SeamlessM4T—Massively Multilingual & Multimodal Machine Translation

Seamless Communication, Loïc Barrault*, Yu-An Chung*, Mariano Cora Meglioli*, David Dale*, Ning Dong*, Paul-Ambroise Duquenne*, Hady Elsahar*, Hongyu Gong*, Kevin Heffernan*, John Hoffman*, Christopher Klaiber*, Pengwei Li*, Daniel Licht*, Jean Maillard*, Alice Rakotoarison*, Kaushik Ram Sadagopan*, Guillaume Wenzek*, Ethan Ye*, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews†, Can Balioglu†, Marta R. Costa-jussà†‡, Onur Celebi†, Maha Elbayad†, Cynthia Gao†, Francisco Guzmán†, Justine Kao†, Ann Lee†, Alexandre Mourachko†, Juan Pino†, Sravya Popuri†, Christophe Ropers†, Safiyyah Saleem†, Holger Schwenk†, Paden Tomasello†, Changhan Wang†, Jeff Wang†, Skyler Wang†,§

Meta AI, §UC Berkeley

Abstract

What does it take to create the Babel Fish, a tool that can help individuals translate speech between any two languages? While recent breakthroughs in text-based models have pushed machine translation coverage beyond 200 languages, unified speech-to-speech translation models have yet to achieve similar strides. More specifically, conventional speech-to-speech translation systems rely on cascaded systems composed of multiple subsystems performing translation progressively, putting scalable and high-performing unified speech translation systems out of reach. To address these gaps, we introduce **SeamlessM4T**—**M**assively **M**ultilingual & **M**ultimodal **M**achine **T**ranslation—a single model that supports speech-to-speech translation, speech-to-text translation, text-to-speech translation, text-to-text translation, and automatic speech recognition for up to 100 languages. To build this, we used 1 million hours of open speech audio data to learn self-supervised speech representations with w2v-BERT 2.0. Subsequently, we created a multimodal corpus of automatically aligned speech translations, dubbed SeamlessAlign. Filtered and combined with human-labeled and pseudo-labeled data (totaling 406,000 hours), we developed the first multilingual system capable of translating from and into English for both speech and text. On Fleurs, SeamlessM4T sets a new standard for translations into multiple target languages, achieving an improvement of 20% BLEU over the previous state-of-the-art in direct speech-to-text translation. Compared to strong cascaded models, SeamlessM4T improves the quality of into-English translation by 1.3 BLEU points in speech-to-text and by 2.6 ASR-BLEU points in speech-to-speech. On CVSS and compared to a 2-stage cascaded model for speech-to-speech translation, SeamlessM4T-Large's performance is stronger by 58%. Preliminary

∗. Equal contribution, alphabetical order

†. Research and engineering leadership, equal contribution, alphabetical order

‡. Corresponding Author. Email: costajussa@meta.com.

human evaluations of speech-to-text translation outputs evinced similarly impressive results; for translations from English, XSTS scores for 24 evaluated languages are consistently above 4 (out of 5). For into English directions, we see significant improvement over Whisper-Large-v2's baseline for 7 out of 24 languages. To further evaluate our system, we developed Blaser 2.0, which enables evaluation across speech and text with similar accuracy compared to its predecessor when it comes to quality estimation. Tested for robustness, our system performs better against background noises and speaker variations in speech-to-text tasks (average improvements of 38% and 49%, respectively) compared to the current state-of-the-art model. Critically, we evaluated SeamlessM4T on gender bias and added toxicity to assess translation safety. Compared to the state-of-the-art, we report up to 63% of reduction in added toxicity in our translation outputs. Finally, all contributions in this work—including models, inference code, finetuning recipes backed by our improved modeling toolkit Fairseq2, and metadata to recreate the unfiltered 470,000 hours of SeamlessAlign —are open-sourced and accessible at `https://github.com/facebookresearch/seamless_communication`.

Contents

1. Introduction

The Hitchhiker's Guide to the Galaxy's Babel Fish, *Star Trek's* Universal Translator, and *Doctor Who's* Tardis Translation Circuit are all variants of the same thing—computational devices that grant the ability to translate between any two languages. Casting aside their chimeric origins, the social need for realizing such visions has never been greater. For one, an increasingly interconnected world calls for the development of technologies that can facilitate and streamline multilingual contact both online and offline. Moreover, the proliferation of mobile devices and the platform economy worldwide provides the vehicle for on-demand speech-to-speech translation (S2ST) to become a staple in most people's lives.

Despite the centrality of speech in everyday communication, machine translation (MT) systems today remain text-centric. Speech support, if and when present, is often seen as cursory to its text-based counterpart. While single, unimodal models such as No Language Left Behind (NLLB; [NLLB Team et al., 2022]) push text-to-text translation (T2TT) coverage to more than 200 languages, unified S2ST models are far from achieving similar scope or performance. This modality-based disparity could be attributed to many causes, but audio data scarcity and modeling constraints remain key obstacles. The very challenge around why speech is harder to tackle from an MT standpoint—that it encodes more information and expressive components—is also why it is superior at conveying intent and forging stronger social bonds between interlocutors.

Bringing the Babel Fish into technical reality hinges on developing foundational speech-to-speech translation (S2ST) systems. Today, existing systems of such kind suffer from three main shortcomings. One, they tend to focus on high-resource languages such as English, Spanish, and French, leaving many low-resource languages behind. Two, they mostly service translations from a source language into English (X–eng) and not vice versa (eng–X). Three, most S2ST systems today rely heavily on cascaded systems composed of multiple subsystems that perform translation progressively—e.g., from automatic speech recognition (ASR) to T2TT, and subsequently text-to-speech (TTS) synthesis in a 3-stage system. Attempts to unify these multiple capabilities under one singular entity have led to early iterations of end-to-end speech translation systems [Lavie et al., 1997; Jia et al., 2019b; Lee et al., 2022a]. However, these systems do not match the performance of their cascaded counterparts [Agarwal et al., 2023], which are more equipped to leverage large-scale multilingual components (e.g., NLLB for T2TT or Whisper for ASR [Radford et al., 2022]) and unsupervised or weakly-supervised data.

To address these limitations, we introduce **SeamlessM4T** (**M**assively **M**ultilingual & **M**ultimodal **M**achine **T**ranslation), a unified system that supports ASR, T2TT, speech-to-text translation (S2TT), text-to-speech translation (T2ST), and S2ST (see Table 1 for an overview). To build this, we used 1 million hours of open speech audio data to learn self-supervised speech representations with w2v-BERT 2.0. Subsequently, we created a multimodal corpus of automatically aligned speech translations of more than 470,000 hours, dubbed SeamlessAlign. We then combined a filtered subset of this corpus with human-labeled and pseudo-labeled data, totaling 406,000 hours. Drawing on this assembled dataset, we developed the first multitasking system that performs S2ST from 100 languages to English (100-eng) and from English to 35 languages (eng-35), S2TT for 100-eng and eng-95 languages, ASR for 96, zero-shot T2ST for 95-eng and eng-35 languages, as well as T2TT for 95-eng and eng-95 (see Table 2 for an overview).

Task	Description
ASR	Automatic Speech Recognition
S2ST	Speech-to-Speech Translation
S2TT	Speech-to-Text Translation
T2ST	Text-to-Speech Translation
T2TT	Text-to-Text Translation
X2T	{Speech,Text}-to-Text Translation (multitasking models translating into text)
Task eng–X	A translation task from English
Task X–eng	A translation task into English
Task X–X	A translation task on non-English-centric direction

Table 1: Notations of tasks in this work.

We find that SeamlessM4T-Large, the larger model of the two we release, outperforms the previous state-of-the-art (SOTA) end-to-end S2TT model (AudioPaLM-2-8B-AST [Rubenstein et al., 2023]) by 4.2 BLEU points on Fleurs [Conneau et al., 2022] when translating into English (i.e., an improvement of 20%). Compared to cascaded models, SeamlessM4T-Large improves translation accuracy by over 2 BLEU points. When translating from English, SeamlessM4T-Large improves on the previous SOTA (XLS-R-2B-S2T [Babu et al., 2022]) by 2.8 BLEU points on CoVoST 2 [Wang et al., 2021c], and its performance is on par with cascaded systems on Fleurs. On the S2ST task, SeamlessM4T-Large outperforms strong 3-stage cascaded models (ASR, T2TT and TTS) by 2.6 ASR-BLEU points on Fleurs. On CVSS, SeamlessM4T-Large outperforms a 2-stage cascaded model (Whisper-Large-v2 + YourTTS [Casanova et al., 2022]) by a large margin of 8.5 ASR-BLEU points (a 50% improvement). Preliminary human evaluations of S2TT outputs evinced similarly impressive results. For translations from English, XSTS scores for 24 evaluated languages are consistently above 4 (out of 5); for into English directions, we see significant improvement over Whisper-Large-v2's baseline for 7 out of 24 languages.

In addition, SeamlessM4T-Large further outperforms Whisper-Large-v2 [Radford et al., 2022] on Fleurs ASR with an average word error rate (WER) reduction of 45% over 77 overlapping languages. When evaluating T2TT on Flores [Goyal et al., 2022], our model matches the performance of NLLB-3.3B [NLLB Team et al., 2022] when translating into English and improves by 1 chrF++ point on average when translating from English. To further evaluate SeamlessM4T's performance in S2TT and S2ST, we developed Blaser 2.0, a language and modality-agnostic evaluation metric for text or speech translation. Blaser 2.0 enables evaluation across speech and text modalities with similar accuracy to its predecessor —BLASER [Chen et al., 2023a]—when it comes to quality estimation. We also evaluated model robustness against background noises and speaker variations by creating open robustness benchmarks based on Fleurs. Result-wise, SeamlessM4T-Large is more robust than Whisper-Large-v2 against background noises and speaker variations with an average improvement of 38% and 49%, respectively.

Model	**size**	**Task Language Coverage**†				
		S2TT	S2ST	ASR	T2TT	T2ST
Proprietary models						
USM [Zhang et al., 2023a]	2B+	21-eng	-	102	-	-
Rubenstein et al. [2023]						
AudioPaLM-2-8B-AST	8.0B	98-eng	-	98	-	-
AudioPaLM-8B-S2ST	8.0B	113-Eng	113-eng	98	-	-
Open models						
NLLB Team et al. [2022]						
NLLB-600M-Distilled	0.6B	-	-	-	202-202	-
NLLB-1.3B	1.3B	-	-	-	202-202	-
NLLB-3.3B	3.3B	-	-	-	202-202	-
Babu et al. [2022]						
XLS-R-2B-S2T	2.6B	21-eng eng-15	-	-	-	
Radford et al. [2022]						
Whisper-Medium	0.8B	96-eng	-	97	-	-
Whisper-Large-v2	1.6B	96-eng	-	97	-	-
MMS [Pratap et al., 2023]						
MMS-L61-noLM-LSAH	1.0B	-	-	61	-	-
MMS-L1107-CCLM-LSAH	1.0B	-	-	1107	-	-
This work (SeamlessM4T)						
SeamlessM4T-Large	2.3B	100-eng eng-95	100-eng eng-35	96	95-eng eng-95	95-eng eng-35
SeamlessM4T-Medium	1.2B	100-eng eng-95	100-eng eng-35	96	95-eng eng-95	95-eng eng-35
SeamlessM4T-NLLB-1.3B	1.3B	-	-	-	95-eng eng-95	-

Table 2: A list of state-of-the-art baseline models and SeamlessM4T models. †Language coverage is estimated based on use of supervised labeled data or evaluated zero-shot languages and directions.

Regarding Responsible AI, we focused on added toxicity and gender bias evaluation. On average, we find a low prevalence of added toxicity, varying between 0.11% and 0.21% across modalities, datasets, and translation directions. We significantly reduce added toxicity in all conditions when compared to state-of-the-art models (ranging from 26% to 63%). The greatest added toxicity reduction is achieved for S2TT when compared to Whisper-Large-v2. Beyond this, we also evaluated for gender bias on the Multilingual HolisticBias datasets and found that SeamlessM4T overgeneralizes to masculine forms when translating from neutral terms (with an average preference of ∼10%) while showing a lack of robustness when varying gender by an amount of ∼3%. For these conditions, SeamlessM4T achieved comparable results to state-of-the-art models. We document these effects to motivate further mitigation efforts.

To spur further research in speech translation and to make our work available to the community, we open-source the following at https://github.com/facebookresearch/seamless_communication:

- SeamlessM4T models, including model weights for SeamlessM4T-Large (2.3B parameters) and SeamlessM4T-Medium (1.2B parameters), as well as their inference code and fine-tuning recipes powered by our new modeling toolkit Fairseq2.[1]

- Tools for creating aligned speech data, including metadata to recreate the unfiltered 470,000 hours of SeamlessAlign, Stopes-based pipelines[2] to create alignments similar to SeamlessAlign, and Sonar for speech encoders in 37 languages and text encoders in 200 languages.[3]

- A text-free S2ST automatic evaluation model, Blaser 2.0, inclusive of model weights and inference scripts.

The rest of the article is structured as follows: Section 2 describes the sociotechnical dimensions of multimodal translation and motivates why speech is an important modality to tackle in the context of MT research. It also includes the list of languages and evaluation metrics that our work covers. Section 3 discusses how we created a corpus of automatically aligned speech translations of more than 470,000 hours by developing an extended speech-language identification system and a new multimodal text embedding space imperative to our data mining process. Section 4 details the various modeling techniques we devised to train a multimodal and multitasking translation model that supports multiple languages for source and target sides in both text and speech. Section 5 documents the automatic and human evaluation of our translation outputs, and the robustness of our models in various settings. Section 6 focuses on our Responsible AI effort, where we evaluated our model outputs for bias and toxicity. Finally, we conclude in Section 7, where we discuss the social impact of our work while reflecting on existing challenges and future possibilities.

2. The Sociotechnical Dimensions of Multimodal Translation

2.1 Why Prioritize Speech in Machine Translation?

As is the case with most technologies within natural language processing (NLP) and other language-based research enterprises, MT reached greater maturity in the modality that affords easier record-keeping, data storage, and dispersion: text. By extension, the abundance of digital text makes it a prime candidate for NLP research. In contrast, the relative paucity of speech data relegates research in this area to secondary importance. More specifically, speech is not just spoken text—the two modalities can differ in grammar, registers, and morphology [Plag et al., 1999]. In most situations, speech may also appear to be a richer modality, possessing prosodic and expressive parameters unmatchable by text [Kraut et al., 1992]. Distinctive in their level of interactivity and sociality, speech directs focus at the speaker or audience, while text spotlights the content of a message [Kraut et al., 1992].

1. https://github.com/facebookresearch/fairseq2
2. https://github.com/facebookresearch/stopes
3. https://github.com/facebookresearch/SONAR

Speech & social bonding Research suggests that compared to text-based exchange, communication through speech creates stronger social bonds between interlocutors. For example, in one study, researchers found that interactions including speech (phone, video call, and voice chat) spurred deeper connections between conversation partners compared to those who communicated via text-based media [Kumar and Epley, 2021, 595]. Juxtaposed against speech, which comes with paralinguistic cues such as volume, intonation, and pace, text-based communication is perceived as more impersonal. Interestingly, seeing another person did not make individuals feel more connected than if they had just spoken with their partners. In another study, hearing an outgroup member explain their views out loud made study participants consider them more thoughtful and emotionally warm than reading an explanation of their views [Schroeder et al., 2017]. Across a variety of settings, research demonstrates that speech appears to be unique in its ability to convey one's human traits and, consequentially, strengthen the connection between those sharing an exchange.

Inclusion & accessibility Speech is not only key to communication from a relational standpoint but is also often the most practical and accessible option. For one, UNESCO estimates that 773 million adults (12.5 percent of all adults) worldwide have not received the education necessary to read or write, thus precluding them from using text to communicate or acquire information [Markelova, 2021]. Another group more reliant on speech than text in their everyday lives is those who are blind or with visual impairments. Globally, approximately 43 million people belong to this former category, and 295 million others have moderate to severe visual impairment [GBD 2019 Blindness and Vision Impairment Collaborators, 2021]. Even though voice assistants, text-to-speech systems, and voice-activated technologies today play an important role in supporting these individuals to accomplish everyday tasks, their access to multilingual speech-based translation or communicative tools remains limited. In a world where the volume of auditory content (i.e., podcasts, audiobooks, short-form videos, etc.) is on the rise, the prohibitive nature of this sociotechnical gap may deprive them of experiences or exchanges that could be meaningful and enriching.

Script variance Beyond these factors, text-based communication or translation is further complicated by script variance. For instance, some languages are written in different scripts on either side of a geopolitical border. Urdu, for example, could be written either in the Arabic or Devanagari script depending on where one lives (i.e., Pakistan or India). In such a context, T2TT outputs into Urdu may be illegible to those shown in a script they are unfamiliar with. S2ST, which produces speech outputs, circumvents this multiscript conundrum. In a few other cases, political instabilities around a language's writing system may also motivate the need for speech-based translation. For example, in the last 1,000 years, Uzbek has changed its writing system five times. Despite the fact that—as of February 2021—Uzbekistan announced Uzbek's official transition from the Cyrillic script to a Latin-based alphabet, the former continues to be widely deployed in the country [Jung and Kim, 2023]. For languages where writing systems are actively negotiated, speech-based technologies and translation systems may provide stabilized access to information as transitions unfold.

Cascaded models for S2TT	
WHISPER-MEDIUM + NLLB-600M-DISTILLED	2-stage cascaded
WHISPER-LARGE-V2 + NLLB-1.3B	2-stage cascaded
Cascaded models for S2ST	
WHISPER-LARGE-V2 + NLLB-1.3B + YOURTTS	3-stage cascaded
WHISPER-LARGE-V2 (S2TT) + YOURTTS	2-stage cascaded
SEAMLESSM4T (this work)	unified

Table 3: Options for 2-stage and 3-stage cascaded systems for S2TT and S2ST. These cascades pair Whisper ASR models [Radford et al., 2022] with NLLB's T2TT models [NLLB Team et al., 2022].

2.2 Speech Translation Today

Cascaded systems Before the emergence of unified speech translation models in recent years, much attention in speech-based research has been directed at cascaded approaches by chaining subsystems that perform disparate tasks such as ASR, T2TT, and TTS [Lavie et al., 1997; Wahlster, 2000; Nakamura et al., 2006]. For example, in a 3-stage S2ST cascaded scenario, speech input is first transcribed into text through an ASR system, followed by T2TT, and finally synthesized into speech using TTS (see Table 3). The main benefit of cascaded systems is that they can take advantage of advancements made in areas associated with each subsystem, such as recently released large-scale multilingual T2TT models [NLLB Team et al., 2022; Siddhant et al., 2022; Fan et al., 2020] and weakly-supervised ASR models [Radford et al., 2022; Zhang et al., 2023a; Pratap et al., 2023].

That said, cascaded systems have their limitations. For one, the output of a 2-stage cascaded S2TT system involving ASR and T2TT does not match the quality achievable by a single large-scale T2TT model. This drop in performance underscores the challenge of transferring and translating meaning across modalities and can be attributed to many factors, including: (1) poor transcriptions by ASR models for non-English languages, particularly for low-resourced ones, (2) an increased likelihood of error propagation from the ASR model to the T2TT model and other subsequent models in the cascade (the accumulation of errors exacerbates performance), and (3) domain mismatches between these separately trained subsystems (for example, if an ASR model trained on Wikipedia is used in conjunction with a T2TT model optimized for conversational data, this formation may lead to a distribution mismatch at the T2TT stage). Beyond these reasons, the overemphasis on text in cascaded systems omits paralinguistic features and may not adequately handle elements such as proper names and nouns [Rubenstein et al., 2023].

Direct S2TT models Early research into end-to-end speech translation started with producing text as output [Chan et al., 2016; Berard et al., 2016; Bérard et al., 2018]. Since the emergence of multilingual end-to-end S2TT models in 2019 [Gangi et al., 2019; Inaguma et al., 2019], S2TT has become an increasingly popular research area, and many existing models today are powered by the emergence of open multilingual speech corpora like MuST-C [Di Gangi et al., 2019], EuroParl-ST [Iranzo-Sánchez et al., 2020], CoVoST 2 [Wang et al., 2021c] and VoxPopuli [Wang et al., 2021b]. End-to-end models today have made significant progress and achieved parity with cascaded models on academic benchmarks in several contexts (e.g., constrained data, in-domain settings, specific language pairs, etc.) [Ansari et al., 2020; Potapczyk and Przybysz, 2020b]

While recent state-of-the-art pre-trained models have seen rapid improvements in language coverage, going from 128 in Babu et al. [2022] to more than 1,400 in Pratap et al. [2023], they only translate into English and not the other way around. Another prominent model, Google's Universal Speech Model [Zhang et al., 2023a], is pre-trained in more than 300 languages and can perform ASR on more than 100 languages. Technically, USM can also be adapted to perform ASR and S2TT tasks in any of the 300+ covered languages once given supervised data (but the model was fine-tuned and evaluated on CoVoST 2, which only covers translations from 21 languages into English).

OpenAI's Whisper [Radford et al., 2022] is another large-scale model that serves translations into English, not vice versa. As a multitasking model, Whisper demonstrates that scaling weakly supervised pre-training is sufficient for achieving SOTA ASR and S2TT results sans self-supervision and self-training techniques. Trained on 680,000 hours of data, Whisper has achieved SOTA translation quality in 82 Fleurs languages into English.

Combining a text-based [Anil et al., 2023] and speech-based language model [Borsos et al., 2023], the most recently released AudioPaLM [Rubenstein et al., 2023] is a large language model designed for joint text and speech processing and generation. Akin to USM, AudioPaLM only evaluates text translation outputs from 101 Fleurs languages into English. Upon the publication of this paper, AudioPaLM is the current SOTA model, outperforming Whisper [Radford et al., 2022] in both ASR and S2TT tasks.

Direct S2ST models Beyond text outputs, recent speech translation research has focused on building models that directly produce target speech representations (i.e., spectrograms, discrete units, etc.). In this area, Translatotron [Jia et al., 2019b] emerged as the first direct S2ST model. When it comes to quality, however, the model lagged behind 2-stage cascaded systems by a large margin. Translatotron-2 [Jia et al., 2022a] significantly improved its predecessor's performance and bridged the gap with cascaded systems by incorporating a two-pass decoding approach. Although Translatotron relied on S2TT as an auxiliary task during training, the target spectrograms were directly generated at inference time. Translatotron-2, on the other hand, relies on the intermediate decoding outputs of phonemes.

Concurrently with Translatotron, Tjandra et al. [2019] proposed S2ST models based on discrete speech representations that do not require text transcriptions in training. These discrete representations or *units* are learned through unsupervised term discovery and a sequence-to-sequence model trained to translate units from one language to another. Relatedly, Lee et al. [2022a] uses HuBERT [Hsu et al., 2021], a pre-trained speech representation model, to encode speech and learn target-side discrete units. S2ST is, thus, decomposed into speech-to-unit (S2U) and subsequently unit-to-speech with a speech re-synthesizer [Polyak et al., 2021].

On coverage and evaluation of S2ST systems To date, the aforementioned AudioPaLM [Rubenstein et al., 2023], which supports both text and speech as input and output, is the current SOTA for S2TT and S2ST. Although the model design suggests that it can support multilingual translation on both source and target sides, its performance is only reported for translating into English. Similarly, although Whisper can transcribe non-English languages, it only supports S2TT into English. To consolidate the current landscape of language coverage and related tasks in speech translation systems, we provide in Table 2 a list of SOTA models in text and speech translation. This language coverage is estimated based on supervised labeled data or evaluated zero-shot languages and directions. We also provide the list of ASR, T2TT, S2TT and S2ST evaluation metrics used by this work in Table 4. For S2ST, our evaluation focuses on the semantic content of the translation. Throughout this paper, we primarily evaluated our models on the following datasets:

- Flores-200 [NLLB Team et al., 2022]: a many-to-many multilingual translation benchmark dataset for 200 languages (we evaluated on devtest).
- Fleurs [Conneau et al., 2022]: an n-way parallel speech and text dataset in 102 languages built on the machine translation Flores-101 benchmark [Goyal et al., 2022]. Fleurs is well suited for several downstream tasks involving speech and text. We evaluate on the test set, except in ablation experiments where we evaluate on the dev set.
- CoVoST 2 [Wang et al., 2021c]: a large-scale multilingual S2TT corpus covering translations from 21 languages into English and from English into 15 languages. We evaluate on the test set.
- CVSS [Jia et al., 2022b]: a multilingual-to-English speech-to-speech translation (S2ST) corpus, covering sentence-level parallel S2ST pairs from 21 languages into English. We evaluate text-based semantic accuracy on CVSS-C for the tasks of S2ST and T2ST. We note that some samples from the evaluation data were missing (in 8 out of 21 languages: Catalan, German, Estonian, French, Italian, Mongolian, Persian and Portuguese).

The overarching goals of this effort In light of the gaps delineated above, our work seeks to advance speech translation in the following ways:

1. Creating a unified large model that can handle the full suite of tasks involved in text and speech translation: S2ST, S2TT, T2ST, T2TT, and ASR. This lays the important groundwork for the next generation of on-device and on-demand multimodal translation, which can be derived from this model.
2. Expanding language coverage both in terms of the number of supported languages and translation directions (i.e., going beyond translations into English by including translation from English). That roughly two dozen languages account for more than half of the world's speaking population means that a relatively small group of languages (out of more than 7,000) produce a disproportionately large linguistic footprint. Whether in the text or speech modality, these languages are deemed high-resource, giving them prioritization in today's AI development. That said, when language technologies are developed primarily with this group in mind, the needs of half the world's population are left behind. Our effort seeks to bridge the translation gap between those who speak high and low-resource languages.
3. Maintaining systematic evaluations of our systems throughout our workflow to ensure safe and robust performance. This allows us to understand how to direct our efforts to make both the current and future iterations of our contribution more equitable and fair across user demographics.

2.3 Languages

Today, broadly accessible speech translation models cover anywhere between 21 [Zhang et al., 2023a] to 113 [Rubenstein et al., 2023] source languages depending on the wide range of tasks involved. However, none of these existing speech-based translation models can also service T2TT. To build a unified, multimodal, and multitask model that can handle both speech and text as source inputs, we set our speech source language goal at 100.

We summarize information about each of our supported languages in Table 5. Further details on the table headers are provided below.

Code We represent each language with a three-letter ISO 639-3 code.

Language There may be multiple ways to refer to the same language; due to formatting limitations, only one of the versions is displayed. The language names have been cross-referenced with major linguistic information platforms such as Ethnologue [Lewis, 2009] and Glottolog [Hammarström et al., 2022].

Family and Subgrouping We provide Language family information for each language based on the Glottolog database [Hammarström et al., 2022].

Script We provide script information in ISO 15924 codes for writing systems.

Resource level We categorize the speech resource level as high, medium, or low depending on the volume of available primary data for S2TT into English (with x the amount of primary data in hours, *high* if $x > 1000$, *medium* if $x \in]500, 1000]$ and *low* if $x \in [0, 500]$).

Primary data is defined as open-source S2TT and pseudo-labeled ASR data. Absent such data, we report the language as zero-shot (when evaluating S2TT into English).

Source. We indicate whether a source language is in the speech (Sp) or text (Tx) modality, or both.

Target. We indicate whether a target language is in the speech (Sp) or text (Tx) modality, or both.

