LIA and ELYADATA systems for the IWSLT 2025 low-resource speech translation shared task

Chaimae Chellaf[*,1,4], Haroun Elleuch[*,1,2], Othman Istaiteh[*,1], Fortuné Kponou[*,1,3]
Fethi Bougares[1,2], Yannick Estève[1], Salima Mdhaffar[1]

[1]LIA (France), [2]Elyadata (Tunisia), [3]UAC/IMSP (Benin), [4]LundiMatin (France)

Correspondence: salima.mdhaffar@univ-avignon.fr

Abstract

In this paper, we present the approach and system setup of our participation in the IWSLT 2025 low-resource speech translation shared task. We submitted systems for three language pairs, namely Tunisian Arabic to English, North Levantine Arabic to English, and Fongbé to French. Both pipeline and end-to-end speech translation systems were explored for Tunisian Arabic to English and Fongbé to French pairs. However, only pipeline approaches were investigated for the North Levantine Arabic–English translation direction. All our submissions are based on the usage of pre-trained models that we further fine-tune with the shared task training data.

1 Introduction

The International Workshop on Spoken Language Translation (IWSLT) is an annual scientific conference dedicated to the study and advancement of spoken language translation technologies. It serves as a platform for researchers and practitioners to present their work on speech translation, encompassing areas such as automatic speech recognition (ASR) and machine translation (MT). IWSLT has played a pivotal role in the advancement of spoken language translation (ST) by providing a structured environment to evaluate and compare different approaches. Its emphasis on real-world challenges, such as low-resource languages and real-time translation, has contributed to the development of more robust and versatile translation systems. IWSLT 2025 (Abdulmumin et al., 2025) proposes two shared tasks: High-resource ST and Low-resource ST. Several language pairs were proposed this year for the low-resource task. In this paper, we focus on Tunisian Arabic-English, North Levantine-English and Fongbé-French language pairs. This paper describes the approach and system setup of the joint participation of LIA and Elyadata in the tasks as mentioned earlier.

For the Tunisian Arabic-English and Fongbé–French tracks, both end-to-end (E2E) and pipeline approaches were explored. In contrast, only pipeline approaches were investigated for the North Levantine Arabic–English track. For E2E approaches, we focus on fine-tuning self-supervised learning (SSL) models and Whisper models (Radford et al., 2023a). All systems are trained with an unconstrained setup, which means that any resource, including pre-trained language models, can be used, except for evaluation sets. We particularly investigate the SAMU-like approach (Khurana et al., 2022) to enrich the SSL speech encoder with semantic information. For pipeline approaches, we focus on fine-tuning large language models (LLMs).

The remaining of the paper is structured as follows: Section 2 presents the related work. Section 3 is dedicated to describe our systems for Tunisian Arabic to English. The experiments for Fongbé to French and for North Levantine to English are presented, respectively, in sections 4 and 5. Section 6 concludes the paper and discusses future work.

2 Related Work

The Speech Translation task has received considerable attention from the research community, and numerous approaches have been proposed. Traditional speech translation (ST) approaches follow a cascade architecture (Matusov et al., 2005; Kumar et al., 2015; Laurent et al., 2023), where an automatic speech recognition (ASR) system is followed by a machine translation (MT) module applied to the ASR output. Recent advances in deep neural networks for both ASR and MT have led to substantial improvements in the overall performance of ST systems.

More recently, end-to-end speech translation

[*]These authors contributed equally to this work

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 274–281

models (Bérard et al., 2018; Duong et al., 2016; Bérard et al., 2016) have gained attention as an alternative to the traditional cascade architecture. These models aim to directly translate speech in a source language into text or speech in the target language without requiring intermediate text transcriptions. End-to-end models reduce latency, avoid error propagation between ASR and MT components, and can be optimized globally for the final translation objective.

