JHU IWSLT 2024 Dialectal and Low-resource System Description

Nathaniel R. Robinson[1] **Kaiser Sun**[1] **Cihan Xiao**[1] **Niyati Bafna**[1] **Weiting Tan**[1]
Haoran Xu[1] **Henry Li Xinyuan**[1] **Ankur Kejriwal**[1] **Sanjeev Khudanpur**[1,2]
Kenton Murray[1,2] **Paul McNamee**[2]
[1]Johns Hopkins University Center for Language and Speech Processing
[2]Human Language Technology Center of Excellence
Baltimore, USA
{nrobin38,hsun74,cxiao7,nbafna1,wtan12,hxu64,xli257,khudanpur,kenton,
mcnamee}@jhu.edu; akejriw2@alumni.jh.edu

Abstract

Johns Hopkins University (JHU) submitted systems for all eight language pairs in the 2024 Low-Resource Language Track. The main effort of this work revolves around fine-tuning large and publicly available models in three proposed systems: i) end-to-end speech translation (ST) fine-tuning of SeamlessM4T v2; ii) ST fine-tuning of Whisper; iii) a cascaded system involving automatic speech recognition with fine-tuned Whisper and machine translation with NLLB. On top of systems above, we conduct a comparative analysis of different training paradigms, such as intra-distillation of NLLB, joint training and curriculum learning of SeamlessM4T v2, and multi-task learning and pseudo-translation with Whisper. Our results show that the best-performing approach differs by language pairs, but that i) fine-tuned SeamlessM4T v2 tends to perform best for source languages on which it was pre-trained, ii) multi-task training helps Whisper fine-tuning, iii) cascaded systems with Whisper and NLLB tend to outperform Whisper alone, and iv) intra-distillation helps NLLB fine-tuning.

1 Introduction

With recent developments in data-driven machine learning and Transformer-based models (Vaswani et al., 2017), speech translation (ST) systems (which accept spoken input in one language and automatically output corresponding text in another) have undergone major strides in performance (Radford et al., 2023; Barrault et al., 2023; Sperber and Paulik, 2020). While these works demonstrate the effectiveness of using large pretrained models for speech translation between high-resource language pairs and establish new state-of-the-art (SOTA) performance in these setups, less attention has been devoted to whether these advances also benefit low-resource language pairs, and how they compare with SOTA systems for these languages.

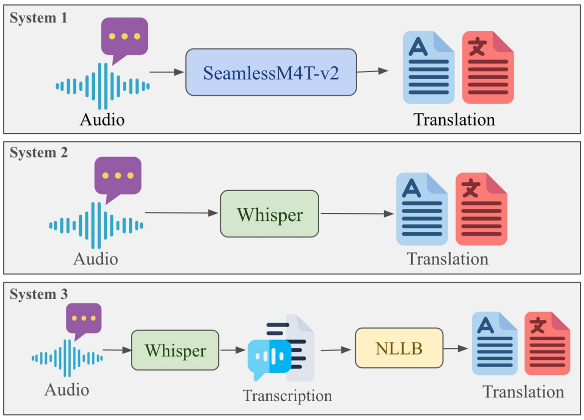


Figure 1: Proposed frameworks for fine-tuning.

Some of the populations with the greatest need for ST tools are those speaking low-resource languages, which typically have less institutional support and funding for the development for NLP and speech tools (He et al., 2024; Kesiraju et al., 2023b; Karakasidis et al., 2023): some speak minority languages in the areas where they live and need translation tools to communicate across a language barrier, or to consume or search for information more effectively online (Neto et al., 2020). Certain populations speaking low-resource languages may also have low literacy rates or limited writing traditions in their native languages, increasing the imperative for speech-based, rather than text-based, translation systems (Besacier et al., 2006).

In this work, we developed ST systems for eight language pairs, as organized in the IWSLT 2024 Dialectal and Low-resource Speech Translation Shared Task. We approached this problem by leveraging systems pre-trained on a large amount of multilingual data and subsequently fine-tuning them for specific tasks: both end-to-end speech ST and cascaded ST (i.e. transcription followed by text-based translation). We compared different approaches and pre-trained models for each language pair, and we experimented with combining data from multi-

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 188–201
August 15-16, 2024

ple related languages into the same train set.

Among the systems introduced, the approaches based on SeamlessM4T v2 (Barrault et al., 2023) outperform others for language pairs that it has seen during pretraining and for which supervised ST data are available (e.g. `mar-hin`, `gle-eng`, `bho-hin`, and `mlt-eng`). In other cases, a cascaded system is the most successful of the proposed approaches, namely, for `apc-eng`, `bem-eng`, `que-spa`, and `tmh-fra`.

2 Prior Work

A number of prior studies introduce methods aiming to address low-resource ST. In IWSLT's evaluation for low-resource and dialectal ST 2023, Agarwal et al. (2023) note three practices that consistently help performance: (1) use of pre-trained models, (2) systems combining both end-to-end and cascaded models, and (3) synthetic data augmentation. These recommendations inform our decisions to fine-tune pre-trained models and experiment with both cascaded and end-to-end approaches.

Williams et al. (2023) used cascaded ST systems for Quechua-to-Spanish ST in IWSLT challenge 2023. Shanbhogue et al. (2023) fine-tuned pre-trained speech models, and E. Ortega et al. (2023); Laurent et al. (2023) leveraged both pre-trained speech and text models in cascaded systems. Deng et al. (2023); Hussein et al. (2023) explored both end-to-end and cascaded ST. The most comparable submission to ours from the 2023 challenge was that of Mbuya and Anastasopoulos (2023), who used pre-trained models and applied them to several language pairs. With the findings and recommendations from prior work, we adapt a similar approach, but fine-tuning SeamlessM4T v2 (Barrault et al., 2023), Whisper (Radford et al., 2023), and NLLB (NLLB Team et al., 2022) instead of self-supervised learning representations (SSLR). Our approach differs from works described above, primarily in that we fine-tune models trained for automatic speech recognition (ASR), machine translation (MT), and ST, rather than fine-tuning representations obtained from language modeling objectives, such as wav2vec2 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), XLS-R (Babu et al., 2022), or mBART (Liu et al., 2020a), for the tasks of ASR, MT, and ST. The findings from our systems shed light on the potential benefits provided by the pretrained multilingual models.