Task	Metric	Type	Area	Details
ASR	WER		Quality Robustness	Text normalization follows Whisper*
T2TT	chrF++[†]	Automatic	Quality	SacreBLEU signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.3.1
	BLEU[‡]	Automatic	Quality	SacreBLEU signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1 Except for cmn, jpn, tha, lao and mya with character-level tokenization: nrefs:1|case:mixed|eff:no|tok:char|smooth:exp|version:2.3.1
	Blaser 2.0	Automatic Model-based	Quality	
S2TT	BLEU	Automatic	Quality Robustness Bias	Similar to T2TT
	Blaser 2.0	Automatic Model-based	Quality	Chen et al. [2023a]
	XSTS	Human	Quality	Licht et al. [2022]
	chrF$_{MS}$	Automatic	Robustness Bias	following Wang et al. [2020], replaced BLEU with chrF for the quality metric SacreBLEU signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.3.1
	CoefVar$_{MS}$	Automatic	Robustness	following Wang et al. [2020], replaced BLEU with chrF for the quality metric SacreBLEU signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.3.1
	ETOX	Automatic	Toxicity	
S2ST	ASR-BLEU	Automatic	Quality	Transcribing English with Whisper-Medium and non-English with Whisper-Large-v2 BLEU on normalized transcriptions following Radford et al. [2022]
	ASR-chrF	Automatic	Bias	Transcribing English with Whisper-Medium and non-English with Whisper-Large-v2 chrF on normalized transcriptions following Radford et al. [2022]
	Blaser 2.0	Automatic Model-based	Quality Bias	
	XSTS	Human	Quality	
	MOS	Human	Naturalness	
	ASR-ETOX	Automatic	Toxicity	Transcribing English with Whisper-Medium and non-English with Whisper-Large-v2 ETOX on normalized transcriptions following Radford et al. [2022]
T2ST	ASR-BLEU	Automatic	Quality	Similar to S2ST

Table 4: The list of automatic and human evaluation metrics used by this work. * `https://github.com/openai/whisper/tree/main/whisper/normalizers` † Popović [2015] ‡ Papineni et al. [2002]

Code	Language name	Family	Subgrouping	Script	Resource	Source	Target
afr	Afrikaans	Indo-European	Germanic	Latn	low	Sp, Tx	Tx
amh	Amharic	Afro-Asiatic	Semitic	Ethi	low	Sp, Tx	Tx
arb	Modern Standard Arabic	Afro-Asiatic	Semitic	Arab	high	Sp, Tx	Sp, Tx
ary	Moroccan Arabic	Afro-Asiatic	Semitic	Arab	low	Sp, Tx	Tx
arz	Egyptian Arabic	Afro-Asiatic	Semitic	Arab	low	Sp, Tx	Tx
asm	Assamese	Indo-European	Indo-Aryan	Beng	low	Sp, Tx	Tx
ast	Asturian	Indo-European	Italic	Latn	zero-shot	Sp	–
azj	North Azerbaijani	Turkic	Common Turkic	Latn	low	Sp, Tx	Tx
bel	Belarusian	Indo-European	Balto-Slavic	Cyrl	high	Sp, Tx	Tx
ben	Bengali	Indo-European	Indo-Aryan	Beng	high	Sp, Tx	Sp, Tx
bos	Bosnian	Indo-European	Balto-Slavic	Latn	low	Sp, Tx	Tx
bul	Bulgarian	Indo-European	Balto-Slavic	Cyrl	low	Sp, Tx	Tx
cat	Catalan	Indo-European	Italic	Latn	high	Sp, Tx	Sp, Tx
ceb	Cebuano	Austronesian	Malayo-Polynesian	Latn	zero-shot	Sp, Tx	Tx
ces	Czech	Indo-European	Balto-Slavic	Latn	high	Sp, Tx	Sp, Tx
ckb	Central Kurdish	Indo-European	Iranian	Arab	low	Sp, Tx	Tx
cmn	Mandarin Chinese	Sino-Tibetan	Sinitic	Hans, Hant	high	Sp, Tx	Sp, Tx
cym	Welsh	Indo-European	Celtic	Latn	medium	Sp, Tx	Sp, Tx
dan	Danish	Indo-European	Germanic	Latn	medium	Sp, Tx	Sp, Tx
deu	German	Indo-European	Germanic	Latn	high	Sp, Tx	Sp, Tx
ell	Greek	Indo-European	Graeco-Phrygian	Grek	medium	Sp, Tx	Tx
eng	English	Indo-European	Germanic	Latn	high	Sp, Tx	Sp, Tx
est	Estonian	Uralic	Finnic	Latn	medium	Sp, Tx	Sp, Tx
eus	Basque	Basque	Basque	Latn	medium	Sp, Tx	Tx
fin	Finnish	Uralic	Finnic	Latn	high	Sp, Tx	Sp, Tx
fra	French	Indo-European	Italic	Latn	high	Sp, Tx	Sp, Tx
gaz	West Central Oromo	Afro-Asiatic	Cushitic	Latn	zero-shot	Sp, Tx	Tx
gle	Irish	Indo-European	Celtic	Latn	low	Sp, Tx	Tx
glg	Galician	Indo-European	Italic	Latn	low	Sp, Tx	Tx
guj	Gujarati	Indo-European	Indo-Aryan	Gujr	low	Sp, Tx	Tx
heb	Hebrew	Afro-Asiatic	Semitic	Hebr	low	Sp, Tx	Tx
hin	Hindi	Indo-European	Indo-Aryan	Deva	medium	Sp, Tx	Sp, Tx
hrv	Croatian	Indo-European	Balto-Slavic	Latn	medium	Sp, Tx	Tx
hun	Hungarian	Uralic	Hungarian	Latn	medium	Sp, Tx	Tx
hye	Armenian	Indo-European	Armenic	Armn	low	Sp, Tx	Tx
ibo	Igbo	Atlantic-Congo	Benue-Congo	Latn	low	Sp, Tx	Tx
ind	Indonesian	Austronesian	Malayo-Polynesian	Latn	medium	Sp, Tx	Sp, Tx
isl	Icelandic	Indo-European	Germanic	Latn	low	Sp, Tx	Tx
ita	Italian	Indo-European	Italic	Latn	high	Sp, Tx	Sp, Tx
jav	Javanese	Austronesian	Malayo-Polynesian	Latn	medium	Sp, Tx	Tx
jpn	Japanese	Japonic	Japanesic	Jpan	high	Sp, Tx	Sp, Tx
kam	Kamba	Atlantic-Congo	Benue-Congo	Latn	zero-shot	Sp	–
kan	Kannada	Dravidian	South Dravidian	Knda	low	Sp, Tx	Tx
kat	Georgian	Kartvelian	Georgian-Zan	Geor	low	Sp, Tx	Tx
kaz	Kazakh	Turkic	Common Turkic	Cyrl	medium	Sp, Tx	Tx
kea	Kabuverdianu	Indo-European	Italic	Latn	zero-shot	Sp	–
khk	Halh Mongolian	Mongolic-Khitan	Mongolic	Cyrl	low	Sp, Tx	Tx
khm	Khmer	Austroasiatic	Khmeric	Khmr	low	Sp, Tx	Tx
kir	Kyrgyz	Turkic	Common Turkic	Cyrl	low	Sp, Tx	Tx
kor	Korean	Koreanic	Korean	Kore	medium	Sp, Tx	Sp, Tx
lao	Lao	Tai-Kadai	Kam-Tai	Laoo	low	Sp, Tx	Tx
lit	Lithuanian	Indo-European	Balto-Slavic	Latn	low	Sp, Tx	Tx
ltz	Luxembourgish	Indo-European	Germanic	Latn	zero-shot	Sp	–
lug	Ganda	Atlantic-Congo	Benue-Congo	Latn	medium	Sp, Tx	Tx
luo	Luo	Nilotic	Western Nilotic	Latn	zero-shot	Sp, Tx	Tx
lvs	Standard Latvian	Indo-European	Balto-Slavic	Latn	low	Sp, Tx	Tx
mai	Maithili	Indo-European	Indo-Aryan	Deva	zero-shot	Sp, Tx	Tx
mal	Malayalam	Dravidian	South Dravidian	Mlym	low	Sp, Tx	Tx
mar	Marathi	Indo-European	Indo-Aryan	Deva	low	Sp, Tx	Tx
mkd	Macedonian	Indo-European	Balto-Slavic	Cyrl	low	Sp, Tx	Tx
mlt	Maltese	Afro-Asiatic	Semitic	Latn	low	Sp, Tx	Sp, Tx
mni	Meitei	Sino-Tibetan	Kuki-Chin-Naga	Beng	zero-shot	Sp, Tx	Tx
mya	Burmese	Sino-Tibetan	Burmo-Qiangic	Mymr	low	Sp, Tx	Tx

Code	Language name	Family	Subgrouping	Script	Resource	Source	Target
nld	Dutch	Indo-European	Germanic	Latn	high	Sp, Tx	Sp, Tx
nno	Norwegian Nynorsk	Indo-European	Germanic	Latn	low	Sp, Tx	Tx
nob	Norwegian Bokmål	Indo-European	Germanic	Latn	low	Sp, Tx	Tx
npi	Nepali	Indo-European	Indo-Aryan	Deva	low	Sp, Tx	Tx
nya	Nyanja	Atlantic-Congo	Benue-Congo	Latn	low	Sp, Tx	Tx
oci	Occitan	Indo-European	Italic	Latn	zero-shot	Sp	–
ory	Odia	Indo-European	Indo-Aryan	Orya	low	Sp, Tx	Tx
pan	Punjabi	Indo-European	Indo-Aryan	Guru	low	Sp, Tx	Tx
pbt	Southern Pashto	Indo-European	Iranian	Arab	medium	Sp, Tx	Tx
pes	Western Persian	Indo-European	Iranian	Arab	low	Sp, Tx	Sp, Tx
pol	Polish	Indo-European	Balto-Slavic	Latn	high	Sp, Tx	Sp, Tx
por	Portuguese	Indo-European	Italic	Latn	medium	Sp, Tx	Sp, Tx
ron	Romanian	Indo-European	Italic	Latn	high	Sp, Tx	Sp, Tx
rus	Russian	Indo-European	Balto-Slavic	Cyrl	medium	Sp, Tx	Sp, Tx
slk	Slovak	Indo-European	Balto-Slavic	Latn	medium	Sp, Tx	Sp, Tx
slv	Slovenian	Indo-European	Balto-Slavic	Latn	low	Sp, Tx	Tx
sna	Shona	Atlantic-Congo	Benue-Congo	Latn	zero-shot	Sp, Tx	Tx
snd	Sindhi	Indo-European	Indo-Aryan	Arab	zero-shot	Sp, Tx	Tx
som	Somali	Afro-Asiatic	Cushitic	Latn	low	Sp, Tx	Tx
spa	Spanish	Indo-European	Italic	Latn	high	Sp, Tx	Sp, Tx
srp	Serbian	Indo-European	Balto-Slavic	Cyrl	low	Sp, Tx	Tx
swe	Swedish	Indo-European	Germanic	Latn	low	Sp, Tx	Sp, Tx
swh	Swahili	Atlantic-Congo	Benue-Congo	Latn	medium	Sp, Tx	Sp, Tx
tam	Tamil	Dravidian	South Dravidian	Taml	medium	Sp, Tx	Tx
tel	Telugu	Dravidian	South Dravidian	Telu	medium	Sp, Tx	Sp, Tx
tgk	Tajik	Indo-European	Iranian	Cyrl	low	Sp, Tx	Tx
tgl	Tagalog	Austronesian	Malayo-Polynesian	Latn	medium	Sp, Tx	Sp, Tx
tha	Thai	Tai-Kadai	Kam-Tai	Thai	medium	Sp, Tx	Sp, Tx
tur	Turkish	Turkic	Common Turkic	Latn	medium	Sp, Tx	Sp, Tx
ukr	Ukrainian	Indo-European	Balto-Slavic	Cyrl	medium	Sp, Tx	Sp, Tx
urd	Urdu	Indo-European	Indo-Aryan	Arab	medium	Sp, Tx	Sp, Tx
uzn	Northern Uzbek	Turkic	Common Turkic	Latn	medium	Sp, Tx	Sp, Tx
vie	Vietnamese	Austroasiatic	Vietic	Latn	medium	Sp, Tx	Sp, Tx
xho	Xhosa	Atlantic-Congo	Benue-Congo	Latn	zero-shot	Sp	–
yor	Yoruba	Atlantic-Congo	Benue-Congo	Latn	low	Sp, Tx	Tx
yue	Cantonese	Sino-Tibetan	Sinitic	Hant	low	Sp, Tx	Tx
zlm	Colloquial Malay	Austronesian	Malayo-Polynesian	Latn	low	Sp	–
zsm	Standard Malay	Austronesian	Malayo-Polynesian	Latn	low	Tx	Tx
zul	Zulu	Atlantic-Congo	Benue-Congo	Latn	low	Sp, Tx	Tx

Table 5: SeamlessM4T languages. We display the language code, name, family, subgroup, and script, as well as the speech resource level and whether the language is supported as a source or a target in the speech and/or text modalities. Zero-shot here refers to S2TT or S2ST tasks with the language in question as source.

3. SeamlessAlign: Automatically Creating Aligned Data for Speech

Developing an effective multilingual and multimodal translation system like SeamlessM4T requires sizable resources across many languages and modalities. Some human-labeled resources for translation are freely available, albeit often limited to a small set of languages or in very specific domains. Well-known examples are parallel text collections such as Europarl [Koehn, 2005] and the United Nations Corpus [Ziemski et al., 2016]. A few human-created collections also involve the speech modality, like CoVoST [Wang et al., 2020, 2021c] and mTEDx [Salesky et al., 2021]. Yet no open dataset currently matches the size of those used in initiatives like Whisper [Radford et al., 2022] or USM [Zhang et al., 2023a], which proved to unlock unprecedented performance.

Parallel data mining emerges as an alternative to using closed data, both in terms of language coverage and corpus size. The dominant approach today is to encode sentences from various languages and modalities into a joint fixed-size embedding space and to find parallel instances based on a similarity metric. Mining is then performed by pairwise comparison over massive monolingual corpora, where sentences with similarity above a certain threshold are considered mutual translations [Schwenk, 2018; Artetxe and Schwenk, 2019a]. This approach was first introduced using the multilingual Laser space [Artetxe and Schwenk, 2019b]. Teacher-student training was then used to scale this approach to 200 languages [Heffernan et al., 2022; NLLB Team et al., 2022] and subsequently, the speech modality [Duquenne et al., 2021, 2023a].

In this section, we describe how we employed parallel data mining to create SeamlessAlign: the largest open dataset for multimodal translation to date, totaling 470,000 hours. The overall workflow is summarized in Figure 1, and builds on the approach deployed in SpeechMatrix [Duquenne et al., 2023a]. Starting with a large collection of raw audio, we chunked files into overlapping segments and applied speech Language Identification (LID). On the text side, we used the same sentence-segmented dataset drawn from NLLB [NLLB Team et al., 2022]. Speech and text corpora were then projected into a common embedding space, in which mining was performed to identify translation pairs with optimal segmentation. Several improvements over the original SpeechMatrix pipeline are introduced:

- an improved and extended speech language identification (LID) model,
- a novel multimodal embedding space,
- increased coverage from 17 to 37 languages,
- increased raw audio amount, totaling 4 million hours.

In the current version, mining was focused on 37 target languages of the SeamlessM4T system. Scaling to all 100 languages will be explored in future iterations of our work.

3.1 Speech-language identification

Language identification (LID) of raw audio data is a critical component of our workflow. Incorrectly labeling speech at this stage can prevent high-quality audio segments from being aligned or, worse, add noise to the resulting paired sets. This can adversely affect the performance of the downstream translation system.

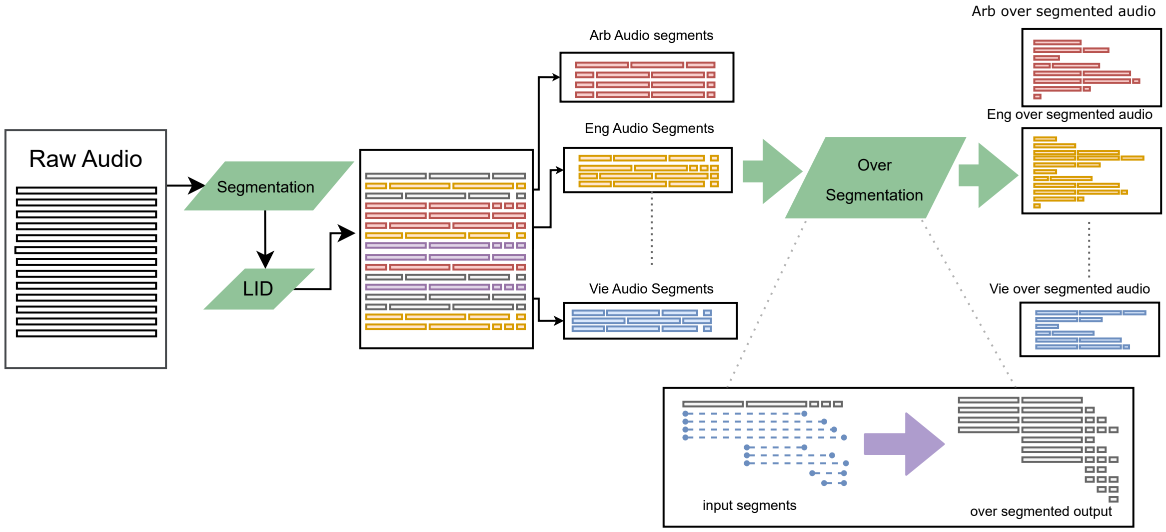

Figure 1: Workflow of speech processing.

While numerous off-the-shelf LID models exist, none could cover our target list of 100 languages.[4] Therefore, we trained our own model, following the ECAPA-TDNN architecture introduced in [Desplanques et al., 2020], for which an open-source model trained on VoxLingua107 [Valk and Alumäe, 2021] is available. The new model adds support for several new languages, including Moroccan Arabic, Egyptian Arabic, Central Kurdish, West Central Oromo, Irish, Igbo, Kyrgyz, Ganda, Maithili, Meitei, Nyanja, Odia, Cantonese, and Zulu.

3.1.1 Training

Baseline We first retrained a system from scratch on VoxLingua107 data to reproduce a baseline. This system, dubbed *VL107 baseline*, achieved a classification error rate of 5.25% on the development set of VoxLingua107 at epoch 30. Comparatively, the open-sourced model available on HuggingFace,[5] referred to as *VL107 HF*, yields an error rate of 7%.

Experimental setup With our training pipeline validated, we finally trained our own model for 40 epochs. This required about 172 hours on 8 GPUs. A total of 17k hours of speech were used, with an average of 171 hours per language, ranging from 1 to 600 hours. The test corpus covers our 100 languages of interest and is composed of the FLEURS test set, the VoxLingua107 development set, and extra test data extracted from VAANI,[6] IIITH [Kumar Vuddagiri et al., 2018] and KENCORPUS[7] [Wanjawa et al., 2022].

Results The F1 scores on the test data for all models are presented in Table 6. The results are given for the 100 SeamlessM4T languages, and the 79 languages in common with VoxLingua107. We can see that training on the additional languages slightly decreases the

4. MMS [Pratap et al., 2023] has recently been released and covers them all, but it was not available when this project started
5. `https://huggingface.co/TalTechNLP/voxlingua107-epaca-tdnn`
6. `http://vaani.iisc.ac.in`
7. `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/6N5V1K`

	Overall		Intersection	
	↑F1-micro *(n=100)*	↑F1-macro *(n=100)*	↑F1-micro *(n=79)*	↑F1-macro *(n=79)*
VL107 HF	82.3%	-	94.1%	92.6%
VL107 baseline	82.5%	-	94.4%	93.0%
LID100	86.0%	81.9%	92.9%	91.1%

Table 6: F1 micro and macro average for the considered LID systems over all SeamlessM4T languages and the intersection of supported languages across models. Dashes are used for models that do not support the full 100 scope.

overall performance for the common set of languages, which is a direct consequence of the presence of a higher number of close languages. For example, Zulu (zul) is very often confused with Nyanja (nya), Igbo (ibo) with Yoruba (yor), and Modern Standard Arabic (arb) with Moroccan Arabic (ary) and Egyptian Arabic (arz). Our model improves classification (F1 difference greater than 5%) on 17 languages with an average gain of 14.6%, not counting the newly covered languages, while decreasing classification for 12 (with an average loss of 9.8%).

3.1.2 Filtering

While it is important to retrieve the maximum amount of data for mining, we must also ensure high quality in LID labeling. Depending on the quantity of data available for a particular language, it may be useful to filter it to retain higher-quality data. We thus estimated the Gaussian distribution of the LID scores per language for correct and incorrect classifications on the development corpus. We selected a threshold per language such that $p(correct|score) > p(incorrect|score)$. By rejecting 8% of the data, we were able to further increase the F1 measure by almost 3%.

	↑**F1 micro**	↑**Coverage**
LID100	86.0%	100%
+filtering	89.5%	92.1%

Table 7: F1 micro average and coverage across 100 languages for the LID100 system with and without filtering.

3.2 Gathering raw audio and text data at scale

Text pre-processing On the text side, we rely entirely on the same dataset deployed in NLLB [NLLB Team et al., 2022]. The same data sources, cleaning, and filtering steps are used and run at scale with our Stopes library.

Audio pre-processing We start with 4 million hours of raw audio originating from a publicly available repository of crawled web data. Table 10 provides statistics on the amount of raw audio for each language. Approximately 1 million hours in this collection are in English. We then applied a series of pre-processing steps to curate and improve the overall speech quality. Firstly, we deduplicated the audio file URLs found in the repository, downloaded the audio files, and resampled at 16KHz. Subsequently, we filtered out the non-speech data with a bespoke audio event detection (AED) model.

Audio segmentation To perform S2TT or S2ST mining, it is desirable to split audio files into smaller chunks that map as closely as possible to self-contained sentences, equivalent to sentences in a text corpus. However, genuine semantic segmentation in speech is an open-ended problem–pauses can be an integral part of a message and can naturally occur differently across languages. For mining purposes, it is impossible to prejudge what specific segments can maximize the overall quality of the mined pairs.

We thus followed the over-segmentation approach drawn from [Duquenne et al., 2021] (as depicted in Figure 1). First, we used an open Voice Activity Detection (VAD) model [Silero, 2021] to split audio files into shorter segments. Subsequently, our speech LID model was used on each file. Finally, we created several possible overlapping splits of each segment and left the choice of the optimal split to the mining algorithm described in the next section. This over-segmentation strategy roughly octuples the number of potential segments considered.

3.3 Speech mining

The overall workflow of our mining process is shown in Figure 2. First, we trained encoders for text (Section 3.3.1) and speech (Section 3.3.2). These are then used to project both modalities into a joint embedding space. We then mined speech segments against text sentences or speech segments in other languages to create large amounts of S2TT and S2ST pairs. These corpora are subsequently combined with other resources to train the SeamlessM4T model.

3.3.1 Sonar text embedding space

Architecture and training setup We developed a novel sentence embedding space, named **S**entence-level multim**O**dal and la**N**guage-**A**gnostic **R**epresentations—in short, Sonar. Sonar substantially outperforms the previous Laser space. It follows the same two-step

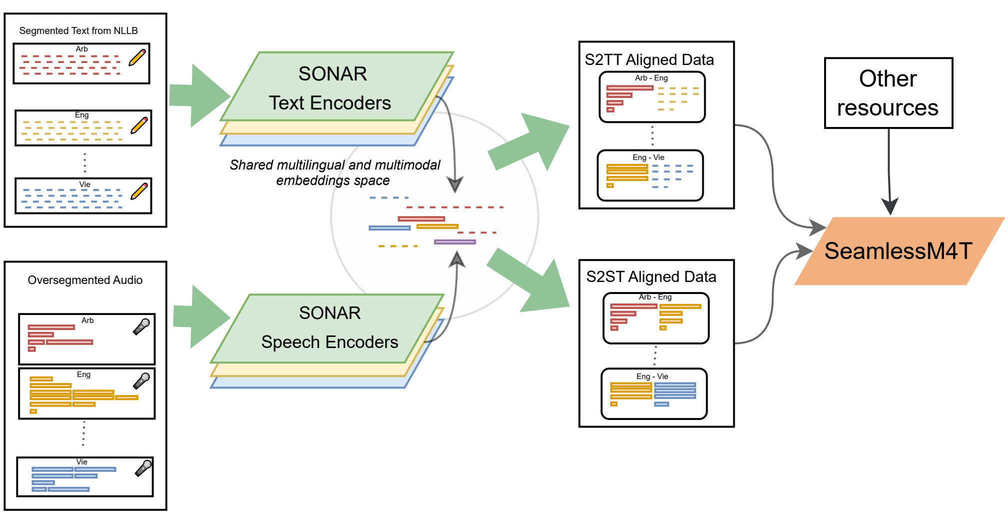

Figure 2: Workflow of the Sonar encoding and mining processes.

Model	↑spBLEU		↑COMET	
	X–eng (*n=200*)	eng–X (*n=200*)	X–eng (*n=89*)	eng–X (*n=89*)
SONAR	32.7	21.6	85.9	84.2
NLLB-1.3B (MT topline)	35.2	24.9	86.5	85.2

Table 8: Average performance on FLORES devtest set over the 200 NLLB languages and 89 languages supported by COMET: translation spBLEU and COMET scores, auto-encoding spBLEU.

approach: we first trained a text embedding space and then relied on a teacher-student training strategy to extend it to the speech modality. Similarly to LASER, the initial text SONAR space uses an encoder-decoder architecture, but is based on the NLLB-1.3B model, capable of translating across 200 languages [NLLB Team et al., 2022]. We replaced the intermediate representation with a fixed-size vector using mean-pooling (i.e., the decoder thus attends to a single vector). This architecture is fine-tuned using all of NLLB's T2TT training data, and we explored several training objectives. A detailed ablation study can be found in Duquenne et al. [2023b]. This yields a powerful, massively multilingual sentence representation that can be decoded into all 200 languages of the NLLB project. Figure 3 provides an illustration of the SONAR architecture and Table 8 summarizes the translation evaluation of the SONAR framework.

Evaluation for mining On pure translation performance, we observe that the fixed-size representation bottleneck leads to a 7% and 13% decrease in BLEU score when translating into English (35.2→32.7) and out of English, respectively (24.9→21.6). This is a rather

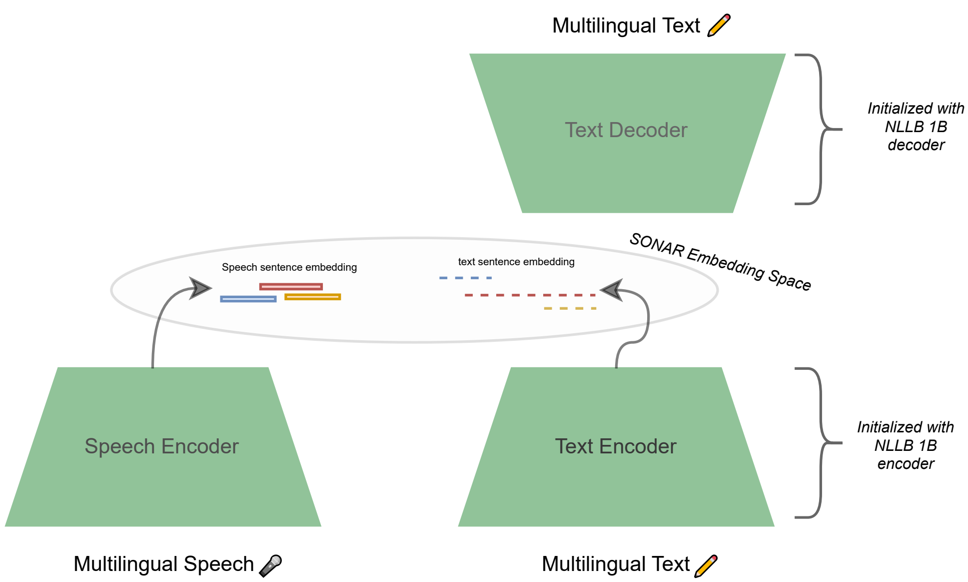

Figure 3: SONAR architecture.

interesting result, given that the use of attention is commonly considered mandatory to achieve reasonable performance.

On mining performance, we rely on the multilingual similarity search `xsim` metric, which measures the percentage of sentences in the FLORES dataset which are not correctly aligned when searching for the closest vector in the embedding space. The improved version `xsim++` [Chen et al., 2023b] added challenging English sentences on the target side. Both of these metrics are a good proxy to the actual T2TT mining task while being much faster to compute.

As summarized in Table 9, SONAR substantially outperforms other popular multilingual sentence representations like LASER3 [Heffernan et al., 2022] or LaBSE [Feng et al., 2022].

	Overall		**Intersection**	
	↓`xsim` *(n=200)*	↓`xsim++` *(n=200)*	↓`xsim` *(n=98)*	↓`xsim++` *(n=98)*
SONAR	1.4	15.2	0.1	9.3
LASER3	5.1	36.4	1.1	27.5
LaBSE	10.7	36.1	1.5	15.4

Table 9: Comparison of similarity search results (error rates) on all 200 FLORES languages, and limited to the intersection of 98 languages on which each model has been trained on.

3.3.2 Training speech encoders

Architecture and training setup As a second step and following [Duquenne et al., 2021], the new SONAR text embedding space is extended to the speech modality through teacher-student training. In that work, a fixed-size speech representation was obtained by taking the BOS output of a pretrained XLS-R model [Babu et al., 2022]. This model was then fine-tuned to maximize the cosine loss between this pooled speech representation and sentence embeddings in the same languages (ASR transcriptions) or in English (speech translations). We improved this initial recipe by doing the following:

- MSE loss instead of a cosine loss was used. This enables us to use the SONAR text decoder on speech input,
- W2V-BERT 2.0 speech front-end instead of XLS-R. W2V-BERT 2.0 was optimized on 143 languages (see Section 4.1 for details),
- Attention-pooling. Instead of the usual pooling methods (i.e., mean or max-pooling), we implemented a 3-layer sequence-to-sequence model to convert the variable length sequence of W2V-BERT 2.0 to a fixed size vector,
- Training on human-performed ASR transcriptions only. We collected at least 100 hours of ASR transcriptions for most of the languages (see Table 10 column *"train"*) and trained the speech encoders exclusively on them,
- Following [Heffernan et al., 2022; NLLB Team et al., 2022], we grouped languages by linguistic families (i.e., Germanic or Indian languages) and trained them together in one speech encoder. Alternative language groupings, which might consider the acoustic characteristics of each language, are left open for future research.