With the emergence of robust transformer-based architectures and multilingual pretraining methods, such as those used in SeamlessM4T (Seamless Communication et al., 2023), speech translation systems have gained momentum, leading to diversity in model architectures and training methods. Meta's SeamlessM4T stands out as a unified multimodal system capable of handling speech-to-text translation across 101 input and 96 output languages. OpenAI's Whisper (Radford et al., 2023a) is an automatic speech recognition (ASR) system that also offers speech-to-text translation capabilities. Trained on 680,000 hours of multilingual and multitask supervised data, Whisper demonstrates robustness to accents, background noise, and technical language. It supports transcription in multiple languages and translation from those languages into English. Of the 680,000 hours of labelled audio used by Whisper, 117,000 hours cover 96 other languages. The dataset also includes 125,000 hours of X→EN translation data. Beyond Whisper and SeamlessM4T, several other models have emerged that employ self-supervised learning (SSL) to enhance performance in speech translation. Wav2vec 2.0 (Baevski et al., 2020), introduced by Facebook AI, is one of the earliest SSL-based models that significantly improved ASR performance. Wav2vec 2.0 is typically coupled with a Transformer decoder for speech translation. Building on this foundation, w2v-BERT (Chung et al., 2021) and HuBERT (Hsu et al., 2021) models have been developed. In this paper, we investigate these recent advances in speech-to-text translation systems to participate in the IWSLT low-resource speech translation shared task.

3 Tunisian Arabic-English Experiments

3.1 Data

The Tunisian Arabic dataset (LDC2022E01) used in our experiments was developed and provided by LDC2 to the IWSLT 2025 participants. It comprises 383h of Tunisian conversational speech with manual transcripts, from which 160h are also translated into English. Thus, it is a three-way parallel corpus, comprising audio, transcript, and translation. This LDC data constitutes the basic condition of the dialect task. Arabic dialects are the informal form of communication in everyday life in the Arab world. Tunisian Arabic is one of several Arabic dialects. There is no standard written Arabic form that all Tunisian speakers share. However, the transcripts of Tunisian conversations of the LDC2022E01 Tunisian Arabic dataset follow the rules of the Tunisian Arabic CODA – Conventional Orthography for Dialectal Arabic (Habash et al., 2012).

3.2 Pipeline ST

3.2.1 ASR systems

Two ASR systems have been trained for the Tunisian dialect. The first ASR system **(Primary)** is based on the w2v_Bert 2.0 (Barrault et al., 2023) speech encoder. In addition to the speech encoder model, we incorporate an extra layer with 1024 neurons and LeakyReLU as the activation function, followed by a fully-connected layer and a final 37-dimensional softmax layer, each dimension corresponding to a character. The weights of these two additional layers were randomly initialized. In contrast, the weights of the speech encoder part for SSL models in the neural architecture were initialized using pre-trained weights. The fine-tuning is done with the LDC2022E01 training set using a character-level CTC loss function. We optimize the loss with an Adam optimizer of learning rate equal to 1×10^{-5} for both the speech encoder and Adadelta with learning rate equal to 1.0 for the linear layer.

The second ASR system **(contrastive 1)** is trained with the same dataset and is based on the Whisper-large-v3 model (Radford et al., 2023b).

We fine-tune this Whisper model for the ASR task with the LDC2022E01 dataset. we used AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of $1e-5$ and weight decay of 0.01. The encoder was left unfrozen throughout the training, which consisted of 10 epochs with a warmup of 500 steps, a patience of two epochs for early stopping, and a maximum gradient norm of 2.0. Training was performed using FP16 precision with a sampling rate of 16 kHz, and the data was randomly sorted. We set the batch size per GPU

to 8 and used 4 H100 80GB GPUs with a gradient accumulation factor of 4, resulting in an effective batch size of 128 (8 × 4 GPUs × 4 accumulation steps). We used a beam size of 8 for decoding, with decode ratios ranging from 0.0 to 1.0. The model was optimized using a negative log-likelihood loss.

We use the SpeechBrain toolkit to train ASR systems (Ravanelli et al., 2024).

3.2.2 MT model

We fine-tuned a machine translation model (**contrastive3**) based on the NLLB-200 1.3B architecture (Costa-Jussà et al., 2022), a multilingual transformer model designed to support high-quality translation across over 200 languages, including many low-resource ones. This model was specifically adapted for the task of translating Tunisian Arabic into English.

Fine-tuning was performed using the LDC2022E01 translation training set, with optimization carried out using the Cross-Entropy loss function. We used the AdamW optimizer with a learning rate of 1×10^{-5} and a batch size of 16, with a beam size of 8 for decoding. We use the HuggingFace framework to train the MT model.

3.3 End-to-end ST

The entire dataset used to train the E2E system includes 160 hours of data with gold translations provided for the task, and 223 hours without translations, which we automatically translated using the MT system described in Section 3.2.2. We filter a portion of the 223 hours of translated data using the BLASER score to improve translation quality.

