in terms of

Taq-Fr BLEU) as our contrastive submission. We then ensembled the three runs as our **primary** submission. Finally, **constrastive 2** is the ensemble model used as primary submission to the *Que-Es* task, which covers both low-resource languages, and combines *XSL-R Large* with *NLLB 3.3B*.

Results. Our primary submission significantly outperforms the previous state of the art of 13.2 BLEU (+7.5 BLEU) on the IWSLT 2022 test set by Khurana et al. (2022).[11] It also ranks first in this year's edition, with +7.7 BLEU over the second best primary submission. Our contrastive submissions rank second and third (beating the second best primary submission by +5.4 and +2.8 BLEU).

4.1.1 Ablation Study

In Appendix Table 18 we compare our **contrastive 1** model (the non-ensembled version of our primary submission) with other architectures trained on the same data to validate our choice of hyperparameters.

Speech features. The wav2vec 2.0 models trained with Tamasheq (*Niger-Mali* and *Tamasheq*) largely outperform the well-known massively multilingual models (*XLSR-53* and *XLS-R*) on *Taq-Fr* (e.g. +2.5 BLEU *Tamasheq* compared to *XLS-R L*). These models are larger and trained on considerably more data, but do not include any Tamasheq speech. Similar to previous works (Pasad et al., 2021; Boito et al., 2022b), when extracting features from wav2vec 2.0 we find that the 8$^{\text{th}}$ layer gives better results than the 11$^{\text{th}}$ (penultimate) layer (+2.5 BLEU for *Niger-Mali*).

For HuBERT, on the contrary, features from the 11$^{\text{th}}$ layer give the best results (+0.2 BLEU compared to 8$^{\text{th}}$ layer). When using the *right layer*, we find that wav2vec 2.0 outperforms HuBERT (+2.7 BLEU *Niger-Mali* compared to *mHuBERT-Taq*).

Finally, *Niger-Mali* is as good on *Taq-Fr* as the *Tamasheq* wav2vec 2.0, but performs considerably better on *Fr-En* (+4.1 BLEU), probably because it was trained with French audio. The best *Fr-En* performance is achieved with *XLS-R L*. We find worse performance on *Fr-En* with *XLS-R XL* (-2.0 BLEU), but this may be due to layer selection.

Pre-trained MT model. The larger the model used for initialization, the better the performance (even more so for *Fr-En*). However, we find that the gain from using NLLB 3.3B over NLLB 1.3B is too small to justify the increase in model size and decoding latency (3 times slower). At the same model size, NLLB 600M performs considerably better than mBART (+1.7 BLEU on *Taq-Fr*, +3.6 BLEU on *Fr-En*).

Trained parameters. Fine-tuning too many encoder layers results in overfitting, which hurts *Taq-Fr* and *Fr-En* performance. On the other hand, fine-tuning just 1 or 2 layers instead of 3 does *not* result in a large BLEU drop. Similarly, adapter modules are not always needed. Disabling decoder adapters does not degrade *Taq-Fr* performance (+0.2 BLEU), but results in a slight drop in *Fr-En* performance (-0.9 BLEU), which could be attributed to a domain adaptation effect (to the mTEDx domain). Disabling encoder adapters has more impact on performance for *Taq-Fr* (-0.8 BLEU), with similar effect on performance for *Fr-En* (-1.0 BLEU). Section 4.3 shows that these adapters are important for domain adaptation.

Convolutions. The number of convolutional layers does not impact performance much (range of 1.1 BLEU on *Taq-Fr* and 3.2 BLEU on *Fr-En* for 0 to 3 layers), but it can have a large impact on decoding speed: each layer divides the input length by a factor of 2 resulting in a roughly 3.5$\times$ speed-up from 0 to 3 layers. Interestingly, even though it was trained on much shorter sequences, the MT model seems to adapt quite well to any input length, even without any convolutions – we achieve a better *Taq-Fr* result without any convolutions, but a worse *Fr-En* result.[12] However, models with fewer convolutional layers seem to converge faster (as shown in Appendix Figure 2).

Stacked layers. While our approach described in Section 2 fine-tunes some parameters of the pre-trained MT model, we can instead plug new Transformer layers at the bottom of the encoder, without changing any existing parameter. These "stacked layers" result in slightly larger models but are conceptually simpler, as they try to map the speech features into the same representation space as the input text embeddings of the MT model. Appendix Table 17 compares this architecture with the one used in our submission to the *Taq-Fr* task. We see

[11] Here we are referencing the model pre-trained using the Niger-Mali dataset that was presented at JSALT 2022: https://www.clsp.jhu.edu/jsalt-2022-closing-presentations/

[12] Without any convolution, the speech feature to target token ratio is **12:1**.

that it performs similarly well (sometimes better) and that it does not add any noticeable decoding latency. We can even reach the same *Taq-Fr* performance as our contrastive submission by just adding a single Transformer layer plus one convolution layer and small adapters (28M trained parameters in total). Finally, disabling all adapters only results in a small BLEU drop, suggesting that it is indeed possible to map the speech features into the text input space, with only one Transformer layer. This is surprising, considering that the input to this layer is 6 times as long as the target sequence on average.