3 Task Description

On the challenge website this year,[1] the organizers stated, "The goal of this shared task is to benchmark and promote speech translation technology for a diverse range of dialects and low-resource languages." To forward this aim, this year's task focuses on ST for eight language pairs: Levantine Arabic to English (`apc-eng`), Bemba to English (`bem-eng`), Bhojpuri to Hindi (`bho-hin`), Irish to English (`gle-eng`), Maltese to English (`mlt-eng`), Marathi to Hindi (`mar-eng`), Quechua to Spanish (`que-spa`), and Tamasheq to French (`tmh-fra`). Levantine is one of the most spoken Arabic dialects, with the majority native-speaking populations in Syria, Lebanon, Palestine, and Jordan. Both Levantine Arabic and Maltese are Semitic languages of the Afroasiatic family. Bemba is a Bantu language of the Niger-Congo family, spoken by over 30% of Zambia's population (Sikasote and Anastasopoulos, 2022). Bhojpuri, Hindi, and Marathi are Indo-Aryan languages; Hindi and Marathi are Scheduled languages in India and have government backing for their support, whereas Bhojpuri, like many other languages on the so-called Hindi Belt, lacks official status, has a much smaller writing tradition, and is only recently gaining attention in NLP (Kumar et al., 2022; Mundotiya et al., 2021; Bafna et al., 2023). Each of the source languages is low-resource, with Tamasheq, Bemba, and Levantine Arabic having the fewest Wikipedia articles overall (Robinson et al., 2023). Despite their low digital support, these languages have a large native speaker base, including Marathi's 83 million, according to Ethnologue.[2]

The organizers provide different varieties of data for each of these language pairs. We used predominantly provided datasets, along with some external data, all of which are outlined in Table 1. We differentiate datasets of four types: **ASR**, indicating source language speech with corresponding transcriptions; **E2E**, indicating source language speech with corresponding target language translations that could supervise end-to-end ST; **MT**, indicating source language text with corresponding target language translations; and **ST**, indicating source language speech with both corresponding transcriptions and target language translations.

Though this year's task accepts both *unconstrained* submissions, allowing the use of external

[1] https://iwslt.org/2024/low-resource
[2] https://www.ethnologue.com/

datasets and pre-trained models, and *constrained* submissions, our submission is limited to the *unconstrained* track, since all of our methods involved fine-tuning pre-trained models.

4 Proposed Methods

We introduce three primary frameworks, which are applied to different language pairs according to the availability of the data: (1) we fine-tune SEAMLESSM4T v2 for end-to-end ST using **E2E** data; (2) we fine-tune Whisper (Radford et al., 2023) for end-to-end ST using **E2E** (and optionally **ASR**) data; (3) to form a cascaded ST system, we fine-tune Whisper for ASR using **ASR** data, then fine-tune NLLB for machine translation (MT) using **MT** data. The fine-tuning approaches are illustrated in Figure 1. Note that each **ST** dataset contains exactly one **E2E**, **ASR**, and **MT** dataset implicitly.

We explore various methodological additions to these methods. We look at joint fine-tuning and curriculum learning with the SEAMLESSM4T v2-based approaches. We investigate several fine-tuning setups for the Whisper-based systems, including pseudo-translation fine-tuning, multitask training with ASR and MT as well as ASR-only and ST-only fine-tuning. We also looked at intra-distillation as a method of enhancing NLLB in MT. These ideas are further detailed below.

4.1 SEAMLESSM4T v2-based systems

Barrault et al. (2023) introduce SEAMLESSM4T v2, a model capable of end-to-end expressive and multilingual translations in a streaming fashion. SEAMLESSM4T v2 supports multilingual input and output in both speech and text modalities, with a dedicated sub-model handling each modality combination. It has 2.3B parameters and is pretrained on 1M hours of unlabeled audio in 143 languages, using the `w2v-BERT XL` architecture (Chung et al., 2021). It is then fine-tuned on text MT into English (`x-eng`) for 95 languages, ASR for 96 languages, ST into English for 89 languages, and speech-to-speech translation into English for 95 languages, and out of English `eng-x` for 35 languages. The pretraining languages of SEAMLESSM4T v2 include English, Irish, Maltese, Hindi, Marathi, and Arabic,[3] but not Quechua, Tamasheq, or Bemba.

[3]We assume that the pretraining corpus also contains some Levantine and Tunisian Arabic, but these languages are not labeled distinctly from each other.

Our Systems We fine-tune SEAMLESSM4T v2 on **E2E** ST data, aiming to leverage the vast pretraining and ASR and ST capabilities of SEAMLESSM4T v2, which we expect to be beneficial in data-scarce scenarios. Although the SEAMLESSM4T v2 models are evaluated mostly on X-Eng/Eng-X directions in Barrault et al., 2023, we hypothesize that they will succeed in X-X directions post-finetuning, due to ASR pretraining in source and target languages. Note that this approach is only applicable to language pairs where **E2E** data are available (`gle-eng`, `mlt-eng`, `aeb-eng`, `bem-eng`, `que-spa`, `tmh-fre`, `mar-hin`, `bho-hin`). We also evaluate the zero-shot performance of SEAMLESSM4T v2 on these language pairs.

Experimental Setup For each language pair, we fine-tune SEAMLESSM4T `v2-large` for four epochs, with a learning rate of 1×10^{-6} and batch size of 32. For `que-spa` translation, we use learning rate 1×10^{-8} for 15 epochs due to its small dataset size. For `bem-eng` and `tmh-fra`, a learning rate of 1×10^{-7} is used for training. The full hyperparameter list and details of hyperparameter tuning are included in Appendix A.1.

4.1.1 Multilingual training

Mixed Data Training For pairs with the same target language (`gle-eng`+`mlt-eng`, `bho-hin`+`mar-hin`), we fine-tune SEAMLESSM4T v2 on the combined dataset created by concatenating and shuffling the data, using the same hyperparameter settings as in Section A.1.

Curriculum Training Tunisian Arabic (`aeb`) and Maltese are both Semitic languages and share close linguistic relationships. We use a 12.6-hour subset of the Tunisian Arabic-to-English (`aeb-eng`) **ST** data used by Hussein et al. (2023) to conduct a curriculum training attempt using Tunisian as an augmentation for Maltese. The model undergoes initial fine-tuning on `aeb-eng` ST for two epochs with a learning rate of 1×10^{-6}, followed by a 5-epoch-fine-tuning on `mlt-eng` at a learning rate of 1×10^{-7}.