SAMU-XLSR (Khurana et al., 2022) is a multilingual multimodal semantic speech representation learning framework where the speech transformer encoder XLS-R is fine-tuned using semantic supervision from the pre-trained multilingual semantic text encoder LaBSE (Feng et al., 2022a). The training and modeling details follow the original paper (Khurana et al., 2022). In this work, we use the same training framework except that we trained our model starting from another speech encoder: w2v_Bert 2.0 (Barrault et al., 2023) and another semantic text encoder BGE M3-Embedding (Chen et al., 2024). We use the CommonVoice-v19 (Ardila et al., 2020) to train this model. In this paper, we refer to this model as SAMU-BGE.

We use the standard encoder-decoder architecture for our translation model. The training of our E2E ST model is divided into three stages. First, we specialize the SAMU-BGE model with the Tunisian ST dataset. Second, we fine-tune the mBart model for text-to-text translation from Tunisian to English. Once our speech encoder (SAMU-BGE) and our decoder (mBart) are fine-tuned, we initialize the encoder and decoder using these models. A feed-forward network projection layer is used to connect the encoder and decoder, bridging the two modules. The described system presents our **Primary** ST system.

For the **contrastive 1** system, we use Whisper-large-v3 to train the ST model. The training of our model is separated into two stages. First, we train an end-to-end ASR model (the ASR model is described in Section 3.2.1). Then, once our ASR model is trained, we fine-tune this Whisper model for the translation task. We used AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 1e-5 and weight decay of 0.01. The encoder was left unfrozen throughout the training, which ran for 10 epochs with a warmup of 500 steps, a patience of two epochs for early stopping, and a max gradient norm of 2.0. Training used FP16 precision with a sampling rate of 16 kHz, and the data was randomly sorted. We set the batch size per GPU to 8 and used 4 H100 80GB GPUs with a gradient accumulation factor of 4, resulting in an effective batch size of 128 (8 × 4 GPUs × 4 accumulation steps). We used a beam size of 8 for decoding, with decode ratios ranging from 0.0 to 1.0. The model was optimized using a negative log-likelihood loss. We combined different augmentations to perform data augmentation: speed perturbation (resample the audio signal at a rate that is similar to the original rate, to achieve a slightly slower or slightly faster signal), frequency drop (randomly drops several frequency bands to zero) and chunk drop (an augmentation strategy that helps a model learn to rely on all parts of the signal, since it can't expect a given part to be present).

For the **contrastive 2** system, we use Whisper-large-v3 to train the ST model without the step of ASR fine-tuning and without data augmentation. We apply the same parameters described for the **contrastive 1** system.

3.4 Results

3.4.1 Results for ASR

The ASR results in terms of word error rate (WER) are reported in Table 1 on the development datasets and the internal test provided by the organisers.

Table 1: WER (%) results for Tunisian dialect ASR.

	Dev	Test Int	Test1	Test2
Primary	36.3	39.63	38.6	40
Contrastive 1	36.78	40.43	39.2	40.3

3.4.2 Results for ST

The ST results in terms of BLEU scores are reported in Table 2 on the development datasets and the internal test provided by the organisers.

Table 2: BLEU results for Tunisian dialect to English translation.

	Dev	Test Int	Test1	Test2
Primary	25.04	21.41	22.3	21
Contrastive 1	24.72	21.12	22	20.3
Contrastive 2	24.63	20.40	21.6	19.2
Contrastive 3	23.77	20.23	21.4	19.6

4 Fongbé-French Experiments

4.1 Data

The dataset used in our experiments comprises a total of 61 hours of speech. For the end-to-end speech translation (ST) task, we used the entire dataset. We also used internal data for the automatic speech recognition (ASR) task, by using Fongbé transcripts we collected for a 36 hour subset. To ensure a fair comparison between the end-to-end and cascade systems, we excluded the validation and test portions of the ST dataset from the ASR training set.