4.2 Quechua-Spanish Results

The test and validation scores of our submissions to the *Que-Es* task are reported in the second half of Table 4 and 11, respectively. Because these models are also trained on *Taq-Fr* data, we additionally report their performance on that task.

System description. As we do not have a speech feature extractor specialized to Quechua speech, our **contrastive 1** submission uses a massively multilingual wav2vec 2.0 model: XLS-R Large (18^{th} layer). Compared to our Tamasheq submission, it is also initialized with a larger MT model (NLLB 3.3B), which we found to perform better in this setting. The training settings are the same as for the Tamasheq models, except that we only fine-tune the bottom 2 encoder layers (instead of 3) and validate every 2 500 updates, since this larger model tends to converge faster. Another difference is that we train on both Tamasheq and Quechua data (in addition to the mTEDx and TED-LIUM data). Like in our Tamasheq submission, we train 3 models with different random seeds and ensemble them as our **primary** submission. Our **constrastive 2** submission uses a single model with the same training settings, but starts from a smaller pre-trained MT model (NLLB 1.3B).

Results. Our primary submission in the *Que-Es* task also ranked first, with 17.7 BLEU on the official test set. The full ranking results were not communicated in time to this camera-ready. They will be made available later through the conference findings paper (Agarwal et al., 2023).

Data contamination. We found shortly after our submission that all the audio files used in the official test and validation sets are also present in the ASR training data shared by the organizers for the unconstrained setting. This means that our *Que-Es* ST models are evaluated in an unrealistic setting, where they are tasked to translate Quechua utterances of which they already know the transcription into Quechua. For this reason, we filtered the ASR data to remove all audio files also present in the validation and test sets for *Que-Es*, and we re-trained models on this filtered data.[13] While our official submission results presented in Table 4 use the "contaminated" dataset for comparison with the other submissions, we think any future comparison to our work should be done with the updated results in Appendix Table 11. Note that similar care should be taken with the results of other participants.

4.3 Results and Analysis in a High-Resource Setting

The results of our ablation studies (Section 4.1.1) seem to indicate that our models are reasonably good on *Fr-En* translation, even though we do early stopping and tune our hyper-parameters based on *Taq-Fr* performance. Here, we further investigate the performance of our approach on high-resource ST by training models in the setting of the IWSLT 2021 Multilingual Task (Anastasopoulos et al., 2021). This task evaluates the performance of multilingual ST models in 4 *training directions*, for which in-domain training data is provided, and 3 *zero-shot directions*, for which no training data is provided.

We use *XLS-R Large* as the speech feature extractor, experiment with both *NLLB 1.3B* and *NLLB 3.3B* as the MT model, and perform early stopping based on the average validation BLEU across the 4 official training directions. We train our models on all the mTEDx language pairs that are not zero-shot, along with TED-LIUM (English ASR) and the Tamasheq and Quechua data (see Table 15). Note that the use of pre-trained models and English ASR means our models fall into the unconstrained setting.

Table 5 presents our results on this task, compared with the best unconstrained submission (FAIR; Tang et al., 2021).[14] We find that both our models outperform FAIR's ensemble submission in the training directions, even though they require substantially less compute and data to train, and they are not ensembled. In the zero-shot direc-

[13] In the updated version, we use NLLB 1.3B by default instead of NLLB 3.3B, like for *Taq-Fr*. Appendix Table 11 presents *uncontaminated* results.

[14] SacreBLEU signature (Post, 2018): `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0`

Model	Total params	Trained params	Training directions				Zero-shot directions		
			Es-En	Fr-En	Fr-Es	Pt-En	Pt-Es	It-En	It-Es
FAIR at IWSLT 2021	700M		40.4	36.4	34.4	29.0	34.4	28.4	34.6
(Tang et al., 2021)	3×700M (ensemble)		42.2	38.7	36.5	31.0	**38.2**	**29.4**	**37.3**
XLS-R + NLLB 1.3B	317M + 1.38B	70M	43.7	39.4	38.0	31.5	35.9	28.9	35.0
XLS-R + NLLB 3.3B	317M + 3.36B	115M	**44.0**	**39.9**	**38.3**	**33.1**	38.1	29.3	36.9
XLS-R + NLLB 1.3B, ASR + MT cascade			41.8	35.6	34.4	29.7	35.8	29.3	35.2

Table 5: Results on the IWSLT 2021 Multilingual task. We report BLEU scores on the IWSLT 2021 test sets. Our NLLB 1.3B and 3.3B models took respectively 34 and 46 h to train on 4 V100 GPUs, while FAIR's models each took 7 days to train on 8 V100 GPUs. Also note that FAIR's models were trained on much larger amounts of data, **including data for the "zero-shot" directions** (which, in their case is only zero-shot w.r.t the in-domain TED data).