4.2 Whisper-based systems

Whisper (Radford et al., 2023) is an end-to-end multi-task speech model based on a transformer-like encoder-decoder architecture. For this study, we focus primarily on its LARGE-V2 variant, which is pre-trained on 680k hours of multilingual ASR

Lang.	Type	Amount	Size	Genre(s)	Sources
apc-eng	ASR	28h	3.2GB	Spontaneous speech	Makhoul et al. (2005)
	MT	120k lines	84MB	Subtitles	Sellat et al. (2023)
bem-eng	ST	180h	21GB	Dialogue description	Sikasote et al. (2023)
	ASR	24h	3.0GB	Read speech	Sikasote and Anastasopoulos (2022)
bho-hin	E2E	25h	2.6GB	News audio	Agarwal et al. (2023)
gle-eng	E2E	11h	2.2GB	Read speech	Agarwal et al. (2023)
mlt-eng	ST	14h	1.6GB	Telephone speech	CV; Hernandez Mena et al. (2020)
	MT	2.1M lines	710MB	Web-crawled	Bañón et al. (2023, 2020)
mar-hin	E2E	30h	3.5GB	News audio	Agarwal et al. (2023)
	ASR	1100h	150GB	Read speech; News	CV; He et al. (2020); Bhogale et al. (2022)
que-spa	ST	1.7h	300MB	Radio	Ortega et al. (2020)
	ASR	48h	5.2GB	Radio	Cardenas et al. (2018)
	MT	26k lines	3.7MB	Mixed; Magazine	Tiedemann (2012); Ortega et al. (2020)
tmh-fra	E2E	19h	2.2GB	Radio	Zanon Boito et al. (2022)

Table 1: Data information. "CV" refers to Common Voice (`https://commonvoice.mozilla.org/`).

and X-to-Eng speech translation data. During pre-training, the model is exposed to over 90 languages, including English, Marathi, Hindi, Maltese, and modern standard Arabic. However, Bemba, Bhojpuri, Quechua, Levantine Arabic, and Tamasheq, are absent from the pre-training data.

To address the gaps in language coverage and enhance model performance across diverse linguistic settings, we fine-tune the model in various ways tailored to specific scenarios. As the original model's pre-training setup, we manipulate the prompt and supervision of the utterances at fine-tuning time to guide the model to perform different tasks, as detailed in the subsequent sections. In addition, for languages previously unseen by the model, we expand its vocabulary and embedding layer to create new language tags for the model to take condition on.

4.2.1 Fine-tuning paradigms

ASR-only Fine-tuning For language pairs with only **ASR** data or a limited amount of **E2E** or **ST** data, such as `apc-eng` and `que-spa`, Whisper is trained with only the ASR objective to serve as an ASR module in a cascaded system. The training and decoding prompt used is the conventional `<|src-lang|><|transcribe|>`. The resulting cascaded system's MT module is an NLLB model described in § 4.3.

E2E-only Fine-tuning We train with Whisper's ST-only objective for the `tmh-fra` pair. However, because Whisper is pre-trained for X-Eng ST only, instead of directly translating into French, we fine-tune the system to translate Tamasheq speech into English text. Specifically, we translate the French labels of the **E2E** data into English using NLLB out of the box to formulate a `tmh-eng` **E2E** dataset. We then fine-tune Whisper with this dataset and utilize the trained model as the ASR module for a cascaded system, whose MT module is also NLLB. Similarly, English-to-French translation is conducted out-of-the-box.

Pseudo-translation For `bho-hin` and `mar-hin` language pairs, due to the absence of 3-way parallel **ST** data, the phylogenetic proximity between the languages, and the non-English-centric translation directions, we explore a novel adaptation of the model which we call *pseudo-translation*. Specifically, to enable Whisper to translate into non-English languages, we prompt the model to "transcribe" the source language speech signals with the target language transcription prompt, i.e. `<|tgt-lang|><|transcribe|>`. Conceptually, this is equivalent to treating Bhojpuri and Marathi as pseudo-Hindi speech and conducting ASR (an approach that is especially linguistically motivated in the case of Bhojpuri, as it is closely related to Hindi). Such design is motivated by the fact that Whisper is pre-trained with weakly supervised data, which implicitly empowers the model's audio-conditioned language model to perform some extent of de-noising. Consequently, we may model the non-English translation process as a noisy transcription task with the proposed prompts.

Multi-task Learning Previous yet unpublished experiments suggest that multi-task learning (MTL) tends to improve the model's performance across downstream metrics. Hence, for bem-eng and mlt-eng, as the 3-way parallel **ST** data is sufficient, we fine-tune Whisper on both the ASR and E2E ST tasks with E2E X-Eng ST being the end goal. In particular, we create the ASR and E2E ST dataset objectives respectively with their corresponding prompts, i.e. <|src-lang|><|transcribe|> and <|src-lang|><|translate|>, and concatenate them to form a multi-task dataset for fine-tuning, allowing the sampler to draw samples with different supervisions stochastically. Kesiraju et al.'s (2023a) use a large amount of Marathi ASR data (He et al., 2020; Bhogale et al., 2022) for Marathi-to-Hindi ST. Therefore, we further extend the idea of constructing data to mar-hin, which has abundant non-parallel ASR and E2E ST data yet no 3-way parallel data. We combine the pseudo-translation technique to perform non-parallel ASR and E2E pseudo-ST multi-task training.[4]

4.2.2 Whisper training details

We employ a range of techniques to expedite the training of Whisper and optimize the utilization of our hardware resources. Specifically, we adopt Low-Rank Adapters (LoRA) (Hu et al., 2021), gradient checkpointing (Chen et al., 2016), and Zero Redundancy Optimizer (ZeRO) (Rajbhandari et al., 2020) to fine-tune all Whisper models. We allow trainable decomposed weight matrices with a rank of 200 for the embedding layer, all the attention layers, and the first feed-forward layer in the transformer blocks, resulting in a total of 289,157,200 trainable parameters, approximately 16% of the original model's parameter count.

We apply conventional speech data augmentation in the fine-tuning process, including SpecAug (Park et al., 2019) and speed perturbation (Ko et al., 2015) with parameters 0.9, 1.0, 1.1.

4.3 NLLB fine-tuning

NLLB Team et al.'s (2022) NLLB is an encoder-decoder framework designed for extensive multilingual translation across more than 200 languages. It incorporates the sparsely gated mixture of experts (Du et al., 2022) to balance enhanced modeling capacity with efficient training and inference. Training of the NLLB model involves three objectives—translation loss, denoising loss, and language modeling loss—all calculated using the negative log-likelihood (NLL) loss function but with distinct datasets. Translation loss utilizes clean parallel texts, while denoising loss employs techniques from denoising auto-encoders (Liu et al., 2020b) that introduce noise into the source text. The language modeling objective of NLLB uses monolingual data to train the decoder.

