SpeechT: Findings of the First Mentorship in Speech Translation

Yasmin Moslem☆   **Juan Julián Cea Morán***   **Mariano Gonzalez-Gomez***

Muhammad Hazim Al Farouq*   **Farah Abdou***   **Satarupa Deb***

Abstract

This work presents the details and findings of the first mentorship in speech translation (SpeechT), which took place in December 2024 and January 2025. To fulfil the mentorship requirements, the participants engaged in key activities, including data preparation, modelling, and advanced research. The participants explored data augmentation techniques and compared end-to-end and cascaded speech translation systems. The projects covered various languages other than English, including Arabic, Bengali, Galician, Indonesian, Japanese, and Spanish.

1 Introduction

At the beginning of the mentorship on speech translation, the participants were provided with the following descriptions and guidelines for each task:

Data: Define, collect, and process bilingual speech data in a chosen language. Your dataset should consist of "train", "dev/validation", and "test" splits. By the end of the task, each participant should share a Hugging Face link to their datasets. The dataset page metadata should include sections for data sources, processing steps you applied in detail, and credits/citations of the original datasets.

Modelling: Choose one of the popular models, e.g. Whisper (Radford et al., 2022) or Wav2Vec (Baevski et al., 2020), and fine-tune it on the data prepared in the first task. Experimenting with different fine-tuning approaches and hyperparameters is encouraged. By the end of the task, the participants should share their fine-tuned models, and evaluation scores on the test dataset.

Advanced Research: Enhance the quality of your model through experimenting with advanced approaches, including creating synthetic data (Lam et al., 2022; Moslem, 2024), comparing end-to-end
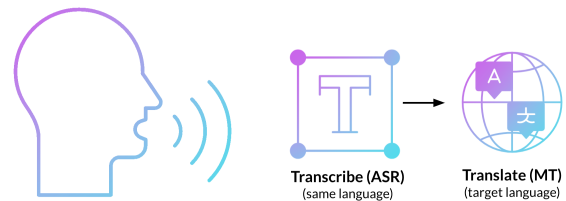

Figure 1: Cascaded Speech-to-Text System: Two models are trained, one for ASR, and one for MT of the transcriptions.
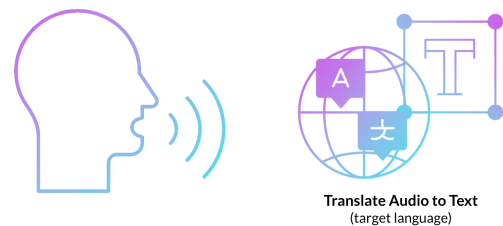

Figure 2: End-to-End Speech-to-Text System: One model is trained to generate the translation directly.

systems to cascaded systems (Agarwal et al., 2023), using language models (e.g. n-grams) (Baevski et al., 2020), domain adaptation (Samarakoon et al., 2018), or any other valid approach. By the end of the task, the participants should share their advanced models. They should also clarify how the advanced approach improved the speech translation quality compared to the original fine-tuned model.

Release & Publication: Write the project details to publish as a paper. Moreover, the outcomes of all the projects are publicly accessible.[1]

2 End-to-End vs. Cascaded systems

Speech translation systems can be (a) "cascaded" systems, or (b) "end-to-end" systems (Agarwal et al., 2023; Ahmad et al., 2024). Cascaded speech translation systems use two models, one for auto-

---

*Correspondence: yasmin[at]machinetranslation.io
☆Participant in the mentorship

[1] https://huggingface.co/SpeechT

matic speech recognition (ASR) and one for textual machine translation (MT) (cf. Figure 1). End-to-end speech translation systems use one model for the whole process (cf. Figure 2).

2.1 Cascaded Speech Translation

Cascaded speech systems involve sequential modules for Automatic Speech Recognition (ASR), Machine Translation (MT), and optionally Text-to-Speech (TTS), simultaneously combined to deliver the output to the end user. The ASR system generates transcriptions from the input audio, and then the MT model translates the transcriptions into the target language. Among the advantages of building "cascaded" systems are:

- Better quality in production.