ISO	Raw audio [h]	Train ASR [h]	X–eng (↑BLEU) Ours	Whisper	Mined audio [h] Sen2Txx	Sxx2Ten	Sxx2Sen
arb	106755	822	28.7	25.5	1568	8072	776
ben	7012	335	18.9	13.2	606	1345	263
cat	43531	1738	35.1	34.2	1570	4411	354
ces	41318	181	29.2	27.8	1454	6905	602
cmn	79772	9320	16.2	18.4	5440	18760	1570
cym	24161	99	14.5	13.0	–	4411	278
dan	34300	115	31.9	32.7	2499	6041	583
deu	490604	3329	32.7	34.6	91715	17634	1921
est	12691	131	23.8	18.7	1022	3346	607
fin	32858	184	22.2	22.1	651	6086	526
fra	282179	2057	31.2	32.2	21523	17380	3337
hin	15118	150	19.2	22.0	1041	2977	530
ind	11559	269	26.5	29.1	1938	2658	510
ita	79480	588	25.3	23.6	4378	6508	817
jpn	75863	17319	17.4	18.9	1973	21287	1141
kan	1451	114	20.0	11.6	–	232	198
kor	37854	316	15.0	21.3	–	8657	640
mlt	2122	106	23.2	13.5	131	130	60
nld	93933	1723	25.5	24.0	3720	6859	1210
pes	43788	386	22.2	19.6	–	7122	693
pol	53662	304	21.1	22.3	1324	9389	757
por	141931	269	35.4	38.1	4853	8696	928
ron	18719	135	32.1	31.5	2770	2878	716
rus	103906	259	25.4	27.8	11296	13509	1252
slk	16954	102	29.5	26.1	1267	3785	491
spa	324086	1511	24.3	23.3	27778	17388	2727
swe	125195	144	33.4	37.02	3438	2620	484
swh	18393	361	22.6	7.2	690	2620	484
tam	100331	245	14.3	9.2	–	1664	867
tel	3303	84	15.8	12.5	–	985	536
tgl	4497	108	13.3	24.4	–	633	266
tha	13421	195	15.3	16.1	2577	3563	542
tur	23275	174	21.0	26.6	1417	6545	426
ukr	6396	105	27.9	29.4	1220	1717	392
urd	16882	185	17.6	17.2	773	3416	652
uzn	8105	115	17.9	6.0	475	1846	157
vie	34336	194	17.8	20.4	1689	7692	868
Total/avr	2529741	43772	23.3	22.5	202796	239767	29161

Table 10: Statistics on speech encoders and amount of mined data. Sen2Txx, Sxx2Ten, and SxxSen correspond to English speech paired with foreign text, foreign speech paired with English Text, and foreign Speech paired with English speech, respectively. Dashes are unmined directions. We provide the amount of raw audio data for mining and the amount of human-provided ASR transcripts to train the speech encoders. The speech encoders are evaluated for S2TT using BLEU on the Fleurs test set. Our model performs zero-shot S2TT. Finally, the last three columns provide the amount of mined data.

Evaluation of speech encoders The trained speech encoders are to be used in S2TT and S2ST mining, and the resulting paired data is to be fed into the SeamlessM4T system (see section 4). Consequently, an ideal evaluation would consist of testing various iterations of each speech encoder by using them in an end-to-end loop: performing mining, then training a S2TT or S2ST translation system on the mined data, and potentially comparing different thresholds of the Sonar score. Unfortunately, this is a very compute-intensive recipe.

Instead, given that the Sonar embedding space comes with a text decoder, we chose to evaluate the individual speech encoders on a S2TT task. That is, following [Duquenne et al., 2022, 2023c], we decoded foreign speech embeddings into English texts. Results are summarized in Table 10, column *"X-eng BLEU"*. For comparison, we also provide the performance of Whisper-Large-v2 [Radford et al., 2022]. It is important to emphasize that the Sonar speech encoders were trained on ASR transcriptions only and the Sonar text decoder has never been exposed to any speech input. Therefore, the reported results correspond to fully zero-shot speech translation.

Despite the zero-shot scenario, the Sonar speech encoders compare favorably to a model like Whisper-Large-v2, which was trained on a massive amount of translated audio. Gaps in BLEU points can be observed in some high resource languages such as German, Russian or Portuguese, However, zero-shot speech translation with our speech encoders outperforms Whisper-Large-v2 on several low-resource languages – particularly for Swahili and several South Asian languages like Bengali, Kannada, Telugu, and Tamil.

3.3.3 Speech mining

Margin setting Mining was performed using a margin criterion with our Stopes data processing library[8] [Andrews et al., 2022]. The overall processing is identical to that developed for T2TT mining in NLLB [NLLB Team et al., 2022]. We performed so-called *global mining*, where all speech segments in one language are compared to all speech segments in another language. *Local mining*, on the contrary, would try to leverage knowledge on longer speech chunks that are likely to contain many parallel segments. A typical example would be documentation on an international event in multiple languages. Such high-level information is very difficult to obtain at scale.

First, the embeddings for all speech segments and text sentences are calculated. These are then indexed with the FAISS library [Johnson et al., 2019], enabling efficient large-scale similarity search on GPUs. Finally, nearest neighbors to all elements in both directions are retrieved, and margin scores are computed following the formula introduced in [Artetxe and Schwenk, 2019a]:

$$\text{score}(x, y) = \text{margin}\left(cos(x, y), \sum_{z \in NN_k(x)} \frac{cos(x, z)}{2k} + \sum_{v \in NN_k(y)} \frac{cos(y, v)}{2k} \right) \tag{1}$$

where x and y are the source and target sentences, and $NN_k(x)$ denotes the k nearest neighbors of x in the other language. We set k to 16.

In past work, a threshold of 1.06 on the margin score was used for bitext mining based on Laser embeddings [Schwenk et al., 2021; NLLB Team et al., 2022]. The Sonar space,

8. `https://github.com/facebookresearch/stopes`

however, displayed different dynamics and the optimal threshold was adapted accordingly. Since full end-to-end evaluation with S2TT or S2ST training is too compute-intensive, we set the new threshold at 1.15 after some human inspection. The statistics reported in Table 10 are based on this threshold.

Mined dataset We performed mining of speech in foreign languages against English texts (column Sxx2Ten in Table 10) and English speech (column Sxx2Sen in Table 10). Given the sheer size of our raw English speech (1 million hours) and foreign text collections (often more than 1 billion sentences), we carried out this operation only for some languages (column Sen2Txx in Table 10). Other directions are left for future work.

Except for Maltese, for which we had access only to a small amount of raw audio, we were able to mine more than 100 hours of speech alignments with English speech for all languages. The alignments with English texts reached a thousand hours for most languages and exceeded ten thousand hours for six (i.e., German, French, Spanish, Japanese, Russian, and Mandarin Chinese). Overall, SeamlessAlign covers 37 languages and a total of 470,000 hours:

- English speech to non-English text (Sen2Txx)—approximately 200,000 hours
- Non-English speech to English text (Sxx2Ten)—approximately 240,000 hours
- Non-English speech to English speech (Sxx2Sen)—approximately 29,000 hours

Adding such huge amounts of data to train a massively multilingual S2ST translation system represents a substantial computational challenge. As described in Section 4, not all of this data was used for modeling, but only a subset with the highest Sonar alignment scores. Since our mined data can help support many different use cases, we are open-sourcing the meta-data for the full amount[9] (i.e., up to a Sonar threshold of 1.15), to allow the community to rebuild SeamlessAlign and use it for their own purposes. The optimal threshold can thus be tuned based on the task, balancing dataset size and alignment quality. Our mining code is also open-sourced in the Stopes library.

3.4 Related work

3.4.1 Speech LID

Spoken language identification has been traditionally approached in a two-stage workflow: a classifier is trained on top of conventional representations like the i-vector or x-vector, extracted from the raw audio signal [Dehak et al., 2011; Snyder et al., 2018]. The same idea has been revisited in end-to-end, integrated neural architectures [Cai et al., 2019; Miao et al., 2019; Wan et al., 2019]. These approaches typically fall short as the input audio goes shorter, which can be an issue with speech recordings involving multiple speakers talking to each other in turn. New methods were developed to tackle this very problem. Lopez-Moreno et al. [2014] show that a simple feed-forward network can outperform i-vectors on this task. More complex architectures such as convolutional neural networks or Bi-LSTMs prove to be more efficient in capturing information from the speech input [Lozano-Diez et al., 2015; Fernando

9. available at `https://github.com/facebookresearch/seamless_communication`

et al., 2017]. Some other approaches try to bridge the gap with models focused on longer segments through teacher-student training [Shen et al., 2018, 2019].

Recent initiatives aimed at increasing language coverage to go beyond a handful of conventionally very high-resource languages. The ECAPA-TDNN architecture introduced in [Desplanques et al., 2020] has proven effective to distinguish between the 107 languages of Voxlingua107 [Valk and Alumäe, 2021]. The XLS-R pretrained model [Babu et al., 2022] is also fine-tuned on a language identification task using the same dataset. WHISPER-LARGE-V2 is another popular model that can perform this task for 99 languages [Radford et al., 2022]. Very recently, the MMS project further broadened language support to 4,000 spoken languages [Pratap et al., 2023].

3.4.2 Speech segmentation

To achieve sentence-like speech segments, a commonly employed method is pause-based segmentation using Voice Activity Detection (VAD). This approach is widely utilized in various applications, including speech mining, ASR, and speech translation. In this work, we adopted the over-segmentation strategy proposed by Duquenne et al. [2021] on top of the segments obtained through VAD segmentation. While this over-segmentation significantly improves the recall of the mining process, it does come with certain drawbacks. Specifically, it leads to a substantial increase (8x) in the number of segments, introducing noise in the embedding space, and raising the computational demand for the mining process. Pause-based segments may not align with semantically coherent sentences; in fact, they tend to be too short because speaker pauses can extend beyond sentence boundaries. Consequently, for speech translation, researchers have put forward more sophisticated segmentation strategies with the potential to deliver higher-quality speech translation results. Gállego et al. [2021] used a pretrained wav2vec 2.0 instead of VAD to detect speech segments. Potapczyk and Przybysz [2020a] proposed a divide-and-conquer (DAC) algorithm that iteratively operates on top of VAD longest detected pauses until all segments become below a max-segment length parameter. Gaido et al. [2021] further builds upon this through a hybrid approach. SHAS [Tsiamas et al., 2022] train a classifier on top of wav2vec 2.0 using optimal segmentation from a manually segmented corpus. Similar to Potapczyk and Przybysz [2020a], it then applies a DAC algorithm on the splitting probabilities of the network to obtain final segmentation decisions. This approach demonstrated significant gains over simple pause-based segmentation and other baselines in speech-to-text translation tasks. These segmentation methods could be promising for speech mining, suggesting exciting avenues for future research.

3.4.3 Multilingual and multimodal representations

Several works have studied how to learn multilingual sentence representations. Well known approaches are LASER [Artetxe and Schwenk, 2019b], LaBSE [Feng et al., 2022], or [Yang et al., 2019; Ramesh et al., 2022]. While LASER was trained with an MT translation objective, a decoder compatible with the LASER embedding space is not freely available. To the best of our knowledge, SONAR is the first sentence embedding space for which an efficient and multilingual decoder is available. Another direction of research is to first train an English sentence representation (e.g., sentence-BERT [Reimers and Gurevych, 2019]) and in a second step, extend it to more languages using teacher-student training [Reimers and Gurevych, 2020]. The same approach was used to extend Laser to 200 languages, named Laser3 [Heffernan et al., 2022].

Learning unsupervised representations of speech is the focus of several works, whether involving monolingual [Baevski et al., 2022] or multilingual speech [Babu et al., 2022; Hsu et al., 2021; Chung et al., 2021]. Examples of joint text and speech pre-trained models are mSLAM [Bapna et al., 2022] and Mu^{2}SLAM [Cheng et al., 2023]. Duquenne et al. [2021] were the first to introduce fixed-size text and speech representations that can be used to perform multimodal mining, followed by [Khurana et al., 2022]

3.4.4 Speech mining

The proof of concept of a joint text/speech representation that can be used to perform text/speech or speech/speech mining was presented by Duquenne et al. [2021]. In follow-up work, this approach was used to align speech in 17 languages in the VoxPopuli corpus to give rise to the SpeechMatrix corpus [Duquenne et al., 2023a]. The authors mined for parallel speech segments in all 136 possible combinations of languages, yielding a total of 418 thousand hours of speech-to-speech alignments, out of which about 46 thousand hours are aligned with English. SpeechMatrix is a large corpus, but the domain is rather limited since the raw audio of the VoxPopuli corpus is derived from European Parliament speeches. The corpus SpeechMatrix is freely available. Khurana et al. [2022] use a joint text/speech embedding space, dubbed Samu-Xlsr, to evaluate the recall of text and speech retrieval in the corpora CoVoST 2, MUST-C, and MTEDx.

There are several works that indirectly create speech-to-speech corpora. One direction of research is to perform speech synthesis on corpora aligned at the text level, (e.g., the CVSS corpus [Jia et al., 2022b] which is based on the CoVoST 2 speech-to-text translation corpus).

4. SeamlessM4T Models

Direct speech-to-text translation models have made significant progress in recent years [Berard et al., 2016; Weiss et al., 2017a; Di Gangi et al., 2019; Agarwal et al., 2023], and achieved parity with cascaded models on academic benchmarks under specific situations (e.g., constrained data, in-domain settings, specific language pairs, etc.). However, with the arrival of massively multilingual translation models [NLLB Team et al., 2022; Siddhant et al., 2022; Fan et al., 2020] and weakly supervised ASR models [Radford et al., 2022; Zhang et al., 2023a; Pratap et al., 2023], which leverage massive quantities of labeled data for training large foundation models, these comparisons have become outdated. To put it simply, direct models now lag significantly behind strong cascaded models.

One of our goals with SeamlessM4T is to bridge the gap between direct and cascaded models for S2TT in large multilingual and multimodal settings by building a stronger direct X2T model (for translating both text and speech into text) that combines a strong speech representation learning model with a massively multilingual T2TT model. Beyond text outputs, our second goal builds on recent speech translation advancements, which have placed much emphasis on building systems that produce speech outputs [Jia et al., 2019b; Lee et al., 2022a; Inaguma et al., 2023]. We enable speech-to-speech translation with UnitY [Inaguma et al., 2023], a two-pass modeling framework that first generates text and subsequently predicts discrete acoustic units. Unlike cascaded models, the different components in UnitY (see Figure 4) can be jointly optimized.[10]

The aforementioned approach alleviates the issue of cascaded error propagation and domain mismatch, while relying on an intermediate semantic representation to mitigate the problem of multi-modal source-target mapping. The vocoders for synthesizing speech are trained separately (see Section 4.3.1). Figure 4 provides an overview of the SeamlessM4T model, including its four building blocks: (1) SeamlessM4T-NLLB a massively multilingual T2TT model, (2) w2v-BERT 2.0, a speech representation learning model that leverages unlabeled speech audio data, (3) T2U, a text-to-unit sequence-to-sequence model, and (4) multilingual HiFi-GAN unit vocoder for synthesizing speech from units.

The SeamlessM4T multitask UnitY model integrates components from the first three building blocks and is fine-tuned in three stages, starting from an X2T model (1,2) with English target only and ending with a full-fledged multitask UnitY (1,2,3) system capable of performing T2TT, S2TT and S2ST, as well as ASR. In what follows, we first describe unsupervised speech pre-training (w2v-BERT 2.0) in Section 4.1. We then introduce the X2T model in Section 4.2, starting with the data preparation pipeline in Section 4.2.1. Section 4.2.2 describes our multilingual T2TT model, and Section 4.2.3 details how the speech encoder and the T2TT model are jointly fine-tuned for X2T with multimodal and multitask capabilities. Next, we look at the S2ST task, starting from the acoustic unit extraction pipeline and vocoder design to map units back to speech waveforms in Section 4.3.1 Then, we describe T2U pre-training in Section 4.3.2. Section 4.3.3 ultimately outlines how all these components come together in the third and final stage of fine-tuning. We evaluated

10. There are two views of what constitutes a direct model in speech-to-speech translation literature: (1) A model that does not use intermediate text representation [Lee et al., 2022a] and (2) A model that directly predicts the target spectrogram [Jia et al., 2022a]

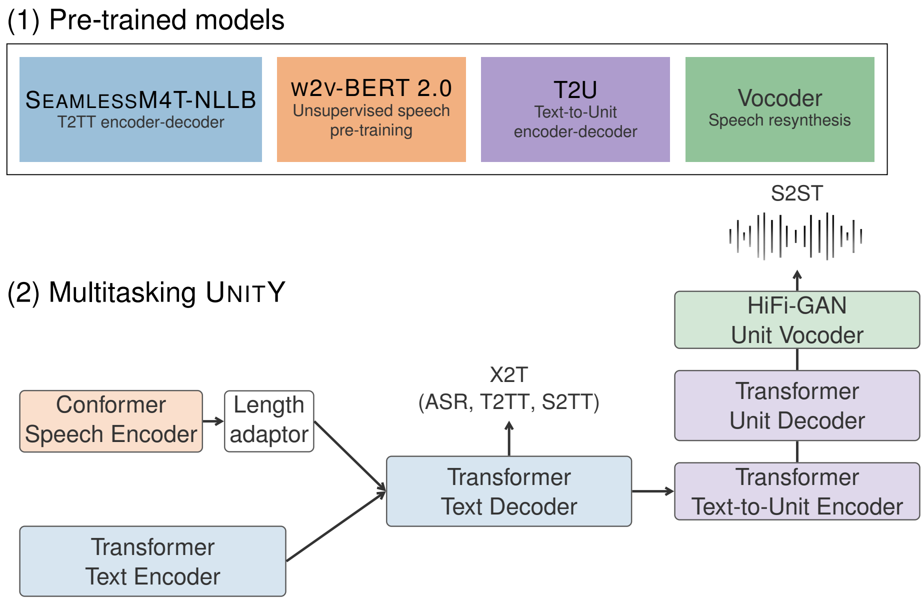

Figure 4: Overview of SeamlessM4T. (1) shows the pre-trained models used when finetuning multitasking UnitY. (2) outlines multitasking UnitY with its two encoders, text decoder, T2U encoder-decoder, and the supporting vocoders for synthesizing output speech in S2ST.

our model using standard automatic metrics in Section 4.4 and compared its performance with state-of-the-art speech translation models.

4.1 Unsupervised Speech Pre-training

Labels for speech recognition and translation tasks are scarce and expensive, especially for low-resource languages. It is challenging to train speech translation models with only limited access to supervision. Self-supervised pre-training with unlabeled speech audio data is, thus, a practical approach for reducing the need for supervision in model training. This method helps achieve the same recognition and translation quality with much less labeled data than models without pre-training. It also helps push the limits of model performance with the same amount of labeled data. The most recent and publicly available state-of-the-art multilingual speech pre-trained model is MMS [Pratap et al., 2023]. It extends its predecessor, XLS-R [Babu et al., 2022], with additional 55K hours of training data and covers more than 1,300 new languages (see Table 11). Besides MMS, USM [Zhang et al., 2023a] is a proprietary SOTA multilingual speech pre-trained model that leverages the latest model architecture (BEST-RQ [Chiu et al., 2022] instead of wav2vec 2.0 [Baevski et al., 2020]), has the largest scale of training data (12M hours), and covers more than 300 languages.

w2v-BERT 2.0 follows w2v-BERT [Chung et al., 2021] to combine contrastive learning and masked prediction learning, and improves w2v-BERT with additional codebooks in both learning objectives. The contrastive learning module is used to learn Gumbel vector quantization (GVQ) codebooks and contextualized representations that are fed into the subsequent masked prediction learning module. The latter refines the contextualized representations by a different learning task of predicting the GVQ codes directly instead of

Model	Languages	Hours	Model type	Open model
XLS-R-2B-S2T	128	0.4M	wav2vec 2.0 [Baevski et al., 2020]	✓
USM	over 300†	12M	BEST-RQ [Chiu et al., 2022]	
MMS	1406	0.5M	wav2vec 2.0 [Baevski et al., 2020]	✓
SeamlessM4T-Large	over 143†	1M	w2v-BERT 2.0	✓

Table 11: A comparison of multilingual speech pre-training in state-of-the-art ASR and S2TT models. †Estimated from the part of data that has language information.

polarizing the prediction probability of correct and incorrect codes at the masked positions. Instead of using a single GVQ codebook, w2v-BERT 2.0 follows Baevski et al. [2020] to use product quantization with two GVQ codebooks. Its contrastive learning loss $\mathcal{L}_c$ is the same as that in w2v-BERT, including a codebook diversity loss to encourage the uniform usage of codes. Following w2v-BERT, we use GVQ codebooks for masked prediction learning and denote the corresponding loss as $\mathcal{L}_{m_{\text{GVQ}}}$. We also created an additional masked prediction task using random projection quantizers [Chiu et al., 2022] (RPQ), for which we denote the corresponding loss as $\mathcal{L}_{m_{\text{RPQ}}}$. The overall w2v-BERT 2.0 training loss $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = w_c \mathcal{L}_c + w_{m_{\text{GVQ}}} \mathcal{L}_{m_{\text{GVQ}}} + w_{m_{\text{RPQ}}} \mathcal{L}_{m_{\text{RPQ}}}, \tag{2}$$

where loss weights w_c, $w_{m_{\text{GVQ}}}$ and $w_{m_{\text{RPQ}}}$ are set to 1.0, 0.5, and 0.5, respectively.

We follow the w2v-BERT XL architecture [Chung et al., 2021] for the w2v-BERT 2.0 pre-trained speech encoder in SeamlessM4T-Large, which has 24 Conformer layers [Gulati et al., 2020] and approximately 600M model parameters. The w2v-BERT 2.0 model is trained on 1 million hours of open speech audio data that covers over 143 languages.

4.2 X2T: Into-Text Translation and Transcription

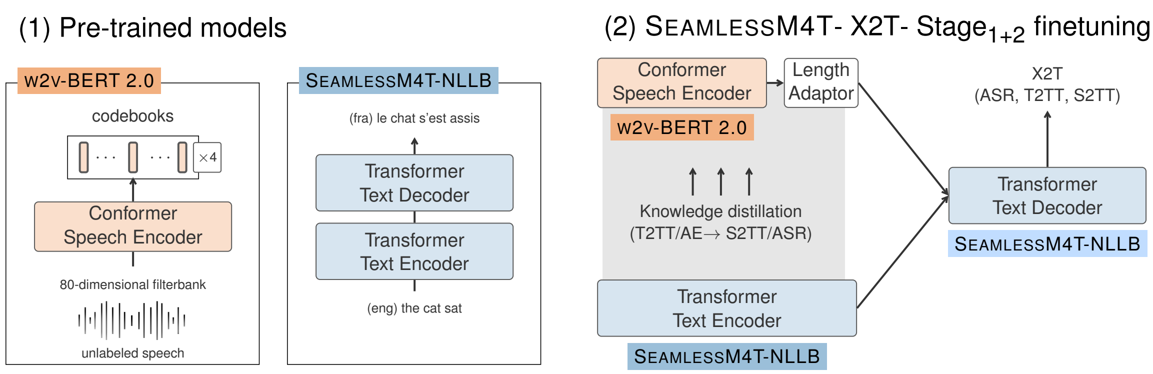

Figure 5: Overview of the SeamlessM4T X2T model. (1) describes the main two building blocks: w2v-BERT 2.0 and SeamlessM4T-NLLB. (2) describes the training of the X2T model. In Stage$_1$, the model is trained on X–eng directions and in Stage$_2$, eng–X directions are added.

The core of our multitask UnitY framework is the X2T model, a multi-encoder sequence-to-sequence models with a Conformer-based encoder [Gulati et al., 2020] for speech input and another for Transformer-based encoder [Vaswani et al., 2017] for text input—both of which are joined with the same text decoder. Our X2T model is trained on S2TT data pairing speech audio in a source language with text in a target language.

4.2.1 Preparing X2T data

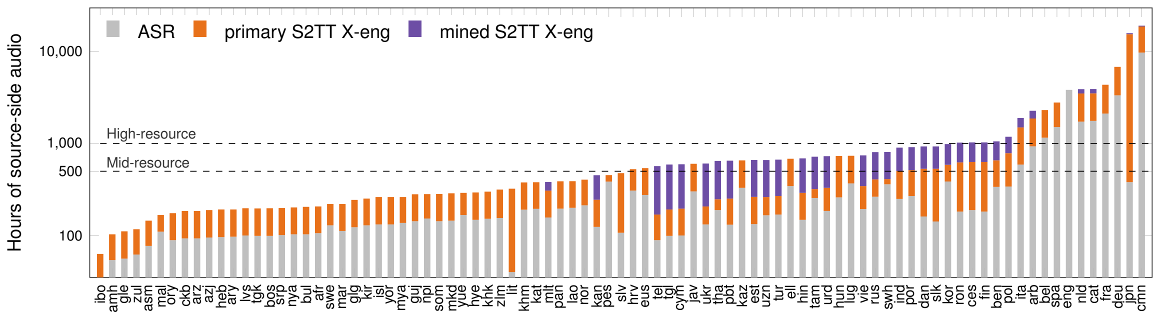

Figure 6: Statistics of ASR and X–eng S2TT data used to train our SeamlessM4T model. We show the data size in hours of speech (log-scale) between ASR, S2TT primary and mined. Languages are sorted in ascending resource-level. For numerical statistics see Table 35

Processing human-labeled data When using human-labeled data, we removed special tokens such as <silence> and <no-speech> from the verbatim transcriptions. We additionally perform length filtering to remove examples exceeding a maximum text length of 100 sub-word tokens (based on the text tokenizer described below) and pairs with a skewed text-to-audio length ratio that exceeds 5 sub-words per second. Doing so improves the batching efficiency when training and eliminates pairs that are likely to be noisy or misaligned.

Pseudo-labeling As with any sequence-to-sequence task, S2TT performance is dependent on the availability of high-quality training data. However, the amount of human-labeled S2TT data is scarce in comparison to its T2TT or ASR counterparts. To address this shortage of labeled data, we resort to pseudo-labeling [Jia et al., 2019a; Pino et al., 2020] the ASR data with a multilingual T2TT model. In this case, we used NLLB-200-3.3B and generated pseudo-labels with the recommended decoding options from NLLB Team et al. [2022]. Hereafter, we refer to human-labeled and pseudo-labeled data as *primary* data.

Parallel data mining Even with pseudo-labeled ASR data, the amount of S2TT data is insignificant compared to the scale of T2TT data. Consider for instance the English-Italian direction, one of the highly resourced pairs in T2TT with over 128M parallel sentences—only 2M pairs of English text paired with Italian audio are available for S2TT. Parallel data mining (see how SeamlessAlign was built in Section 3) is another strategy we draw upon to collect more training data. This kind of mining, however, tends to produce noisy alignments and requires some filtering. We use the top 400 hours of SeamlessAlign (see Section 3) in each of 33 X–eng directions and the top 200 hours in each of 29 eng–X directions based on Sonar alignment scores. This amounts to an additional 18.3K hours of speech audio. We show in Section 4.5.3 that these select amounts of mined data lead to a good trade-off between performance boosts and computational costs of training.

Filtering We perform additional filtering on the combined pool of *primary* and *mined* data. Following NLLB Team et al. [2022], we implemented a toxicity filter. This removes pairs that have *toxicity imbalance*, (i.e., when the difference in the number of toxic items detected in the source and target is above a certain threshold). For S2TT data, transcriptions are used as a proxy for the speech input when counting toxic items. We set the imbalance threshold at 1. In addition, we also applied a length filter. We removed pairs in which the utterance is shorter than 0.1 seconds or longer than 50 seconds. We also removed pairs in which the text is longer than 250 sub-words (based on the tokenizer described below). Lastly, we removed pairs in which the text contains more than 20% of emojis, more than 50% of punctuations, or more than 50% of spaces.

Figure 6 shows the distribution of filtered X–eng S2TT data used to train SeamlessM4T models. Based on the total amount of speech audio hours in each language, we assessed its resource level: *high-resource* are languages with more than 1000 hours of supervision, *mid-resource* are those between 500 and 1000 hours, and *low-resource* are those with less than 500 hours.

