The KIT Speech Translation Systems for IWSLT 2024 Dialectal and Low-resource Track

Zhaolin Li, Enes Yavuz Ugan, Danni Liu, Carlos Mullov,
Tu Anh Dinh, Sai Koneru, Alexander Waibel, Jan Niehues
Karlsruhe Institute of Technology
firstname.lastname@kit.edu

Abstract

This paper presents KIT's submissions to the IWSLT 2024 dialectal and low-resource track. In this work, we build systems for translating into English from speech in Maltese, Bemba, and two Arabic dialects Tunisian and North Levantine. Under the unconstrained condition, we leverage the pre-trained multilingual models by fine-tuning them for the target language pairs to address data scarcity problems in this track. We build cascaded and end-to-end speech translation systems for different language pairs and show the cascaded system brings slightly better overall performance. Besides, we find utilizing additional data resources boosts speech recognition performance but slightly harms machine translation performance in cascaded systems. Lastly, we show that Minimum Bayes Risk is effective in improving speech translation performance by combining the cascaded and end-to-end systems, bringing a consistent improvement of around 1 BLUE point.

1 Introduction

In this paper, we describe KIT's systems submitted to IWSLT 2024 Dialectal and Low-resource Track. We focus on three language pairs: Bemba (ISO code: bem) to English, Maltese (ISO code: mlt) to English, and Dialectal Arabic to English. The Dialectal Arabic language pair evaluates the performance of two Arabic vernaculars, namely Tunisian (ISO code: aeb) and North Levantine (ISO-3 code: apc). Maltese and Tunisian language pairs are available in IWSLT2023 (Agarwal et al., 2023), and the others are newly included this year. The submissions are under Unconstrained Conditions to leverage pre-trained models and additional data resources.

Recent advancements in dialectal and low-resource speech translation show the benefits of utilizing pre-trained models (Gow-Smith et al., 2022; Laurent et al., 2023; Hussein et al., 2023; Deng et al., 2023). Nowadays, the capacities of pre-trained models are expanded by incorporating more extensive data and expanding language coverage. This work leverages the state-of-the-art pre-trained models, including SeamelssM4T (Barrault et al., 2023), MMS (Pratap et al., 2023), and NLLB (NLLB Team et al., 2022).

Cascaded and End-to-End (E2E) are popular Speech Translation (ST) systems. The Cascaded system consists of Automatic Speech Recognition (ASR) and Machine Translation (MT) models, while the E2E systems integrate both functions into one model. Recent work shows the E2E system shows comparable performance to the cascaded system in speech translation(Liu et al., 2023; Zhou et al., 2023; Huang et al., 2023; Hrinchuk et al., 2023), while there needs research to show which system performs better on dialectal and low-resource scenarios (Deng et al., 2023; Laurent et al., 2023; Kesiraju et al., 2023; Shanbhogue et al., 2023; E. Ortega et al., 2023; Hussein et al., 2023).

Building ST systems for low-resource datasets always suffers from data limitations. Accordingly, we collect available training resources and investigate the training strategies for using them. Although datasets other than the development data might introduce domain differences that could potentially model performance, we explore the benefits of using extra-supervised data. Furthermore, we investigate adapter fine-tuning training to address data scarcity. By freezing the pre-trained parameters and only fine-tuning the adapter parameters, this approach decreases the number of trainable parameters.

In addition to building ST systems, this work explores the decoding approach Minimum Bayes Risk(MBR) to re-rank the candidate translation (Kumar and Byrne, 2004; Hussein et al., 2023) from the built systems. We explore the combination of individual systems and across systems, and our findings suggest combining translations from the

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 269–276

cascaded and the E2E systems is effective for all language pairs.

2 Data Description

2.1 Development and test data

The organizers provide the development data for each language pair, which is from the same dataset of the test data for evaluating systems. The development data was released at the beginning, and test data was released when the evaluation period started for the final comparison of submissions. As shown in Table 1, the development data of North Levantine has only a validation split, indicating the importance of transferring knowledge from other data resources, such as standard Arabic. We report the system performance on Tunisian development data, although we have no submission for it due to the unexpected unavailability of the test data at the end of the evaluation period. The Maltese language pair includes two datasets, and we report scores only on the Masri dataset because we use the train split of CV development for training. We evaluated the Bemba systems on the test split of development data, but we later found the test split was the same as the test data.

