The SETU-DCU Submissions to IWSLT 2024 Low-Resource Speech-to-Text Translation Tasks

Maria Zafar[†], **Antonio Castaldo**[δ], **Neha Gajakos, Prashanth Nayak, Rejwanul Haque**[†], **Andy Way**
[†]South East Technological University, Carlow, Ireland
[δ]Università di Pisa, Italy
ADAPT Centre, Dublin City University, Ireland
C00304029@setu.ie,antonio.castaldo@phd.unipi.it,neha.gajakos@adaptcentre.ie
prashanth.nayak@adaptcentre.ie,rejwanul.haque@setu.ie,andy.way@adaptcentre.ie

Abstract

Natural Language Processing (NLP) research and development has experienced rapid progression in the recent times due to advances in deep learning. The introduction of pre-trained large language models (LLMs) is at the core of this transformation, significantly enhancing the performance of machine translation (MT) and speech technologies. This development has also led to fundamental changes in modern translation and speech tools and their methodologies. However, there remain challenges when extending this progress to underrepresented dialects and low-resource languages, primarily due to the need for more data.

This paper details our submissions to the IWSLT speech translation (ST) tasks. We used the Whisper model for the automatic speech recognition (ASR) component. We then used mBART and NLLB as cascaded systems for utilising their MT capabilities. Our research primarily focused on exploring various dialects of low-resource languages and harnessing existing resources from linguistically related languages. We conducted our experiments for two morphologically diverse language pairs: Irish-to-English and Maltese-to-English. We used BLEU, chrF and COMET for evaluating our MT models.

1 Introduction

The introduction of the Transformer architecture (Vaswani et al., 2017) is considered to be a groundbreaking development in NLP. This innovation has led to the rise of LLMs, which have become the catalyst for the AI revolution we are presently witnessing. LLMs have consistently pushed the boundaries of research by improving upon the state-of-the-art across various NLP tasks. The variants of Transformer such as the Transducer (Chen et al., 2021), Conformer (Nguyen et al., 2021), and ESPnet (Watanabe et al., 2018) have become essential to the success observed in a range of speech tasks, including both text-to-speech and speech-to-text MT.

This study explores low-resource speech-to-text translation, focusing on Irish-to-English and Maltese-to-English language pairs. We focused on developing our ST systems following two standard approaches:

- End-to-End (E2E) system: An E2E system in ST performs translation from one language to another without any intermediate steps. This process uses a single model to manage the entire translation process.
- Cascaded System: A cascaded system in ST uses a two-step process. First, it converts speech into text using ASR, and then it translates that text into another language. This process uses separate models in each step.

The rest of the paper is organised as follows: Section 2 describes our related work. Our datasets are explained in Section 3. Section 4 describes the models we used. In Section 5, we discuss our experiments and results. We conclude with avenues for future work in Section 6.

2 Related Work

This section discusses some foremost papers related to our work. Hussein et al. (2023) recently utilise LLMs for MT, such as mBART (Liu et al., 2020) and NLLB-200 (Team et al., 2022), which they used within both E2E and cascaded ST frameworks. Furthermore, enhancements in ASR were achieved by employing pseudo-labeling for data augmentation and adjusting for channel variations in telephone speech data. Additionally, they employed Minimum Bayes-Risk decoding to optimise the integration of their E2E and cascaded ST systems. The proposed framework led to impressive results.

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 128–133
August 15-16, 2024 ©2024 Association for Computational Linguistics

Ortega et al. (2023) utilised the Fairseq S2T framework[1], where they used *log mel-scale filter bank* (Ortega et al., 2023) features for audio representation and Transformer for translation. Their systems integrated ASR and MT into the framework sequentially. The system's ASR component was powered by a pre-trained XLS-R model (Babu et al., 2021), enhanced with a fine-tuning step. At the same time, translations were performed using an MT system developed from a fine-tuned LLM. They found that in low-resource settings, like Quechua-to-Spanish, direct ST methods (combining ASR and MT) tended to outperform standalone LLM applications.