- Each component (ASR, MT, TTS) can be individually optimized.

- Domain-specific (e.g. legal or medical) MT can be easily integrated.

2.2 End-to-End (E2E) Speech Translation

In end-to-end (E2E) speech systems, one model produces the whole process. E2E systems can also be extended with "cascaded" components. Among the advantages of building E2E systems are:

- Simpler deployment

- Better performance (lower latency)

3 Approaches to synthetic data

When the data is limited for the language or domain, synthetic data can be used to augment the authentic data. Synthetic data for speech translation systems can be generated in diverse methods, including:

- Using TTS models to generate synthetic source audio for authentic translations (Moslem, 2024)

- Using MT models to generate translations of audio transcriptions

- Sampling, translating, recombining: Lam et al. (2022) used an advanced approach to create synthetic data, by first chunking segments and transcriptions, creating a memory of prefix-suffix chunks based on part-of-speech tagging. Then they retrieve chunks from the memory to augment prefix chunks with similar suffix

chunks. Finally, they translate the new transcription with MT. Tools such as WhisperX (Bain et al., 2023) (based on Whisper) can be used for creating alignments based on word-level timestamps.

4 Projects

Most of the projects used a mix of data augmentation of authentic data with synthetic data, fine-tuning models, and comparing the performance of "end-to-end" speech systems to "cascaded" systems (cf. Section 2).

Participants used the Hugging Face Transformers library to fine-tune pretrained models. They fine-tuned Whisper (Radford et al., 2022) for "end-to-end" speech translation, and for ASR in the "cascaded" system. Moreover, they fine-tuned NLLB-200 (Costa-jussà et al., 2022) for text-to-text translation as part of "cascaded" speech translation systems. For evaluation, they used the sacreBLEU library (Post, 2018) to obtain BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2017) scores. In addition, one of the participants calculated COMET scores (Rei et al., 2020). For inference, they either used the Transformers library or Faster-Whisper (based on CTranslate2 (Klein et al., 2020)) for audio translation and transcription with Whisper. For text-to-text translation with OPUS and NLLB-200 models, some of them used the Transformer library directly while others used CTranslate2 with *float16* quantization, which is more efficient. For synthetic data generation, they used ChatGPT (OpenAI, 2023) and OPUS (Tiedemann and Thottingal, 2020) models.

Given that each participant chose a language pair, we dedicate a section for each project based on the language pair, including Galician-to-English, Indonesian-to-English, Spanish-to-Japanese, Arabic-to-English, and Bengali-to-English. Each language section describes data, modelling, and evaluation of each project.

4.1 Galician-to-English

4.1.1 Data [GL-EN]

In this project, two different Galician-to-English Speech Translation datasets have been curated. First, we compiled the dataset *OpenHQ-SpeechT-GL-EN* from the *crowdsourced high-quality Galician speech data set* by Kjartansson et al. (2020). After deduplicating the Galician audio-transcription pairs, we have applied a machine translation step to generate the corresponding English translations. More specifically, we have used

Language Pair	Train	Dev	Test	Dataset
AR-EN	2,228	278	279	*farahabdou/FLEURS-AR-EN-split*
BN-EN	41,984	9,000	1,000	*satarupa22/indic-en-bn*
ES-JA	9,972	1,440	1,345	*Marianoleiras/voxpopuli_es-ja*
GL-EN	4,798	507	282	*juanjucm/OpenHQ-SpeechT-GL-EN*
GL-EN	2,742	496	212	*juanjucm/FLEURS-SpeechT-GL-EN*
ID-EN	1,243	792	844	*cobrayyxx/COVOST2_ID-EN*

Table 1: Data Statistics

GPT-4o (Brown et al., 2020; OpenAI, 2023) with the following prompt:

> [{"role":"system", "content": "You are a helpful assistant that translates Galician (gl-ES) to English (en-XX).", },
>
> {"role": "user", "content": {source_text}}]

Given the absence of reference translation, we assessed the translation quality using CometKiwi (*wmt23-cometkiwi-da-xl*) (Rei et al., 2023), measuring an average score of 0.75. In total, this dataset contains approximately ten hours and twenty minutes of audio.