Model	New params	Taq-Fr
Joint training	0	**21.06**
Adapters 64 (all)	6.4M	17.60
Adapters 256 (all)	15.9M	18.18
Adapters 256 (bottom)	1.6M	19.24
Conv + Adapters 256 (bottom)	2.5M	19.13

Table 6: BLEU scores on the *Taq-Fr* validation set, when training jointly with IWSLT 2021 and Tamasheq data; versus incremental (2-stage) training. The "New params" columns give the number of Tamasheq-specific parameters added.

tions, our NLLB 1.3B version performs worse than FAIR's ensemble, which is not surprising since they used training data for the zero-shot language directions (from other datasets), whilst we do not.[15] We find that using the larger NLLB 3.3B model for initialization considerably improves our zero-shot results.

4.3.1 Incremental Learning

A limitation of our approach for low-resource ST is that we need to know in advance (when training the multilingual ST model) the set of low-resource languages to cover. Here, we show that it is possible to add a new low-resource language into an existing model without re-training it, similar to what has been previously done by Berard (2021) for text-to-text MT. We train a model following the IWSLT 2021 setting presented above, but without any Tamasheq or Quechua data. Then, we attempt to adapt it to *Taq-Fr* using four different approaches: **1)** adding adapters of dimension 64 in the bottom layers and training all adapters (including in the decoder layers and top encoder layers); **2)** adding adapters of dimension 256 in the bottom layers and fine-tuning all adapters; **3)** adding adapters of dimension 256 in the bottom layers and training only those; **4)** adding adapters of dimension 256 in the bottom layers and training both those and the convolutional layer.

We keep the same training settings as before, except that: we train on *Taq-Fr* data only; we train only the parameters mentioned above; we validate more often (every 1 000 updates); and we disable checkpoint averaging. Table 6 shows the performance of these four incremental training methods, compared to training on the entire language set from scratch. Even though incremental training does not perform quite as well, it appears to be a viable option that can achieve decent results. Lastly, we highlight that our experiments were limited to these four incremental learning settings (without hyper-parameter search), and that better results may be obtained with other parameter-efficient adaptation methods, or with more regularization.

4.3.2 Multimodality and Domain Transfer

Since our systems are initialized with an MT model, of which just a few encoder layers are modified, it is straightforward to use our ST models for text-to-text translation: we just need to store both the MT and ST bottom layers and route tokens through the MT ones (see Figure 1). However, one question that remains is whether the ST adapters can be used for text-to-text decoding.

As an investigation of this, Appendix Table 19 measures the MT performance (NLLB 1.3B) on the IWSLT 2021 test sets (same domain as the mTEDx training data) with and without the ST adapters. Surprisingly, we see that not only can we use these adapters for both text and speech modalities, but they actually improve the MT scores (+2.7 BLEU on average), even though they were only trained with ST and ASR data. This suggests that the fine-tuned bottom layers are able to fully map the speech representations into the text represen-

[15] NLLB has been pretrained on these language pairs for MT, but we do not train on ST data for them.

Adapter Size	Encoder Adapters	Decoder Adapters	Taq-Fr BLEU	Taq-En BLEU	Taq-Ko BLEU	Taq-Fr chrF	Taq-En chrF	Taq-Ko chrF
64	✓	✓	19.1	**17.1**	12.6	44.2	40.8	18.2
128	✓	✓	19.2	16.7	9.6	**44.7**	40.3	14.5
64	✓	✗	**19.3**	16.8	14.6	44.4	**42.4**	21.5
✗	✗	✗	17.5	16.2	14.4	43.0	40.8	21.5
ST (contrastive 1) + MT (NLLB 1.3B) cascade			✗	15.0	**15.7**	✗	38.6	**22.2**

Table 7: BLEU and chrF results for *Taq-{Fr, En, Ko}* using **contrastive 1** and its variants (models trained without adapters or with larger adapters), on the IWSLT 2022 *Taq-Fr* test set or silver-standard Korean and English references obtained with MT. The last row is a cascade of speech translation followed by text translation (Taq→Fr→X).

tation space and that the adapters further improve performance by allowing domain adaptation of the MT model (which is hard to do at the very bottom layers). Note that the encoder adapters seem to be the most important ones, which is consistent with the findings of Cooper Stickland et al. (2021) that adapting the encoder is the most effective strategy for domain adaptation. Lastly, we highlight that adapting the MT model directly with MT data (mTEDx's transcriptions and translations) gives even better results (+4.6 BLEU on average), but this cross-modality domain transfer is an interesting by-product of our parameter-efficient approach.

5 Zero-Shot Capabilities

Throughout this paper we have argued that one advantage of the multilingual models we propose is their potential for zero-shot translation, a setting in which a system produces translation in an unseen language pair by leveraging its existing knowledge of both languages. In Section 4.3 we showed that our models are competitive with the best submission to IWSLT 2021 on the three zero-shot high-resource language pairs, despite the fact that these pairs were not truly zero-shot for that system. In this section, we further illustrate the zero-shot capabilities of our models by translating Tamasheq speech in two settings: **1)** target language seen during both MT pre-training and ST adaptation (English); **2)** target language only seen during MT pre-training (Korean).