Mbuya and Anastasopoulos (2023) used self-supervised pre-trained speech models to improve translation performance in specific applications. Their study utilised self-supervised models such as Wav2vec 2.0 (Baevski et al., 2020), XLSR-53, and Hubert (Hsu et al., 2021). Their findings indicated that the Wav2vec 2.0 and Hubert models achieved similar performance levels in tasks involving low-resource languages and dialects. Moreover, they found that the Wav2vec 2.0 model performed better after removing its top three layers, a modification that the Hubert model did not require. In contrast, the XLSR-53 model showed weaker results in low-resource contexts but excelled in translating dialects, outperforming both Wav2vec 2.0 and Hubert in those scenarios.

Vakharia et al. (2023) investigated a novel approach termed "style embedding intervention" for low-resource formality control in spoken language translation. By assigning distinct style vectors to individual input tokens their proposed method comprehended and managed the subtleties of translating between formal and informal styles. They found that their approach surpasses previous "additive style intervention" techniques, particularly for the English-to-Korean translation task, enhancing average matched accuracy. After analysing their "style embedding intervention" model, they found that most of the style information was acquired in the <bos> (beginning of the sentence) token, further improving the average matched accuracy.

In their study, Radhakrishnan et al. (2023) employed a basic E2E framework based on Transformers and explored various techniques such as replacing encoder blocks with Conformer and pre-training the encoder. Their approach resulted in a substantial improvement in translation quality.

Williams et al. (2023) utilised a cascaded approach for their ST systems. For the ASR component, they used the XLS-R model. The MT component was based on mBART-50. They conducted experiments for English-to-Maltese language pairs, with the approach showing significant improvement over their baseline systems.

Experiments by Kesiraju et al. (2023) used E2E translation framework based on a bilingual ASR system. The model was jointly trained using Connectionist Temporal Classification and attention mechanisms. Furthermore, they employed techniques such as speed perturbation for data augmentation and re-scoring the top hypotheses using an external language model. They also introduced a cascaded system that utilised the same bilingual ASR and MT systems. Their experiments demonstrated significant improvements over the baseline for the Hindi-to-Marathi language pair.

The systems submitted to the previous year's IWLST offline and low-resource speech translation tracks employed various strategies for improving the performance of E2E or cascaded systems. As for ASR, several submissions adopted a mix of Transformer and conformer models (Zhang et al., 2022; Nguyen et al., 2021) or fine-tuned existing models (Zhang and Ao, 2022; Zanon Boito et al., 2022; Denisov et al., 2021). These efforts resulted in improved ASR performance through techniques such as training ASR on synthetic data with added punctuation, noise-filtering, and domain-specific fine-tuning (Zhang and Ao, 2022; Zhang et al., 2022), or integrating an intermediate model to refine the ASR output concerning casing and punctuation (Nguyen et al., 2021). As for MT, they predominantly relied on Transformer-based architectures (Zhang et al., 2022; Nguyen et al., 2021) or fine-tuning on preexisting LLMs (Zhang and Ao, 2022). Additionally, methods employed to improve MT performance included multi-task learning (Denisov et al., 2021), training the MT component robustly on noisy ASR output data (Nguyen et al., 2021), and re-ranking and de-noising techniques (Ding and Tao, 2021).

While there have been extensive and rapid developments in ST, the field of low-resource and dialect ST still needs to be explored. In this paper, we discuss our submissions to the IWSLT ST task. We conducted our experiments for two low-resource language pairs: Maltese-to-English and

[1]Fairseq: https://github.com/facebookresearch/fairseq/tree/main

Irish-to-English.

3 Datasets

We utilised the data provided by IWSLT for our experiments. The data statistics are detailed in Table 1.

Irish-to-English		
	Audios	Sentences
Train	7,478	7,478
Dev	1,120	1,120
Test	347	347
Maltese-to-English		
Train + Dev	7,542	7,542
Test	1,864	1,864

Table 1: Statistics of the datasets used.