The second dataset is *FLEURS-SpeechT-GL-EN*. This is a subset of the *FLEURS* (Conneau et al., 2023) dataset, which contains two thousand parallel audio-transcription pairs in a hundred and two languages. For assembling our dataset, each Galician audio-transcription pair has been aligned with the corresponding English text. For this dataset, we used the same method for measuring translation quality, achieving an average score of 0.76. After cleaning and deduplication, this dataset contains around ten hours of audio. Table 1 shows more details about the data.

4.1.2 Modelling [GL-EN]

We first employed Whisper to train an "end-to-end" speech translation system. Whisper is a set of strong automatic speech recognition (ASR) architectures, trained on multilingual and multitask audio data. They can be further fine-tuned for speech translation. It supports Galician audio and text, making it a good choice for our data. Given our compute limitations, we experimented with two different backbones: *whisper-small* and *whisper-large-v3-turbo*, a simplified architecture of *whisper-large* with fewer parameters in the decoder section. We fine-tuned both models over our two datasets (cf. Section 4.1.1).

To further improve our "end-to-end" results, we trained a "cascaded" system which splits the speech translation task into two consecutive steps (cf. Section 2). Intuitively, this separation allows each model to specialise in a specific step of the pipeline, while adding one extra level of explainability to the whole process. The first module consists of a *whisper-large-v3-turbo*, this time in transcription mode, for generating Galician text given the input audio. Thereafter, on the same train split, we fine-tuned the MT model *NLLB-200-distilled-600M* on Galician-to-English text translation.

Inference was performed using the Transformers library. More specifically, we used its pipeline functionality to encapsulate pre-processing and post-processing steps. Training and inference were run on one RTX 4090 GPU.

4.1.3 Evaluation [GL-EN]

For the *FLEURS-SpeechT-GL-EN* dataset, the most performant "end-to-end" system was based on *whisper-small*, achieving a BLEU score of 22.62 and a ChrF++ score of 46.11. For the *OpenHQ-SpeechT-GL-EN* dataset, *whisper-large-v3-turbo* was better, with a BLEU score of 55.65 and a ChrF++ score of 72.19. Regarding our cascaded system for *FLEURS-SpeechT-GL-EN*, after using the MT model to translate the transcription generated by the ASR model, we obtained a BLEU score of 37.19 and a ChrF++ score of 61.33. For *OpenHQ-SpeechT-GL-EN*, the cascaded approach resulted in a BLEU score of 66.05 and a ChrF++ score of 79.58. Hence, the cascaded approach, despite being more computationally demanding, allows for a better specialization for each part of the system, hence generating significantly better results (cf. Table 2).

4.2 Indonesian-to-English

4.2.1 Data [ID-EN]

The dataset was compiled by extracting the English and Indonesian datasets from CoVoST2 (Wang et al., 2021b), a speech dataset in 21 languages, including Indonesian. Columns besides the index,

Indonesian audio with its transcription, and English transcription were removed. The next preprocessing step was checking duplicate indices within each split and identifying overlapping indices across the splits. This dataset was first used to train an "end-to-end" speech-translation system. For speech translation using a "cascaded" system, two models were trained: an automatic speech recognition (ASR) model and a machine translation (MT) model. Hence, the audio and transcription columns were used to train the ASR model, while textual source and target columns were used to train the text-to-text MT model.

4.2.2 Modelling [ID-EN]

We employed different approaches for the speech-translation tasks, an "end-to-end" system and a "cascaded" system (cf. Section 2). The pretrained model *whisper-small* was used for training the "end-to-end" system. We fine-tuned the model with the Indonesian audio and English transcription directly. Meanwhile, in the "cascaded" system, the model was fine-tuned to predict the audio transcription in the same language, which is Indonesian. As a "cascaded" system requires an MT model for translating Indonesian transcription into English, we fine-tuned *nllb-200-distilled-600M*, with batch size of 2 and gradient accumulation steps of 8 to simulate the effect of larger batch sizes. The model was trained for 10 epochs, saving the best epoch in the end.