Evaluation settings. To score BLEU and chrF[16] in the chosen target languages, we use a commercial translation service to translate the French side of the IWSLT 2022 test set to English and Korean. Note that this is only a *silver-standard* made of synthetic data, and thus the evaluation will inevitably be biased.[17] Our goal is solely to assess whether our systems have *some* zero-shot ST abilities. We evaluate our *Taq-Fr* **contrastive 1** system, and variants of this system with fewer or larger adapters. We compare with a *cascade* baseline, in which we first perform *Taq-Fr* ST, followed by *Fr-En* or *Fr-Ko* MT using the text-to-text path from Figure 1. In this setting, the adapters are disabled during MT.

Results. In Table 7, we measure the zero-shot translation capabilities of our approach on this silver-standard test set. We evaluate four models: our **contrastive 1** submission presented in Section 4.1, and variants of this model with increased adapter size, adapters only in the encoder, or no adapters. We compare against a cascade baseline that is not zero-shot, which consists in translating the Tamasheq speech into French text and then translating this text into English or Korean.

We observe that, in the case of English, which was seen during ST adaptation, adapters can be helpful (+2 BLEU over the cascade baseline). On the other hand, for Korean, unseen during ST adaptation, systems with adapters in the decoder (first two rows) perform worse, as they likely bring some degree of *language confusion*. Results are even worse with larger adapters, with over 40% of output sentences being in the wrong language. In this setting, the best results are achieved with only encoder adapters or no adapters at all (-1 BLEU compared to the baseline).

Appendix Table 13 measures the percentage of output sentences in the correct language and the percentage of Hangul versus Latin character in each system's outputs. We find that models with

[16]SacreBLEU signature: `nrefs:1|case:mixed|eff:no|tok:X|smooth:exp|version:2.3.1`, (En: `X=13a`, Ko: `X=ko-mecab-0.996/ko-0.9.2-KO`). chrF signature: `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1`

[17]For instance, we observe that these generated translations contain both the Korean transliteration in Hangul of named entities and the original version in the Latin script. This will likely penalize our produced translation during scoring.

Utterance id	Target	Content
2016-11-23_id_7	**Ref**	Chers auditeurs, rappelez-vous que vous écoutez **Studio Kalangou** en ce moment.
	Fr	Chers auditeurs, n'oubliez pas que vous êtes avec le **Studio Kalangou**.
	En	Well, listeners, don't forget that you are with **Studio Kalangou** right now.
	Ko	청취자 여러분, 지금 **Studio Kalangou**와 함께 있는 것을 잊지 마세요.
2016-06-27_id_5	**Ref**	Les examens du **BEPC** sont terminés et les corrections ont commencé hier après-midi dans la ville de Niamey.
	Fr	Les examens du **BEPC** sont terminés et sur toute l'étendue du territoire, les travaux de leur suivi ont débuté hier après-midi à Niamey.
	En	The **BEPC** exams are over and throughout the country, the monitoring activities started yesterday afternoon in Niamey.
	Ko	**BEPC** 시험은 끝났습니다. 전국에서 검사 작업은 어제 오후 Niamey에서 시작되었습니다.
2016-10-27_id_39	**Ref**	D'autres informations que nous apportons aujourd'hui concernent un projet appelé **aniamey.com** qui informe que l'État du Nigéria a refoulé des Nigériens, au nombre de 53, qui arrivent (), qui habitent dans la ville de Mina sur le territoire du Niger ou Neja.
	Fr	D'autres informations que nous apportons aujourd'hui concernent les informations apportées par un programme dénommé **Niamey Point Com** qui a apporté des informations selon lesquelles le Nigeria a accueilli 53 Nigériens qui habitent la ville de Mena qui se trouve sur le territoire du Niger ou le Niger.
	En	Today, we're going to talk about the information about a program called **Niamey Point Com**, which reports that Nigeria has brought back 53 Nigerians who live in the town of Mena in Niger.
	Ko	우리게임의 오늘 기사에서는 **Niamey Point Com**라는 프로그램으로 나이지리아가 미네에 거주하는 53명의 니그르인을 귀환시켰다는 소식이 있습니다.

Table 8: Some decoding examples for *Taq-Fr*, *Taq-En* and *Taq-Ko* language pairs, accompanied by the French reference (Ref). Utterance id corresponds to the suffix of the audio files in the IWSLT 2022 test set.

adapters in the decoder (first two rows) generate more Latin characters. Note that the ideal translation is not necessarily 100% Hangul, as it might sometimes be best to keep the foreign named entities in the Latin alphabet. Table 8 illustrates this with a few examples of translations from our **contrastive 1** system.