Table 3: ST and ASR dataset description

Experiments	Split	Hours	Sentences
ASR	Train	29	19.9k
ASR	Valid	3.54	2.4k
ASR	Test	3.93	2.5k
ST	Train	48	29.5k
ST	Valid	6.1	4.1k
ST	Test	5.9	3.9k

4.2 Pipeline ST

4.2.1 ASR system

We conducted three automatic speech recognition (ASR) experiments for ASR. In the first experiment, Fongbé transcripts containing diacritics were used to establish a baseline, referred to as ASR With Diacritics. The second experiment was performed using transcripts without diacritics (ASR Without Diacritics). In the third, we introduced a novel diacritic substitution strategy: monosyllabic words containing diacritics were systematically replaced by their base syllables appended with a unique numerical identifier (ASR with Sub). This method was designed to retain key linguistic distinctions while modifying the representation of diacritics, potentially improving the model's ability to generalize across phonetically similar patterns. For each setting, we trained a separate *SentencePiece* tokenizer (Kudo and Richardson, 2018) at the character level using the combined training and validation sets. The resulting vocabulary sizes were 62, 44, and 36 for the diacritics, no-diacritics, and substitution settings, respectively. All ASR models shared the same architecture, consisting of an AfriHuBERT speech encoder followed by three fully connected layers with 1024 dimensions. Training was performed using Connectionist Temporal Classification (CTC) loss over 50 epochs. The ASR model trained without diacritics achieved the lowest WER of 17.02%, outperforming both the model trained with diacritics (21.98%) and the substitution-based model (22.18%), as detailed in Table 4.

Table 4: WER (%) results for Fongbé ASR

	Dev	Test
ASR with Diacritics	17.25	21.89
ASR without Diacritics	12.71	17.02
ASR with Sub	24.63	22.18

4.2.2 MT model

We fine-tuned three versions of the NLLB-200 1.3B model on Fongbé manual transcriptions: one with diacritics, the second without diacritics, and a third using diacritic-substituted sentences, as described in the previous section. The model trained on diacritic transcriptions achieved the best performance, followed by the substitution-based model. The results show in Table 5 highlight the importance of diacritics in Fongbé for translation quality, while also demonstrating that the substitution approach offers a competitive alternative positioned between the performance of models trained with and without diacritics.

Table 5: BLEU (%) results for Fongbé MT

	Dev	Test
MT With Diacritics	58.9	55.41
MT Without Diacritics	47.39	44.95
MT with Sub	57.56	53.88

Table 6: BLEU results for Fongbé to French translation.

	Dev	Test
Primary	59.24	39.6
Contrastive 1	54.87	37.23
Contrastive 2	48.39	32.76
Contrastive 3	41.60	28.32

4.3 Fongbe Speech Translation system

We explored the use of various speech encoders—specifically HuBERT-147, AfriHuBERT, and XLS-R-1B in combination with different text decoders, including mBART and NLLB. All experimental results are presented in Table 6.

For the cascade experiments, we paired each ASR system with its corresponding machine translation (MT) system. The best-performing cascade system combines ASR with diacritics and MT with diacritics, and is designated as the **Primary** system. The second-best system, referred to as **Contrastive 1**, used both ASR and MT models trained on diacritic-substituted data. The third system, **Contrastive 2**, employed ASR and MT models trained on data without diacritics.

In the end-to-end setting, for experiments involving XLS-R-1B, we applied a semantic alignment strategy inspired by the method proposed in Khurana et al. (2022), using translated labels. SAMU, which builds on XLS-R-1B, integrates a frozen Language-Agnostic BERT Sentence Encoder (LaBSE) (Feng et al., 2022b) as the master model to align Fongbé speech embeddings and French text embeddings in a shared XLS-R representation space.

We also investigated the impact of several data augmentation techniques, including speed perturbation, frequency drop, and chunk drop. Our best end-to-end systems combined the AfriHuBERT encoder with the NLLB decoder, and the SAMU model with NLLB, both enhanced by these augmentations. Among them, the SAMU-NLLB system achieved the highest performance in the end-to-end speech translation task, ranking fourth overall among all submitted systems. Consequently, we selected the SAMU-NLLB end-to-end system as the **Contrastive 3** submission.

4.4 Results

Overall, for the Fongbé Speech Translation task, we proposed both cascade and end-to-end systems. All cascade systems outperformed the end-to-end approach, with a gap of approximately ~11 BLEU points between the best-performing cascade system and the submitted end-to-end system (SAMU-NLLB + Data Augmentation). This performance difference highlights the potential for improving end-to-end models through more effective encoder adaptation techniques for the decoder, aiming to narrow the gap between end-to-end and cascade performance. The superiority of cascade systems can be attributed, in part, to the use of in-domain ASR data for fine-tuning the decoder, which provides a more aligned and semantically rich input for the translation model.