Vanilla Fine-tuning We fine-tune the open-source NLLB model[5] with the released MT corpora for apc-eng, bem-eng, and que-spa. Specifically, we use the distilled 600M-parameter NLLB model as the base model and fine-tune the model with NLL loss. Following NLLB Team et al. (2022), we append language tokens on both source and target sequences during training and force decode the target language token during inference. We use a learning rate of 1×10^{-4} and set the maximum number of target tokens per batch to 1600. We train all translation models on a single V100 machine and accumulate gradient updates every 4 steps.

Fine-tuning with Intra-distillation We also fine-tune with intra-distillation (ID), which is an effective task-agnostic training method, aiming to encourage all parameters to contribute equally (Xu et al., 2022, 2023). Given an input batch, ID needs to forward pass the model K times to obtain K outputs and each time a random subset of parameters is zeroed out. The core idea of ID is to minimize the difference of these K outputs to approximately minimizing the contribution gap of the parameters that are zeroed-out, because the K outputs are forced to be the same with different zeroed parameters. Let $\{p_1, \cdots, p_i, \cdots, p_K\}$ denote the K outputs. The ID loss is then formulated by the X-divergence (Xu et al., 2022) to minimize the difference of K outputs as

$$\mathcal{L}_{id} = \frac{1}{K}\sum_{i=1}^{K} \mathbb{KL}(p_i \parallel \bar{p}) + \mathbb{KL}(\bar{p} \parallel p_i)$$

$$\text{where } \bar{p} = \frac{1}{K}\sum_{i=1}^{K} p_i$$

Let the original task loss be $\mathcal{L}_i$ for the i^{th} pass. Then, the total loss is a combination of the original task loss and ID loss, given as

$$\min \frac{1}{K}\sum_{i=1}^{K} \mathcal{L}_i + \alpha\mathcal{L}_{id}$$

[4] Note that in this case, since the ST data is used for pseudo-translation, only <|translate|> tags are used.

[5] Available at: https://huggingface.co/docs/transformers/en/model_doc/nllb

where α is a hyper-parameter to control the strength of ID.

5 Results and Discussion

Table 2 displays the results for all of our MT systems. We calculate scores using the same BLEU (Papineni et al., 2002) configuration as the task organizers.[6] We include scores from internal **Dev** and **Test** sets when available, as well as the official **Eval** scores. Details of data splitting are in Appendix A.2. The results show that SEAMLESSM4T v2 systems perform best for half of the language pairs: `bho-hin`, `gle-eng`, `mar-hin`, and `mlt-eng`. Cascaded systems employing Whisper and NLLB for MT performed best for the others: `apc-eng`, `bem-eng`, `que-spa`, and `tmh-fra`. (Note these first three language pairs employed Whisper for ASR and a fine-tuned NLLB model for MT, while `tmh-fra` employed Whisper for X-Eng ST and NLLB out of the box for MT into French.)

5.1 End-to-end ST

The SEAMLESSM4T v2 models' poor performance on `bem-eng`, `que-spa`, and `tmh-fra` is likely due to the absence of Bemba, Quechua, or Tamasheq in its pre-training corpus. We include zero-shot results for SEAMLESSM4T v2 out of the box in Table 3, which illustrate that the pre-trained model already performs well on `mlt-eng` and `gle-eng`,[7] but poorly on unseen language pairs.

We remark that our fine-tuning process brings notable improvements for `bho-hin`, `mar-hin`, and `mlt-eng`. In particular, SEAMLESSM4T v2 is successful for `bho-hin` despite not being pre-trained explicitly on Bhojpuri data, possibly because the Hindi pretraining data contains some Bhojpuri, or because SEAMLESSM4T v2 is capable of extrapolating fairly well to Bhojpuri given its high linguistic similarity to Hindi. Interestingly, the mixed data training (comb.) for language pairs sharing a target language does not significantly improve performance for either source language, though we expected it to benefit the lower-resource pair. In the case of `gle,mlt-eng`, there are domain differences (read speech vs. telephonic speech) between the fine-tuning corpora, possibly resulting in unhelpful or negative interference; Irish and Maltese are also not linguistically related, limiting cross-lingual transfer. On the other hand, with `bho,mar-hin`, Marathi and Bhojpuri both belong to the Indic subfamily of languages, and the speech translation data for both respective language pairs is from the news domain, averaging about 7 seconds each. The lack of success of joint fine-tuning for both these setups resonates with the findings of Sun et al. (2023), which presents several experiments showing that multilingual training for speech translation may not always benefit low-resource languages. We also note that curriculum training likewise did not improve performance for `mlt-eng`.

In our evaluation of Whisper systems, we emphasize two significant observations. Firstly, as anticipated, the BLEU scores for the `mar-hin` and `bho-hin` language pairs validate the efficacy of the proposed pseudo-translation method. This finding not only demonstrates that the model is capable of handling non-English translations with minimal fine-tuning, but also underscores its adaptability to linguistically similar language pairs. Secondly, the consistent performance gain observed with Whisper MTL over Whisper E2E as illustrated by the `mar-hin` results underscores the advantages of multi-task learning. This method treats fine-tuning on multiple tasks as involving one primary task and several auxiliary tasks, which collectively contribute to enhanced outcomes on all tasks involved.

5.2 Cascaded ST

Cascaded ST via fine-tuned Whisper for ASR and fine-tuned NLLB for MT is our best-performing approach for `apc-eng`, `bem-eng`, and `que-spa`, though it is much better for `apc-eng` and `bem-eng` than for `que-spa`. The relatively low performance of `que-spa` can be possibly attributed to it being a non-English-centric translation direction.

Table 4 presents the ASR performance of the fine-tuned Whisper models on 5 language pairs with different objectives. Those trained with the ASR-only objective are used solely as the ASR module in cascaded systems, while the systems trained with the multi-task learning objective are used for both direct translation and ASR for cascaded systems. Interestingly, we observe that for Bemba, the CERs (25.1 for **dev** and 17.9 for the **test1** set) are significantly lower than the WERs. We find through manual inspection that the model

[6] With sacrebleu signature `nrefs:1 | case:lc | eff:no | tok:13a | smooth:exp | version:2.0.0`.

[7] There is a considerable discrepancy between the `gle-eng` dev and test scores from IWSLT 2023, with the latter being suspiciously high. Mbuya and Anastasopoulos (2023) suggest that the inflated test scores may be due to overlap between train and test sets.