For inference, we used Faster-Whisper for both translation and transcription with Whisper after converting the model into the CTranslate2 formate with float16 quantization, with a batch size 5 and the VAD filter enabled.[2] Similarly, for textual translation with NLLB-200, we used CTranslate2 with float16 quantization. Training was run on the T4 GPU from Google Colab, while inference used an RTX 2000 Ada GPU.

4.2.3 Evaluation [ID-EN]

The evaluation result of the "cascaded" system outperforms the "end-to-end" system on the *CoVoST2* test set. The "end-to-end" system achieved a BLEU score of 37.02 and ChrF++ score of 56.04 after fine-tuning Whisper Small, considerably improving the baseline (whose scores were BLEU 25.87 and ChrF++ 43.79). The "cascaded" system which fine-tuned both Whisper for transcription and NLLB-200 for translation achieved 48.60 BLEU score and

65.10 ChrF++ score, which outperforms both the baseline (BLEU 38.24 and ChrF++ 56.88) and the fine-tuned end-to-end model (cf. Table 2).

4.3 Spanish-to-Japanese

4.3.1 Data [ES-JA]

The foundational dataset is *VoxPopuli* (Wang et al., 2021a), from which we extracted audio and Spanish transcriptions. We generated Japanese translations using OPUS models (Tiedemann and Thottingal, 2020), initially translating from Spanish to English and then from English to Japanese. While multilingual options existed, this two-step approach was chosen due to the strong performance of high-resource language pairs. Post-processing was necessary to refine the dataset. First, we removed blank spaces, which are not typical in Japanese writing, ensuring proper formatting and consistency. Then, we eliminated empty texts and employed quality estimation with a threshold of 0.7 to filter out low-quality translations, using the CometKiwi (*wmt23-cometkiwi-da-xl*) model. This process helped maintain alignment between the audio, transcriptions, and translations, resulting in a final dataset of approximately 12.7k rows. Regarding content, the dataset consists of European Parliament event recordings featuring various Spanish accents. As a result, models trained on this data are likely to perform better in similar parliamentary or formal discourse scenarios (cf. Table 1).

4.3.2 Modelling [ES-JA]

We built two systems for the Spanish-to-Japanese (ES-JA) speech translation task, an "end-to-end" system and a "cascaded" system (cf. Section 2). The backbone of the "end-to-end" model is *whisper-small*, which has been trained on the ES-JA VoxPopuli dataset 4.3.1. This *whisper-small* model has been fine-tuned specifically for direct speech-to-text translation, meaning that the Spanish audio is encoded and directly decoded into Japanese, without requiring any intermediate transcription step. This approach offers a simpler architecture and a lower computational cost, since only one model is used, training and inference are more efficient.

On the contrary, the "cascaded" approach involves two separate models, (i) the *whisper-small* for transcribing Spanish audio into text, and (ii) the *nllb-200-distilled-600M* for translating the transcribed Spanish text into Japanese. While this method is more resource-intensive, it allows independent optimization of each component.

---

[2]Voice Audio Detection (VAD) removes low-amplitude samples from an audio signal, which might represent silence or noise.

For inference, both approaches process Spanish audio inputs into Japanese text output. In the "end-to-end" approach, the model directly translates Spanish speech into Japanese in a single step (only one model is executed, taking less time and resources). However, in the "cascaded" approach there is a sequential process: The output of the model that transcribes Spanish into text is the input to the model that translates Spanish into Japanese (Two models are used, making it possible to optimize each of them but using more resources), providing a higher quality in terms of translation quality metrics. For this, we used the Hugging Face Transformers library pipelines: "automatic-speech-recognition" and "translation". As for infrastructure, we conducted both training and inference of the models on one RTX 4090 GPU.

4.3.3 Evaluation [ES-JA]

The evaluation of the Spanish-to-Japanese translation models reveals a performance gap between the "end-to-end" and "cascaded" approaches. The "end-to-end" model scores on the test split indicate room for improvement, achieving a BLEU score of 20.86, a ChrF++ score of 23.36, and a COMET score of 77.7. This suggests that while the translations maintain some coherence, they lack the precision and fluency. In contrast, the "cascaded" approach outperforms the "end-to-end" model across all metrics. This system reaches a BLEU score of 35.32, a ChrF++ score of 32.82, and a COMET score of 89.86, demonstrating superior lexical and syntactic alignment with reference translations (cf. Table 2).