Training a Text Tokenizer. The tokenizer used in NLLB-200 [NLLB Team et al., 2022] is trained with *SentencePiece* [Kudo and Richardson, 2018] using the BPE algorithm [Gage, 1994; Sennrich et al., 2016]. These multilingual tokenizers, with their underlying vocabularies, are trained by sampling data from each language. Due to artifacts of sampling and the much larger number of unique symbols in logo-graphic writing systems, the result of this is that many key Chinese characters are missing from the original NLLB-200 vocabulary. To address this issue, we force the inclusion of these characters even in cases where they may not appear in the sampled *SentencePiece* training data. In order to decide which characters to include, we looked at the MTSU list[11] and similar character frequency lists obtained from mined data in order to select the top 5000 Simplified Chinese characters, Traditional Chinese characters, and Japanese kanji characters. We then forced their inclusion, as long as they appeared at least 15 times in our training data to guarantee that the model would be able to learn how to embed these tokens.

We re-trained a 256K-sized *SentencePiece* vocabulary on NLLB data [NLLB Team et al., 2022] for SeamlessM4T. The resulting tokenizer improves the coverage of the MTSU top 5K Chinese characters from 54% to 84%.

4.2.2 Training a Large-Scale Multilingual Text-to-Text Translation Model

We follow the same data preparation and training pipelines from NLLB Team et al. [2022] using Stopes [Andrews et al., 2022]. Having a smaller language coverage (100 instead of NLLB's 200 languages) allowed us to significantly decrease the size of the model. Whereas the full NLLB-200 model with mixture-of-experts is made up of 54.5B parameters (a number which can later be decreased via distillation), we opted for one of the smaller architectures proposed in NLLB Team et al. [2022], the 1.3B dense model. We limited the NLLB-200 training data to the 95 SeamlessM4T languages to be supported as target text. We additionally included over 75M bitexts from open-source T2TT datasets that were not included in NLLB Team et al. [2022]. These concern Modern Standard Arabic (arb), Mandarin Chinese (cmn), French (fra), Russian (rus), and Spanish (spa).

11. `https://lingua.mtsu.edu/chinese-computing/statistics/index.html`

Model	T2TT (↑chrF++)	
	X–eng *(n=95)*	eng–X *(n=95)*
NLLB Team et al. [2022]		
- 3.3B	60.6	**49.6**
- 1.3B	59.3	48.2
- 1.3B-distil.	59.5	48.8
SeamlessM4T-NLLB-1.3B	**60.7**	**49.6**

Table 12: Average Flores devtest chrF++over the 95 supported languages.

We compare in Table 12 the performance of SeamlessM4T-NLLB to that of comparably-sized NLLB models on Flores, averaging over our 95 languages when translating from English (eng–X) and into English (X–eng). The model outperforms both smaller models from NLLB-200 (1.3B and 1.3B-distil) and is on par with the larger 3.3B model.

4.2.3 Multimodal & multitask into target text

In SeamlessM4T, we leveraged foundational models either pre-trained on unlabeled data (w2v-BERT 2.0 for speech encoder pre-training) or trained on supervised high-resource tasks (NLLB model for T2TT) to improve the quality of transfer tasks (speech-to-text and speech-to-speech). To fuse these pre-trained components and enable meaning transfer through multiple multimodal tasks, we trained an end-to-end model with (a) a speech encoder (w2v-BERT 2.0) postfixed with a length adapter, (b) text encoder (NLLB encoder), and (c) a text decoder (NLLB decoder). For the length adaptor, we used a modified version of M-adaptor [Zhao et al., 2022], where we replaced the 3 independent pooling modules for Q, K, and V with a shared pooling module to improve efficiency.

The model is fine-tuned to jointly optimize the following objective functions:

$$\mathcal{L}_{\text{S2TT}} = -\sum_{t=1}^{|y|} \log p(y_t^{\text{text}} | y_{<t}^{\text{text}}, x^{\text{speech}}), \tag{3}$$

$$\mathcal{L}_{\text{T2TT}} = -\sum_{t=1}^{|y|} \log p(y_t^{\text{text}} | y_{<t}^{\text{text}}, x^{\text{text}}), \tag{4}$$

where x^{text} and x^{speech} are the source text and speech in the source language $<\ell_s>$ and y^{text} is the target text in the target language $<\ell_t>$. We additionally optimize an auxiliary objective function in the form of token-level knowledge distillation ($\mathcal{L}_{\text{KD}}$), to further transfer knowledge from the strong MT model to the student speech translation task (S2TT).

$$\mathcal{L}_{\text{KD}} = \sum_{t=1}^{|y|} D_{\text{KL}} \left[p(.|y_{<t}^{\text{text}}, x^{\text{text}}) \,\|\, p(.|y_{<t}^{\text{text}}, x^{\text{speech}}) \right]. \tag{5}$$

The final loss is a weighted sum of all three losses: $\mathcal{L} = \alpha\mathcal{L}_{\text{S2TT}} + \beta\mathcal{L}_{\text{T2TT}} + \gamma\mathcal{L}_{\text{KD}}$, where α, β, γ are scalar hyperparameters tuned on the development data. When the task does not fit the design of data triplets, we then replaced translation tasks with auto-encoding—for example, on ASR y^{text} is replaced by x^{text} in which case the teacher distribution is from auto-encoding $(p(.|x^{\text{text}}_{<t}, x^{\text{text}}))$.

We trained our X2T model in two stages. Stage$_1$ targeted training on supervised English ASR and into English S2TT data. We find that this step is necessary not only for improving the quality of X–eng translations but also eng–X translations. In fact, we hypothesized that allowing the model to focus on one target language while fine-tuning multilingual speech representations shields it from the interference that can propagate back from the target side. In Stage$_2$, we add supervised eng–X S2TT and non-English ASR data to the mix.

4.3 Speech-to-Speech Translation

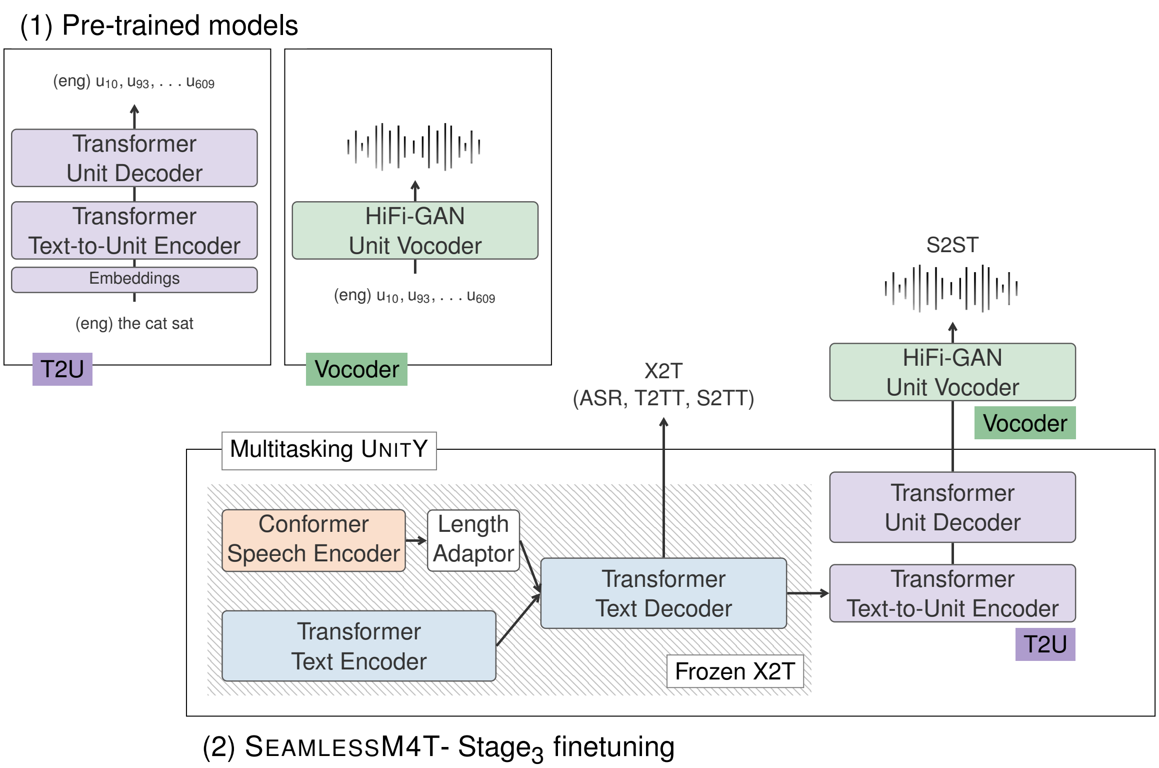

Figure 7: Overview of the SeamlessM4T multitask UnitY model. (1) describes the additional two building blocks on top of X2T: T2U encoder-decoder and unit vocoder. (2) describes the training of the UnitY model. In Stage$_3$, the model is trained on S2ST data.

The key to our proposed speech-to-speech translation model is the use of self-supervised discrete acoustic units to represent target speech, thereby decomposing the S2ST problem into a speech-to-unit translation (S2UT) step and a unit-to-speech (U2S) conversion step. For S2UT, the SeamlessM4T model depicted in Figure 4 uses UnitY as a two-pass decoding framework which first generates text and subsequently predicts discrete acoustic units. Compared to the vanilla UnitY model [Inaguma et al., 2023], (1) the core S2TT model initialized from scratch is replaced with an X2T model pre-trained to jointly optimize T2TT, S2TT, and ASR, (2) the shallow T2U model (referred to as T2U unit encoder and second-pass unit decoder in Inaguma et al. [2023]) is replaced with a deeper Transformer-based encoder-decoder model with 6 transformer layers, (3) the T2U model is also pre-trained on the T2U task rather than trained from scratch. The pre-training of X2T yields a stronger speech encoder and a higher quality first-pass text decoder, while the scaling and pre-training of the T2U model allowed us to better handle multilingual unit generation without interference.

4.3.1 Preparing S2ST data

Discrete acoustic units Recent works have achieved SOTA translation performance by using self-supervised discrete acoustic units as targets for building direct speech translation models [Tjandra et al., 2019; Lee et al., 2022a,b; Zhang et al., 2022; Chen et al., 2023c]. We extracted features from the 35^{th} layer of XLS-R-1B [Babu et al., 2022] for continuous speech representations at a 50Hz frame rate. The mapping from XLS-R continuous representation space to discrete categories is required to map target speech into a sequence of discrete tokens. We randomly selected and encoded 10K unlabeled audio samples from each language of the 35 supported target languages. We then applied a k-means algorithm on these representations to estimate K cluster centroids [Lakhotia et al., 2021; Polyak et al., 2021; Lee et al., 2022a]. These centroids resemble a codebook that is used to map a sequence of XLS-R speech representations into a sequence of centroid indices or acoustic units. Experiments with different numbers of centroids ($K \in \{1000, 2000, 5000, 10000\}$) show that K=10000 with features from the 35^{th} layer of XLS-R-1B achieves the best speech re-synthesis WER [Polyak et al., 2021].

XLS-R has a broader language coverage than existing HuBERT [Hsu et al., 2021] models, and we found it provided similar speech re-synthesis performance to HuBERT on overlapping languages. We also experimented with w2v-BERT 2.0, which showed inferior performance. This can be attributed to w2v-BERT training with contrastive and MLM objectives, encouraging the model to only learn about semantic tokens rather than acoustic ones.

Synthesizing multilingual units with HiFi-GAN Following Gong et al. [2023], we built the multilingual vocoder for speech synthesis from the learned units. The HiFi-GAN vocoder [Kong et al., 2020] is equipped with language embedding to model the language-specific acoustic information. Moreover, to mitigate cross-lingual interference, language identification is used as an auxiliary loss in multilingual training. We used a combination of commissioned and publicly available datasets, including single-speaker and multi-speaker TTS datasets, to train the multilingual vocoder on 36 target languages capable of converting the discrete units predicted by our S2UT model into waveforms. Compared to monolingual vocoders, we increased the model capacity by doubling the embedding dimension for both the duration predictor and the speech-language identification (LID) classifier to reach 1280.

Pseudo-labeling with text-to-unit The insufficient amount of parallel speech-to-speech training data significantly limits the training of high-quality S2UT models. To overcome this data scarcity, it is common practice to use TTS models to convert text from speech-to-text datasets (see Section 4.2.1) into synthetic speech [Jia et al., 2019b; Lee et al., 2022a]. This synthetic speech is in turn converted into units using the previously described unit extraction pipeline. This two-step unit extraction process is a slow process and is harder to scale

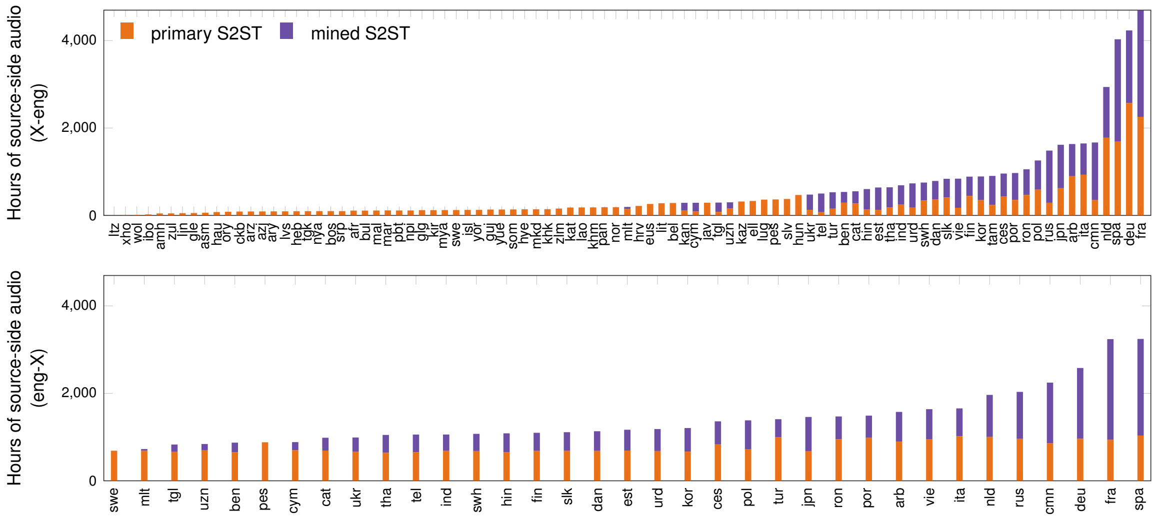

Figure 8: Statistics of S2ST data used in Stage_3 of training SeamlessM4T model. We show the data size in hours of speech between primary and mined. Languages are sorted in ascending resource-level. For numerical statistics see Table 36

given the dependencies on TTS models. High-quality off-the-shelf TTS models are hard to come by for all languages, especially for low-resource ones. Training reliable monolingual or multilingual in-house TTS models is also not scalable given the challenges around gathering high-quality clean speech data. To overcome these challenges, we circumvented the need for synthesizing speech and trained multilingual text-to-unit (T2U) models on all the 36 target speech languages. These models can directly convert the text into target discrete units and can be trained on ASR datasets that are readily available. The multilingual training benefits from cross-lingual transfer between high-resource and low-resource languages, thereby also improving the quality of the pseudo-labeled data. To remove outlier samples from our paired data, we filtered based on the number of seconds of audio generated per text token length ratio, discarding any pair with a ratio exceeding 0.5.

Parallel data mining: SeamlessAlign We added up to 2,500 hours of mined speech-to-speech data from SeamlessAlign per language direction depending on its availability (see Section 3) . We used the XLSR-based unit extraction pipeline for extracting discrete acoustic units for target speech from the mined data. An in-house ASR model is then deployed to generate text transcriptions for the first pass decoder based on the target speech.

Figure 8 shows the distribution of all S2ST data used to train SeamlessM4T models between the primary and mined data.

4.3.2 T2U modeling

The T2U model is a Transformer-based encoder-decoder model trained on aligned text units from ASR data. We trained T2U models for two purposes: (1) performing pseudo-labeling (Section 4.3.1) and (2) initializing the T2U component in UnitY. For (1), we trained a model with 12 encoder and 12 decoder layers. For (2), we trained a smaller T2U model with 6 encoder and 6 decoder layers. Initial experiments showed that, although the smaller T2U model is of a lower quality than the larger one, fine-tuning the smaller T2U in UnitY with labels from the larger one (i.e., distilling knowledge from the stronger T2U) can bridge the gap while being parameter-efficient.

4.3.3 Stage$_3$ Finetuning for S2ST

In the last stage of fine-tuning, we initialized the multitask UnitY model (see figure 4) with (1) the pre-trained X2T model and (2) the pre-trained T2U model and fine-tuned on a combination of X–eng and eng–X S2ST translation data totaling 121K hours (see breakdown in figure 8). We froze the model weights corresponding to the X2T model and only fine-tuned the T2U component. This is to ensure that the performance of the model on tasks from the previous stages of fine-tuning remains unchanged.

4.4 The SeamlessM4T Models

With all the components laid out in the previous sections, we trained the SeamlessM4T-Large model in the outlined three stages. SeamlessM4T-Large has 2.3B parameters and is fine-tuned on T2TT for 95 languages paired with English, on ASR for 96 languages, on S2TT for 89 languages paired with English, and on S2ST for 95 directions into English and 35 target languages out of English. The amount of supervised data per direction is detailed in tables 35 and 36. This means that, for some source languages, our models are evaluated zero-shot to reach the coverage described in table 2 of 100-eng.

To provide a reasonably sized model, we followed the same recipe to train SeamlessM4T-Medium. This model has 57% fewer parameters than SeamlessM4T-Large and is intended to be an accessible test bed to either fine-tune, improve on, or engage in analysis with. SeamlessM4T-Medium has the same coverage as SeamlessM4T-Large but builds on smaller and more parameter-efficient components (see Figure 4). We pre-trained a smaller w2v-BERT 2.0 with 300M parameters and used the distilled model from NLLB Team et al. [2022] (NLLB-600M-Distilled) to initialize the T2TT modules of the multitask UnitY. See a comparison between SeamlessM4T-Large and SeamlessM4T-Medium in Table 13.

	w2v-BERT 2.0*	**T2TT**	**T2U**	**Total**
SeamlessM4T-Large	669M	1370M	287M	2326M
SeamlessM4T-Medium	366M	615M	170M	1151M

Table 13: #parameters of the building components used in SeamlessM4T models.
*: includes the parameters of the length adaptor .

We evaluated our models on all four supervised tasks (T2TT, ASR, S2TT, and S2ST) as well as the zero-shot task of text-to-speech translation (T2ST, also referred to as cross-lingual text to speech synthesis [Zhang et al., 2023b]). To generate text hypotheses, we decoded with beam-search (width=5). We scored with chrF++for T2TT and SacreBLEU for S2TT (default 13a tokenizer and character-level tokenizer for Mandarin Chinese (cmn), Japanese (jpn), Thai (tha), Lao (lao), and Burmese (mya); see signatures in Table 4). For ASR, we scored with WER on normalized transcriptions and references following Radford et al. [2022].

During S2ST and T2ST inference, we performed two-pass beam-search decoding— the best hypothesis out of the first-pass decoding is embedded with the text decoder and is sent to T2U to search for the best unit sequence hypothesis. We use a beam-width of 5 for both searches. We evaluated S2ST and T2ST accuracy with ASR-BLEU [Lee et al., 2022a] with Whisper-Large-v2 as the underlying ASR model for eng–X directions and with Whisper-Medium for X–eng directions. We set the decoding temperature of Whisper at zero and used greedy decoding to ensure a deterministic behavior of the ASR model. The transcribed hypotheses, as well as the references, are normalized following Radford et al. [2022] before computing BLEU scores in the same manner we did for S2TT.

4.4.1 Comparison to cascaded approaches.

On the set of languages supported by both SeamlessM4T and Whisper, we compare in Table 14 the performance of our direct S2TT model to that of cascaded models, namely combinations of Whisper ASR models and NLLB T2TT models. SeamlessM4T-Large surpasses the cascaded models with less than 3B of parameters in X–eng directions by 2 BLEU points and in eng–X directions by 0.5 BLEU points. We also add to the comparison in Table 14 cascaded models with the large NLLB-3.3B T2TT model. These models exceed 4B parameters and only outperform SeamlessM4T-Large in eng–X directions. SeamlessM4T-Large improves on Whisper-Large-v2 +NLLB-3.3B by 1.3 BLEU points on average in X–eng directions.

Table 15 compares S2ST between SeamlessM4T-Large and cascaded models. For S2ST, we look at two options for cascading: (1) 3-stage with ASR, T2TT, and TTS and (2) 2-stage with S2TT and TTS. Our SeamlessM4T-Large outperforms 2-stage cascaded models on Fleurs X–eng directions by 9 ASR-BLEU points. It also outperforms stronger 3-stage cascaded models (Whisper-Large-v2 + NLLB-3.3B + YourTTS) by 2.6 ASR-BLEU points. On CVSS, SeamlessM4T-Large outperforms the 2-stage cascaded model (Whisper-Large-v2 + YourTTS) by a large margin of 14 ASR-BLEU points. On Fleurs eng–X directions, SeamlessM4T-Large has an average ASR-BLEU of 21.5 on 32 X–eng directions excluding target languages where Whisper-Large-v2 (the ASR model used for ASR-BLEU) has a WER higher than 100. Comparably, the medium-size model (SeamlessM4T-Medium) scores an average ASR-BLEU of 15.4 on S2ST eng–X directions.

4.4.2 Multitasking X2T results.

We report in Table 16 results on the Fleurs benchmark for the tasks of ASR and S2TT (X–eng and eng–X), and the related Flores benchmark for T2TT (X–eng and eng–X). We also report results on the evaluation test set of CoVoST 2 (X–eng and eng–X) The SeamlessM4T model outperforms the previous direct SOTA model (AudioPaLM-2 8B AST [Rubenstein et al., 2023]) by 4.2 BLEU points in S2TTX–eng directions (i.e., an improvement of 20%). In CoVoST 2 eng–X, SeamlessM4T-Large improves upon the previous SOTA (XLS-R) by 2.8 BLEU points. However, in X–eng, it lags behind AudioPaLM by 3.7 BLEU points. For ASR, SeamlessM4T outperforms Whisper [Radford et al., 2022]

12. Scoring Whisper-Large-v2, using `https://github.com/openai/whisper` with the recommended decoding options, results in BLEU scores lower by 0.3 BLEU points on average than what is reported in Radford et al. [2022].

Model	type	size	S2TT (↑BLEU) X–eng (n=81)	S2TT (↑BLEU) eng–X (n=88)
Whisper-Medium (ASR) + NLLB-1.3B	cascaded	2B	19.7	20.5
Whisper-Medium (ASR) + NLLB-3.3B	cascaded	4B	20.4	21.8
Whisper-Large-v2 (ASR)+ NLLB-1.3B	cascaded	2.8B	22.0	21.0
Whisper-Large-v2 (ASR)+ NLLB-3.3B	cascaded	4.8B	22.7	**22.2**
Whisper-Large-v2	direct	1.5B	17.9	-
AudioPaLM-2-8B-AST	direct	8B	19.7	-
SeamlessM4T-Medium	direct	1B	20.9	19.2
SeamlessM4T-Large	direct	2B	**24.0**	21.5

Table 14: Comparison against cascaded ASR +T2TT models on Fleurs S2TT.

Model	type	size	S2ST X–eng (↑ASR-BLEU) Fleurs (n=81)	S2ST X–eng (↑ASR-BLEU) CVSS (n=21)
YourTTS [Casanova et al., 2022]				
+Whisper-Large-v2 (S2TT)	2-stage cascaded	1.6B	17.3	22.6
+Whisper-Medium (ASR) + NLLB-1.3B	3-stage cascaded	2.1B	19.9	
+Whisper-Medium (ASR) + NLLB-3.3B	3-stage cascaded	4.1B	20.6	
+Whisper-Large-v2 (ASR)+ NLLB-1.3B	3-stage cascaded	2.9B	22.1	
+Whisper-Large-v2 (ASR)+ NLLB-3.3B	3-stage cascaded	4.9B	23.2	
SeamlessM4T-Medium	unified	1.2B	20.4	28.1
SeamlessM4T-Large	unified	2.3B	**25.8**	**36.5**

Table 15: Comparison against 2/3-stage cascaded models on Fleurs and CVSS S2ST X–eng.

on the overlapping 77 supported languages with a WER reduction of 45%. We additionally compared against MMS [Pratap et al., 2023] on Fleurs-54, a subset of Fleurs languages that both MMS and Whisper support. SeamlessM4T-Large outperforms the MMS variants evaluated with CTC by more than 6% WER, but it is surpassed by the variants that leverage monolingual n-gram language models (5% WER better).

In the T2TT support task, our SeamlessM4T model matches the performance of NLLB-3.3B [NLLB Team et al., 2022] in X–eng directions and improves on eng–X directions by 1 chrF++point. To further understand where the improvements in Fleurs S2TT X–eng directions are coming from, we bucket languages by resource-level (see the exact list of languages in Table 35) and report average BLEU scores per resource-level in Table 17. The results show that SeamlessM4T-Large strongly improves the quality of translating from low-resourced languages with an improvement of +7.4 BLEU (i.e., 40% improvement over AudioPaLM-2-8B-AST). We also average in column low† over low-resource directions that are supervised in AudioPaLM-2-8B-AST—the gain of +5 BLEU in that subset of

Model	size	S2TT (↑BLEU) Fleurs X–eng (n=81)	Fleurs eng–X (n=88)	CoVoST 2 X–eng (n=21)	CoVoST 2 eng–X (n=15)
XLS-R-2B-S2T	2.6B		x	22.1	27.8
Whisper-Large-v2	1.5B	17.9	x	29.1	x
AudioPaLM-2-8B-AST	8.0B	19.7	x	**37.8**	x
SeamlessM4T-Medium	1.2B	20.9	19.2	29.8	26.6
SeamlessM4T-Large	2.3B	**24.0**	**21.5**	34.1	**30.6**

Model	size	ASR (↓WER) Fleurs (n=77)	Fleurs-54 (n=54)	T2TT (↑chrF++) Flores X–eng (n=95)	Flores eng–X (n=95)
NLLB-3.3B	3.3B	x	x	60.7	49.6
Whisper-Large-v2	1.5B	41.7	43.7	x	x
MMS-L61-noLM-LSAH	1.0B	x	31.0	x	x
MMS-L1107-CCLM-LSAH	1.0B*	x	**18.7**	x	x
SeamlessM4T-Medium	1.2B	**21.9**	22.0	55.4	48.4
SeamlessM4T-Large	2.3B	23.1	23.7	**60.8**	**50.9**

Table 16: Multitasking X2T results. Performance of SeamlessM4T-Large on X2T tasks (S2TT, ASR and T2TT) compared to SOTA direct translation models. For Fleurs S2TT X–eng, we report the average BLEU scores over languages Whisper supports. For Fleurs ASR, we report the average normalized WER over languages supported by both SeamlessM4T and Whisper. For MT, we average chrF++ scores over the supported written languages in SeamlessM4T. *: MMS is CTC-based, and this version decodes with an n-gram language model for each language. Note that for all external models included in this comparison, we lifted the results reported in their respective papers and matched their evaluation and scoring pipeline for a fair comparison.[12]

Model	Fleurs S2TT X–eng (↑BLEU) High (n=15)	Medium (n=25)	Low (n=34)	Low† (n=23)
Whisper-Large-v2	24.2	19.4	16.1	18.1
AudioPaLM-2-8B-AST	**27.9**	20.9	18.0	22.0
SeamlessM4T-Medium	23.9	21.8	22.2	23.5
SeamlessM4T-Large	26.9	**25.2**	**25.4**	**27**

Table 17: Fleurs S2TT X–eng by resource-level. In each resource-level (high, medium and low), we average over languages that are covered by all 3 models. In low†, we exclude low-resource languages that are evaluated as zero-shot by AudioPaLM-2-8B-AST.

directions suggests that this improvement goes beyond sheer supervision, but instead should be attributed to the quality of supervised data and the training recipes.

Model	size	Fleurs X–eng (n=81)			Fleurs eng–X (n=88)		
		↑BLEU	↑spBLEU	↑Blaser 2.0	↑BLEU	↑spBLEU	↑Blaser 2.0
Whisper-Large-v2	1.5B	17.9	19.9	3.29	x	x	x
SeamlessM4T-Medium	1.2B	20.9	23.1	3.56	19.2	26.0	3.68
SeamlessM4T-Large	2.3B	**24.0**	**26.4**	**3.66**	**21.5**	**28.9**	**3.71**

Table 18: S2TT results with spBLEU and Blaser 2.0 we report here the performance of Whisper-Large-v2 and SeamlessM4T-Large measured with spBLEU & Blaser 2.0. Note that unlike BLEU scores copied from Radford et al. [2022], the spBLEU and Blaser 2.0 scores are based on our evaluation using `https://github.com/openai/whisper` with the recommended decoding options.