	Development			Test
Lang.	Train	Valid	Test	
apc	-	1126	-	974
aeb	202k	3833	4204	-
mlt_masri	4962	648	-	668
mlt_cv	3923	1235	-	1224
bem	82k	2782	2779	2779
bem_asr1	-	-	-	977
bem_asr2	-	-	-	3756

Table 1: Statistic on development and test data. Lang is the language code of the source language. The value indicates the number of sample. One sample of the datasets consists of the audio, transcript, and translation in English.

2.2 Additional data resource

Under the unconstrained condition, we collected additional datasets of the language pairs and explored leveraging these resources to improve model performance. The ASR data resources are publicly available except for the SyKIT and MINI dataset, which is the in-house dataset in the conversational domain. SyKIT is a dataset that consists of people from Syria conversing in dialogues on various topics via a Zoom setup. The MINI dataset is read speech and is based on an electronic version of the M.I.N.I. (International Neuropsychiatric Interview). The MT data resources are all from OPUS collection (Tiedemann, 2009).

Lang.	Corpus	Type	#Hour/#Sent.
apc	LDC2005S08	ASR	60h
	LDC2006S29	ASR	250h
	SyKIT	ASR	50h
	Tatoeba	MT	20
aeb	SRL46	ASR	12h
	GNOME	MT	646
ara	SLR148	ASR	111h
	MGB	ASR	1200h
	MINI	ASR	10h
	CCMatrix	MT	5M
	NLLB	MT	5M
	OpenSubtitles	MT	3M
bem	BembaSpech	ASR	24h
	NLLB	MT	427k
mlt	MASRI-Headset v2	ASR	7h
	MASRI-Farfield	ASR	10h
	MASRI-Booths	ASR	2h
	MASRI-MEP	ASR	1h
	MASRI-COMVO	ASR	7h
	MASRI-TUBE	ASR	13h
	NLLB	MT	14M
	DGT	MT	3.5M
	TildeMODEL	MT	2M

Table 2: Overview of the additional data resources.

2.3 Pre-processing

Due to computational limitations, the ASR and ST training data over 15 seconds is removed. Although the training scenario is low-resourced, statistics show only a very small portion of training samples are removed. Afterwards, we introduce data augmentation with Gaussian noise, time stretch, time mask, and frequency mask [1].

3 Method

We conduct preliminary evaluations on Tunisian dialects to assess systems performance and then apply the promising approaches to other languages for effective analysis. The motivation is that the Tunisian language pair has effective systems from IWSLT 2023 (Agarwal et al., 2023) for approach analysis.

3.1 Cascaded Systems

The cascaded system is composed of ASR and MT modules and allows each component to be optimized independently. We explore the ASR and MT modules individually to mitigate the requirement on the supervised ST data, aiming to leverage the supervised ASR and MT data individually.

[1] https://github.com/asteroid-team/torchaudiomentations

3.1.1 ASR

We build two ASR systems with MMS and SeamlessM4T to leverage pre-trained multilingual models. The MMS system is the encoder-only model with the CTC training loss, and the SeamelssM4T model is the encoder-decoder model with cross-entropy training loss. We build the MMS system because the MMS model is pre-trained with more than 1,400 languages, including Maltese and Bemba. The motivation for using SeamlessM4T is its capacity for multilingual generation as an encoder-decoder model.

Our initial findings indicated that the Seamlessm4t system exhibited superior performance on Tunisian and North Levantine data over the MMS system. Consequently, we directed our efforts toward enhancing this particular model.

Given the scarcity of supervised ASR data we explore training strategies of using only the development data or mixing all available training resources. Using all available data increases the amount of supervised data while bringing domain differences that might lead to performance degradation. Consequently, we explore the two-step fine-tuning serving as knowledge transfer. This entails initially fine-tuning the pre-trained model using all available ASR data, followed by training the fine-tuned model solely with the target data.

The amount of supervised data might be insufficient to fine-tune the parameters of the SeamlessM4T model fully. To address this, we explore the parameter-efficient fine-tuning approach Low-Rank Adapters (LORA) by adding and only fine-tuning the LORA adapter (Hu et al., 2021).

3.1.2 MT

The pre-trained SeamlessM4T is a multitask model that supports both audio and text inputs. Besides ASR, we also explore its capacity for MT. Note that Bemba and Maltest are covered in the pre-trained SeamelssM4T model while the Arabic dialects are not.