Lang.	System	Submisson	Dev	Test	Eval	Lang.	System	Submisson	Dev	Test	Eval
`apc-eng`	Whisper+NLLB+ID	primary	-	**32.0**	**16.0**	`tmh-fra`	Whisper+NLLB	primary	**8.0**	**7.0**	**6.1**
	Whisper+NLLB	contrastive1	-	30.2	14.7		Seamless	contrastive1	0.3	1.3	0.5
`bem-eng`	Whisper+NLLB+ID	primary	**26.3**	**30.4**	**32.6**	`mar-hin`	Seamless	primary	**32.1**	**40.9**	**37.7**
	Whisper+NLLB	contrastive1	22.6	29.0	27.0		Seamless comb.	contrastive1	31.0	39.4	37.3
	Whisper MTL	contrastive2	23.5	27.8	26.7		Whisper MTL	contrastive2	26.3	34.9	28.5
	Seamless	-	6.6	15.4	-		Whisper E2E	-	24.4	32.8	-
`bho-hin`	Seamless	primary	**34.9**	-	**24.4**	`que-spa`	Whisper+NLLB+ID	primary	**15.7**	**11.7**	**12.5**
	Seamless comb.	contrastive1	34.5	-	23.9		Whisper+NLLB	contrastive1	6.9	6.1	6.4
	Whisper E2E	contrastive2	28.6	-	12.2		Seamless	contrastive2	1.8	0.9	0.9
`mlt-eng`	Seamless	primary	**52.9**	**54.2**	-						
	Seamless curr.	contrastive1	47.3	47.1	-	`gle-eng`	Seamless	primary	25.2	**52.7**	15.3
	Whisper MTL	contrastive2	34.5	35.1	-		Seamless comb.	contrastive1	**27.6**	51.6	**16.0**
	Seamless comb.	-	51.6	53.1	-						

Table 2: BLEU scores for each system. **Dev** and **Test** denote our internal tuning and test sets, when available. **Eval** denotes the official evaluation. `apc-eng` **Test** scores are from text-only MT, since our data had no source speech-to-translation alignments for ST evaluation. "ID" indicates use of intra-distillation with NLLB fine-tuning. "Comb." refers to mixed data training, and "curr." refers to curriculum training.

Lang.	Dev$_{zero}$	Dev$_{ft}$
`bem-eng`	0.9	**6.6**
`gle-eng`	**27.7**	25.2
`mar-hin`	0.0	**32.1**
`mlt-eng`	47.8	**52.9**
`que-spa`	**1.9**	1.8
`tmh-fra`	0.4	**8.0**

Table 3: Zero-shot and fine-tuned performance of SeamlessM4T v2 on dev set. Model generally improves after fine-tuning, except for `que-spa` and `gle-eng`.

Lang.	Objective	Dev	Test
`apc-eng`	ASR-only	11.5	10.4
`que-eng`	ASR-only	34.4	34.5
`bem-eng`	MTL	57.3	47.3
`mar-hin`	MTL	37.2	37.3
`mlt-eng`	MTL	23.8	-

Table 4: WER of the Whisper model fine-tuned on each language. *ASR-only* suggests that the model is trained to perform ASR-only to serve as an ASR module for a cascaded system, whereas *MTL* suggests that the model is trained to perform E2E ST and ASR.

tends to make minor spelling errors, presumably due to its unfamiliarity with the language's writing system, as suggested by the decent proficiency in its translation performance. This may cause error propagation in cascaded ST.

In our MT module, we implemented intra-distillation to enhance ST results by balancing the contributions of the model parameters. Consistent with prior studies Xu et al. (2022, 2023), intra-distillation consistently improves performance across all evaluated translation directions, with the most significant enhancement observed for `que-spa`. MT performance was reasonably high for the three language pairs for which we employed cascaded ST. The cascaded approach for `mlt-eng` performs poorly, likely because our Maltese bitexts were noisy. Additionaly, NLLB has already been pre-trained on Maltese and may not benefit further from the noisy post-training.

6 Conclusion and Future Work

In this work, we describe our submitted systems for all eight language pairs in the IWSLT 2024 Low-Resource Language Track. We explore various fine-tuning approaches for large publicly available pre-trained models, compare end-to-end and cascaded systems, as well as investigate the benefits of joint and curriculum training, multitask learning, as well as intra-distillation. We find that the best-performing strategy is language-pair dependent, with fine-tuned SeamlessM4T v2 generally performing best on languages that are included in its pretraining corpus. Fine-tuned Whisper generally performed better with multi-task fine-tuning than standard fine-tuning, and better still when employed in a cascaded system with fine-tuned NLLB (with best results employing intra-distillation).

For future improvements, augmenting MT fine-tuning data with ASR hypotheses, as in Gow-Smith et al. (2023), could equip NLLB better for cascaded ST. Future work could also employ data augmentation of text and speech data, as in Shanbhogue et al. (2023), via textual back-translation (Sennrich et al., 2016), speech synthesis for augmentation (Rossenbach et al., 2020; Robinson et al., 2022), or other methods. Lastly, future research could employ the use of SSLR, or employ the large amounts of raw audio available—particularly for Tamasheq—to train SSLR systems, following Gow-Smith et al. (2023).

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, 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, Xutai 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 Kr. Ojha, 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)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale.

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.

Niyati Bafna, Cristina España-Bonet, Josef Van Genabith, Benoît Sagot, and Rachel Bawden. 2023. Cross-lingual strategies for low-resource language modeling: A study on five Indic dialects. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, pages 28–42, Paris, France. ATALA.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Marta Bañón, Malina Chichirau, Miquel Esplà-Gomis, Mikel L. Forcada, Aarón Galiano-Jiménez, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, and Jaume Zaragoza-Bernabeu. 2023. Maltese-english parallel corpus MaCoCu-mt-en 2.0. Slovenian language resource repository CLARIN.SI.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Laurent Besacier, Bowen Zhou, and Yuqing Gao. 2006. Towards speech translation of non written languages. In *2006 IEEE Spoken Language Technology Workshop*, pages 222–225. IEEE.

Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages. *arXiv preprint*.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *Preprint*, arXiv:1604.06174.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Pan Deng, Shihao Chen, Weitai Zhang, Jie Zhang, and Lirong Dai. 2023. The USTC's dialect speech translation system for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 102–112, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2022. Glam: Efficient scaling of language models with mixture-of-experts. *Preprint*, arXiv:2112.06905.

John E. Ortega, Rodolfo Zevallos, and William Chen. 2023. QUESPA submission for the IWSLT 2023 dialect and low-resource speech translation tasks. In

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 261–268, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Edward Gow-Smith, Alexandre Berard, Marcely Zanon Boito, and Ioan Calapodescu. 2023. NAVER LABS Europe's multilingual speech translation systems for the IWSLT 2023 low-resource track. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 144–158, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. 2020. Open-source multi-speaker speech corpora for building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu speech synthesis systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6494–6503, Marseille, France. European Language Resources Association.