Model	S2ST (↑ASR-BLEU)				S2ST (↑Blaser 2.0)		
	Fleurs X–eng (n=101)	Fleurs X–eng (n=82)	Fleurs eng–X (n=35)	Fleurs eng–X (n=32)	Fleurs X–eng (n=82)	Fleurs eng–X (n=35)	Fleurs eng–X (n=32)
SeamlessM4T-Medium	17.9	20.8	14.3	15.4	3.62	3.63	3.63
SeamlessM4T-Large	22.7	26.3	19.8	21.5	3.85	3.94	3.95

Table 19: S2ST results with ASR-BLEU and Blaser 2.0 we report here the performance of SeamlessM4T-Large and SeamlessM4T-Medium measured with ASR-BLEU & Blaser 2.0.

4.4.3 Zero-shot Text-to-Speech Translation

We evaluate Fleurs S2TT on the reverse task of T2ST. We report in Table 20 the average ASR-BLEU scores on 87 X–eng directions (the overlap between Fleurs and the languages supported by SeamlessM4T text encoders). We also report the average ASR-BLEU on 32 eng–X directions (excluding Bengali, Telugu and Northern Uzbek where Whisper-Large-v2 ASR WER is above 100). The X–eng average ASR-BLEU is higher than the ASR-BLEU of S2ST X–eng (34.9 vs. 24.6) where the eng–X average is similar to that of S2ST (22.5 vs. 21.5). This result demonstrates that (1) the quality of SeamlessM4T on zero-shot T2ST is on-par with the supervised tasks and (2) that non-English speech source is the most challenging input to translate with our model.

Model	Fleurs T2ST (↑ASR-BLEU)		
	X–eng (n=88)	eng–X (n=35)	eng–X (n=32)
SeamlessM4T-Large	34.9	20.7	22.5

Table 20: zero-shot Fleurs T2ST we report the average ASR-BLEU of SeamlessM4T-Large on Fleurs T2ST.

4.4.4 Evaluation with spBLEU and Blaser 2.0.

To avoid expanding the set of special case languages evaluated with character-level tokenization, we evaluated with spBLEU using the Flores-200 sentence piece tokenizer. Table 18 reports spBLEU scores on Fleurs S2TT X–eng and eng–X. We also report in the same table the average Blaser 2.0 scores (for more on Blaser 2.0 see Section 5.1). Since Blaser 2.0 is modality-agnostic, we can also score the task of S2ST with Blaser 2.0. Table 19 provides the average Blaser 2.0 scores of SeamlessM4T-Large and SeamlessM4T-Medium on S2ST X–eng and eng–X directions. Since Blaser 2.0 supports 83 languages (including English), we average over 82 X–eng directions. For eng–X, we show averages of 35 languages, then averages excluding 3 languages with a WER exceeding 100%. Since Blaser 2.0 supports all 35 target languages the scores are more reliable and less affected by the noisiness of the ASR model underlying ASR-BLEU (a difference of -1.7ASR-BLEU points with the addition of 3 directions). The full results and metrics per evaluation direction can be found at `https://github.com/facebookresearch/seamless_communication`.

4.4.5 Evaluation of X–X directions with spBLEU.

Since SeamlessM4T models support multiple languages on both the source and target sides, we can evaluate non-English centric directions (labeled X–X) in a zero-shot manner.

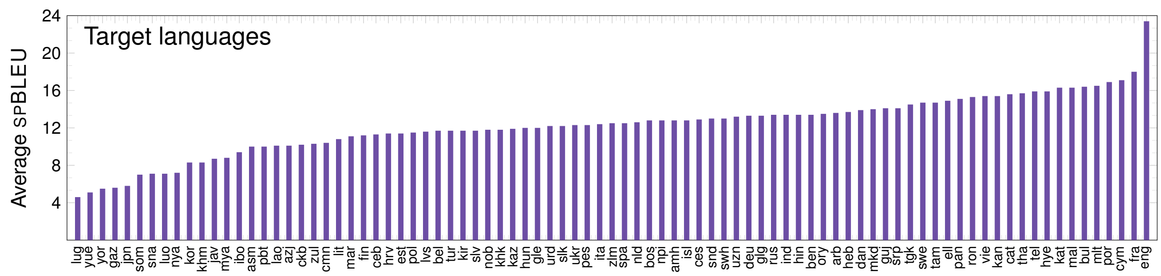

Figure 9: S2TT Fleurs X–X results. We evaluate X–X directions from Fleurs and average spBLEU scores. For a given target text language, we average scores over 100 source languages.

4.5 Analysis and Ablations

4.5.1 Unsupervised speech pre-training

We explored various techniques to enhance the quality of our encoders' representations, including algorithm-wise improvements and pre-training data scaling.

Experimental setup In our ablation, we aimed to evaluate the w2v-BERT variants by their performance on the downstream S2TT task. All pre-trained w2v-BERT speech encoders are composed of 24 Conformer layers [Gulati et al., 2020] with approximately 600M of parameters. Each speech encoder was used to initialize an S2TT model. The text decoder was initialized with the decoder from NLLB-1.3B, a large multilingual machine translation model covering 200 languages [NLLB Team et al., 2022] with 1.3B parameters. We fine-tuned the S2TT models on the task of speech translation into English (X–eng S2TT) on 67 languages. We fine-tuned all the speech encoder parameters and only fine-tuned

ID	Configuration	**Fleurs X–eng (↑BLEU)**
A	w2v-BERT baseline with updated XLS-R data (400K hrs, 143 langs)	12.4
B	A + product quantization with 2 GVQ codebooks	12.5
C	B + increased open training data from 400K hours to 1M hours	12.7
D	C + 2 RPQ codebooks for masked prediction objective	**12.8**

Table 21: Ablation on w2v-BERT variants and training data scaling.

LayerNorms and Self-attention in the text decoder (LNA-D [Li et al., 2021a]). Our learning rate increased up to 3e-4 through 4000 warm-up updates and subsequently followed the inverse square root learning rate schedule. We trained on 32 GPUs with a batch size of 960K frames in each for 100K updates. We report BLEU scores (SacreBLEU[13] [Post, 2018]) evaluated on the test set of all 101 X–eng directions from Fleurs [Conneau et al., 2022]. Given the coverage of our training data, this means that 34 of the directions were evaluated as zero-shot.

Results We summarize our ablation results in Table 21. We see that product quantization with 2 GVQ codebooks outperforms normal quantization with a single GVQ codebook (A vs. B). Scaling training data leads to performance gains (B vs. C). Adding additional masked prediction learning objectives with 2 RPQ codebooks helps improve performance (C vs. D).

Language (code)	**ASR**	**S2TT**				**S2ST**			
		X–eng		eng–X		X–eng		eng–X	
	Primary	Primary	Mined	Primary	Mined	Primary	Mined	Primary	Mined
arb	934	942	600	1,959	600	899	736	895	681
ben	338	320	600	1,987	499	292	246	652	221
eng	3,845	-	-	-	-	-	-	-	-
hin	148	143	600	2,066	600	138	466	656	430
ind	250	254	600	1,818	596	248	443	684	375
ita	591	910	600	2,279	600	930	716	1,020	636
jpn	381	15,141	600	1,798	259	624	993	681	779
por	269	246	600	2,250	600	355	606	983	508
rus	264	144	600	2,161	600	290	1,093	959	1,075
spa	1,515	1,285	-	2,505	574	1,694	2,335	1,035	2,209
swh	361	50	600	1,930	596	342	411	682	392
tha	190	59	600	1,941	101	184	462	641	408
tur	169	100	600	2,135	600	156	375	998	411
urd	185	145	600	1,844	507	179	555	682	502
vie	194	151	600	2,396	600	176	666	954	684
Total	9,633	19,890	7,800	29,068	2,701	6,508	10,103	11,523	9,312

Table 22: Hours of data in the ablation dataset for the tasks of ASR, S2TT and S2ST, split between eng–X and X–eng when relevant. For each task, we report hours of training data between primary and mined. By default, S2TT mined data is capped at 400 hours in X–eng and at 200 hours in eng–X.

13. see Table 4

4.5.2 Multimodal & multitasking X2T

Ablation dataset To iterate on different multitasking recipes, we constructed a smaller multilingual speech translation benchmark with 14 languages paired with English. The supervised S2TT data comes from two sources: primary (open-source or licensed) and mined, whereas the ASR data is either from open-sourced or licensed datasets. The T2TT data we used in our multitasking fine-tuning is limited to bitexts produced in the pseudo-labeling process, i.e., translated transcriptions in the ASR datasets (see Section 4.2.1). For a breakdown of the ablation dataset, see Table 22.

Experimental setup We fine-tuned multilingual translation models on our ablation dataset with different mixes of tasks. As a baseline, we only trained on primary S2TT data (eng–X + X–eng), optimizing L1: $\mathcal{L}_{\text{S2TT}}$ exclusively. With the data fixed, we experimented with two other objectives to optimize: (L2) with joint optimization of T2TT and S2TT ($\mathcal{L}_{\text{S2TT}} + \mathcal{L}_{\text{T2TT}}$) and (L3) with the additional knowledge distillation objective with T2TT as the teacher and S2TT as the student. We then added more data, namely ASR data and mined data respectively, and compared the performance of models trained with different objectives in the three data setups.

We initialized X2T models with our w2v-BERT 2.0 speech encoder and SeamlessM4T-NLLBT2TT model. We fine-tuned all parameters in the speech encoder and text encoder, while only fine-tuning LayerNorms and Self-attentions in the text encoder (LNA [Li et al., 2021b]). We trained all models for 100K updates (corresponding to 5-7 epochs). To regularize our models, we applied LayerDrop (p=0.1) to the speech encoder with masking (p=0.1). For the text encoder-decoder, we applied regular dropouts ($p = 0.1$). We evaluated the last checkpoint on development data and evaluated BLEU scores on Fleurs dev for translation tasks (including T2TT) and Whisper-style normalized WER for ASR.

Results Within each data setup (D1, D2 or D3), we see in Table 23 that adding T2TT to the multitasking loss, as expected, helps the performance on T2TT (+1.8 BLEU on average D1,2,3). Without adding this loss, fine-tuning exclusively on S2TT leads to catastrophic forgetting of the pre-training T2TT task (comparing L1 to L2). However, the accuracy of S2TT is seldom affected by this joint training with T2TT. Knowledge distillation is proving to be a necessary ingredient to leverage joint fine-tuning with T2TT. After adding knowledge distillation (L1 to L3), S2TT's performance improves by 0.6 BLEU points on average (D1,2,3).

If we compare the three different data setups, adding ASR data is crucial to supporting the ASR task as evaluating it as zero-shot leads to 3× higher error rates. Joint fine-tuning with T2TT and the auxiliary knowledge distillation loss has no negative effect on ASR given that for ASR data, the teacher task is auto-encoding (see Section 4.2.3). Adding mined S2TT data for which the source text is not available for T2TT to teach S2TT, still helps S2TT in the M3 task mix. We note, however, that the accuracy of T2TT drops as we add more speech-text only data (ASR and mined S2TT) without the aligned text-text data.

4.5.3 Leveraging Mined Speech-Text Data

Experimental setup We fine-tuned S2TT models on increasing amounts of mined data from SeamlessAlign. On top of the primary S2TT data, in the first model, we add 200

Data	D1: S2TT data			D2: D1+ASR data			D3: D2+Mined data		
Task	S2TT *(n=28)*	ASR* *(n=15)*	T2TT *(n=28)*	S2TT *(n=28)*	ASR *(n=15)*	T2TT *(n=28)*	S2TT *(n=28)*	ASR *(n=15)*	T2TT *(n=28)*
Metric	↑BLEU	↓WER	↑BLEU	↑ BLEU	↓WER	↑BLEU	↑BLEU	↓WER	↑BLEU
L1: $\mathcal{L}_{\text{S2TT}}$	26.5	36.5	34.1	26.7	16.4	34.2	27.6	**15.8**	34.7
L2: L1 + $\mathcal{L}_{\text{T2TT}}$	26.6	36.4	**36.8**	26.7	16.8	**36.1**	27.6	16.3	**35.4**
L3: L2 + $\mathcal{L}_{\text{KD}}$	**27.1**	**35.9**	36.7	**27.2**	**16.2**	**36.1**	**28.3**	**15.8**	35.3

Table 23: Ablations on multitasking objectives in three different data setups. Results are reported on Fleurs dev.

hours of mined data in each direction, 400 hours in the second and 600 hours in the last. SeamlessAlign is ranked based on Sonar scores and we selected the top ranking pairs up to the desired amount of additional data.

Results Table 24 reports the results of models trained with increasing amounts of mined data. The model trained with at most 400 hours in each direction achieves the best average BLEU score. This signals that some filtering of SeamlessAlign —e.g., based on Sonar similarity scores—can improve the quality of the model's translations without inflating the computational cost of training.

	X–eng *(n=14)*		eng–X *(n=14)*	
Data setting	↑BLEU	Δ	↑BLEU	Δ
Baseline	23.9		29.0	
+ 200H mined	25.9	+2.0	29.4	+0.4
+ 400H mined	**26.6**	**+2.7**	**29.8**	**+0.8**
+ 600H mined	26.0	+2.1	29.5	+0.5

Table 24: Ablations on the use of mined data. Results are reported on Fleurs dev.

4.5.4 T2U pre-training in UnitY

Experimental setup Similar to the ablation dataset described in Section 4.5.2, we built an S2ST ablation dataset with pseudo-labeled S2ST data (eng–X + X–eng) to fine-tune multilingual UnitY models. With the data fixed, we compare two options for using pre-trained components when fine-tuning UnitY. In the first (M1), we initialized the speech encoder with its adaptor and the first pass decoder with a pre-trained X2T model. In the second (M2), we additionally initialized the T2U of UnitY with a pre-trained T2U model. In both setups, we only fine-tuned the weights of the T2U model on S2ST data.

Results We evaluated our models on Fleurs dev for S2ST and report ASR-BLEU scores in Table 25. We note that T2U pre-training is beneficial for the fine-tuning of UnitY as it converges faster (comparing ASR-BLEU scores after 10K updates) and is, therefore, more computationally efficient.

4.5.5 Leveraging mined speech-to-speech data

To measure the impact of adding mined S2ST data to Stage$_3$ of UnitY fine-tuning, we compared model M2 from Section 4.5.4 to a model trained following the same training procedure, but with more mined data from SeamlessAlign (see amounts of additional data per direction in Table 22.

Results The results in Table 25 show that adding mined data improves eng–X translation accuracy by 0.5 ASR-BLEU points, but it decreases that of X–eng by 0.2. However, we do notice slight improvements in the quality of the speech generated and hence add SeamlessAlign for the final model training.

			Fleurs S2ST (↑ASR-BLEU)	
	Model	**updates**	X–eng	eng–X
M1	T2U from scratch	10K	6.9	1.8
		20K	23.3	12.4
M2	pre-trained T2U	10K	18.1	8.8
		20K	24.2	15.2
		50K†	**26.5**	18.6
M3	pre-trained T2U + Mined data	80K†	26.3	**19.1**

Table 25: Ablations on pre-training UnitY's T2U and use of S2ST mined data. Results are reported on Fleurs dev. † 80K and 50K correspond to 2 epoches in the two different data settings.

4.6 Related work

Two-pass sequence generation Two-pass decoding has the advantage of maintaining end-to-end optimization capability while inheriting the benefits of a cascading approach. Xia et al. [2017] and Hu et al. [2020] incorporate an additional search process to find a better output. Dalmia et al. [2021] re-ranks the intermediate hypotheses using an external module such as a language model. Zhao et al. [2019] injects specific information in the intermediate decoder to bias the output toward the desired domain. Sainath et al. [2019] provides an intermediate output to users before generating the final output for streaming applications. The two-pass approach makes the optimization tractable and results in better speech translation performance [Sung et al., 2019; Anastasopoulos and Chiang, 2018].

Codec-based audio modeling In contrast to acoustic units extracted from SSL-based audio representation models (e.g., XLS-R in this work), recent advances in quantized, audio codec auto-encoders enabled successful research combining large, autoregressive language models and audio data. Open-source EnCodec [D'efossez et al., 2022] and proprietary SoundStream [Zeghidour et al., 2022] models are widely known examples of quantized audio auto-encoders. One advantage of codec-based units is that they can be converted back to the waveform without needing an externally trained vocoder.

In speech translation research, VaLLE [Wang et al., 2023a] introduced the conditional autoregressive modeling of EnCodec-based audio data to perform text-to-speech synthesis.

VaLLE-X [Zhang et al., 2023b] subsequently built upon VaLLE to scale language coverage and enable language translation using a model cascade. VIOLA [Wang et al., 2023c] later explored the ability of decoder-only codec-based LM to translate without cascades.

Multimodality & multitask for speech & text Multimodality and multitask on the source side are orthogonal to multitask learning with two-pass decoding, where the goal is to provide the second task with higher-level representations produced from the first task decoder Anastasopoulos and Chiang [2018].

In general, multitask learning aims to improve generalization by leveraging domain-specific information contained in the training signals of related tasks [Caruana, 1997; Vandenhende et al., 2021]. Compared with single tasks, multitasking has the potential to improve performance by sharing complementary information or acting as a regularizer. Maninis et al. [2019], Liu et al. [2019], and Pfeiffer et al. [2020] introduced task-dependent components to enhance individual task performance. Weiss et al. [2017b] explored different multitask training strategies for speech translation, and they find the one-to-many strategy, in which an encoder is shared between the speech translation and ASR tasks, is more effective. Bahar et al. [2019] and Tang et al. [2021] compared different multitask strategies for S2TT, and confirmed the effectiveness of many-to-one training, in which T2TT and S2TT are trained together and the decoder is shared between two tasks.

Recent works have also trained multitask and multimodal encoders by learning joint representations of multiple modalities. The motivation is that the learned features will be richer and that inter-modal tasks can benefit from such joint training. These techniques were explored in audio [Chen et al., 2022; Bapna et al., 2022; Zhang et al., 2023a; Rubenstein et al., 2023], in vision [Chen et al., 2020; Gan et al., 2020; Fu et al., 2021], as well as audiovisual [Shi et al., 2022; Anwar et al., 2023].

5. Automatic and Human Evaluation

Up to this point, to evaluate our model, we have used standard automatic evaluation metrics for each particular task as reported in Table 4. In this section, for the tasks of S2TT and S2ST, we extend beyond these standard automatic metrics to include additional automatic and human evaluations to further assess our contributions. Automatic evaluations in this section reflect the robustness of our models in terms of noise and domains. Human assessment focuses on the preservation of speaker intention, as well as the subjective quality of the audio generated. To start, we introduce BLASER 2.0, a new, modality-agnostic evaluation metric that enables quality estimation for both speech and text.

5.1 Modality-Agnostic Automatic Metric: BLASER 2.0

Description BLASER 2.0 is the new version of BLASER [Chen et al., 2023a], which works with both speech and text modalities—hence being modality-agnostic. Like the first version, our approach leverages the similarity between input and output sentence embeddings. The new version uses SONAR embeddings (3.3.1), supports 83 languages in the speech modality and 200 in text, and is extendable to future encoders for new languages or modalities that share the same embedding space. For the purposes of evaluating speech outputs (and unlike ASR-based metrics), BLASER offers the advantage of being text-free.

More specifically, in BLASER 2.0, we take the source input, the translated output from any S2ST, S2TT, or T2TT model, and the reference speech segment or text, and convert them into SONAR embedding vectors (h_{src}, h_{mt}, and h_{ref}, respectively). For the supervised version of BLASER 2.0, these embeddings are combined and fed into a small, dense neural network that predicts an XSTS score for each translation output. For the unsupervised version, we use, similar to Chen et al. [2023a], the average of source-translation and reference-translation cosine similarities.

In addition, we trained a reference-free version of the system called BLASER 2.0-QE (for Quality Estimation). BLASER 2.0-QE is a supervised model trained only with source and translation embeddings. It can be applied in settings where reference translations are missing or if their quality is questionable.

Data The supervised version of BLASER 2.0 was trained on the XSTS-annotated data (Licht et al. [2022]), which includes the same S2ST annotations as in the original BLASER (Chen et al. [2023a]). Additional S2ST, S2TT, and T2ST annotations come from a variety of other internal studies, including NLLB human evaluations NLLB Team et al. [2022], and T2TT annotations are drawn from NLLB (NLLB Team et al. [2022]). We filtered out all audio longer than 30 seconds because the SONAR encoders were not trained on long audio.

For the original BLASER data, train/test splits were reused. The other datasets were split randomly in 80/20 proportion so that the same source audio or text always goes to the same partition. Details on the data are reported in Table 26.

Data part	**test size**	**train size**	**systems**	**langs**	$\rho_{\textbf{unsup.}}$	$\rho_{\textbf{sup.}}$	$\rho_{\textbf{QE}}$
BLASER 1.0 S2ST data	9804	10690	10	9	0.51	0.56	0.53
Other S2ST data	5453	15904	8	13	0.47	0.48	0.38
S2TT and T2ST data	5205	10246	7	8	0.49	0.54	0.51
T2TT data	20311	86776	2	59	0.49	0.61	0.60
All data	40773	123616	24	62	0.51	0.59	0.56

Table 26: The data for BLASER 2.0: test and train size, number of systems and languages, Spearman correlation of unsupervised, supervised, and reference-free BLASER 2.0 scores with XSTS labels on the test subset.

Training For the supervised model, the architecture is the same as for the BLASER 1.0 model: a 3-layer perceptron with *tanh* activations on top of 6 concatenated vectors of normalized embeddings and their derivatives: $[h_{ref}; h_{mt}; h_{src} \odot h_{mt}; |h_{src} - h_{mt}|; h_{ref} \odot h_{mt}; |h_{ref} - h_{mt}|]$. For the QE version, we used the same settings but with reference-free inputs: $[h_{src}; h_{mt}; h_{src} \odot h_{mt}; |h_{src} - h_{mt}|]$.

We used the training code for BLASER 1.0 with a few modifications in the hyperparameters intended to mitigate overfitting: 50% dropout, 0.1 weight decay, batch size of 1024, and full linear decay of the learning rate by the end of the training. To compensate for the increased batch size, we trained for 50 instead of 20 epochs.

Results Table 27 presents the performance of unsupervised and supervised BLASER on the BLASER 1.0 test data. The unsupervised 2.0 model slightly outperforms its predecessor. The

Model	↑**Pearson Correlation**						**average**
	eng-deu	eng-spa	eng-fra	spa-eng	fra-eng	rus-eng	
BLASER 1.0 unsup	0.32	0.58	0.64	0.50	0.48	0.43	0.49
BLASER 2.0 unsup	**0.37**	0.75	0.71	0.59	0.57	0.49	0.58
BLASER 2.0 QE	0.34	0.73	0.71	0.54	0.48	0.45	0.54
BLASER 1.0 sup	0.33	0.75	0.71	0.58	**0.57**	**0.53**	0.58
BLASER 2.0 sup	0.36	**0.75**	**0.73**	**0.58**	0.56	0.50	**0.58**

Table 27: Pearson correlations of unsupervised and supervised BLASER models with XSTS scores on the BLASER 1.0 test data.

supervised v1.0 and v2.0 models have the same average correlation with human judgments. Because BLASER 2.0 supports more languages, we used this for evaluations.

The last three columns in Table 26 present correlations of the 2.0 model's predictions with XSTS scores for all data partitions. Based on the results, the supervised model outperforms the unsupervised on each partition. The reference-free model scores between them in most cases, but for the new S2ST data, its performance is below that of the unsupervised model. We hypothesized that on this subset, references sometimes diverge from the sources, either due to errors of speech segmentation or synthesis, or due to non-literal translation that makes sense only in the context. A manual examination of a few samples corroborates this hypothesis, but more analysis of the role of reference in BLASER models is required in the future. Full BLASER 2.0 scores for SeamlessM4T models are reported in table 18. Additionally, the next section 5.2 reports the corresponding correlations of BLASER 2.0 scores with human scores.

5.2 Human Evaluation

Human evaluation is a vital tool in assessing the quality of our systems. We first briefly describe related work in the area, followed by a detailed description of the entire human evaluation process, including protocols, data, and calibration.

Related Work. Human evaluation has been widely applied to machine translation in the scientific community. Two of the most popular models of human evaluation are deployed within the context of International Evaluation Campaigns. The WMT conference [Kocmi et al., 2022] asks participants to evaluate the outputs of translation systems using a predefined protocol, typically that of Direct Assessment [Graham et al., 2013]. Beyond this text-based evaluation, the IWSLT Evaluation Campaign covers speech translation. As an example, the speech-to-speech track[14] evaluates speech output quality in four dimensions. The first one is translation quality, which focuses on capturing meaning, and annotators rank target audio between 1 and 5. The rest of the dimensions cover naturalness, including voice and pronunciation, clarity of speech for understandability, and sound quality, which takes into account noise and other artifacts. These criteria are used as an alternative to the Mean Opinion Score (MOS).

14. https://iwslt.org/2023/s2s

5.2.1 Human Evaluation Protocols

Similar to the related work aforementioned, for S2TT evaluation, we used the XSTS protocol to assess translation quality. We defer discussion of S2ST results to a later update, but we do evaluate S2ST using two protocols: XSTS for translation quality, and MOS to assess naturalness. We defer discussion of the MOS protocol to a later paper update.

XSTS. XSTS [Licht et al., 2022] evaluates translation quality in terms of semantic meaning preservation, and has previously been used to evaluate the NLLB models [NLLB Team et al., 2022]. While XSTS was originally designed to evaluate text, the protocol is effectively modality agnostic, and we required only small adaptations in order to support S2ST and S2TT tasks. For instance, the S2ST and S2TT versions of the protocol required additional instructions for annotators regarding the treatment of non-speech tags (e.g. `<laugh>`)—which the annotators were instructed to ignore—and how they should consider pauses and non-speech noises (they are instructed to ignore these as well). On the execution logistics side, conversations with vendors used for our evaluation work indicated that the evaluation of S2ST translations seemed to require a higher cognitive load for the annotators than T2TT (as a result of not being able to experience the source and target simultaneously), and thus was slower to conduct.

XSTS annotation and calibration process. During annotation, 3 annotators examined each source-target audio pair (or audio-text pair) and evaluated the item for semantic similarity using the XSTS protocol. Prior to annotating, all annotators went through a set of monolingual English 'practice' evaluations with score justifications. To expedite evaluation, more than 3 (up to 24) annotators were used per language pair; each evaluated sentence pair was shown to 3 annotators, assigned essentially randomly, and with calibration set items intermixed in the evaluation. In cases where the 3 annotators had a disagreement of score values of 2 or more, 2 additional annotators evaluated the same item again, bringing the total to 5 evaluator scores for those items. The median score over annotators of the same audio pair was then taken for each evaluation sentence pair; the median is used for robustness. The process was the same for both S2ST and S2TT evaluations. For overall direction scores, we report the mean of this median score (or some aggregate, such as the fraction of sentences with a median XSTS score above a given threshold) across all evaluated items in the dataset generated by a particular system in a language direction. Calibration set items received the exact same treatment, resulting in 1 score per sentence pair per annotator pool, and language-level scores were calibrated using the mean score on the calibration set for the crew of annotators evaluating a given language direction; the calibration set and methodology is described below.

To enable interlanguage comparison of model quality, a mono-lingual "cross-lingual calibration set" [Licht et al., 2022] was generated and included in the evaluation, and scores were calibrated using the 'moderated calibration' methodology established previously [Licht et al., 2022; NLLB Team et al., 2022]. The calibration process was found to reduce language-level annotator biases and has been shown to improve correlation with automatic metrics as a result. Running a calibration set or 'gold set' of items with a known score, even one much-reduced in size (e.g., 50–100 items instead of the 500 here) is useful as a diagnostic tool for ensuring annotation quality, even if one is not intending on doing interlanguage calibration. Annotator crews sufficiently 'out of calibration' can be identified, and their results excluded, or additional training can be conducted to improve their performance.

5.2.2 Evaluation Framework

Dataset Human evaluations were conducted utilizing the 'test' partition of the Fleurs dataset [Conneau et al., 2022]. The Fleurs 'test' partition provides up to 350 sentences sourced from the FLORES-101 dataset [Goyal et al., 2022] for each supported language (Fleurs supports 102 languages). Each sentence has up to 3 recorded audios spoken by different speakers (depending on which recordings passed quality review), along with the associated FLORES-101 text. The quality review requirement means that each language may not have a recording for all the 350 sentences, and that for those sentences that do have recordings, not all three speaker recordings may be present.

When conducting an evaluation of a translation system for a particular language direction, we filtered down the Fleurs data to a subset of sentences that have recordings in both languages in order to have a common, bidirectional evaluation set per-language pair. We do this to ensure S2TT and S2ST evaluations both use an identical set of sentences. Because the coverage of Fleurs varies per language, the subset of items present in the evaluation set varies per language and thus also per language pair; though there is a majority of items common across languages, and we believe the scores to be largely comparable as they were drawn from the same domain.