4.4 Arabic-to-English & Bengali-to-English

Due to the similarity of the projects of the Arabic-to-English and Bengali-to-English language pairs, we combine them in one section. Unlike the aforementioned projects that fine-tuned models for all systems, these two projects fine-tuned models for the "end-to-end" system. In addition, the Bengali-to-English project fine-tuned Whisper for the "cascaded" system. However, both project used the baseline of NLLB-200 600M without fine-tuning.

4.4.1 Data [AR-EN & BN-EN]

The dataset used in the Arabic-to-English project is a subset of the FLEURS dataset (Conneau et al., 2023), while the Bengali-to-English project used the IndicVoices dataset after filtering out segments whose mining scores are less than 0.7 (Jain et al., 2024; Javed et al., 2024). The data is split into training and test sets to facilitate model training and

evaluation. As the datasets include both the transcriptions and translations, it is useful for "end-to-end" speech translation tasks, as well as "cascaded" systems that involve separate speech recognition and machine translation models. Table 1 illustrates more details about the used data.

4.4.2 Modelling [AR-EN & BN-EN]

Two approaches were employed for the Arabic-to-English and Bengali-to-English translation tasks:

End-to-End Model: The model utilizes whisper-small model, which is a pre-trained speech-to-text model capable of handling "end-to-end" speech translation. This model directly translates Arabic or Bengali speech into English text without intermediate steps. While the Arabic model was fine-tuned on the FLEURS dataset, the Bengali models were fine-tuned with the IndicVoices dataset.

Cascaded Model: This approach combines two models: (i) Automatic Speech Recognition (ASR) using the Whisper model to transcribe Arabic speech into Arabic text, and (ii) Machine Translation (MT) using NLLB-200 to translate the transcribed Arabic or Bengali text into English.

For Arabic-to-English inference, the Hugging Face Transformers library was used for both speech-to-text transcription and text translation tasks, as well as "end-to-end" speech translation. For Bengali-to-English "end-to-end" translation, the FasterWhisper library (based on CTranslate2) was used after converting the model with float16 quantization, while translation with NLLB-200 600M used CTranslate2. Training and inference utilized Google Colab, as well as GPU P100 on Kaggle and a multi-GPU setup comprising two NVIDIA T4 GPUs on Kaggle.

4.4.3 Evaluation [AR-EN & BN-EN]

As in the case of other projects, the results of English-to-Arabic and Bengali-to-English speech translation indicate that the "cascaded" model outperforms the "end-to-end" model in terms of translation quality (cf. Table 2).

5 Conclusions

The SpeechT mentorship brought together several practitioners and students from diverse companies and institutions across the world to explore speech translation. The participants have diverse backgrounds, ranging from generic software knowledge to text-to-text MT experience. Ultimately, five participants have made successful submissions and contributed to this work (cf. Section 6).