Taiqi He, Kwanghee Choi, Lindia Tjuatja, Nathaniel R. Robinson, Jiatong Shi, Shinji Watanabe, Graham Neubig, David R. Mortensen, and Lori Levin. 2024. Wav2gloss: Generating interlinear glossed text from speech. *Preprint*, arXiv:2403.13169.

Carlos Daniel Hernandez Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, and Ian Padovani. 2020. MASRI-HEADSET: A Maltese corpus for speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6381–6388, Marseille, France. European Language Resources Association.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Amir Hussein, Cihan Xiao, Neha Verma, Thomas Thebaud, Matthew Wiesner, and Sanjeev Khudanpur. 2023. JHU IWSLT 2023 dialect speech translation system description. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 283–290, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Georgios Karakasidis, Nathaniel Robinson, Yaroslav Getman, Atieno Ogayo, Ragheb Al-Ghezi, Ananya Ayasi, Shinji Watanabe, David R. Mortensen, and Mikko Kurimo. 2023. Multilingual tts accent impressions for accented asr. In *Text, Speech, and Dialogue*, pages 317–327, Cham. Springer Nature Switzerland.

Santosh Kesiraju, Karel Beneš, Maksim Tikhonov, and Jan Černocký. 2023a. BUT systems for IWSLT 2023 Marathi - Hindi low resource speech translation task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 227–234, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Santosh Kesiraju, Marek Sarvaš, Tomáš Pavlíček, Cécile Macaire, and Alejandro Ciuba. 2023b. Strategies for improving low resource speech to text translation relying on pre-trained asr models. In *INTERSPEECH 2023*. ISCA.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proc. Interspeech 2015*, pages 3586–3589.

Ritesh Kumar, Siddharth Singh, Shyam Ratan, Mohit Raj, Sonal Sinha, Bornini Lahiri, Vivek Seshadri, Kalika Bali, and Atul Kr Ojha. 2022. Annotated speech corpus for low resource indian languages: Awadhi, bhojpuri, braj and magahi. *arXiv preprint arXiv:2206.12931*.

Antoine Laurent, Souhir Gahbiche, Ha Nguyen, Haroun Elleuch, Fethi Bougares, Antoine Thiol, Hugo Riguidel, Salima Mdhaffar, Gaëlle Laperrière, Lucas Maison, Sameer Khurana, and Yannick Estève. 2023. ON-TRAC consortium systems for the IWSLT 2023 dialectal and low-resource speech translation tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 219–226, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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

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

John Makhoul, Bushra Zawaydeh, Frederick Choi, and David Stallard. 2005. Bbn/aub darpa babylon levantine arabic speech and transcripts. *Linguistic Data Consortium (LDC), LDC Catalog No.: LDC2005S08*.

Jonathan Mbuya and Antonios Anastasopoulos. 2023. GMU systems for the IWSLT 2023 dialect and low-resource speech translation tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 269–276, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Rajesh Kumar Mundotiya, Manish Kumar Singh, Rahul Kapur, Swasti Mishra, and Anil Kumar Singh. 2021. Linguistic resources for bhojpuri, magahi, and maithili: statistics about them, their similarity estimates, and baselines for three applications. *Transactions on Asian and Low-Resource Language Information Processing*, 20(6):1–37.

Antonio Carvalho Neto, Fernanda Versiani, Kelly Pellizari, Carolina Mota-Santos, and Gustavo Abreu. 2020. Latin american, african and asian immigrants working in brazilian organizations: facing the language barrier. *Revista Economia & Gestão*, 20(55).

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Nathaniel Robinson, Perez Ogayo, Swetha Gangu, David R Mortensen, and Shinji Watanabe. 2022. When is tts augmentation through a pivot language useful? In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2022, pages 3538–3542.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2020. Generating synthetic audio data for attention-based speech recognition systems. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7069–7073.

Hashem Sellat, Shadi Saleh, Mateusz Krubiński, Adam Pospíšil, Petr Zemánek, and Pavel Pecina. 2023. UFAL parallel corpus of north levantine 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Akshaya Vishnu Kudlu Shanbhogue, Ran Xue, Soumya Saha, Daniel Zhang, and Ashwinkumar Ganesan. 2023. Improving low resource speech translation with data augmentation and ensemble strategies. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 241–250, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Claytone Sikasote and Antonios Anastasopoulos. 2022. BembaSpeech: A speech recognition corpus for the Bemba language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

Claytone Sikasote, Eunice Mukonde, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. BIG-C: a multimodal multi-purpose dataset for Bemba. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2062–2078, Toronto, Canada. Association for Computational Linguistics.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

Haoran Sun, Xiaohu Zhao, Yikun Lei, Shaolin Zhu, and Deyi Xiong. 2023. Towards a deep understanding of multilingual end-to-end speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14332–14348, Singapore. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

Aiden Williams, Kurt Abela, Rishu Kumar, Martin Bär, Hannah Billinghurst, Kurt Micallef, Ahnaf Mozib Samin, Andrea DeMarco, Lonneke van der Plas, and Claudia Borg. 2023. UM-DFKI Maltese speech translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 433–441, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Haoran Xu, Philipp Koehn, and Kenton Murray. 2022. The importance of being parameters: An intra-distillation method for serious gains. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 170–183, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haoran Xu, Jean Maillard, and Vedanuj Goswami. 2023. Language-aware multilingual machine translation with self-supervised learning. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 526–539, Dubrovnik, Croatia. Association for Computational Linguistics.

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier, and Yannick Estève. 2022. 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.

A Additional Experimental Details

A.1 SeamlessM4T v2 hyperparameters

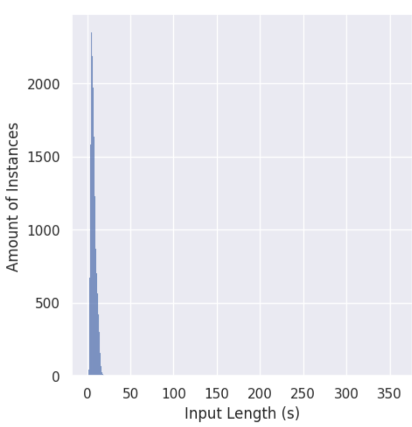
For SeamlessM4T v2 models, the longest audio length is truncated at 30 seconds. To ensure full reproducibility of the result, a random seed of 42 is deployed. We perform a minimum hyperparameter search for each language pair between the learning rate of $\{10^{-5}, 10^{-6}, 10^{-7}\}$. For each language pair, we fine-tune a SeamlessM4T v2-large for four epochs, with a learning rate of 1×10^{-6} and batch size of 32. For Quecha-to-Spanish (`que-spa`) translation, a learning rate of 1×10^{-8} is used for training 15 epochs due to its small dataset size. For all the training trials, a constant learning rate scheduler and a warm-up step of 50 is used. During inference, the maximum generation length is constrained to 256 tokens with greedy decoding.