When preparing Fleurs to be used as a human reference set, pairings had to be made between distinct readers in the source language and readers of the equivalent Fleurs item in the target language. When possible, these pairings were made to match user gender (53% of the time over the entirety of the Fleurs test partition, varying significantly between languages paired with English), and mixed-gender matches had to be made for the remaining 47% of items. We elected to limit human evaluation to 2 unique readings per Fleurs sentence at most.

Language directions, modalities, and systems evaluated We list the languages and modalities evaluated with each protocol in Table 28.

Language selections were made by balancing a mix of resource availability for human annotations, a language sample that captured a large human population while also representing a mix of high and mid-resource languages.

For S2TT, we have XSTS evaluations of 22 languages in the X–eng direction for both the SeamlessM4T-Large and Whisper-Large-v2 models, where generations from the SeamlessM4T-Large model were made using a slightly earlier version via `fairseq` (instead of Fairseq2) but S2TT performance has less than 0.5 BLEU average difference. For eng–X, we have evaluations for the same languages but only for the SeamlessM4T-Large model (with `fairseq` generations). Additionally, we have evaluations for a human reference system (i.e. the Fleurs data itself) for all languages in each direction.

We only evaluated direct models for S2TT and plan to extend the benchmarking to 2 stage cascaded systems in future work to also include benchmarks for eng–X. For S2ST, we do not evaluate eng–X benchmarks due to the complexity involved in running monolingual TTS models for all the target directions. However, in the future, we plan to build such baselines using systems like MMS-TTS Pratap et al. [2023]; we have experimented using these

Modality	Protocol	Direction	Systems	Languages
S2TT	XSTS	X–eng	Whisper-Large-v2 SeamlessM4T-Large *[1]	22[2]
S2TT	XSTS	eng–X	SeamlessM4T-Large *	22
(S2ST)	(XSTS)	X–eng	Whisper-Large-v2 +YourTTS SeamlessM4T-Large	22
(S2ST)	(XSTS)	eng–X	SeamlessM4T-Large	22
(S2ST)	(MOS[3])	X–eng	Whisper-Large-v2 +YourTTS SeamlessM4T-Large	8 (arb, cmn, fra, hin, rus, spa, tel, tur)
(S2ST)	(MOS)	eng–X	SeamlessM4T-Large	22

[1] SeamlessM4T-Large * refers to the SeamlessM4T-Large model using `fairseq` for generations instead of Fairseq2, but S2TT performance was on average within 0.5 BLEU between the two.

[2] Bengali, Catalan, Dutch, Finnish, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Mandarin Chinese, Modern Standard Arabic, Portuguese, Romanian, Russian, Spanish, Swahili, Thai, Turkish, Urdu, Vietnamese

[3] MOS refers to the Mean Opinion Score protocol; more details will follow in a later update containing S2ST evaluations.

Table 28: Summary of evaluations: languages, modalities, models, and protocols used in human evaluations. Modalities and protocols in parentheses are not presented in this paper but will be shared in a later update.

same systems in other sections of this paper, e.g. for the purpose of extending text-based responsible AI datasets to speech (Section 6.3.2).

5.2.3 Preliminary Human Evaluation Results

XSTS results for the S2TT task We present results for the S2TT modality using the XSTS protocol (see Table 28).[15] Figure 10 shows calibrated XSTS scores on the language level for all models and languages evaluated (both X–eng and eng–X). We see that for X–eng language directions, SeamlessM4T-Large quality was consistent with being above an XSTS score of 3 for all 22 evaluated language directions. For eng–X language directions, SeamlessM4T-Large was consistent with being above an XSTS score of 4 for all 22 evaluated language directions.

Notably, in the X–eng direction, we see that SeamlessM4T-Large improves translation quality considerably over the Whisper-Large-v2 baseline for Swahili (an XSTS improvement close to 2.5) and Bengali (an XSTS improvement above 1). SeamlessM4T-Large has significant improvements in language quality over Whisper-Large-v2 for 7 out of the 22 languages evaluated X–eng with regressions in 8; all regressions are smaller than 0.5 XSTS except for Japanese, which has a slightly larger regression.

15. The current section contains only S2TT results. An update containing results for all protocols and modalities enumerated in Table 28, including evaluations of the S2ST tasks, will be provided at a later date along with further analysis.

Direction	System	Avg XSTS	Avg. Items	% 3+	% 4+
X–eng	Human reference	4.67	450.6	99.3	95.3
	Whisper-Large-v2	4.09	450.6	84.6	71.7
	SeamlessM4T-Large	**4.21**	447.0	**88.2**	**73.9**
eng–X	Human reference	4.69	450.6	99.4	96.0
	SeamlessM4T-Large	4.53	445.6	95.9	87.5

Table 29: Overall average XSTS human evaluation results into and out of English, over all 22 evaluated languages. Results were computed for each language direction (see Table 30 for full language-level results). %3+ and %4+ refer to the percent of a language's evaluated sentences with median scores equal to or greater than 3 and 4 respectively.

When averaging over language directions, SeamlessM4T-Large demonstrated superior performance on both average XSTS score and % of sentences above XSTS thresholds of 3 and 4 compared to the Whisper-Large-v2 baseline on X–eng (see Table 29).

We also note generally higher performance in the eng–X direction compared to the X–eng direction. From the automatic results in section 4.4.2, we observe that the higher performance in one direction or the other varies depending on the task (S2TT, S2ST, T2TT or T2TT). For S2TT and in terms of spBLEU and Blaser 2.0 (see Table 18), even if averaging in a different set of languages, this out-performance of eng–X compared to X–eng holds. We hypothesize a few possible explanations for this phenomenon. For example, speech encoding may be a more complicated task than speech or text decoding. If this is the case, having a better performance in English speech encoding could contribute to having a higher performance in the eng–X direction. Data-wise, a plausible explanation could be a difference in audio quality of Fleurs recordings for different languages (e.g. English source sentence audio quality may have been higher, inflating the eng–X scores).

5.2.4 Limitations

Test set limitations The Fleurs [Conneau et al., 2022] test set used for evaluation has limitations in that different language pairs will be evaluated on slightly different sets of sentences, and due to limitations in both the dataset (which contains a maximum of 3 speakers) and timing and cost considerations on the human evaluation front (we evaluate a maximum of two speakers per sentence), we have a lack of diversity in our speaker set per language, which may introduce bias relative to a test set with a larger number of speakers.

Limited sample size of human annotators per language We only have a maximum of 5 (but typically 3) annotator evaluations per sentence for each language in our XSTS evaluations. Relatively small samples of annotators mean annotator bias is important to consider. We try to mitigate this by (1) using the median score per sentence for each language to be robust to outliers, (2) using bootstrap re-sampling of annotator scores to estimate language score uncertainty due to finite annotators, and (3) approximate and correct annotator bias with a cross-lingual calibration set.

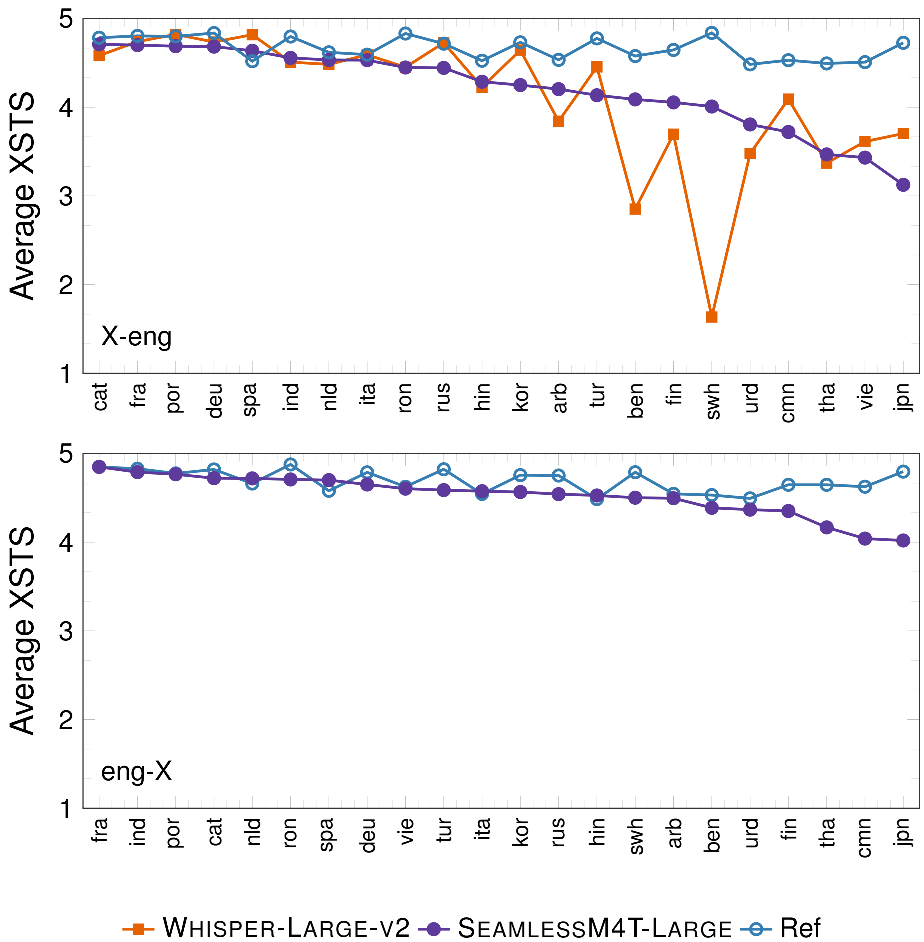

Figure 10: Language Direction level mean XSTS scores per direction for S2TT modality, after calibration. Bootstrapped 95% CI is typically within ±0.12.

Direction	Lang	Seamless[1]	Whisper[2]	Human[3]	Items	%3+	%4+
X–eng	arb	4.2	3.8	4.5	283	92	80
	ben	4.1	2.9	4.6	269	93	78
	cat	4.7	4.6	4.8	638	97	89
	cmn	3.7	4.1	4.5	349	75	56
	deu	4.7	4.7	4.8	347	96	89
	fin	4.1	3.7	4.6	632	78	61
	fra	4.7	4.7	4.8	332	97	89
	hin	4.3	4.2	4.5	388	94	87
	ind	4.6	4.5	4.8	544	94	88
	ita	4.5	4.6	4.6	612	98	94
	jpn	3.1	3.7	4.7	321	49	30
	kor	4.2	4.6	4.7	356	92	72
	nld	4.5	4.5	4.6	251	88	80
	por	4.7	4.8	4.8	632	97	92
	ron	4.4	4.5	4.8	619	91	77
	rus	4.4	4.7	4.7	344	88	78
	spa	4.6	4.8	4.5	348	97	87
	swh	4.0	1.6	4.8	466	87	60
	tha	3.5	3.4	4.5	643	73	47
	tur	4.1	4.5	4.8	566	92	70
	urd	3.8	3.5	4.5	283	84	67
	vie	3.4	3.6	4.5	611	75	46
eng–X	arb	4.5	—	4.5	283	99	92
	ben	4.4	—	4.5	238	99	91
	cat	4.7	—	4.8	638	97	90
	cmn	4.0	—	4.6	349	87	69
	deu	4.7	—	4.8	347	96	89
	fin	4.4	—	4.6	632	89	75
	fra	4.8	—	4.8	332	99	95
	hin	4.5	—	4.5	388	100	98
	ind	4.8	—	4.8	544	98	97
	ita	4.6	—	4.5	612	99	97
	jpn	4.0	—	4.8	321	78	64
	kor	4.6	—	4.8	356	97	85
	nld	4.7	—	4.7	251	97	88
	por	4.8	—	4.8	632	98	95
	ron	4.7	—	4.9	619	97	89
	rus	4.5	—	4.8	344	92	81
	spa	4.7	—	4.6	348	98	89
	swh	4.5	—	4.8	466	95	81
	tha	4.2	—	4.6	643	91	78
	tur	4.6	—	4.8	566	98	88
	urd	4.4	—	4.5	283	96	91
	vie	4.6	—	4.6	611	98	92

[1] SeamlessM4T-Large using `fairseq` generations
[2] Whisper-Large-v2
[3] Human reference

Table 30: Full calibrated XSTS S2TT results; bootstrapped 95% CI widths are ±0.12 on average. %3+ and %4+ refer to the percent of a language's evaluated sentences with median scores equal to or greater than 3 and 4 respectively, and are not calibrated (calibration is only performed on the language level).

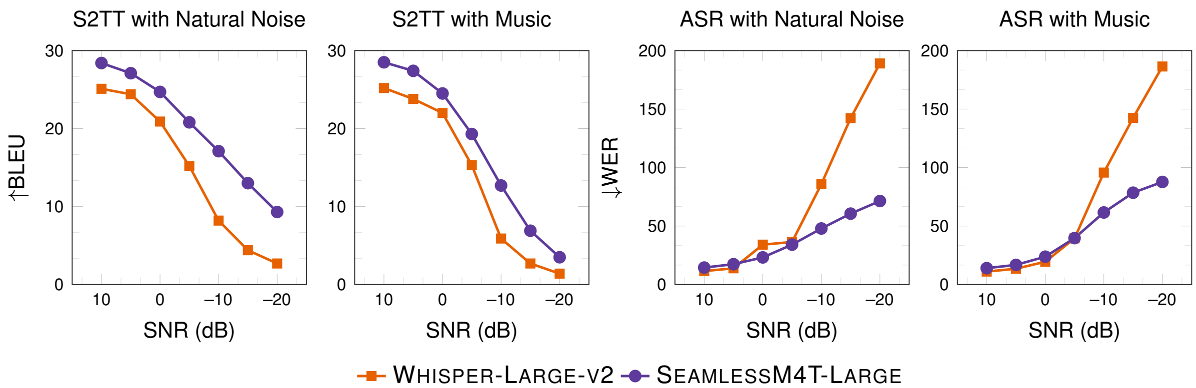

Figure 11: Evaluation results of model robustness against background noises. We report average test BLEU and test WER over 4 languages (3 language families) for X–eng S2TT and ASR on Fleurs with low-to-high input noise level (high-to-low SNR). Simulated noises are sampled from MUSAN [Snyder et al., 2015] on the "noise" and "music" categories.

5.3 Automatic Robustness Evaluation

We evaluate model robustness against non-linguistic perturbations in the real-world speech inputs, including background noises and speaker variations. As reported in several other sections, we compare our model to Whisper-Large-v2.

5.3.1 Robustness Against Background Noises

Related work The analysis of speech model robustness across different background noise levels has been conducted in prior work [Wang et al., 2022; Zhu et al., 2022; Radford et al., 2022] on simulated noisy audios. However, existing simulation-based evaluations are either limited by the noise types (e.g., simple white noise), task coverage (e.g., ASR only), language coverage (e.g., English only), or the replicability of benchmark data. This calls for an open, versatile benchmark to overcome these limitations.

Experimental framework We build a replicable noise-robustness evaluation benchmark based on Fleurs ("noisy Fleurs"), which covers 102 languages, 2 speech tasks (S2TT and ASR), and various noise types (natural noises and music). To create simulated noisy audios, we sampled audio clips from MUSAN [Snyder et al., 2015] on the "noise" and "music" categories, and mixed them with the original Fleurs speech audios under different signal-to-noise ratio (SNR): 10, 5, 0, -5, -10, -15 and -20. We compare models by BLEU-SNR curves (for S2TT) or WER-SNR curves (for ASR), which illustrate the degree of model performance degradation when the noise level of speech inputs increases (i.e., when SNR decreases). Both SeamlessM4T-Large and Whisper-Large-v2 achieve high performance mostly in high-resource languages, where stress tests in the noisy speech setup are more necessary and informative. On low-resource languages, the clean speech setup is already challenging, let alone the noisy one. We hence focus on 4 high-resource languages (French, Spanish, Modern Standard Arabic, and Russian) from 3 different language families for our noise-robustness analysis on SeamlessM4T-Large and Whisper-Large-v2.

Results Figure 11 shows the average test BLEU and test WER over the 4 languages for X–eng S2TT and ASR on FLEURS with low-to-high input noise level (high-to-low SNR). We see that both BLEU-SNR curves for SEAMLESSM4T-LARGE are consistently above those for WHISPER-LARGE-V2. Similarly, SEAMLESSM4T-LARGE's WER-SNR curves are consistently below WHISPER-LARGE-V2's ones. These suggest the superior robustness of SEAMLESSM4T-LARGE in noisy speaking environments. SEAMLESSM4T-LARGE outperforms WHISPER-LARGE-V2 by an average of 33.3% and 42.2% over various noise types and noise levels for X–eng S2TT and ASR, respectively.

5.3.2 ROBUSTNESS AGAINST SPEAKER VARIATIONS

Related work ASR and S2TT systems are expected to minimize the effects of speaker variations which are irrelevant to the input content of interest. Fairness of ASR systems to different speaker subgroups (by race, gender, country, etc.) has been studied in prior work [Liu et al., 2022; Dheram et al., 2022], which requires the availability of accurate speaker demographics labels [Hazirbas et al., 2021; Porgali et al., 2023] for speaker grouping and group-wise scoring. However, these labels are rare in existing ASR benchmarks, limiting the applications of such analysis. To overcome label scarcity, Wang et al. [2020] proposed a set of label-free metrics that do not rely on speaker grouping for analyzing the effects of speaker variations.

Experimental setup We follow Wang et al. [2020] to evaluate model robustness against speaker variations by calculating average by-group mean score and by-group coefficient of variation of an utterance-level quality metric. Instead of using BLEU as the quality metric, we used chrF, which has better stability at the utterance level. The calculation of both robustness metrics does not require explicit speaker subgroup labels. We grouped evaluation samples and corresponding utterance-level chrF scores by content (transcript), and then calculated the average by-group mean score chrF_{MS} and average by-group coefficient of variation $\mathrm{CoefVar}_{MS}$ defined as follows:

$$\mathrm{chrF}_{MS} = \frac{1}{|G|}\sum_{g\in G}\mathrm{Mean}(g)$$

$$\mathrm{CoefVar}_{MS} = \frac{1}{|G'|}\sum_{g\in G'}\frac{\mathrm{StandardDeviation}(g)}{\mathrm{Mean}(g)}$$

where G is the set of sentence-level chrF scores grouped by content (transcript) and $G' = \{g|g \in G, |g| > 1, \mathrm{Mean}(g) > 0\}$. The two metrics are complementary: chrF_{MS} provides a normalized quality metric that, unlike conventional corpus-level metrics, takes speaker variations into consideration, while $\mathrm{CoefVar}_{MS}$ provides a standardized measure of quality variance under speaker variations. For robustness analysis of SEAMLESSM4T-LARGE and WHISPER-LARGE-V2, we conducted an out-of-domain evaluation on FLEURS on all its languages that have at least 40 content groups in the test sets.

Results Table 31 shows the chrF_{MS} and $\mathrm{CoefVar}_{MS}$ scores of SEAMLESSM4T-LARGE and WHISPER-LARGE-V2 on FLEURS X–eng S2TT and ASR test sets. We see that SEAMLESSM4T-LARGE outperforms WHISPER-LARGE-V2 on $\mathrm{CoefVar}_{MS}$ by an average of 49.4% over the 2 tasks. Moreover, SEAMLESSM4T-LARGE outperforms WHISPER-LARGE-V2

Languages	Average #	Whisper-Large-v2		SeamlessM4T-Large	
(≥ 40 content groups)	cont. groups	chrF$_{MS}$↑	CoefVar$_{MS}$↓	chrF$_{MS}$↑	CoefVar$_{MS}$↓
X–eng S2TT for 77 langs	278	40.8	13.7	**45.3**	**9.1**
ASR for 78 langs	280	58.7	17.0	**72.5**	**6.4**

Table 31: Evaluation results of model robustness against speaker variations. We report average by-group mean chrF (chrF$_{MS}$) and average by-group coefficient of variation on chrF (CoefVar$_{MS}$) on FleursX–engS2TT and ASR test sets.

on chrF$_{MS}$ by an average of 18.3%. These suggest the superior robustness of SeamlessM4T-Large when it comes to speaker variations.

6. Responsible AI

In line with our expectations to build systems responsibly, we focus our efforts on the evaluation of added toxicity and bias. Both of these dimensions of responsible AI have drawn significant scientific attention in recent times (e.g., [Kiritchenko et al., 2021; Bender et al., 2021; Costa-jussà, 2019]). Moreover, the occurrence of these unintended errors or translation faults could adversely impact user experiences. Sustained attention devoted to such issues is, thus, vital to the safe deployment of our systems.

Beyond these dimensions, we are also concerned with the concept of fairness. In contrast to the idea of robustness (as conceptualized in section 5.3.2), where the focus is on whether our system performance is affected by the varying qualities of a speaker's voice, fairness in this section is more concerned about the *content* of the translation outputs. Fair outputs do not preference or skew towards particular demographics and tend to treat different groups somewhat equitably. We document the results of these evaluations to better direct mitigation efforts.

6.1 Definitions

We begin by detailing how we define errors that arise from *added toxicity* and *gender bias*.

Toxicity. In their taxonomy of critical machine translation errors, [Sharou and Specia, 2022] define "deviation in toxicity" as "instances where the translation may incite hate, violence, profanity, or abuse against an individual or a group (a religion, race, gender, etc.) due to incorrect translations," which "covers cases where toxicity is introduced into the translation when it is not in the source, deleted in the translation when it is in the source, mistranslated into different (toxic or not) words, or not translated at all (i.e., the toxicity remains in the source language or transliterated)." Our definition of *added toxicity* departs slightly from theirs in that it does not cover instances of untranslated toxic source content or of toxic source content deleted in the translation. To put it simplistically, added toxicity is the introduction of toxic elements not present in a source utterance.

Gender Bias. Another error with which responsible AI is concerned lies in the propagation and amplification of gender bias. In machine translation, gender bias is observed when translations show errors in linguistic gender determination despite the fact that there are sufficient gender clues in the source content for a system to infer the correct gendered forms. To illustrate this phenomenon, sentence (1) below does not contain enough linguistic clues for a translation system to decide which gendered form should be used when translating into a language where the word for *doctor* is gendered. Sentence (2), however, includes a gendered pronoun which most likely has the word *doctor* as its antecedent.

1. I didn't feel well, so I made an appointment with my doctor.
2. My doctor is very attentive to **her** patients' needs.

Gender bias is observed when the system produces the wrong gendered form when translating sentence (2) into a language that uses distinct gendered forms for the word *doctor*. A single error in the translation of an utterance the like of sentence (1) would not be sufficient to conclude that gender bias exists in the model; doing so would take consistently observing one linguistic gender over another. It has previously been hypothesized that one possible source of gender bias is gender representation imbalance in large training and evaluation data sets, e.g. [Costa-jussà et al., 2022; Qian et al., 2022].

6.2 Toxicity

Warning: this section contains examples that may be offensive to some.

6.2.1 Motivation

Context As mentioned above, added toxicity means introducing toxicity in the translation output not present in the input. This can be classified as a critical error; one that could lead users to distrust a translation system. As such, it is important to quantify how much toxicity our models add. We are also interested in combining added toxicity analysis with demographic bias analysis to determine whether added toxicity is generated more in certain demographic axes than in others.

Related work While related research in speech toxicity detection is quite limited [Iskhakova et al., 2020; Yousefi and Emmanouilidou, 2021], toxicity detection for text-based approaches has been widely explored in different contexts. Many examples of these efforts can be found in large evaluations like JigSaw Series Kaggle Competitions[16] or WMT Critical Error detection [Specia et al., 2021]. Recently, in the context of T2TT, there has been a substantial push to scale toxicity detection by using a word-list-based detection method for models such as NLLB [NLLB Team et al., 2022], which further spurred research into analyzing toxicity at scale [Costa-jussà et al., 2023] and mitigation strategies [Gilabert et al., 2023]. Using a dataset that covers different demographic axes can allow for further analysis of which demographic axes are most sensitive to toxicity [Costa-jussà et al., 2023]. So far, datasets that cover a wide range of demographic axes mostly focus on text and more attention needs to be directed at speech (an example of a text data is HolisticBias [Smith et al., 2022]).

Proposed methodology Inspired by ASR-BLEU, this work proposes using ASR-ETOX as a new metric to detect added toxicity in speech and evaluate added toxicity for SeamlessM4T's S2ST capability. Essentially, this metric follows a cascaded framework by first

16. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

deploying a standard ASR module (i.e., the same that it is used for ASR-BLEU as defined in Table 4), then the toxicity detection module, ETOX [Costa-jussà et al., 2023], which uses the Toxicity-200 word lists. For S2TT, the translated output can be directly evaluated with ETOX. In both cases (S2ST and S2TT), we measure added toxicity at the utterance/sentence level. We first compute toxicity detection for each input in the evaluation dataset and the corresponding output. Then we compare them and count a case as containing added toxicity only when the output value exceeds the one displayed by the input.

6.2.2 Experimental Framework

Language directions and modalities Similarly to the previous human evaluation framework in Section 5.2, we evaluated S2ST and S2TT on Fleurs. Distinctive from human evaluation, we extended toxicity evaluation to cover all languages for which we provide translations for as summarized in Table 5. Igbo, Burmese, Nepali, and Assamese have issues related to segmentation and consistencies in the toxicity word lists.With these problems, these languages tend to over-detect toxicity and we consider them to be outliers. Therefore, we excluded them from the analysis and results.

Datasets We used two datasets to analyze added toxicity. One, we used Fleurs to better align with our human evaluation effort and other evaluative components of this work. In addition, we used the English-only HolisticBias framework [Smith et al., 2022], which has been shown to trigger true added toxicity in previous studies [Costa-jussà et al., 2023]. HolisticBias comprises 26 templates, encompassing more than 600 descriptors across 13 demographic axes, along with 30 nouns. The dataset consists of over 472K English sentences utilized in the context of two-person conversations. Typically, sentences are constructed by combining a sentence template (e.g., "I am a [NOUN PHRASE]."), a noun (e.g., parent), and a descriptor (e.g., disabled). The nearly 600 descriptors cover various demographic aspects, including ability, race/ethnicity, and gender/sex. The nouns may indicate a specific gender (e.g., woman, man) or avoid gender references (e.g., child, kid). Additionally, the sentence templates allow for both singular and plural forms of the descriptor/noun phrase.

In this work, we extend HolisticBias to speech by applying the default "en" transformer-tts model from fairseq Sˆ2 ([Wang et al., 2021a]). It first converts input texts into IPA phonemes, then passes them to a mel spectrogram generator transformer model , and finally feeds the outputs to a HiFi-Gan vocoder to create the waveform.

Models As a baseline system for S2TT X–eng, we employ Whisper-Large-v2 [Radford et al., 2022]. As for S2ST X–eng, we apply the Casanova et al. [2022] to generate synthesized speech from the output of Whisper-Large-v2 S2TT. For S2TT eng–X, we employ the cascade system of Whisper-Large-v2 + NLLB-3.3B [NLLB Team et al., 2022]. Below, we report results for SeamlessM4T-Large.

Evaluation We use the Github implementation of ETOX[17] For languages without spaces, we use the spm tokenization option in the tool. For ASR, we use the same implementation framework used for ASR-BLEU as reported in Table 4.

17. https://github.com/facebookresearch/stopes/tree/main/demo/toxicity-alti-hb/ETOX

6.2.3 Results

Automatic toxicity detection on Fleurs We evaluated the output of SeamlessM4T-Large on the Fleurs dataset. Figure 12 presents results from S2TT and S2ST for X–eng and eng–X directions, where we show the number of sentences that contain added toxicity. When looking at the amount of added toxicity per sentence, less than 5% of the cases contain more than 1 added toxicity token per sentence. Overall, Fleurs shows a relatively low prevalence of added toxicity of 0.15%, averaging across languages, tasks, and translation directions.

For S2TT in X–eng (Figure 12 (left)), added toxicity is 0.11% averaged across languages, and 27 language pairs contain some added toxicity. For S2ST (Figure 12 (right)), added toxicity is 0.12% averaged across languages, and 35 language pairs contain added toxicity.

For S2TT in eng–X, (Figure 12 (left)), added toxicity is 0.21% averaged across languages, and 32 language pairs contain added toxicity. For S2ST (Figure 12 (right)), added toxicity is 0.16% averaged across languages, and 16 language pairs contain added toxicity. The main difference across modalities is the reduced amount of added toxicity in S2ST for the eng–X translation direction. We comment on this difference alongside the results from the HolisticBias dataset later in this section.

By comparison, for S2TT in X–eng, Whisper-Large-v2's added toxicity is 0.31% averaged across languages and is prevalent in 58 languages. For overlapping languages in Whisper-Large-v2 and SeamlessM4T-Large, the latter shows an added toxicity reduction of 63%. For S2ST in X–eng, Whisper-Large-v2 + YourTTS's added toxicity lies at 0.27% averaged across languages and is prevalent in 52 languages.Again, for overlapped languages in this cascaded S2ST system and SeamlessM4T-Large, ours show a reduction of toxic tokens by 62%. For S2TT in eng–X, the Whisper-Large-v2 + NLLB-3.3B cascaded combination adds toxicity by 31% averaged in languages and added toxicity is prevalent in 39 languages. For overlapping languages, SeamlessM4T-Large reduces this amount by 26%. The filtering of imbalanced toxicity in the training data as reported in Section 4.2.1 may have contributed to this improvement.