Language Pair	System	Model	Type	Dataset	BLEU	ChrF++
GL-EN	End-to-End	Whisper Small	Baseline	Fleurs	16.01	44.99
	End-to-End	Whisper Large Turbo	Baseline	Fleurs	5.09	26.59
	Cascaded	+ NLLB-200 600M	Baseline	Fleurs	34.47	59.29
	End-to-End	Whisper Small	Fine-tuned	Fleurs	22.62	46.11
	End-to-End	Whisper Large Turbo	Fine-tuned	Fleurs	18.96	46.00
	Cascaded	+ NLLB-200 600M	Fine-tuned	Fleurs	**37.19**	**61.33**
	End-to-End	Whisper Small	Baseline	OpenHQ	21.46	41.12
	End-to-End	Whisper Large Turbo	Baseline	OpenHQ	3.38	21.82
	Cascaded	+ NLLB-200 600M	Baseline	OpenHQ	43.01	64.52
	End-to-End	Whisper Small	Fine-tuned	OpenHQ	50.96	69.24
	End-to-End	Whisper Large Turbo	Fine-tuned	OpenHQ	55.64	72.19
	Cascaded	+ NLLB-200 600M	Fine-tuned	OpenHQ	**66.05**	**79.58**
ID-EN	End-to-End	Whisper Small	Baseline	CoVoST2	25.87	43.79
	Cascaded	+ NLLB-200 600M	Baseline	CoVoST2	38.24	56.88
	End-to-End	Whisper Small	Fine-tuned	CoVoST2	37.02	56.04
	Cascaded	+ NLLB-200 600M	Fine-tuned	CoVoST2	**48.60**	**65.10**
ES-JA	End-to-End	Whisper Small	Baseline	VoxPopuli	0.48	3.18
	Cascaded	+ NLLB-200 600M	Baseline	VoxPopuli	21.34	23.21
	End-to-End	Whisper Small	Fine-tuned	VoxPopuli	20.86	23.36
	Cascaded	+ NLLB-200 600M	Fine-tuned	VoxPopuli	**35.32**	**32.82**
AR-EN	End-to-End	Whisper Small	Baseline	Fleurs	5.65	31.75
	End-to-End	Whisper Small	Fine-tuned	Fleurs	15.06	39.03
	Cascaded	+ NLLB-200 600M	Baseline	Fleurs	**24.38**	**51.79**
BN-EN	End-to-End	Whisper Small	Baseline	IndicVoices	6.33	24.60
	End-to-End	Whisper Small	Fine-tuned	IndicVoices	10.08	30.97
	Cascaded	+ NLLB-200 600M	Baseline	IndicVoices	**20.42**	**42.51**

Table 2: Results: Cascaded systems outperform end-to-end systems in speech translation across all language pairs.

Successful submissions incorporated a range of techniques. In particular, participants experimented with synthetic data generation with large language models (e.g. GPT4) and MT models (e.g. OPUS). The focus of most of the experiments was comparing the speech translation performance of "end-to-end" systems with "cascaded" systems (cf. Section 2). For this purpose, the participants fine-tuned pretrained models, including Whisper and NLLB-200. While the "end-to-end" systems fine-tuned Whisper for direct speech translation, building the "cascaded" systems involved two steps, namely fine-tuning Whisper for ASR, and then employing an MT model (e.g. NLLB) for translation of the generated transcription. As Table 2 illustrates, "cascaded" systems outperformed "end-to-end" across all language pairs. In conclusion, this mentorship has enabled the participants to experiment with various system designs and fine-tuning strategies, deepening their understanding of the speech translation area through hands-on practice.

6 Contributions

- **Yasmin Moslem:** Organizer and mentor of *SpeechT* mentorship in Speech Translation

Participants (alphabetically ordered)

- **Farah Abdou:** Participant, Arabic-to-English Speech Translation
- **Juan Julián Cea Morán:** Participant, Galician-to-English Speech Translation
- **Mariano Gonzalez-Gomez:** Participant, Spanish-to-Japanese Speech Translation
- **Muhammad Hazim Al Farouq:** Participant, Indonesian-to-English Speech Translation
- **Satarupa Deb:** Participant, Bengali-to-English Speech Translation

References

Agarwal, M., Agrawal, S., Anastasopoulos, A., Bentivogli, L., Bojar, O., Borg, C., Carpuat, M., Cattoni, R., Cettolo, M., Chen, M., Chen, W., Choukri, K., Chronopoulou, A., Currey, A., Declerck, T., Dong, Q., Duh, K., Estève, Y., Federico, M., Gahbiche, S., Haddow, B., Hsu, B., Mon Htut, P., Inaguma, H., Javorský, D., Judge, J., Kano, Y., Ko, T., Kumar, R., Li, P., Ma, X., Mathur, P., Matusov, E., McNamee, P., P. McCrae, J., Murray, K., Nadejde, M., Nakamura, S., Negri, M., Nguyen, H., Niehues, J., Niu, X., Kr. Ojha, A., E. Ortega, J., Pal, P., Pino, J., van der Plas, L., Polák, P., Rippeth, E., Salesky, E., Shi, J., Sperber, M., Stüker, S., Sudoh, K., Tang, Y., Thompson, B., Tran, K., Turchi, M., Waibel, A., Wang, M., Watanabe, S., and Zevallos, R. (2023). FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In Salesky, E., Federico, M., and Carpuat, M., editors, *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.