Apart from SeamlessM4T, we also fine-tune NLLB (NLLB Team et al., 2022) because the pre-trained model covers more language pairs, including all three language pairs of this paper. Given the large vocabulary size of 256K, we freeze the word embedding to save memory. We also follow the recommendations of Cooper Stickland et al. (2021) regarding fine-tuning pre-trained MT models on many-to-English directions and freezing the decoder apart from cross-attention.

Given the extremely limited MT data on the two Arabic dialects (apc and aeb; Table 2), we fine-tune SeamlessM4T or NLLB jointly on these languages along with modern standard Arabic (ara), resulting in a many-to-English system for {apc, aeb, ara}→eng.

3.2 End-to-End Systems

The E2E system mitigates the error propagation issue in the cascaded system. We develop the E2E model with pre-trained SeamlessM4T consisting of a speech encoder and a text decoder. Since we don't have extra supervised data for ST, we focus on using the development data for our E2E exploration. In addition, we also investigate the effectiveness of fine-tuning with adapters using LORA.

3.3 System Combination

In addition to building ST systems, we explore combining the developed systems using Minimum Bayes Risk (MBR) decoding. MBR decoding is a method used to rerank the candidate translation output. Given a pool of hypothesis translations, MBR uses a utility metric to score each hypothesis against a set of pseudo-references. The hypothesis with the highest average score is then selected as the final translation.

Since the main evaluation metric is the BLEU score, we choose the utility metric as BLEU. For the end-to-end system, we generate 50 hypotheses using epsilon sampling (Hewitt et al., 2022) with temperature 1.0 and epsilon threshold 0.02. For the cascaded system, we generate 50 hypotheses using sampling with a temperature of 0.75. We then combine the hypotheses from both systems, resulting in a hypothesis pool of 100 samples. We use this same hypothesis pool as the pseudo-references to score each individual hypothesis.

4 Experiments and Results

4.1 Model Configuration

ASR We use the pre-trained MMS model with 300M parameters to build the CTC-based ASR system [2]. Compared with other configurations, it has fewer parameters to train and, therefore, fits better to this track. As for the encoder-decoder-based ASR system, we use the pre-trained SeamlessM4T model of the latest version with the large configuration [3]. To reduce the memory footprint, we

[2]https://huggingface.co/facebook/mms-300m

[3]https://huggingface.co/facebook/seamless-m4t-v2-large

use the dedicated model of SeamlessM4T for the speech-to-text task.

MT For the MT systems with SeamlessM4T, we use the same pre-trained model as for ASR but a dedicated model architecture for the text-to-text task [3]. Our finetuned NLLB models are based on the 1B distilled model (NLLB Team et al., 2022). Although the 3B variant gave better initial performance when used out-of-the-box, we could not directly finetune it due to memory constraints. When finetuning, we partially freeze the model as described in §3.1.2.

E2E ST For the ST systems, the pre-trained SeamlessM4T model is the same as for ASR and MT. Here, we use the dedicated SeamlessM4T model for the speech-to-text task[3].

Adapter This work investigates fine-tuning the adapters of LORA with SeamlessM4T models to reduce trainable parameters. We add adapters to all transformer layers of the encoder and decoder. The details regarding our implementation can be found in Appendix A

4.2 Evaluation

As the final evaluation uses lowercase and no punctuation, we follow the setup [4] to process the prediction and reference in the evaluation of this work. Specifically, we process the ASR predictions and references of Tunisian and North Levantine with *arabic_filter* and the other predictions and references with *english_fiter* in evaluation.

For the ASR task, we evaluate with Character Error Rate (CER) and Word Error Rate (WER) using package jiwer[5]. We evaluate MT and ST tasks with BLEU and chrF++ with package sacreBLEU[6].

4.3 ASR

As Table 3 shows, we explore two ASR systems: the encoder-only system with pre-trained MMS (A1) and the encoder-decoder system with pre-trained SeamlessM4T (A2). A2 outperforms A1 for Maltese and Tunisian and is comparable to A1 for Bemba. Considering the pre-trained languages of MMS cover Maltese and Bemba while those of SeamlessM4T only cover Maltese, we regard A2 with SeamlessM4T as a stronger ASR system for this track and explore enhancing this system

With training data in addition to the development data, we investigate training with all supervised ASR data, including the development data. We find using all data boosts Maltese with 5.1 WER points, and gains Bemba with 3.5 WER points. For Tunisian, we gain 3.8 WER points on the validation split but loss 5.2 WER points on the test split. The overfitting to the validation split indicates the importance of improving model robustness. We notice a clear decrease in comparing the scores between A2 and A3 for North Levantine, and we assume the dialect and domain differences are the main causes.