6 Conclusion

In this paper we presented our parameter-efficient multilingual systems as submissions to the IWSLT 2023 Low-Resource Task in the Tamasheq-French and Quechua-Spanish language pairs. The architecture we propose has several advantages: it is computationally and data efficient, it allows the same model to do both speech-to-text and text-to-text translation (or transcription), it maximizes knowledge transfer to improve low-resource performance, and it has good zero-shot translation capabilities. Our submissions reach a new state of the art performance, winning both speech translation challenges, especially for Tamasheq-French, where we outperform the previous state of the art by more than 7 BLEU points.

Future work will include a comprehensive evaluation of the ASR capabilities of our architecture, and the investigation of adapters inside the speech representation model. Moreover, when the speech representation model is frozen, a more in-depth analysis of the optimal layer is needed.

Acknowledgements

This work was partially funded by the European Horizon 2022 project UTTER (Unified Transcription and Translation for Extended Reality), under grant agreement No 101070631.

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.

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.

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

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. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

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.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Alexandre Berard. 2021. Continual learning in multilingual NMT via language-specific embeddings. In *Proceedings of the Sixth Conference on Machine Translation*, pages 542–565, Online. Association for Computational Linguistics.

Alexandre Berard, Dain Lee, Stephane Clinchant, Kweonwoo Jung, and Vassilina Nikoulina. 2021. Efficient inference for multilingual neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8563–8583, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Steven Bird. 2011. Bootstrapping the language archive: New prospects for natural language processing in preserving linguistic heritage. *Linguistic Issues in Language Technology*, 6(4).

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier, and Yannick Estève. 2022a. Speech resources in the Tamasheq language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2066–2071, Marseille, France. European Language Resources Association.

Marcely Zanon Boito, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani, Ha Nguyen, Florentin Barbier, Souhir Gahbiche, and Yannick Estève. 2022b. 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.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. *ISI-NLP 2*, page 21.

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

Asa Cooper Stickland, Alexandre Berard, and Vassilina Nikoulina. 2021. Multilingual domain adaptation for NMT: Decoupling language and domain information with adapters. In *Proceedings of the Sixth Conference on Machine Translation*, pages 578–598, Online. Association for Computational Linguistics.

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*.

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.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the 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*, 16(6):1493–1504.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2014. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3935–3939, Reykjavik, Iceland. European Language Resources Association (ELRA).

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. Multilingual tedx corpus for speech recognition and translation. In *Proceedings of Interspeech*.

Yun Tang, Hongyu Gong, Xian Li, Changhan Wang, Juan Pino, Holger Schwenk, and Naman Goyal. 2021. FST: the FAIR speech translation system for the IWSLT21 multilingual shared task. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 131–137, Bangkok, Thailand (online). Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

A Appendix

A.1 Hyperparameters

Hyper-parameter	Value
Batch size	4 000
Data-parallel GPUs	4
Update freq	2
Max learning rate	0.0005
Initial LR	10^{-7}
Schedule	inverse square root
Warmup steps	10 000
Adam betas	0.9, 0.999
Mixed precision	True
Label smoothing	0.2
Weight decay	0.0
Dropout	$0.3^{\dagger}$
Attention dropout	0.1
Gradient clipping	none
1D Convolutions	1
Conv channels	$80^{\star}$
Conv kernel size	5
Conv stride	2
Embed scaling factor	$\sqrt{1\,024}$
Positional encoding	sinusoidal$^{\alpha}$
Encoder layers	24
Decoder layers	24
Embed dim	$1\,024^{\ddagger}$
FFN dim	8 192
Activation	ReLU
Attention heads	16
Pre-norm	True
Adapter dim	64
Vocab size	250k
Lang-pair temperature	3
Heterogeneous batches	True
Valid freq	5 000
Checkpoint averaging	3
Patience	5
Early stopping metric	BLEU
Beam size	5

Table 9: Hyper-parameters used to train our models.
$\star$: a linear layer followed by a ReLU activation is trained to project the input features (of dimension 768 or 1 024) to the input dimension of the CNN (80).
†: dropout is also applied to the source and target embeddings (after the convolutions and positional encoding) and FFN activations.
‡: 2 048 when the pre-trained MT model is NLLB 3.3B.
α: learned positional embeddings in the decoder when the pre-trained model is mBART.