Ahmad, I. S., Anastasopoulos, A., Bojar, O., Borg, C., Carpuat, M., Cattoni, R., Cettolo, M., Chen, W., Dong, Q., Federico, M., Haddow, B., Javorský, D., Krubiński, M., Kim Lam, T., Ma, X., Mathur, P., Matusov, E., Maurya, C., McCrae, J., Murray, K., Nakamura, S., Negri, M., Niehues, J., Niu, X., Ojha, A. K., Ortega, J., Papi, S., Polák, P., Pospíšil, A., Pecina, P., Salesky, E., Sethiya, N., Sarkar, B., Shi, J., Sikasote, C., Sperber, M., Stüker, S., Sudoh, K., Thompson, B., Waibel, A., Watanabe, S., Wilken, P., Zemánek, P., and Zevallos, R. (2024). FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Stroudsburg, PA, USA. Association for Computational Linguistics.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). Wav2Vec 2.0: A Framework for Self-supervised Learning of Speech Representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, number Article 1044 in NIPS '20, pages 12449–12460, Red Hook, NY, USA. Curran Associates Inc.

Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). WhisperX: Time-Accurate Speech Transcription of Long-form Audio. In *Proceedings of Interspeech 2023, the 24th Annual Conference of the International Speech Communication Association*, pages 4489–4493.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 1877–1901, Virtual. Curran Associates, Inc.

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., and Bapna, A. (2023). FLEURS: FEW-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No Language Left Behind: Scaling human-centered machine translation. arXiv [cs.CL].

Jain, S., Sankar, A., Choudhary, D., Suman, D., Narasimhan, N., Khan, M. S. U. R., Kunchukuttan, A., Khapra, M. M., and Dabre, R. (2024). BhasaAnuvaad: A Speech Translation Dataset for 13 Indian Languages. arXiv [cs.CL].

Javed, T., Nawale, J. A., George, E. I., Joshi, S., Bhogale, K. S., Mehendale, D., Sethi, I. V., Ananthanarayanan, A., Faquih, H., Palit, P., Ravishankar, S., Sukumaran, S., Panchagnula, T., Murali, S., Gandhi, K. S., R, Ambujavalli, M, Manickam K, Vaijayanthi, C. V., Karunganni, K. S. R., Kumar, P., and Khapra, M. M. (2024). IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages. arXiv [cs.CL].

Kjartansson, O., Gutkin, A., Butryna, A., Demirsahin, I., and Rivera, C. (2020). Open-Source High Quality Speech Datasets for Basque, Catalan and Galician. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 21–27.

Klein, G., Zhang, D., Chouteau, C., Crego, J., and Senellart, J. (2020). Efficient and high-quality neural machine translation with OpenNMT. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 211–217, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lam, T. K., Schamoni, S., and Riezler, S. (2022). Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 245–254, Stroudsburg, PA, USA. Association for Computational Linguistics.

Moslem, Y. (2024). Leveraging Synthetic Audio Data for End-to-End Low-Resource Speech Translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 265–273.

OpenAI (2023). GPT-4 Technical Report. arXiv [cs.CL].

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Popović, M. (2017). chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. arXiv [eess.AS].

Rei, R., Guerreiro, N. M., Pombal, J., van Stigt, D., Treviso, M., Coheur, L., C. de Souza, J. G., and Martins, A. (2023). Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (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.

Samarakoon, L., Mak, B., and Lam, A. Y. S. (2018). Domain adaptation of end-to-end speech recognition in low-resource settings. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 382–388. IEEE.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. (2021a). VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003.

Wang, C., Wu, A., and Pino, J. (2021b). CoVoST 2 and Massively Multilingual Speech-to-Text Translation. In *Proceedings of Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association*, pages 2247–2251.