Building on A3, we investigate knowledge transfer from all training datasets to the target dataset with the second step of fine-tuning. Here, we explore full training (A4), which is the same as previous experiments, and adapter training with LORA (A5) as described in subsection 4.1. We find knowledge transfer is effective for North Levantine and Tunisian while not for Maltese and Bemba. The potential reason is the dialects have clear differences from other training datasets, and a second step of fine-tuning enables the model to be specialized on the target dataset. While all training datasets of Maltese or Bemba are from the same languages, the second step of fully fine-tuning (A4) fails to keep the knowledge learned in the first step of fine-tuning and causes performance degradation because of less supervised training data. On the contrary, we observe the knowledge transfer with adapter fine-tuning (A5) works on memorizing the knowledge in the first step but leads to no improvement over A3.

As described in subsection 2.1, the North Levantine has only the valid split in development data, so we implement different training strategies with details in Appendix C. Besides, the training for Tunisian A3 has modifications to other languages, and details are available in Appendix B.

In Table 3, we report the CER and WER scores with normalization for North Levantine and Tunisian, same as (Hussein et al., 2023), for comparison with systems of previous years. The normalization is performed on both the predictions and references and implemented with the camel_tools package [7]. The ASR results without normalization are in Appendix D. There are no scores for others

[4] https://github.com/kevinduh/iwslt22-dialect/blob/main/1_prepare_stm.py

[5] https://github.com/jitsi/jiwer

[6] https://github.com/mjpost/sacrebleu

[7] https://github.com/CAMeL-Lab/camel_tools

	Model	apc_valid	aeb_valid	aeb_test	mlt_masri_valid	bem_test
A1	wav2vec-mms	-	26.2/59.3	29.1/63.6	19.2/61.5	10.0/37.3
A2	SeamlessM4T development data	**39.1/55.3**	21.5/46.5	24.5/45.7	7.2/21.8	10.0/36.6
A3	SeamlessM4T all data	48.0/72.8	21.0/42.7	26.7/ 50.9	**5.7/16.7**	**9.3/33.1**
A4	A3 + transfer	44.6/68.7	**16.8/33.7**	**23.0/43.8**	8.6/24.0	9.6/33.6
A5	A3 + transfer LORA	-	20.9/42.1	25.7/49.1	5.9/17.6	9.3/33.1
	2023 best ASR	-	-/36.5	-/41.7		
B1	NLLB all MT data	**24.9/53.6**	**30.4/52.6**	**26.8/50.2**	31.2/53.7	**28.4/52.1**
B2	SeamlessM4T all MT data	17.9/44.8	16.9/37.9	13.2/34.9	41.6/63.8	28.0/52.9
B3	SeamelssM4T development data	-	5.3/24.3	4.7/23.8	**52.6/72.6**	**28.4/52.8**
	2023 best MT	-	30.5/-	26.4/-	-	-
C1	Best ASR + B1	16.1/40.3	24.7/47.7	20.2/43.9	-	**2**7.5/51.6
C2	Best ASR + B3	-	-	-	47.1/69.1	27.0/52.0
D1	SeamlessM4T	-	22.3/44.9	19.3/42.7	47.2/69.2	27.7/51.3
D2	SeamelssM4T LORA	-	8.2/27.5	6.9/26.4	44.3/66.9	14.1/35.3
E1	Best Cascaded	-	24.4/47.1	20.6/43.6	47.3/69.3	27.6/51.6
E2	Best E2E	-	22.6/44.5	19.9/42.2	48.0/69.5	27.1/49.6
E3	Best Cascaded & E2E	-	**25.5/47.9**	**21.3/44.3**	**50.6/71.2**	**29.3/52.3**
	2023 Best ST	-	24.9/-	22.2/-	-	-

Table 3: Experimental results on development dataset. **A, B, C, D, and E** indicates the ASR, MT, cascaded ST, E2E ST, and MBR systems. The results for ASR are in the format of CER/WER, and those for MT and ST are in the format of BLEU/chrF++. The best ASR, MT and ST systems of 2023 IWSLT are both from (Hussein et al., 2023)

as they are new language pairs this year.