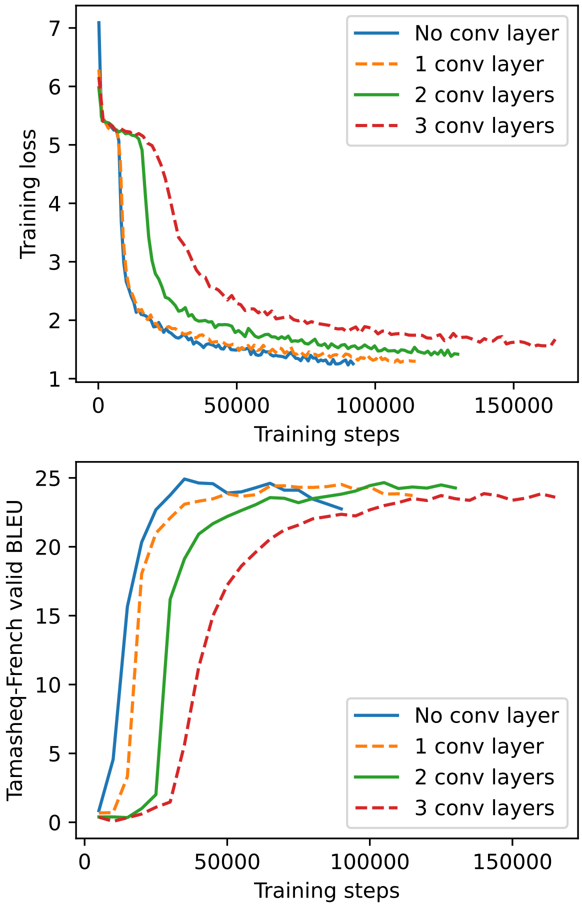

Figure 2: Training loss and *Taq-Fr* validation BLEU of variants of our *contrastive 1* model, that have 0 to 3 convolutional layers (1 by default).

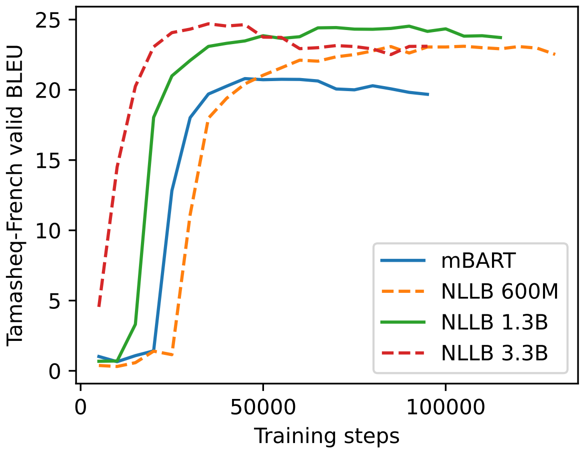

Figure 3: *Taq-Fr* validation BLEU of variants of our *contrastive 1* model that are initialized with various MT models (NLLB 1.3B by default).

A.2 Additional Results

Task	Source	Target	hours:minutes	# utterances
ASR	French	French	218:59	117,081
ASR	Italian	Italian	118:39	50,895
ASR	Portuguese	Portuguese	179:33	91,257
ASR	Spanish	Spanish	214:15	103,076
ST	French	English	57:39	31,207
ST	French	Spanish	42:14	21,862
ST	French	Portuguese	26:53	14,322
ST	Portuguese	English	63:13	31,868
ST	Spanish	French	9:34	4,568
ST	Spanish	English	79:37	37,168
ST	Spanish	Italian	11:50	5,616
ST	Spanish	Portuguese	47:01	22,012

Table 10: Statistics for all the mTEDx languages (train+valid) seen by our systems for the IWSLT 2021 evaluation setup described in Section 4.3.

		Taq-Fr valid	Que-Es valid	Que-Es test
Taq-Fr	primary	**26.13**	✗	✗
	contrastive 1	24.53	✗	✗
	contrastive 2	22.88	**20.29**	**17.74**
Que-Es	primary	22.88	**20.29**	**17.74**
	contrastive 1	20.81	19.03	15.67
	contrastive 2	21.31	16.78	15.25
Que-Es (updated)	primary	22.36	16.52	**15.70**
	contrastive 1	20.97	15.15	15.55
	contrastive 2	20.31	16.30	13.17

Table 11: Validation and test results on the IWSLT 2023 low-resource track. Lines 3 and 4 correspond to the same model. The "*Que-Es* (updated)" results correspond to new models trained on filtered Quechua ASR data, where we removed audio files that are also in the ST valid and test sets. In this updated version, **primary** and **contrastive 1** use NLLB 1.3B and **contrastive 2** uses NLLB 3.3B.

		Taq-Fr test	Que-Es valid
Taq-Fr	contrastive 1	19.13 ± 0.06	✗
	contrastive 2	16.89 ± 0.18	18.34 ± 0.59
Que-Es	contrastive 1	16.89 ± 0.18	18.34 ± 0.59
Que-Es (updated)	contrastive 1	16.51 ± 1.12	14.98 ± 0.16
	contrastive 2	16.56 ± 0.30	15.66 ± 0.60

Table 12: Statistics (BLEU average and standard deviation) for the submitted models which have 3 runs with different seeds. The *Taq-Fr* and *Que-Es* BLEU scores are respectively over the IWSLT 2022 test set and the IWSLT 2023 validation set.