4 Cascaded System

This section describes the architecture of our cascaded system. Like standard cascaded ST systems, our ASR and MT models are interconnected, i.e. the output from the ASR model serves as the input to the MT system. For this experiment, we selected the OpenAI Whisper model[2] (Radford et al., 2022) as our ASR system. We fine-tuned the OpenAI Whisper small model on Maltese speech to optimise its ASR capabilities. As for the MT component, we used two different pre-trained models, mBART-50[3] (Liu et al., 2020) and NLLB-200-distilled-600M[4] (Team et al., 2022). Both models were fine-tuned on the Maltese-to-English bilingual data.

As pointed out above, we used the OpenAI Whisper model as our ASR system. We aligned the data format with the model's input requirements to prepare our data. This involved removing unnecessary data chunks from the dataset, eliminating special characters, and converting the sentences to lowercase. Since our input audio was sampled at 48kHz, we downsampled it to 16kHz before passing it to the OpenAI Whisper feature extractor, as 16kHz is the sampling rate expected by the model. Additionally, we adjusted the audio inputs to the correct sampling rate using the "cast column" method. This operation does not alter the audio files directly but instead instructs the dataset to resample the audio samples on-the-fly the first time they are loaded.

We empirically identified that the following hyperparameter settings provided us the best results: batch size of 16, learning rate of 1e-5, 500 warmup steps, 30,000 max steps, per-device eval batch size of 8, generation max length of 225, and intervals of 1,000 steps for saving and evaluating, and 25 steps for logging.

4.1 The MT systems

As previously discussed, we choose mBART-50 and NLLB-200-distilled-600M as the choice of our MT models. We fine-tuned these models on the Maltese-to-English bilingual data (cf. Table 1). For the purpose of our evaluation, we used the BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), and ChrF (Popović, 2015) metrics.

5 End-to-end System

Our submission for the English-Irish language pair comprises a fine-tuned version of OpenAI Whisper Small[5] to perform direct ST. Note that in this experiment, the audio files are resampled at 16kHz. This experiments were carried out for Irish-to-English only. In terms of hyperparameters selection, for our E2E experimentation, we identified that the following settings provided us the best results: batch size of 16, learning rate of 1e-5, 500 warmup steps, 1 gradient accumulation steps, generation max length of 225, and intervals of 500 steps for saving and evaluating. The model was fine-tuned over three epochs. We also integrated early stopping with `Patience = 2`. The data preprocessing pipeline was the same as the one used for the cascaded system (see Section 4).

6 Results

This section discusses the results that we obtained from our experiments. Table 2 shows the results obtained by evaluating our models on the evaluation test set while Table 3 shows the results obtained on blind test set provided by IWSLT. We can see from Table 2 that our models are reasonably good in the Maltese-to-English translation task. Our primary submission for Maltese-to-English was based on cascaded setup (Whisper + NLLB fine-tuning). For this setup, we obtained 52.60 BLEU, 72.12 chrF and 0.831 COMET points on the IWSLT 2023 evaluation test set. Our contrastive system is also

[2]Whisper: https://openai.com/research/whisper

[3]mBART-50: https://huggingface.co/facebook/mbart-large-50

[4]NLLB-200: https://ai.meta.com/research/no-language-left-behind/

[5]Whisper: https://openai.com/research/whisper

Model	BLEU	ChrF	COMET
Maltese-English			
Whisper-small	56.67	81.92	0.84
NLLB-200-600M	52.6	72.12	0.83
mBART-50	44.7	65.53	0.79
Irish-English			
Whisper-small (E2E)	0.14	33.05	-

Table 2: Results for our translation systems on evaluation test set.

a cascaded system; however, this time, we used mBART-50 as the decoder. This setup provided us 44.70 BLEU, 65.53 chrF, 0.796 COMET points on the evaluation test set. Fine-tuning OpenAI's Whisper model for the English-to-Irish language pair has led to a performance improvement, despite the fact that the language is unsupported and unavailable in the model's training data. Unfortunately, the BLEU score remains low, probably due to instances of overgeneration and undergeneration. The ChrF score, which measures character-based similarity, is higher but still indicates room for improvement as far as translation quality is concerned.