Automatic Toxicity Detection on HolisticBias Dataset Figure 13 (left) shows results for S2TT languages with the highest added toxicity when translating HolisticBias from eng–X (note that HolisticBias is only available in English).Here, we observe a slightly higher amount of added toxicity compared to Fleurs for S2TT and a slightly lower amount for S2ST. Overall, HolisticBias shows a prevalence of added toxicity of 0.19% for S2TT and 0.13% for S2ST, averaged across languages. For S2TT, there are 84 languages that are affected by added toxicity. When looking at added toxicity per sentence, less than 0.003 % of the outputs contain more than one added toxicity token. Figure 14 (left) shows results for S2ST languages when translating the HolisticBias dataset. In total, there are 34 languages with added toxicity.

Through manual inspection, when comparing toxic words being detected in S2TT translation but not in S2ST, we observe that toxic words are similar with minor differences. We hypothesize that using ASR before toxicity detection tends to cause false negatives, which would explain the high decrease in added toxicity from S2TT to S2ST (from 0.19% to 0.13%), which also happened in Fleurs (from 0.21% to 0.16%). For example, in the case of English to Catalan, the word "merda" in the S2ST output is usually written as "mereda",

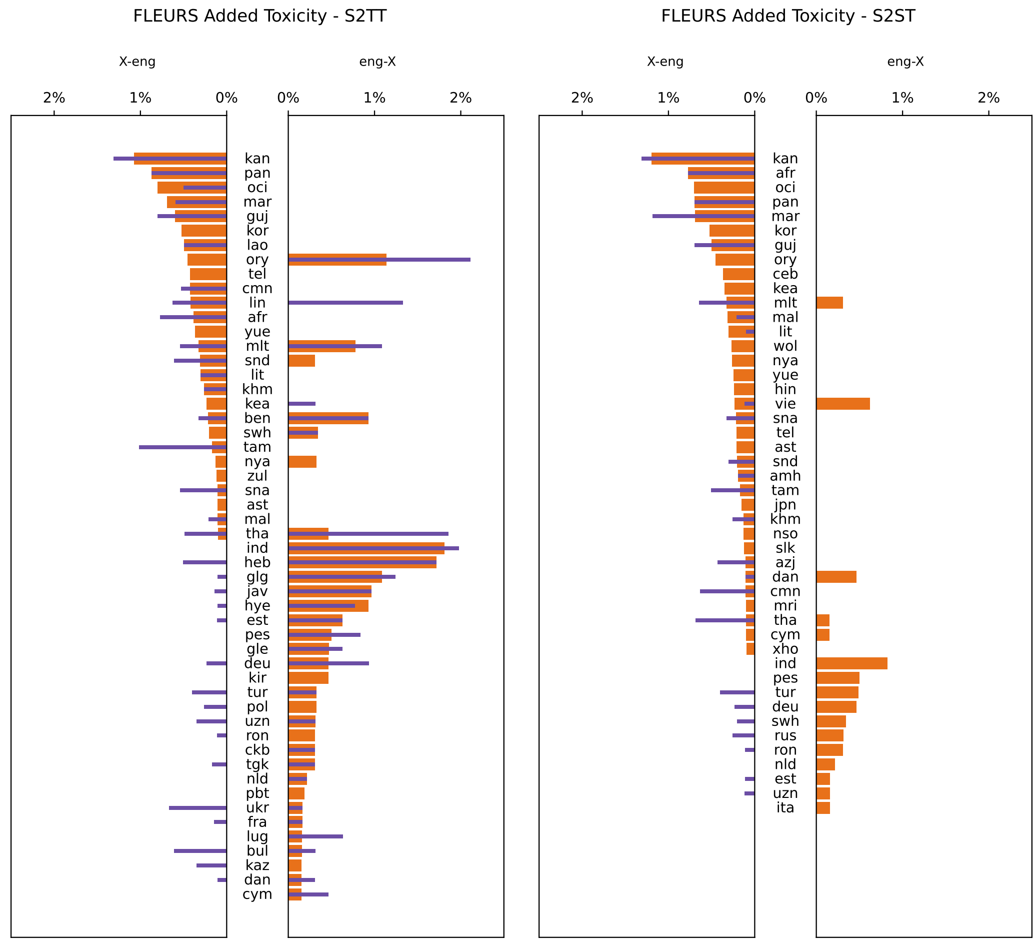

Figure 12: Added toxicity for X–eng and eng–X for S2TT (left) and S2ST (right) in Fleurs. The figure shows the number of outputs with added toxicity per language both for SeamlessM4T-Large (blue) and Whisper-Large-v2 and Whisper-Large-v2 + YourTTSsystems when available (orange).

and therefore not identified by ETOX. This type of example brings light to the limitations presented by detection based on tokens in a blacklist.

Following previous work [Costa-jussà et al., 2023], we perform an analysis of toxicity per HolisticBias' axes and report them in Figures 13 and 14 (right). Figures show the distribution of toxic translations per category and how they vary per language. We see that different languages differ in their distributions of toxic terms as a function of demographic axes. For most languages, the toxicity distribution across an axis is proportional to the axis' overall share. For instance, the main category in terms of volume is 'body type', representing 25% of the dataset. This same category tends to accumulate a larger amount of toxicity as well. However, for some languages the toxic sentences appear to be highly concentrated in a particular axis—such is the case for Bengali (80% socio-economic status), Nyanja (66% characteristics), and Kyrgyz (94% cultural) to name a few.

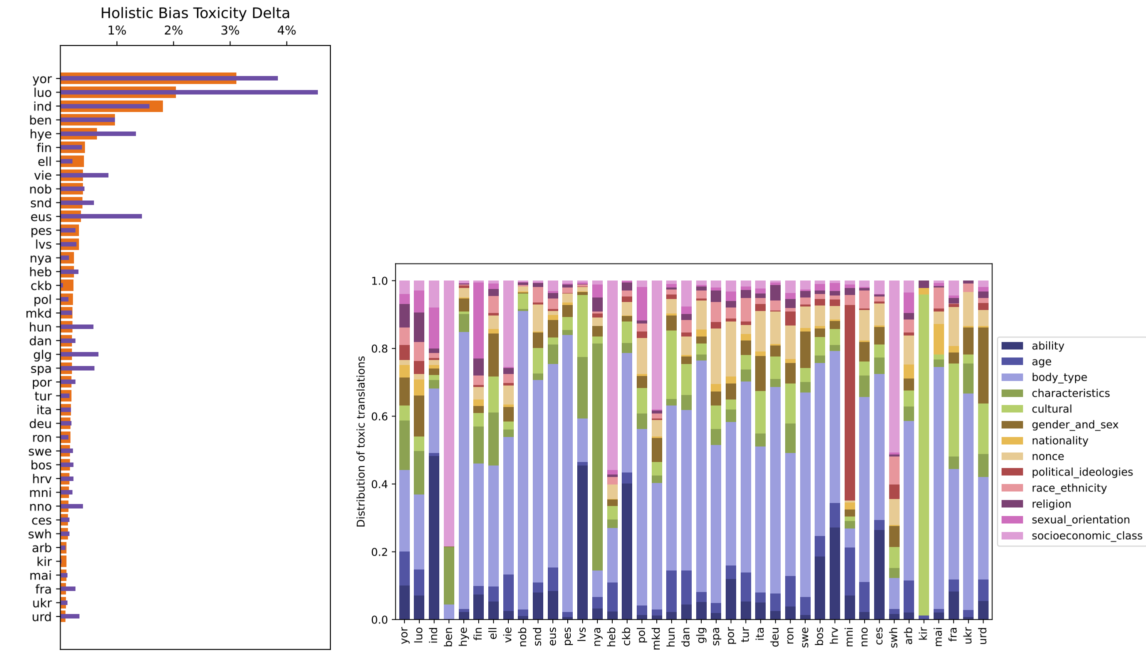

Figure 13: (left) Added toxicity for eng–X, S2TT in HolisticBias. Showing top 40 languages. The plotted languages are above 500 samples of added toxicity—0.1% of the dataset. (right) Different languages differ in distributions of toxic terms as a function of demographic axes, with some languages' toxicity being dominated by only one or two axes.

The categories that have a higher concentration of toxicity for S2TT and S2ST are nonce (0.79% and 0.46%) and sexual orientation (0.62% and 0.35%). Nonce category (nonsense) is a bit of an outlier as far as terms are concerned because they do not specifically refer to any demographic groups. In terms of categories for least added toxicity, those would be age for S2TT (0.37%), and political ideologies for S2ST.

6.2.4 Toxicity key findings and contributions

To summarize, our key findings and contributions include: (1) proposing a metric for speech toxicity detection for languages at scale (ASR-ETOX), (2) showing that while levels and types of added toxicity vary significantly as a function of language and dataset, added toxicity in our systems has a relatively low prevalence (varying from 0.11% to 0.21% across modalities, language directions, and datasets), and (3) our evaluation against the state-of-the-art shows that SeamlessM4T-Large reduces toxicity by 51% across modalities and language directions in Fleurs and by 34% in HolisticBias for eng–X in S2TT.

6.3 Bias

6.3.1 Motivation

Unequal training datasets can lead to demographic and representational biases that affect our models and their generated outputs. These biases can adversely impact users by perpetuating allocation biases when used in situated contexts. In recent years, the MT field has made

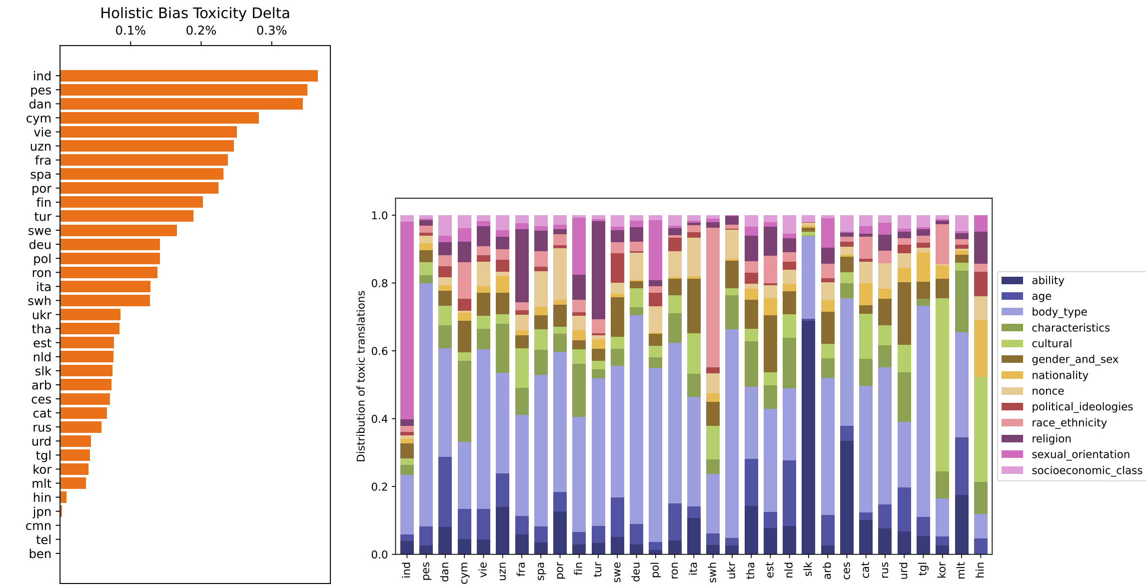

Figure 14: (left) Added toxicity for eng–X, S2ST in HolisticBias. Showing all target languages. (right) Similarly to S2TT, different languages differ in distributions of toxic terms as a function of the demographic axis, with some languages' toxicity being dominated by only one or two axes.

significant progress in uncovering [Prates et al., 2020], evaluating [Stanovsky et al., 2019; Renduchintala et al., 2021; Costa-jussà et al., 2022; Bentivogli et al., 2020], or even mitigating many of these forms of biases [Renduchintala and Williams, 2022]. However, much work lies ahead of us when it comes to this domain of research.

Related work Multilingual HolisticBias dataset [Costa-jussà et al., 2023] consists of an extension to HolisticBias. It contains translations for three different patterns and 118 descriptors, available in 50 different languages. Depending on whether gender inflection exists in a language, each language has one or two references. Each translated sentence includes the masculine, neutral and, when applicable, a feminine iteration. The dataset enables quantification of gender biases across demographic aspects for T2TT and has the highest language coverage at the time of writing. Previous work on this matter is mostly in text [Stanovsky et al., 2019; Renduchintala et al., 2021; Levy et al., 2021; Costa-jussà et al., 2022; Renduchintala and Williams, 2022] and tend to be English-centric, with few demographic axes and multilingual references. Similar efforts for the speech modality remain sparse [Costa-jussà et al., 2022; Bentivogli et al., 2020].

Contributions. In this work, we used Multilingual HolisticBias and its speech extension (described in the following section) to compare the performance of S2TT and S2ST. The eng–X direction allows comparing performance in the presence of masculine or feminine references, and the X–eng direction enables robustness comparisons in translations when we alter gender inflection. A typical example of the language pair of English-Spanish would be "I'm a homemaker" and the corresponding translations "Soy amo de casa" and "Soy ama de casa" in Spanish. When translating from English to Spanish, we can measure if the system overgeneralizes to one gender, while in the other direction, we can evaluate the robustness of the translation to gender inflection.

6.3.2 Bias Experimental Framework

Dataset: Speech Extension of Multilingual HolisticBias In order to compare the performances across modalities (S2ST and S2TT), we begin by extending the Multilingual HolisticBias dataset from text to speech by using the TTS model[18] provided by Pratap et al. [2023]. Due to the limitations of this TTS model in correctly generating speech for numbers, we manually converted all numerical numbers to words for each language. For instance, the sentence "I have friends who are 50 years old." is transformed into "I have friends who are fifty years old." After processing through TTS, we obtained the synthesized speech for 325 sentences across 19 languages. These languages are supported both by MMS-TTS and the Multilingual HolisticBias[19] dataset [Costa-jussà et al., 2023]. For each of these languages (except English), we generated two speeches, one for each set of gendered texts.

Language directions and modalities We use this generated TTS data as input for S2TT and S2ST and as a reference for S2ST. We conducted the translations in two directions—eng–X and X–eng. Concretely, in X–eng, we translated both masculine and feminine versions of the speech. It's worth noting that some target languages are not available in the SeamlessM4T S2ST model, so we performed translations on only 17 languages for the S2ST task in the eng–X direction. For S2TT in eng–X, we have all languages included in the Multilingual HolisticBias dataset (n=25). For reference, the complete language list used in our experiments can be found in Table 32.

<table>
<tr><th></th><th>X–eng</th><th>eng–X</th></tr>
<tr><td>S2ST</td><td rowspan="2">arb,bul,cat,deu,ell,fra,lvs,mar,nld, por,ron,rus,spa,swe,tam,tha,ukr,urd</td><td>arb,cat,ces,dan,deu,fra,ita,nld,por,ron, rus,slk,spa,swe,tha,ukr,urd</td></tr>
<tr><td>S2TT</td><td>arb,bel,bul,cat,ces,dan,deu,ell,fra,ita, lit,lvs,mar,nld,por,ron,rus,slk,slv,spa, swe,tam,tha,ukr,urd</td></tr>
</table>

Table 32: List of language codes in the bias evaluation experiments, organized by task and language direction.

Evaluation In terms of evaluation metrics for S2TT, we used chrF as reported in Table 4, except that nw:2 was changed to nw:0. Instead of using BLEU as the quality metric, we used chrF because it is more equipped to handle shorter utterances, which better suits the evaluation of the Multilingual HolisticBias dataset. This dataset is relatively small (325 utterances) and with short sentences (on average, 6 words per utterance) [Costa-jussà et al., 2023]. In this context, we find chrF more adequate for comparison [Ma et al., 2019], since BLEU quickly drops when not enough lengthy n-grams are matched. For S2ST,

18. https://github.com/facebookresearch/fairseq/tree/main/examples/mms#tts-1
19. Arabic, Belarusian, Bulgarian, Catalan, Czech, Danish, German, Greek, French, Italian, Lithuanian, Latvian, Marathi, Dutch, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, Tamil, Thai, Ukrainian, Urdu.

we used ASR-chrF.[20] and Blaser 2.0 proposed in this work. It is worth noting that when evaluating Blaser 2.0, we included only 14 languages (including English)[21] for the eng–X direction (overlaps between the languages from the generated TTS data and the languages available in our S2ST model). Additionally, since MMS-TTS generations are not deterministic, we repeated the measurements three times for both S2ST and S2TT. The final metric values are then averaged to ensure robustness and accuracy in our evaluations.

Models We used the SeamlessM4T-Large model and several different baselines. For X–eng S2TT, we employed Whisper-Large-v2 [Radford et al., 2022]. As for X–eng S2ST, we used YourTTS [Casanova et al., 2022] to generate synthesized speech from the output of Whisper-Large-v2 S2TT. For eng–X S2TT, we utilized a cascade system: ASR from Whisper Large-v2 [Radford et al., 2022], followed by T2TT via NLLB-3.3B [NLLB Team et al., 2022]. For SeamlessM4T-Large S2TT, we used a beam size of ten. For SeamlessM4T-Large S2ST, we set the beam size to five for both the first pass decoder and the second pass decoder. As for the baseline, we set the beam size to five for NLLB-3.3B and used the default values for Whisper-Large-v2 and YourTTS.

6.3.3 Bias evaluation results

This section focuses on analyzing gendered translations when using neutral inputs (eng–X) and the gap in translation performance between inputs that only differ in gender (X–eng).

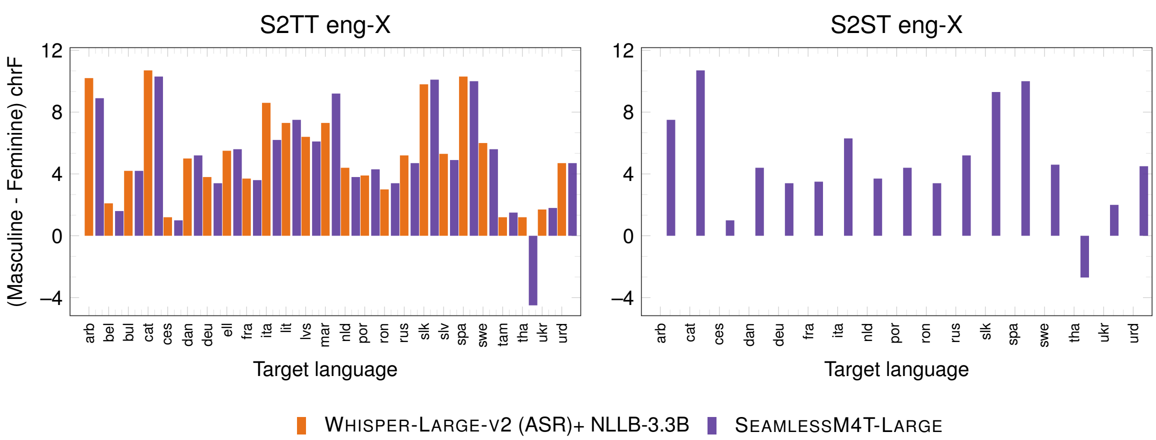

Figure 15: Left: The chrF points difference between masculine and feminine forms for eng–X S2TT using English speech as source and X text translation (masculine or feminine) as reference. Right: The ASR-chrF points difference between masculine and feminine forms for eng–X S2ST using English speech as source and X text translation (masculine or feminine) as reference.

eng–X. In our analysis, we utilize the masculine or the feminine human translations of the non-English languageas references. The source for this analysis is the English (eng) Multilingual HolisticBias dataset, comprising a collection of unique sentences with

20. The transcription is done by Whisper-Large-v2 and Whisper-Medium [Radford et al., 2022] for eng–X and X–eng respectively. chrF has been calculated the same way as S2TT except that in S2ST the text from both prediction and reference are normalized.

21. The list of language codes for these 14 languages: arb,cat,deu,eng,fra,nld,por,ron,rus,spa,swe,tha,ukr,urd.

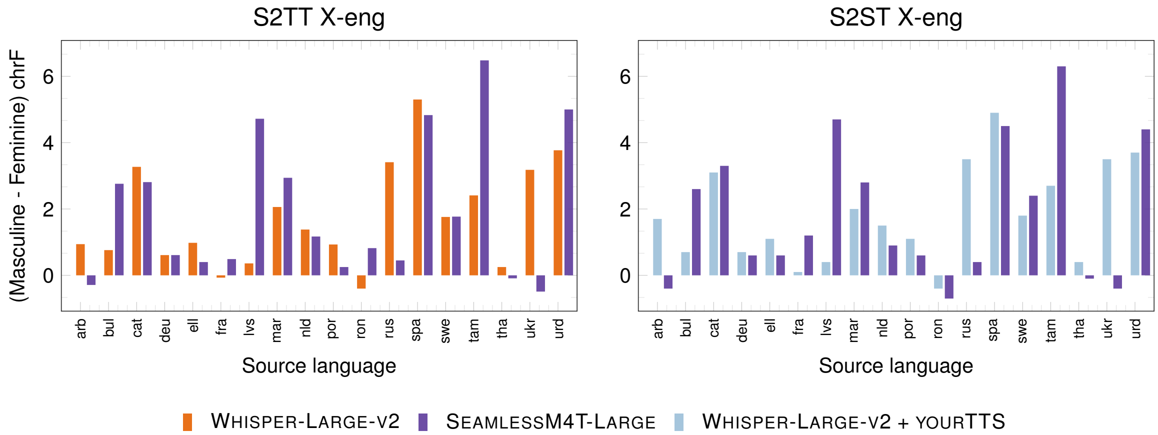

Figure 16: (left) The chrF points difference between masculine and feminine for X–eng S2TT using X speech synthesized by the masculine or feminine version of the text and English text as a reference. (right) The ASR-chrF points difference between masculine and feminine forms for X–eng S2ST using X speech synthesized by the masculine or feminine version of text and English text as reference.

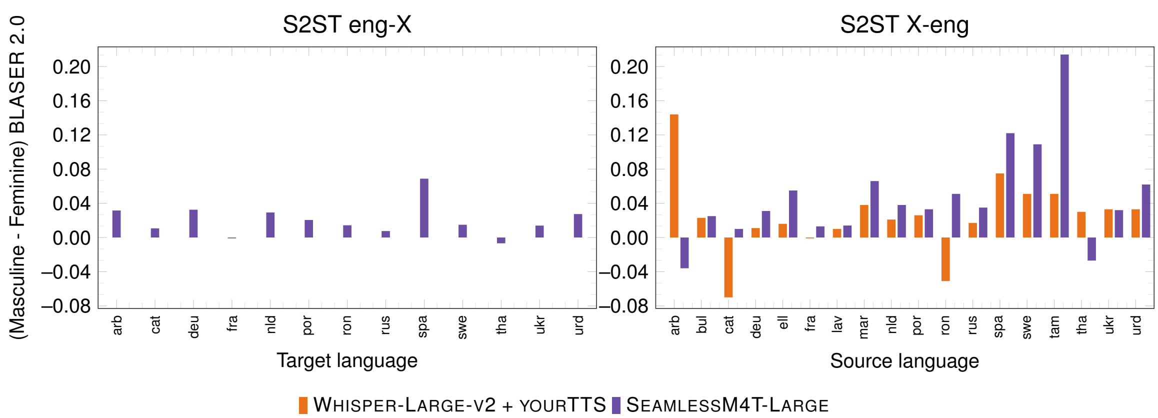

Figure 17: (left) The supervised Blaser 2.0 points difference between masculine and feminine forms for eng–X S2ST using English speech as the source and X text translation (masculine and feminine) as reference. The results are averaged from three experiments. (right) The supervised Blaser 2.0 points difference for X–eng S2ST using X speech synthesized by the masculine or feminine version of text and English text as reference.

ambiguous gender. Figure 15 shows the results per target language, evincing the following patterns:

- In SeamlessM4T-LargeS2TT, the translation quality deteriorates for all the languages except Thai when using the feminine reference, and is especially noticeable in languages like Catalan (with a significant 10.3 chrF points difference), Slovak (10.1), and Spanish (10.0). For the Whisper-Large-v2 + NLLB-3.3B combination, a decline in translation quality is observed across all languages. The highest differences are found in Catalan (10.7), Spanish (10.3), and Arabic (10.2). It's worth mentioning that the biases' distribution over languages is similar between SeamlessM4T-Large

	S2TT				
Axis	Masculine	Feminine	Average	Count	Diff
Cultural	11.4	9.5	10.4	350	1.9
Body type	14.2	12.9	13.6	3750	1.2
Socioeconomic class	14.6	13.3	13.9	400	1.3
Religion	15.5	13.7	14.6	1800	1.8
Gender and sex	16.0	15.1	15.5	1800	1.0
Ability	16.6	15.2	15.9	3300	1.3
Race ethnicity	17.4	15.7	16.5	900	1.7
Characteristics	18.2	16.2	17.2	1900	2.0
Nationality	18.1	16.7	17.4	300	1.4
Sexual orientation	18.5	16.7	17.6	700	1.8
Age	18.6	16.6	17.6	900	1.9
	S2ST				
Axis	Masculine	Feminine	Average	Count	Diff
Cultural	12.2	10.3	11.3	238	1.9
Body type	14.2	13.0	13.6	2550	1.2
Socioeconomic class	14.4	13.1	13.7	272	1.3
Religion	16.3	14.5	15.4	1224	1.9
Gender and sex	16.7	15.7	16.2	1224	1.0
Ability	16.9	15.5	16.2	2244	1.4
Age	17.7	15.8	16.7	612	1.9
Characteristics	17.7	15.9	16.8	1292	1.8
Race ethnicity	18.0	16.4	17.2	612	1.7
Sexual orientation	18.4	16.9	17.7	476	1.5
Nationality	18.7	17.3	18.0	204	1.3

Table 33: Results on mean per axis (across descriptor, template, and language): chrF on S2TT (top) and ASR-chrF on S2ST (bottom) results. Columns (from left to right): masculine references, feminine references, average between the two, the total number of measurements (Count) and the difference between masculine and feminine (Diff). The rows are sorted in ascending order by the average chrF for S2ST and S2TT, respectively. The axes are defined in HolisticBias —for more details, refer to Table 5 in the original paper [Smith et al., 2022].

and the Whisper-Large-v2 + NLLB-3.3B combination, with Thai being the only exception.

- In S2ST, we noticed similar trends in relation to S2TT, where translation quality is lowered in all languages (except Thai) when assessing with the feminine reference. The highest differences are with Catalan (10.7 ASR-chrF points difference), Spanish (10.0), and Slovak (9.3).

The left panel of Figure 17 shows the results for automatic speech evaluation by way of Blaser 2.0. We observe similar trends in the ASR-chrF metric. The translation quality deteriorates by an average of 0.02 supervised Blaser 2.0 points across languages when evaluating with the feminine reference for all languages except Thai. Interestingly, the evaluation for French reveals a negligible difference. The highest differences are found in Spanish (0.07), followed by German (0.03).

These differences show that when no gender information is available in the source sentence, the model will prefer to translate to the masculine form in the target language. Note that for some languages (like Spanish or French), the plural masculine form is indistinguishable from the plural generic form.

X–eng. Our main objective is to assess the translation quality when starting from a gendered sentence and translating it into English. As such, we aim to measure the model's robustness with regard to gender bias and its ability to handle translations between languages that mark grammatical gender towards English. Figure 16 shows the results per source language for SeamlessM4T-Large and Whisper-Large-v2 or Whisper-Large-v2 + YourTTS. We observe that:

- In S2TT, the performance is better when translating from the masculine reference for most languages (15 out of 18 for SeamlessM4T-Large and 16 out of 18 for the Whisper-Large-v2). However, they have different biases towards different languages. The highest differences between the masculine and feminine forms in SeamlessM4T-Large are with Tamil (6.4 chrF points difference) and Urdu (5.0).[22] On the other hand, the highest differences in Whisper-Large-v2 are with Spanish (5.3), Urdu (3.8), and Russian (3.4).
- In S2ST, we observe similar outcomes to those in S2TT. The model quality is mostly better when translating from masculine cases, as evident in 14 out of 18 languages for SeamlessM4T-Large and 17 out of 18 for the Whisper-Large-v2 + YourTTScombination. The most significant differences between masculine and feminine sources in SeamlessM4T-Large are found in Tamil (with an ASR-chrF point difference of 6.3) and Spanish (4.5). The highest differences in Whisper-Large-v2 are in Spanish (4.9), Urdu (3.7), and Ukrainian (3.5).