A.2 Split details

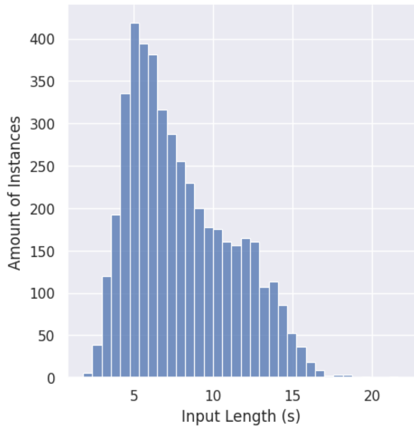
We split data into train, dev, and test when possible, for tuning and internal evaluation. We split Makhoul et al.'s (2005) Levantine Arabic **ASR** data, Sikasote et al.'s (2023) Bemba **ST** data, He et al.'s (2020) Marathi **ASR** data, Cardenas et al.'s (2018) Quechua **ASR** data, and Tiedemann's (2012) `que-spa` **MT** bitext ourselves using a 90-5-5 split. We split Sellat et al.'s (2023) `apc-eng` **MT** bitext ourselves with a 90-5-5 split but then performed our internal test on a 1000-line subset of the held out data. For the large `mlt-eng` **MT** bitexts from Bañón et al. (2023, 2020), we split the data ourselves with a 99-0.5-0.5 and a 98-1-1 split, respectively. We also split Bhogale et al.'s (2022) large Marathi **ASR** dataset ourselves with a 99-0.5-0.5 split. We used the creator's own splits for Sikasote and Anastasopoulos's (2022) Bemba **ASR** data, Agarwal et al.'s (2023) `mar-hin` **E2E** data, Tiedemann's (2012) `que-spa` **MT** bitext, Zanon Boito et al.'s (2022) `tmh-fra` **E2E** data, and the Hindi **ASR** data from Common Voice. We did the same with Agarwal et al.'s (2023) `gle-eng` **E2E** data, using the test set from the 2023 challenge as our internal test set. For the `mlt-eng` **ST** data from Common Voice and Hernandez Mena et al. (2020) and the `que-spa` **ST** data from Ortega et al. (2020), we used their own train and dev splits and then split the dev set in half to create an internal test set. We used Agarwal et al.'s (2023) own train and dev splits without creating an internal test set.

B Instance Length Distribution

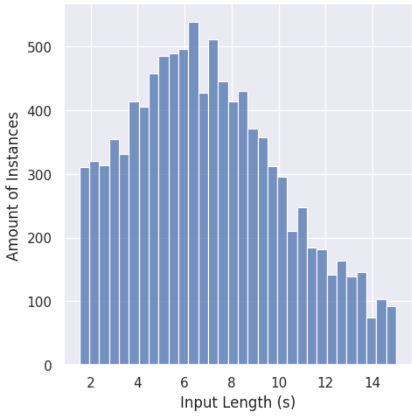
We show the length distribution in Figure 2 and Figure 3. Overall, most datasets show a normal distribution with a slightly skewed tail except for `que-spa`, the amount of instances for which is the smallest. However, we identify some extraordinarily long instances in `bem-eng` training set. These outlier instances can lead to out-of-memory instances if left untreated. Therefore, we truncate the instances that are over 30 seconds when training SEAMLESSM4T v2 and limit the generation length to 256 new tokens.