Adapter Size	Encoder Adapters	Decoder Adapters	Taq-En Lang ID	Taq-Ko Lang ID	Hangul Percentage
64	✓	✓	100%	97%	88%
128	✓	✓	99%	84%	59%
64	✓	✗	100%	100%	95%
✗	✗	✗	100%	100%	96%
✗	✗	✗	100%	100%	93%

Table 13: Percentage of output sentences in the correct language according to the NLLB language ID (Costa-jussà et al., 2022). The last column shows the percentage of output characters that are in the Korean alphabet.

Train vocab	Inference vocab	Inference params	Taq-Fr BLEU	Fr-En BLEU	Speed
Full (256k)	Full (256k)	1.38B	19.1	**36.6**	12.5×
	Filtered (35k)	1.19B	18.9	35.8	**13.0**×
Filtered (35k)	Filtered (35k)	1.19B	**20.0**	35.5	**13.0**×

Table 14: Speech Translation performance on the IWSLT 2022 *Taq-Fr* and mTEDx *Fr-En* test sets of our contrastive *Taq-Fr* submission (non-ensemble version of our primary submission) with several vocabulary filtering strategies: no filtering (first row, corresponds to our submission); inference-time filtering (second row); or training-time filtering (third row). See Table 18 for an explanation of the "speed" column.

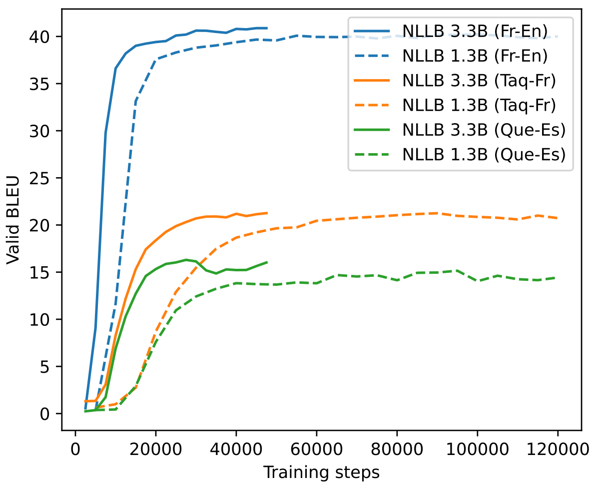

Figure 4: Validation BLEU by language direction (*Fr-En*, *Taq-Fr* and *Que-Es*) of a multilingual model (XLS-R + NLLB 1.3B) which includes both Tamasheq and Quechua (our *updated constrastive 1* submission).

Submission	IWSLT 2023 Taq-Fr	Que-Es	Que-Que	TED-LIUM v2 En-En	mTEDx ASR Fr-Fr	Es-Es	It-It	Pt-Pt	mTEDx ST Fr-En	Fr-Es	Es-Fr	Es-En	Fr-Pt	Pt-En	Es-It	Es-Pt
Taq-Fr primary	✓	✗	✗	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗
Taq-Fr contrastive 1	✓	✗	✗	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗
Taq-Fr contrastive 2	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗
Que-Es primary	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗
Que-Es contrastive 1	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗
Que-Es contrastive 2	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗
IWSLT 2021 setup	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 15: Extensive list of datasets used for training (✓) each system presented in this paper.

Model	URL
mHuBERT-Tamasheq	Unavailable
Tamasheq	`https://huggingface.co/LIA-AvignonUniversity/IWSLT2022-tamasheq-only`
Niger-Mali	`https://huggingface.co/LIA-AvignonUniversity/IWSLT2022-Niger-Mali`
XLSR-53	`https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec`
XLS-R large and xlarge	`https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec/xlsr`

Table 16: Downloading sources for the speech representation models checkpoints used in our experiments.

Stacked layers	FT layers	Adapters	Total params	Trained params	Taq-Fr BLEU	Fr-En BLEU	Speed
1	0	enc+dec (64)	1.40B	28M	**19.2**	35.0	12.0×
1	0	none	1.39B	22M	17.9	33.8	12.2×
0	1	enc+dec (64)	1.38B	28M	18.2	35.1	12.0×
0	1	none	1.37B	22M	17.5	33.3	12.6×
2	0	enc+dec (64)	1.42B	49M	**19.2**	35.1	11.9×
2	0	none	1.41B	43M	18.4	35.0	12.5×
0	2	enc+dec (64)	1.38B	49M	19.0	36.2	12.0×
0	3	enc+dec (64)	1.38B	70M	19.1	**36.6**	12.5×

Table 17: Training stacked layers (i.e. adding and training new bottom encoder layers) versus fine-tuning the existing bottom layers; with or without adapters. The other hyper-parameters are identical to our constrastive submission (underlined scores).