4.4 MT

As Table 3 shows, the system with pre-trained NLLB (B1) suppresses the system with SeamlessM4T (B2) models for North Levantine and Tunisian, and we assume the reason is that NLLB is pre-trained with datasets of North Levantine and Tunisian while SeamlessM4T not. In addition, we notice B1 gives inferior performance for Maltese and shows comparable performance for Bemba compared with B2, although both models cover these two languages in pre-training. We assume the difference in pre-training datasets leads to inconsistent findings for these language pairs because SeamlessM4t and NLLB have similar architectures and model sizes.

Rather than using all available training data, we explore training with only the development data to reduce the effects of domain differences (B3). We notice B3 brings a significant performance decline for Tunisian because its MT data is much less than that for B1 (see Table 2). On the contrary, we observe improvements for Maltese with 11.0 BLEU and 8.8 chrF points. We don't build an MT system (B3) for North Levantine as the supervised MT data is too little.

4.5 ST

We build the cascaded systems from the best ASR models, which are A2 for North Levantine, A4 for Tunisian, and A3 for Maltese and Bemba. The MT models for Arabic dialects are B1, and that for Maltese is B3. We investigate both B1 and B3 for Bemba as they show comparable performance as MT models, and we observe a slight improvement in using B3 on BLEU. We explore building a dedicated cascaded system with the normalized transcriptions for Tunisian, while it gives inferior results than the one without normalization.

Regarding E2E systems, we explore training SeamlessM4T with full fine-tuning and adapter fine-tuning. Full training shows clear advantages over adapter training for all languages in the low-resourced scenario, although more parameters need to be trained. Therefore, we assume only adapting the parameters of LORA is insufficient to fine-tune the SeamlessM4T models on the target language pairs.

4.6 Systems Combination

As can be seen from Table 3, when applying MBR decoding on the output of a single system (E1 and E2), the changes in BLEU and chrF scores are minor. However, when applying MBR decoding on the combined output of the best cascaded and the best end-to-end systems (Row E3), we observe consistent improvement of ≈ 1 BLEU point and ≈ 1 chrF point. This emphasizes the importance of output diversity when using ensembing methods like MBR decoding.

4.7 Submissions and Results

As for the final submission, we chose the MBR of combining the best cascaded and E2E systems as primary, and we chose cascaded as the contrastive1

system and E2E as the contrastive2 system. In addition, we submit the best ASR systems for evaluating the errors in acoustic recognition, which are described in subsection 4.5. The evaluation scores performed by the organizers are shown in Table 4. We notice the primary and contrastive 1 systems for North Levantine clearly outperform the contrastive 2 system, indicating the contributions of the multilingual MT model. We notice the ASR and ST systems achieve very high scores for Maltese, especially the CV partition. We guess one of the potential reasons is the pre-trained models touch the test data because the CommonVoice dataset is widely used in pre-training.

systems	apc	bem	mlt masri	mlt cv
ASR	-	33.2	19.3	2.4
ST primary	20.9	28.8	50.5	67.4
ST contrastive 1	19.7	27.0	46.3	64.2
ST contrastive 2	11.9	28.1	46.7	65.7

Table 4: Evaluation results on test data. The ASR system is evaluated with WER and the ST system is evaluated with BLEU

5 Conclusion

In this work, we develop the cascaded and E2E ST systems with pre-trained multilingual models. The cascaded system outperforms E2E systems for North Levantine and Tunisian and demonstrates comparable performance for Maltese and Bemba. While building the cascaded system, we find performance improvement by involving additional resources in ASR but observe performance degradation with that in MT. Furthermore, we demonstrate combining the cascaded and E2E system with MBR increases model performance for all language pairs. Comparing our system with previous systems for Tunisian, we note superior performance in the validation split but lagging results in the test split, suggesting the need for future investigations to enhance model robustness.

Acknowledgement This work is partly supported by the Helmholtz Programme-oriented Funding, with project number 46.24.01, named AI for Language Technologies, funding from the pilot program Core-Informatics of the Helmholtz Association (HGF). It also received partial support from the Federal Ministry of Education and Research (BMBF) of Germany under the number 01EF1803B (RELATER). The work was partly performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research.