5 North Levantine-English Experiments

5.1 Data

5.1.1 ASR dataset

The training data consisted of the Babylon Levantine corpus (LDC2005S08) and the Levantine Arabic QT (LDC2006T07), both provided by LDC, along with an additional 23 hours of Levantine speech automatically extracted from the QASR dataset using the best performing dialect identification model from Elleuch et al. (2025) [1]. QASR is the largest publicly available Arabic speech recognition dataset, consisting of 2,000 hours of transcribed speech collected from the broadcast domain. It includes both dialectal and Modern Standard Arabic (MSA) speech, as well as code-switching (Mubarak et al., 2021).

5.1.2 MT dataset

The training data for the North Levantine to English machine translation task consisted of two distinct corpora. The first is the **UFAL Parallel Corpus of North Levantine 1.0**, provided to participants in the IWSLT 2025 shared task. This corpus

[1]Whisper-large-v3 encoder trained on the ADI-20-53 dataset for Arabic dialect identification. This dataset comprises 53 hours of speech for 20 country-level dialects. The Levantine subset included speech segments identified as Jordanian, Palestinian, Syrian, and Lebanese.

comprises approximately 120,000 lines of parallel North Levantine, MSA, and English textual data.

The second corpus is **Levanti**[2], which includes 500,000 sentence pairs in Levantine colloquial Arabic (covering Palestinian, Syrian, Lebanese, Jordanian, and Egyptian dialects) and their English translations. Levanti comprises 42,000 real sentences that have been manually translated and validated. Additionally, it includes 466,000 high-quality synthetic sentence pairs, carefully generated using Claude Sonnet 3.5 (Anthropic, 2024). These synthetic examples were created based on diverse dictionary entries and carefully curated examples to enhance the semantic and lexical diversity of the corpus.

5.2 Pipeline ST

5.2.1 ASR systems

We submitted two ASR systems for the North Levantine Arabic to English track, employing the Whisper-large-v3 in an encoder-decoder configuration, trained on a dataset that combines dialectal and Modern Standard Arabic (MSA) transcribed speech. The first system, **contrastive 1**, augmented the Levantine dataset with an equal number of MSA utterances, while the second system, **primary**, further fine-tuned **contrastive 1** solely on the Levantine datasets (LDC2005S08 and LDC2006T07) to specialize the model for Levantine dialects.

5.2.2 MT model

We trained two machine translation (MT) models using the HuggingFace framework, both based on the NLLB-200 1.3B model. The first MT model was fine-tuned on the entire Levanti dataset using a learning rate of 1×10^{-5} and a batch size of 8. The second MT model was fine-tuned on the UFAL Parallel Corpus and the non-synthetic portion of the Levanti dataset, using the same learning rate 1×10^{-5} and a batch size of 6, with a beam size of 5 for decoding.

Following extensive experimentation on the development and test sets, we selected the second MT model (trained on the UFAL Corpus and non-synthetic Levantine data) for the final Levantine-to-English translation task, as it outperformed the first model in terms of translation quality.

We then constructed two cascaded systems (**contrastive 1** and **contrastive 2**) using the trained ASR systems.

[2] https://huggingface.co/datasets/guymorlan/levanti

5.3 ST candidates selection

To evaluate and rank outputs from different ASR-MT combinations, we used the BLASER-REF quality estimation model (Dale and Costa-jussà, 2024; Seamless Communication et al., 2023). BLASER-REF is a reference-based model that estimates translation quality using SONAR embeddings (Duquenne et al., 2023), which map both speech and text from different languages into a shared latent space, making the model inherently language and modality-agnostic.

The model takes three inputs: the original speech signal, a system-generated translation, and a reference translation. As human reference translations were unavailable, we used the transcription as the reference. The speech input was encoded using the SONAR Arabic speech encoder, which was trained on Modern Standard Arabic (MSA); we applied it to Levantine speech due to the lack of a Levantine-specific encoder. The transcription and translation were encoded using the SONAR text encoder, which supports the source (North Levantine Arabic) and target (English) languages.