Speech features	MT model	Conv. layers	FT layers	Adapters	Total params	Trained params	Taq-Fr BLEU	Fr-En BLEU	Speed
Tamasheq (layer 11)	NLLB 1.3B	1	3	enc+dec (64)	1.38B	70M	16.8	32.5	11.6×
Tamasheq (layer 8)					1.38B	70M	19.3	31.6	12.0×
mHuBERT-Taq (layer 11)					1.38B	70M	16.4	37.1	12.1×
mHuBERT-Taq (layer 8)					1.38B	70M	16.2	36.7	12.1×
Niger-Mali (layer 11)					1.38B	70M	16.6	34.6	11.8×
Niger-Mali (layer 8)					1.38B	70M	19.1	36.6	12.5×
XLSR-53 (layer 18)					1.38B	70M	15.9	38.0	12.4×
XLS-R L (layer 18)					1.38B	70M	16.8	**39.4**	12.7×
XLS-R XL (layer 46)					1.38B	70M	15.4	37.4	11.7×
Niger-Mali (layer 8)	mBART (600M)	1	3	enc+dec (64)	0.61B	41M	16.3	28.9	22.9×
	NLLB (600M)				0.62B	41M	18.0	32.5	24.2×
	NLLB (1.3B)				1.38B	70M	19.1	36.6	12.5×
	NLLB (3.3B)				3.36B	165M	19.3	37.3	4.5×
Niger-Mali (layer 8)	NLLB 1.3B	3	3	enc+dec (64)	1.38B	70M	18.5	33.4	**25.5**×
		2			1.38B	70M	19.4	35.4	19.5×
		1			1.38B	70M	19.1	36.6	12.5×
		0			1.38B	70M	**19.6**	34.4	7.1×
Niger-Mali (layer 8)	NLLB 1.3B	1	24	enc+dec (64)	1.37B	508M	16.7	30.7	11.9×
			4		1.38B	91M	19.6	36.8	12.3×
			3		1.38B	70M	19.1	36.6	12.5×
			2		1.38B	49M	19.0	36.2	12.0×
			1		1.38B	28M	18.2	35.1	12.0×
Niger-Mali (layer 8)	NLLB 1.3B	1	1	enc (64)	1.37B	25M	19.1	34.2	12.4×
				none	1.37B	**22M**	17.5	33.3	12.6×
			3	enc+dec (256)	1.40B	88M	18.8	35.8	12.2×
				enc+dec (128)	1.38B	76M	19.2	36.3	12.1×
				enc+dec (64)	1.38B	70M	19.1	36.6	12.5×
				enc (64)	1.37B	67M	19.3	35.7	12.7×
				none	1.37B	64M	18.3	35.6	13.1×

Table 18: Ablation study on *Taq-Fr* ST, with various speech feature extractors, pre-trained MT models used for initialization, and trained parameters. The total parameter counts do not include the parameters of the speech feature extractors. The BLEU scores reported are on the IWSLT 2022 *Taq-Fr* and mTEDx *Fr-En* test sets. The speed metric is relative to real time (i.e., seconds in the test set divided by seconds spent decoding) and does not include feature extraction time. It is obtained by decoding the *Taq-Fr* test set on a single T4 with a batch size of 10 utterances (averaged over 3 decoding runs). The underlined numbers all correspond to the same model, which is our first contrastive submission to the task (the non-ensemble version of our primary submission). All of these models are trained with the same data (see Table 15) and early stopping is done based on *Taq-Fr* valid BLEU scores. The numbers inside parentheses in the *Adapters* column correspond to the bottleneck dimension of the trained adapter modules. Adapters are not added in the encoder layers that are being fine-tuned. These models took between 15 and 47 h each to train on 4 V100 GPUs, with an average training time of 26 h.

Task	Model	Adapters	Training directions				Zero-shot directions		
			Es-En	Fr-En	Fr-Es	Pt-En	Pt-Es	It-En	It-Es
ST	NLLB 3.3B	enc+dec	44.0	**39.9**	38.3	**33.1**	38.1	29.3	36.9
ST	NLLB 1.3B	enc+dec	43.7	39.4	38.0	31.5	35.9	28.9	35.0
		none	36.7	35.0	31.7	23.8	30.5	25.2	31.3
		enc	41.4	38.3	36.0	30.8	36.2	26.2	35.1
		dec	39.1	38.2	33.1	26.9	31.9	27.9	32.9
MT	NLLB 3.3B	none	47.4	39.5	39.2	39.8	48.6	34.0	42.4
MT	NLLB 1.3B	none	47.9	38.9	39.6	39.8	48.5	33.8	41.9
		enc+dec	50.2	40.7	42.2	42.1	51.0	37.6	45.2
		enc	49.9	41.3	42.6	41.9	50.6	36.5	44.9
		dec	48.8	39.2	41.0	41.1	49.7	35.6	43.9
MT	NLLB 1.3B (DA)	enc+dec	**51.3**	**43.2**	**45.2**	**44.7**	**53.2**	**37.8**	**47.1**

Table 19: **Top half:** Speech translation BLEU scores on the IWSLT 2021 test sets, when deactivating encoder adapters, decoder adapters, or both in an ST model at inference time. The ST model is the same one as in Table 5, trained with encoder and decoder adapters. **Bottom half:** Text-to-text MT BLEU scores when using the ST adapters in the initial model and disabling the ST bottom layers and convolutions.