The results obtained on the blind test set are shown in Table 3. For our primary submission for Maltese-to-English we obtained 56.67 BLEU and 81.92 chrF2 points on the IWSLT 2024 blind test sets. Like above, our contrastive systems were cascaded systems; the first and second contrastive systems provided us 52.6 BLEU and 72.12 chrF2 and 44.70 BLEU and 65.53 chrF2 points, respectively. For our primary submission for English-to-Irish language pair we obtained 0.6 BLEU and 15.4 ChrF2 points on the test set.

As shown in Table 2 and Table 3, our best performing system is cascaded system with whisper-small and NLLB-200-600M. However, E2E are better than cascaded system due to the fact that in cascaded systems errors from the ASR can severely impact the performance of the subsequent component (MT). In contarst, E2E models can learn to directly map source language speech to the target language text. Their ability to process input in a single pass can significantly reduce latency compared to cascaded systems that involve multiple stages of processing, thereby avoiding intermediate errors. Our team secured second position for the Maltese-to-English translation task in this competition.

	BLEU	ChrF2
Maltese-English		
Primary	56.67	81.92
Constrastive1-Data1	52.6	72.12
Constrastive1-Data2	44.7	65.53
Irish-English		
Primary	0.6	15.4

Table 3: Official results for our translation systems on blind set.

7 Conclusion

In this study, we discussed our ST models for the IWSLT 2024 Low-Resource Task for both Irish-English and Maltese-English language pairs. Our proposed architecture offers numerous benefits: it is both computationally and data-efficient, supports both speech-to-text and text-to-text translations (including transcription), enhances knowledge transfer which boosts performance in low-resource languages, and exhibits robust translation capabilities.

Future investigations will focus on a detailed assessment of our architecture's ASR functionality and explore the use of adapters within the speech representation model. Additionally, a thorough examination of the optimal layers will be necessary when the speech representation model is not updated.

References

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. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *Preprint*, arXiv:2111.09296.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Xie Chen, Yu Wu, Zhenghao Wang, Shujie Liu, and Jinyu Li. 2021. Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. *Preprint*, arXiv:2010.11395.

Pavel Denisov, Manuel Mager, and Ngoc Thang Vu. 2021. IMS' systems for the IWSLT 2021 low-resource speech translation task. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 175–181, Bangkok, Thailand (online). Association for Computational Linguistics.

Liang Ding and Dacheng Tao. 2021. The USYD-JD speech translation system for IWSLT2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 182–191, Bangkok, Thailand (online). Association for Computational Linguistics.

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. *Preprint*, arXiv:2106.07447.

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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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.

Tuan Nam Nguyen, Thai Son Nguyen, Christian Huber, Ngoc-Quan Pham, Thanh-Le Ha, Felix Schneider, and Sebastian Stüker. 2021. KIT's IWSLT 2021 offline speech translation system. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 125–130, Bangkok, Thailand (online). Association for Computational Linguistics.

John 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.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Balaji Radhakrishnan, Saurabh Agrawal, Raj Prakash Gohil, Kiran Praveen, Advait Vinay Dhopeshwarkar, and Abhishek Pandey. 2023. Sri-b's systems for iwslt 2023 dialectal and low-resource track: Marathi-hindi speech translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 449–454.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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. *Preprint*, arXiv:2207.04672.

Priyesh Vakharia, Pranjali Basmatkar, et al. 2023. Low-resource formality controlled nmt using pre-trained lm. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 321–329.

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.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. *Preprint*, arXiv:1804.00015.

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.

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

Weitai Zhang, Zhongyi Ye, Haitao Tang, Xiaoxi Li, Xinyuan Zhou, Jing Yang, Jianwei Cui, Pan Deng, Mohan Shi, Yifan Song, Dan Liu, Junhua Liu, and Lirong Dai. 2022. The USTC-NELSLIP offline speech translation systems for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 198–207, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Ziqiang Zhang and Junyi Ao. 2022. The YiTrans speech translation system for IWSLT 2022 offline shared task. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 158–168, Dublin, Ireland (in-person and online). Association for Computational Linguistics.