For each utterance, we generated 10 candidate outputs from five ASR and two MT models—the same systems described in Sections 5.2.1 and 5.2.2 with additional variants—and selected the output with the highest BLASER-REF score on a scale from 1 to 5, where higher scores indicate better quality; this combination is considered our **primary** system for ST.

5.4 Results

5.4.1 Results for ASR

The ASR results in terms of word error rate (WER) are reported in Table 7 on the development set 2024 (Dev), test set 2024 (Test), and test set 2025 (Test 2025) provided by the organisers. The **Primary** system outperformed the **Contrastive 1** system on both Dev and Test sets.

Table 7: WER (%) results for North Levantine dialect ASR.

	Dev	Test
Primary	38.43	41.06
Contrastive 1	38.92	42.86

5.4.2 Results for ST

The ST results in terms of BLEU score (BLEU) are reported in Table 8 on the development set 2024

(Dev), test set 2024 (Test), and test set 2025 (Test 2025) provided by the organisers.

Table 8: BLEU results for North Levantine dialect to English translation.

	Dev	Test	Test 2025
Primary	29.64	28.02	22.56
Contrastive 1	28.74	26.87	21.02
Contrastive 2	28.88	26.61	21.45

6 Conclusion

This paper describes the translation systems developed by LIA and ELYADATA for three tracks of the IWSLT 2025 Evaluation Campaign, focusing on low-resource speech translation. The targeted language pairs are Tunisian Arabic–English, North Levantine Arabic–English, and Fongbé–French.

Acknowledgments

This work was funded by the French Research Agency (ANR) through the TRADEF project. It used HPC resources from GENCI-IDRIS: grants AD011012551R3, AD011015051R1, AD011012108R3, AD011014814R1, and AD011015509.

References

Idris Abdulmumin, Victor Agostinelli, Tanel AlumÃ€e, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, OndÅej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick EstÃšve, Marcello Federico, Marco Gaido, DÃ¡vid JavorskÃœ, Marek Kasztelnik, and 30 others. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austia (in-person and online). Association for Computational Linguistics. To appear.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1:1.

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

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12449–12460.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228. IEEE.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Yu-An Chung, Wei-Ning Hsu, Hank Liao Tang, and James Glass. 2021. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2445–2449.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

David Dale and Marta Costa-jussà. 2024. Blaser 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16075–16085.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 949–959.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Haroun Elleuch, Salima Mdhaffar, Yannick Estève, and Fethi Bougares. 2025. ADI-20: Arabic Dialect Identification dataset and models. In *Proceedings of Interspeech 2025*, Rotterdam, The Netherlands. Submitted to Interspeech 2025.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022a. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022b. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Nizar Habash, Mona Diab, and Owen Rambow. 2012. Conventional orthography for dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 711–718, Istanbul, Turkey. European Language Resources Association (ELRA).

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. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1474–1478.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1493–1504.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Gaurav Kumar, Graeme Blackwood, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur. 2015. A coarse-grained model for optimal coupling of asr and smt systems for speech translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1902–1907.

Antoine Laurent, Souhir Gahbiche, Ha Nguyen, Haroun Elleuch, Fethi Bougares, Antoine Thiol, Hugo Riguidel, Salima Mdhaffar, Gaëlle Laperrière, Lucas Maison, and 1 others. 2023. On-trac consortium systems for the iwslt 2023 dialectal and low-resource speech translation tasks. In *International Conference on Spoken Language Translation (IWSLT) 2023*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Evgeny Matusov, Stephan Kanthak, and Hermann Ney. 2005. On the integration of speech recognition and statistical machine translation. In *Interspeech*, pages 3177–3180.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. Qasr: Qcri aljazeera speech resource a large scale annotated arabic speech corpus. 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 2274–2285.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023a. 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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023b. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, and 1 others. 2024. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333):1–11.

Seamless Communication, Barrault, Loïc, Chung, Yu-An, Meglioli, Mariano Cora, Dale, David, Dong, Ning, Duquenne, Paul-Ambroise, Elsahar, Hady, Gong, Hongyu, Heffernan, Kevin, Hoffman, John, Klaiber, Christopher, Li, Pengwei, Licht, Daniel, Maillard, Jean, Rakotoarison, Alice, Sadagopan, Kaushik Ram, Wenzek, Guillaume, Ye, Ethan, and 49 others. 2023. Seamlessm4t—massively multilingual & multimodal machine translation. *ArXiv*.