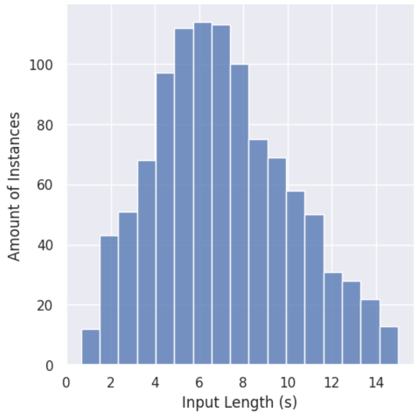
(a) bem-eng TRAINING SET



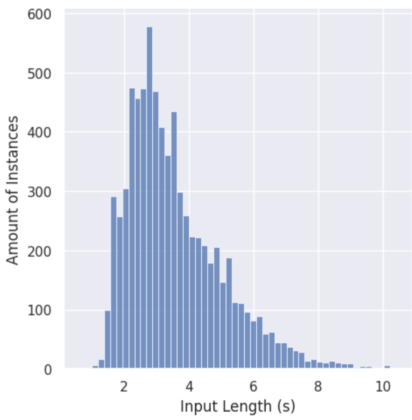
(b) bem-eng DEVELOPMENT SET



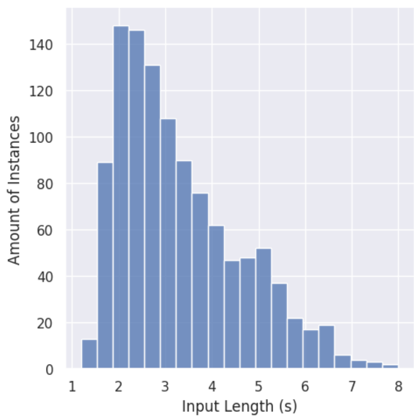
(c) bho-hin TRAINING SET



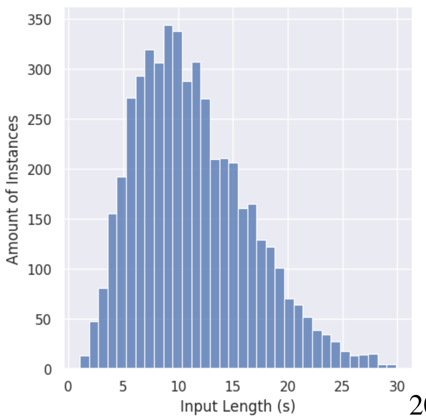
(d) bho-hin DEVELOPMENT SET



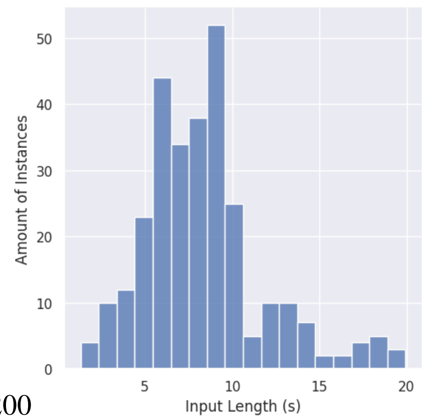
(e) gle-eng TRAINING SET



(f) gle-eng DEVELOPMENT SET

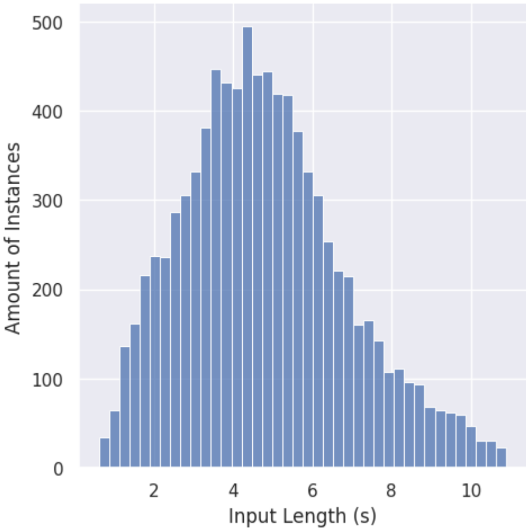


(g) tmh-fra TRAINING SET

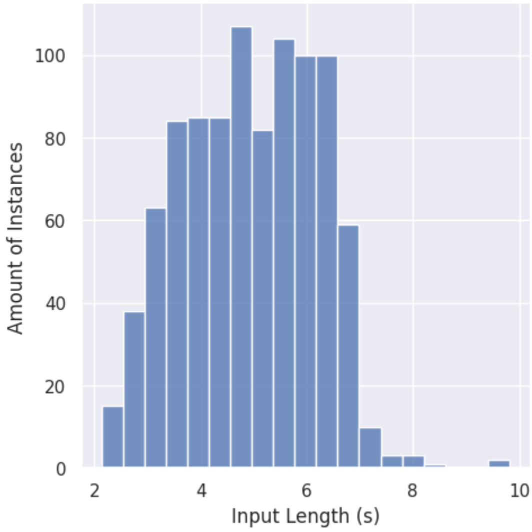


(h) tmh-fra DEVELOPMENT SET

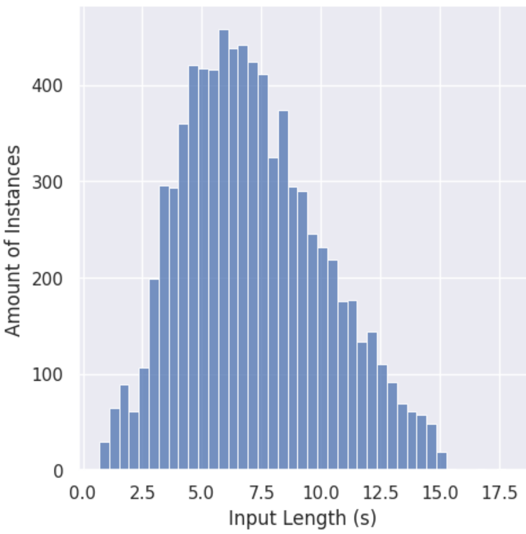
Figure 2: Length distribution (seconds) for each language pair.



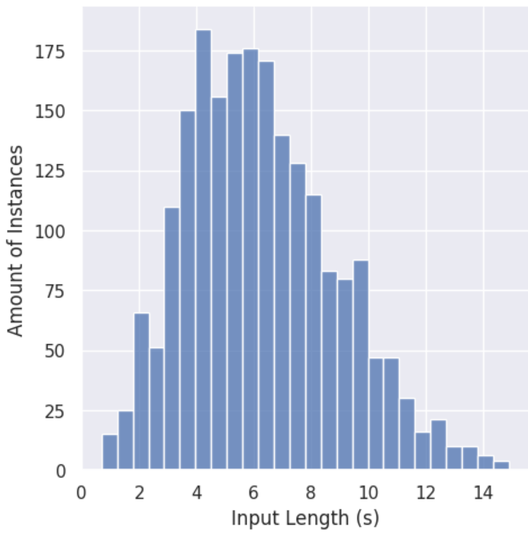
(a) mlt-eng TRAINING SET



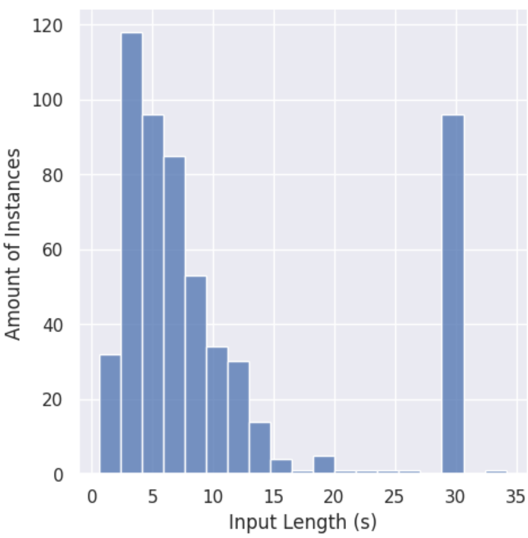
(b) mlt-eng DEVELOPMENT SET



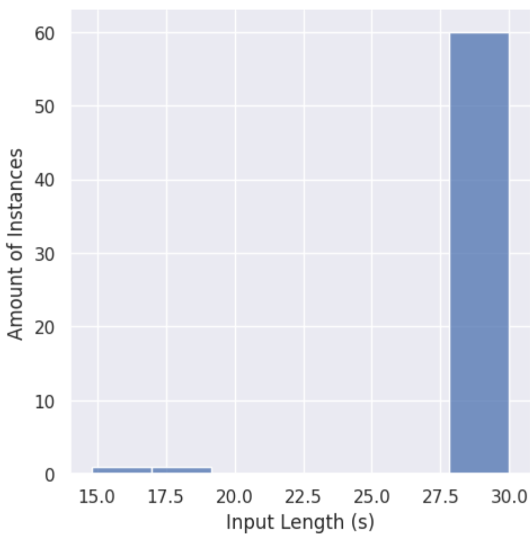
(c) mar-hin TRAINING SET



(d) mar-hin DEVELOPMENT SET



(e) que-spa TRAINING SET



(f) que-spa DEVELOPMENT SET

Figure 3: Length distribution (seconds) for each language pair (continued).