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.

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*.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2021. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453, Online. Association for Computational Linguistics.

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.

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, Mark McConville, William Gillies, Jade Scott, and Roibeard Ó Maolalaigh. 2022. Use of transformer-based models for word-level transliteration of the book of the dean of lismore. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 94–98, Marseille, France. European Language Resources Association.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Oleksii Hrinchuk, Vladimir Bataev, Evelina Bakhturina, and Boris Ginsburg. 2023. NVIDIA NeMo offline speech translation systems for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 442–448, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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.

Wuwei Huang, Mengge Liu, Xiang Li, Yanzhi Tian, Fengyu Yang, Wen Zhang, Jian Luan, Bin Wang, Yuhang Guo, and Jinsong Su. 2023. The xiaomi AI lab's speech translation systems for IWSLT 2023 offline task, simultaneous task and speech-to-speech task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 411–419, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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.

Santosh Kesiraju, Karel Beneš, Maksim Tikhonov, and Jan Černocký. 2023. 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.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

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.

Danni Liu, Thai Binh Nguyen, Sai Koneru, Enes Yavuz Ugan, Ngoc-Quan Pham, Tuan Nam Nguyen, Tu Anh Dinh, Carlos Mullov, Alexander Waibel, and Jan Niehues. 2023. KIT's multilingual speech translation system for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

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 Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc 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. *CoRR*, abs/2207.04672.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages.

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.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

Xinyuan Zhou, Jianwei Cui, Zhongyi Ye, Yichi Wang, Luzhen Xu, Hanyi Zhang, Weitai Zhang, and Lirong Dai. 2023. Submission of USTC's system for the IWSLT 2023 - offline speech translation track. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 194–201, Toronto, Canada (in-person and online). Association for Computational Linguistics.

A Adapter fine-tuning

We implement the adapters of LORA with package PEFT (Mangrulkar et al., 2022). We set the hyperparameters to rank 8, alpha 32, dropout 0.1, and bias as 'lora_only'. To add adapters for all layers in the encoder and decoders of SeamlessM4T, the target modules are "q_proj, v_proj, linear_q, linear_v".

B Tunisian ASR training

In our first endeavor, we gathered all available Arabic data to fine-tune our model. The dataset used for training is detailed in Table 2. To augment the availability of dialectal data for training, we adopted two approaches: utilizing default validation splits or selecting 0.15% of the training data for validation. Subsequently, we combined the validation sets, retaining only 1,500 utterances for validation, and incorporating the remainder into our training data. We applied the same methodology to other Arabic datasets. Thus, our consolidated validation sets comprised a total of 3,000 utterances, with 50% representing dialectal speech. This model underwent training with early stopping set to five epochs, with results documented as A3 in Table 3.

Subsequently, we implemented various strategies further to enhance the model's performance on dialectal speech. In iteration A4, we conducted additional fine-tuning using solely dialectal data. We experimented with further fine-tuning the A3 model with exclusive Tunisian dialectal data and a LORA module in A5. However, given the lack of promising results and Tunisian's exclusion from the challenge, we discontinued further investigation into this approach.

C North Levantine training

For the A2 North Levantine ASR model, we continued fine-tuning the entire model from A3. We assume starting from the fine-tuned ASR models could alleviate the need for training data. As we only have the validation set, fine-tuning utilizes 90% of the validation set for training and reserves the remaining for validating and early stopping. Upon achieving convergence at a training epoch number, we use the same hyperparameters to conduct a new fine-tuning from A3, utilizing the whole validation set for training and stopping with the same epoch number. This approach brings a risk of overfitting to the validation set but could make full use of the available data for training.

For the E2E ST system, we implement the same training strategy as the ASR systems but start from the pre-trained SeamlessM4T model.

	apc_valid	aeb_valid	aeb_test
A1	-	27.4/62.9	31.1/68.4
A2	39.9/56.9	23.7/46.5	27.6/53.6
A3	49.6/75.7	23.1/47.4	29.6/58.9
A4	46.4/72.7	18.6/38.3	26.1/52.2
A5	-	23.1/47.0	28.7/57.0

Table 5: ASR results without normalization

D Tunisian and North Levantine ASR scores without normalization

For comparison with ASR systems from previous years, we report ASR scores with normalization in Table 3d. Here, we report the scores with normalization in Table 5.