The right panel of Figure 17 demonstrates the performance comparison using Blaser 2.0. Like the findings in ASR-chrF, the translation quality generally improves when translating from masculine cases, which is observed in 16 out of 18 languages and 15 out of 18 languages for SeamlessM4T-Large and Whisper-Large-v2 + YourTTS respectively. The highest differences for SeamlessM4T-Large are with Tamil (0.21 supervised Blaser 2.0 points), Spanish (0.12), and Swedish (0.11). For Whisper-Large-v2 + YourTTS, the highest differences are found in Arabic (0.14), Spanish (0.075), and Tamil (0.05).

Average comparison across directions and modalities Table 34 presents the average scores per gender and the comparison with the corresponding baseline.[23] Δ corresponds to the relative variation between genders computed as follows:

$$\Delta = \omega(M - F)/\omega(min(M, F)), \omega \in \{\text{chrF}, \text{ASR-chrF}, \text{Blaser 2.0}\}$$

22. We find that in our experiment, Arabic shows the bias toward the cases when translated from feminine version, which contrasts with the findings in the Multilingual HolisticBias [Costa-jussà et al., 2023] where Arabic exhibited significantly higher performance when translating from the masculine version. We hypothesize that this difference is attributed to our use of a different language code "ara" instead of "arb" when applying the MMS-TTS.
23. For eng–X S2ST, we report only the performance for the SeamlessM4T-Large in absence of baseline.

As mentioned, in eng–X, we evaluated translations from neutral to gendered forms and observed the overgeneralization towards one gender, whereas in X–eng, we evaluated the robustness of translating content that only differs in their gender inflection. Focusing sorely on the results of SeamlessM4T-Large, we noticed that, except for the evaluation outcomes in Blaser 2.0, the difference in performance between the masculine and feminine forms is more pronounced for overgeneralization than for robustness. Turning our attention to the performance comparison, we find that when it comes to overgeneralization, SeamlessM4T-Large slightly outperforms Whisper-Large-v2 + NLLB-3.3B. As for the outcome related to the robustness, SeamlessM4T-Large falls short against Whisper-Large-v2 in S2TT but outperforms Whisper-Large-v2 + YourTTSin S2ST. We further noticed a higher percentage gap in ASR-chrF than for Blaser 2.0. This may imply that ASR (from ASR-chrF) adds some extra biases.

		eng–X SeamlessM4T/Whisper-Large-v2 + NLLB-3.3B		
		Feminine Reference	Masculine Reference	Δ %
S2TT	chrF	45.0/**47.4**	49.9/**52.7**	**10.9**/11.2
S2ST	ASR-chrF	44.9	49.7	10.6
	Blaser 2.0	3.6	3.7	0.6
		X–eng SeamlessM4T/Whisper-Large-v2 (+ YourTTS)		
		Feminine Source	Masculine Source	Δ %
S2TT	chrF	**52.4**/50.4	**54.3**/52.1	3.7/**3.4**
S2ST	ASR-chrF	**53.1**/52.2	**55.0**/54.0	**3.5**/**3.5**
	Blaser 2.0	**3.5**/2.7	**3.6**/2.8	**2.9**/3.7

Table 34: The averaged points across modalities and genders for assessing the overgeneralization (eng–X) and the robustness (X–eng). Δ represents the relative difference between masculine and feminine ($\Delta = \omega(M - F)/\omega(min(M, F)), \omega \in \{chrF, \text{ASR-chrF}, \text{Blaser 2.0}\}$).

Demographic analysis We conducted a similar analysis to that in Costa-jussà et al. [2023]. Table 33 shows the mean chrF or ASR-chrF at the sentence level on the Multilingual HolisticBias axes translations, averaged across descriptors, templates, languages, and masculine vs. feminine references. Among all the axes, we find that cultural, body type, socioeconomic class, and religion are most sensitive to quality disruptions. Furthermore, when considering the difference between masculine and feminine references and the number of effective samples, we observe that both S2ST and S2TT show the highest bias in the ability, body type, religion, and characteristics axes. These observations align with the findings reported in Costa-jussà et al. [2023] pertaining to T2TT.

6.3.4 Gender data representation

Based on our concurrent work [Muller et al., 2023], we discuss the representation bias of several datasets by focusing on how different genders are represented using lexical matching. The closest work that studies gender representation in data is Choubey et al. [2021], where the authors took on this research question using a synthetic dataset. The authors, however, did not share the details of the lexical nouns used to extract this representation.

HolisticBias [Smith et al., 2022] provides a list of gendered nouns and pronouns. We rely on this list to track how many sentences in our data sets contain gendered markers. Since our analysis is only in English, we tokenize for word boundaries using python word-boundary regular expression (\b). As lexical terms, we limited the vocabulary to make our approach scalable to multiple languages [Muller et al., 2023]. This vocabulary includes: 11 masculine nouns[24]; 4 masculine pronouns,[25]; 10 feminine nouns[26] and 4 feminine pronouns[27]. We are matching single words, hence we report the number of words out of the total number of words in the dataset. Figure 18 summarizes the results of the gender representations for several English evaluation and training datasets. Results show that masculine representation is predominant in most of the datasets. Extremely low representations of gender (i.e. low matching of gendered words based on our selected vocabulary) are found in EuroParl, Fleurs, and Flores datasets. However, this low representation is the trade-off to make our approach scalable to multiple languages, as we mentioned. This scalable effort on data characterization could potentially be used for the purpose of balancing datasets to mitigate gender biases.

6.3.5 Bias Key Findings

In this section, we conducted a set of comprehensive evaluations on translation biases for S2TT and S2ST. We demonstrate the following: (1) in the absence of gender information, SeamlessM4T-Large exhibits an average preference of ∼10% towards translating to the masculine form (for both modalities); (2) utilizing feminine form as the source input leads to lower quality English translations compared to its masculine counterpart, showing a lack of robustness against gender inflection by ∼3% ; (3) SeamlessM4T-Large gets comparable bias results to the state-of-the-art, and (4) our gender representation analysis reveals an overrepresentation of masculine lexica compared to feminine in the analyzed datasets. More importantly, these findings pave the way towards standardizing the bias evaluation of speech translation at a massive scale.

6.4 Limitations

Due to the lack of available model-based techniques that could be applied to added toxicity or gender imbalance detection in this multimodal and massively multilingual setting, we used string-matching techniques that present known limitations.

First, the use of toxicity lists with ETOX for added toxicity detection shares the same limitations with other word list-based detection techniques, which were previously discussed at length in NLLB Team et al. [2022] and Costa-jussà et al. [2023]. Briefly, the two main limitations of word list-based detectors are (1) their tendency to over-detect terms that are only toxic in specific contexts, and (2) their reliance on precise tokenization, which is more difficult to achieve in non-segmenting or highly agglutinative languages. When dealing with

24. man, men, bro, bros, guy, guys, boy, boys, father, fathers, dad, dads, son, sons, husband, husbands, grandfather, grandfathers, grandpa, grandpas, brother, brothers.
25. he, him, his, himself.
26. woman, women, lady, ladies, girl, girls, mother, mothers, mom, moms, daughter, daughters, wife, wives, grandmother, grandmothers, grandma, grandmas, sister, sisters
27. she, her, hers, herself.

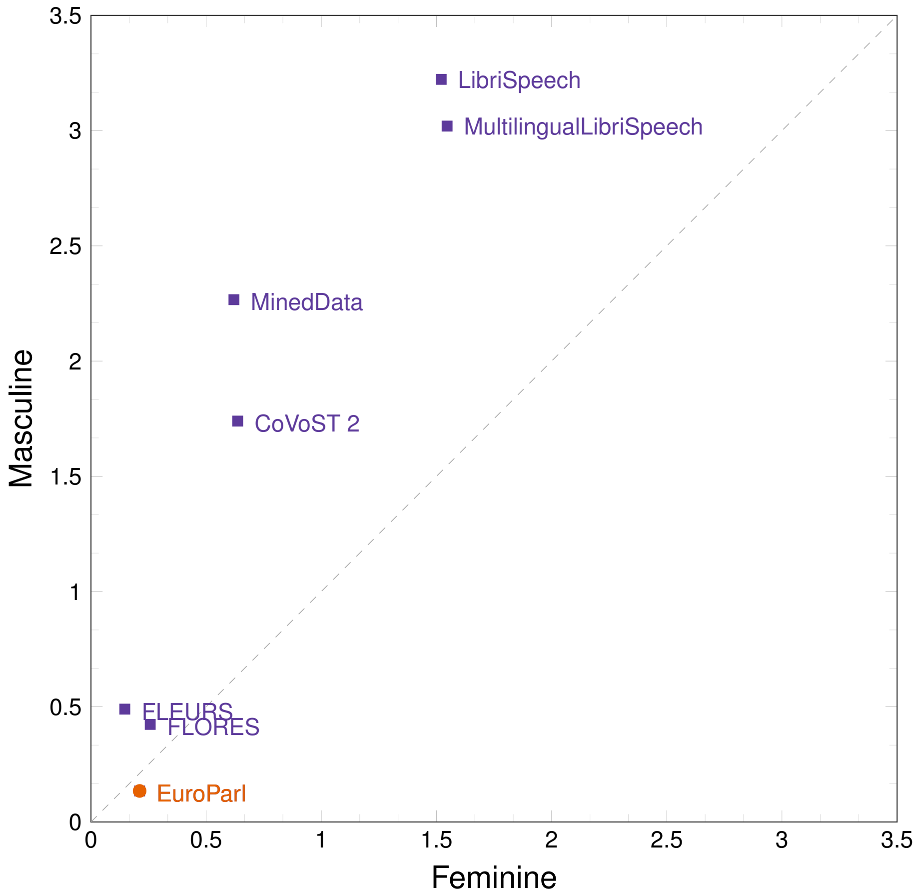

Figure 18: Gender representation of English evaluation datasets (EuroParl, Flores, Fleurs, CoVoST 2, LibriSpeech and MultilingualLibriSpeech), and training mined data (SeamlessAlign). Vertical axis show the percentage of masculine representation and horizontal axis show the percentage of feminine representation.

speech outputs, the process of using ASR before lexical matching adds one more source of error, which tends to lead to false negatives. This particularly affects the directions of eng–X, since ASR tends to be of lower quality for non-English languages.

Second, the use of noun lists for the detection of linguistic gender imbalance in large datasets shares all of the limitations of word list-based techniques previously stated, along with the added difficulty of relying on linguistic gender clues as a proxy for overall gender representation. Indeed, linguistic gender assignment does not follow the same pattern across all languages that mark gender, especially when it comes to inclusive plural forms (i.e., plural forms referring to groups that include more than one gender). In addition to general limitations, the use of a specific and limited set of 30 nouns (selected to mirror those used to build the HolisticBias dataset) does not guarantee that results can be generalized to all other sets of nouns that could be used to investigate gender representation (e.g., occupation nouns).

7. Social Impact & Conclusion

Human communication is multisensorial—we take in sensory input from several modalities to process information in a dynamic way [Holler and Levinson, 2019]. In multilingual contexts, advancements in text-based machine translation have given rise to tools that help individuals communicate and learn in languages where proficiency is low [Lee, 2023]. That said, while foundational models such as NLLB [NLLB Team et al., 2022] push T2TT beyond 200 languages, direct speech translation has yet to achieve similar strides. To bridge this gap, we created a massively multilingual and multimodal machine translation system that paves the way for the next generation of speech translation technologies.

Using novel data and modeling approaches to combine S2ST, S2TT, T2TT, and ASR in a single model, our main contributions are as follows. First, we built a new LID model aligned with our language coverage and conducted speech mining with the help of the newly conceived SONAR—a multilingual and multimodal sentence embedding space—to create a corpus of automatically aligned speech translations of more than 470,000 hours. By fusing four building blocks, (1) SeamlessM4T-NLLB, a massively multilingual T2TT model, (2) w2v-BERT 2.0, a speech representation learning model pre-trained on unlabeled speech audio data, (3) T2U, a text-to-unit sequence-to-sequence model, and (4) HiFi-GAN—a multilingual vocoder for synthesizing speech from units, we built a unified model that covers S2ST from 100 languages to English (100-eng), English to 35 languages (eng-35), and S2TT for 100-eng and eng-95 languages. Notably, compared to previous work on S2ST, which primarily serves translations into English and not vice versa, SeamlessM4T is capable of performing translation from English towards 35 directions. When it comes to S2TT, SeamlessM4T achieves an improvement of 20% BLEU over the previous state-of-the-art in S2TT translation. Preliminary human evaluations of S2TT outputs evinced similarly impressive results; for translations from English, XSTS scores for 24 evaluated languages are consistently above 4 (out of 5). For into English directions, we see significant improvement over Whisper-Large-v2's baseline for 7 out of 24 languages. We then evaluated our model for robustness, revealing that SeamlessM4T is more robust than [Radford et al., 2022] when it comes to background noises and speaker variations. By also including results of the level of added toxicity and gender bias, we hope to motivate future work targeting mitigation efforts.

Made with the goal of promoting accessibility, we open-source all contributions of our work, including two sizes of our model to ensure that even researchers with limited computing resources can use our work. In the section below, we discuss the potential social impact of SeamlessM4T by focusing on its downstream possibilities.

7.1 Augmenting world-readiness

The world we live in has never been more interconnected—the global proliferation of the internet, mobile devices, communicative platforms, and social media exposes individuals to more multilingual content than ever before [Zuckerman, 2008]. The current social order places a demand on a person's "world-readiness" [ACTFL, 2023], a measure of how competent a person is to take on the polyglot world. Initially developed in the context of language learning, world-readiness underscores the importance of being able to communicate in languages beyond one's mother tongue for both instrumental (i.e., employment or schooling) and cultural reasons (i.e., to become a global citizen). That said, while we believe that language acquisition should remain a key mechanism for boosting one's world-readiness, we acknowledge that doing so requires mental and material resources many people may not possess.

The downstream applications that SeamlessM4T supports could allow on-demand access to world-readiness by streamlining multilingual exchange across various contexts. Akin to what T2TT has accomplished for bridging the comprehension of multilingual texts, SeamlessM4T could have the same effects for speech. Research shows that contrary to one's native language, where speech is more organically acquired than reading or writing [Liberman, 1992], this tendency is flipped when it comes to foreign languages [Cheng et al., 1999]. In other words, speech is often deemed more challenging than reading or writing in a foreign language context. SeamlessM4T-supported applications could act as a co-piloting mechanism that supports users in multilingual conversations and boost their confidence in speech-heavy interactions. As speech-based interfaces (i.e., audio assistants, voice memos, live transcriptions, etc.) and auditory content (i.e., podcasts, audiobooks, short-form videos, etc.) become ever more present in people's lives, downstream applications enabled by SeamlessM4T could allow a greater variety of multilingual experiences and in ways that feel more natural and dynamic than its text-based counterparts.

From an inclusion standpoint, SeamlessM4T 's focus on multimodality could make a meaningful difference in augmenting the world-readiness of those with accessibility needs and those whose languages contain multiple writing systems (as aforementioned in 2. For many who lack reading or writing skills, or are unable to rely on sight (i.e., people who are blind or with visual impairment), voice-assisted technologies are essential to how they communicate and stay connected [Belekar et al., 2020]. The ability to translate speech not only gives these groups more comprehensive access to information beyond their native languages, but also in a manner that is better suited for their communicative needs. Additionally, recognizing that some languages may have script variance, SeamlessM4T 's offers up affordances that help circumvent the multiscript conundrum. For languages that do not have standardized writing systems, investments in speech recognition and translation may be instrumental in preventing endangerment. We hope that our effort can help contribute to this important movement.

7.2 Future work

As is the case with most technologies, the distribution of benefits varies based on user demographics and social situation [Wang et al., 2023b]. While we make the case that SeamlessM4T could augment world-readiness by lowering the barriers in cross-lingual communication, some users may experience more difficulties using our work than others. For instance, like many other speech technologies, SeamlessM4T 's ASR performance may vary based on gender, race, accent, or language [Koenecke et al., 2020; Ngueajio and Washington, 2022]. Moreover, our system's performance when it comes to translating slang or proper nouns may also be inconsistent across high and low-resource languages.

Another challenge for S2ST is that speech hinges on immediate reception and feedback compared to written language. In other words, a speaker is limited in their ability to ascertain the quality of an output or make "edits" in a live conversation. Without the ability to plan and revise with the help of back-translation or a native speaker, S2ST may carry higher degrees of interactional risks when it comes to mistranslations or toxicity. We urge researchers and developers who fine-tune or build products using SeamlessM4T to think critically about design features that could help users circumvent these potential obstacles. On a related note, we believe that SeamlessM4T-fueled applications should best be viewed as an augmentation device that assists in translation rather than a tool that replaces the need for language learning or reliable human interpreters. This reminder is especially pertinent in high-stakes situations involving legal or medical decision-making.

Finally, speech is not spoken text—it encompasses a suite of prosodic (i.e., rhythm, stress, and intonation) and emotional components that deserve further research [Elbow, 1985]. To create S2ST systems that feel organic and natural, more research should be directed at output generation that preserves expressivity [Trilla and Alias, 2012]. In addition, the consummate realization of the Babel Fish requires deeper investments into research on low-latency speech translation. Developing systems that enable streaming (i.e., incrementally translating an input sentence as it is being presented) may increase the adoption of such systems in industry or educational contexts [Iranzo-Sánchez et al., 2022; Rybakov et al., 2022]. We hope that SeamlessM4T opens up new possibilities for both of these research areas.

Acknowledgments

We want to extend our gratitude to those who made this work possible. Thanks to Sergey Edunov and Angela Fan for helping shape the earlier stages of the project; Shruti Bhosale, Vedanuj Goswami, Fernando Hernandez and Yun Tang for their help to build better models; Mingda Chen for his contributions to Blaser 1.0; Kiryl Klyushkin for his help to build better experiences; Artyom Kozhevnikov for his contributions to Fairseq2 and Sonar inference; Zhaoheng Ni and Xiaohui Zhang for benchmarking audio denoising models; Neil Seejoor and Mark Duppenthaler for their help in setting up the demo; Vedanuj Goswami, Samuel Hsia, Bilge Acun-Uyan and Carole-Jean Wu who helped with efficiency optimizations; Belen Alastruey, Mohamed Anwar, Heng-Jui Chang, HyoJung Han, Chao-Wei Huang, Hui Lu, Siqi Ouyang, Yifan Peng, Phillip Rust, Jiatong Shi, Neha Verma, Sung-Lin Yeh, and all of our interns and residents for the energy and candid discussions they brought to the team; Mike Clark, Lauren Cohen, Jennifer Pak, Harrison Rudolph for their guidance; Emily Astbury, Lydia Baillergeau, Dana Beaty, Jeffrey Bennett, Jon Carvill, Anne Davidson, Aiman Farooq, Ashley Gabriel, Gopika Jhala, Christopher Johnson, Steph Miles, Ana Paula Kirschner Mofarrej, Raghu Nayani, Alyssa Newcomb, Tamara Piksa, Michelle Restrepo, Noha Rizk, Adébissy Tharinger, who helped our research reach new audiences; Geeta Chauhan, Ankit Gunapal, Caleb Ho, Dinesh Kannappan, Apostolos Kokolis, Teng Li, Matthias Reso, Shubho Sengupta, Hamid Shojanazeri, Xinyuan Zhang for assisting us with compute resources and infrastructure; Emmanuel Dupoux and Eric Michael Smith for their feedback on the paper; Chris Moghbel, Manohar Paluri, Joelle Pineau, Laurens van der Maaten, and Mary Williamson for their continued support of the project.

References

ACTFL. World-readiness standards for learning languages, 2023. URL https://www.actfl.org/educator-resources/world-readiness-standards-for-learning-languages.

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. 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), July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.iwslt-1.1.

Antonios Anastasopoulos and David Chiang. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1008. URL https://aclanthology.org/N18-1008.

Pierre Andrews, Guillaume Wenzek, Kevin Heffernan, Onur Çelebi, Anna Sun, Ammar Kamran, Yingzhe Guo, Alexandre Mourachko, Holger Schwenk, and Angela Fan. stopes - modular machine translation pipelines. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 258–265, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-demos.26.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei

Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.1. URL `https://aclanthology.org/2020.iwslt-1.1`.

Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. *arXiv preprint arXiv:2303.00628*, 2023.

Mikel Artetxe and Holger Schwenk. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1309. URL `https://aclanthology.org/P19-1309`.

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019b. doi: 10.1162/tacl_a_00288. URL `https://aclanthology.org/Q19-1038`.

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. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282, 2022. doi: 10.21437/Interspeech.2022-143.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf`.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language, 2022. URL `https://arxiv.org/abs/2202.03555`.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. A comparative study on end-to-end speech to text translation. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 792–799, 2019. URL `https://api.semanticscholar.org/CorpusID:204791999`.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. mslam: Massively multilingual joint pre-training for speech and text. *arXiv preprint arXiv:2202.01374*, 2022.

Aishwarya Belekar, Shivani Sunka, Neha Bhawar, and Sudhir Bagade. Voice based e-mail for the visually impaired. *International Journal of Computer Applications*, 175(16):8–12, 2020.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.619. URL `https://aclanthology.org/2020.acl-main.619`.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NeurIPS Workshop on End-to-end Learning for Speech and Audio Processing.*, 2016.

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

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409.

Weicheng Cai, Danwei Cai, Shen Huang, and Ming Li. Utterance-level end-to-end language identification using attention-based cnn-blstm. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5991–5995. IEEE, 2019.

Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR, 2022.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964, 2016. doi: 10.1109/ICASSP.2016.7472621.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. BLASER: A text-free speech-to-speech translation evaluation metric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9064–9079, Toronto, Canada, July 2023a. Association for Computational Linguistics. URL `https://aclanthology.org/2023.acl-long.504`.

Mingda Chen, Kevin Heffernan, Onur Çelebi, Alexandre Mourachko, and Holger Schwenk. xSIM++: An improved proxy to bitext mining performance for low-resource languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–109, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.10. URL `https://aclanthology.org/2023.acl-short.10`.

Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, Hirofumi Inaguma, Sravya Popuri, Changhan Wang, Juan Pino, Wei-Ning Hsu, and Ann Lee. Speech-to-speech translation for a real-world unwritten language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4969–4983, Toronto, Canada, July 2023c. Association for Computational Linguistics. URL `https://aclanthology.org/2023.findings-acl.307`.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro J. Moreno, Ankur Bapna, and Heiga Zen. Maestro: Matched speech text representations through modality matching. In *Interspeech*, 2022. URL `https://api.semanticscholar.org/CorpusID:248006130`.

Yong Cheng, Yu Zhang, Melvin Johnson, Wolfgang Macherey, and Ankur Bapna. Mu^{2}SLAM: Multitask, multilingual speech and language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5504–5520. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/cheng23e.html`.

Yuh-show Cheng, Elaine K Horwitz, and Diane L Schallert. Language anxiety: Differentiating writing and speaking components. *Language learning*, 49(3):417–446, 1999.

Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*, pages 3915–3924. PMLR, 2022.

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. GFST: Gender-filtered self-training for more accurate gender in translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1640–1654, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.123. URL `https://aclanthology.org/2021.emnlp-main.123`.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250, 2021. doi: 10.1109/ASRU51503.2021.9688253.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*, 2022. URL `https://arxiv.org/abs/2205.12446`.

Marta Costa-jussà, Christine Basta, Oriol Domingo, and André Rubungo. Occgen: Selection of real-world multilingual parallel data balanced in gender within occupations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 1445–1457. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/09933f07ae2ccbca7212bb4e43de8db0-Paper-Datasets_and_Benchmarks.pdf`.

Marta R. Costa-jussà, Christine Basta, and Gerard I. Gállego. Evaluating gender bias in speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2141–2147, Marseille, France, June 2022. European Language Resources Association. URL `https://aclanthology.org/2022.lrec-1.230`.

Marta R Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. *arXiv preprint arXiv:2305.13198*, 2023.

Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. Interpreting gender bias in neural machine translation: Multilingual architecture matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (11):11855–11863, Jun. 2022. doi: 10.1609/aaai.v36i11.21442. URL `https://ojs.aaai.org/index.php/AAAI/article/view/21442`.

Marta R. Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. Toxicity in multilingual machine translation at scale, 2023.

M.R. Costa-jussà. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, pages 495—-496, 2019. doi: 10.1038/s42256-019-0105-5.

Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, and Shinji Watanabe. Searchable hidden intermediates for end-to-end models of decomposable sequence tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1882–1896, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.151. URL `https://aclanthology.org/2021.naacl-main.151`.

Alexandre D'efossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *ArXiv*, abs/2210.13438, 2022. URL `https://api.semanticscholar.org/CorpusID:253097788`.

Najim Dehak, Pedro A Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak. Language recognition via i-vectors and dimensionality reduction. In *Twelfth annual conference of the international speech communication association*, 2011.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 3830–3834. ISCA, 2020. doi: 10.21437/Interspeech.2020-2650. URL `https://doi.org/10.21437/Interspeech.2020-2650`.

Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. Toward fairness in speech recognition: Discovery and mitigation of performance disparities. *arXiv preprint arXiv:2207.11345*, 2022.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1202. URL `https://aclanthology.org/N19-1202`.

Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk. Multimodal and multilingual embeddings for large-scale speech mining. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15748–15761. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/8466f9ace6a9acbe71f75762ffc890f1-Paper.pdf`.

Paul-Ambroise Duquenne, Hongyu Gong, Benoît Sagot, and Holger Schwenk. T-modules: Translation modules for zero-shot cross-modal machine translation. In *Proceedings of* *the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5794–5806, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.391. URL https://aclanthology.org/2022.emnlp-main.391.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswami, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. SpeechMatrix: A large-scale mined corpus of multilingual speech-to-speech translations. In *ACL (long paper)*, pages 16251–16269, 2023a. URL https://aclanthology.org/2023.acl-long.899.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. SONAR: sentence-level multimodal and language-agnostic representations, 2023b. URL https://arxiv.org/abs/unk.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. Modular speech-to-text translation for zero-shot cross-modal transfer. In *Interspeech*, 2023c.

Peter Elbow. The shifting relationships between speech and writing. *College composition and communication*, 36(3):283–303, 1985.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 2020.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 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, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62.

Sarith Fernando, Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps. Bidirectional modelling for short duration language identification. In *Interspeech*, pages 2809–2813, 2017.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021.

Philip Gage. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994.

Marco Gaido, Matteo Negri, Mauro Cettolo, and Marco Turchi. Beyond voice activity detection: Hybrid audio segmentation for direct speech translation. In Mourad Abbas and Abed Alhakim Freihat, editors, *4th International Conference on Natural Language and Speech Processing, Trento, Italy, November 12-13, 2021*, pages 87–94. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.icnlsp-1.7.

Gerard I. Gállego, Ioannis Tsiamas, Carlos Escolano, José A. R. Fonollosa, and Marta R. Costa-jussà. End-to-end speech translation with pre-trained models and adapters: UPC at IWSLT 2021. In Marcello Federico, Alex Waibel, Marta R. Costa-jussà, Jan Niehues, Sebastian Stüker, and Elizabeth Salesky, editors, *Proceedings of the 18th International Conference on Spoken Language Translation, IWSLT 2021, Bangkok, Thailand (online), August 5-6, 2021*, pages 110–119. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.iwslt-1.11. URL `https://doi.org/10.18653/v1/2021.iwslt-1.11`.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020.

Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. One-to-many multilingual end-to-end speech translation. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 585–592, 2019. URL `https://api.semanticscholar.org/CorpusID:203905407`.

GBD 2019 Blindness and Vision Impairment Collaborators. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the global burden of disease study. *The Lancet global health*, 9(2):e130–e143, 2021.

Javier García Gilabert, Carlos Escolano, and Marta R. Costa-Jussà. Resetox: Re-learning attention weights for toxicity mitigation in machine translation, 2023.

Hongyu Gong, Ning Dong, Sravya Popuri, Vedanuj Goswami, Ann Lee, and Juan Pino. Multilingual speech-to-speech translation into multiple target languages. *arXiv preprint arXiv:2307.08655*, 2023.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL `https://aclanthology.org/2022.tacl-1.30`.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `https://aclanthology.org/W13-2305`.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. Glottolog database 4.6, 2022.

Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. Casual conversations: A dataset for measuring fairness in ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2289–2293, 2021.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.154. URL `https://aclanthology.org/2022.findings-emnlp.154`.

Judith Holler and Stephen C Levinson. Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652, 2019.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Ke Hu, Tara N. Sainath, Ruoming Pang, and Rohit Prabhavalkar. Deliberation model based two-pass end-to-end speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7799–7803, 2020. doi: 10.1109/ICASSP40776.2020.9053606.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577. IEEE, 2019.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. UnitY: Two-pass direct speech-to-speech translation with discrete units. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15655–15680, Toronto, Canada, July 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.acl-long.872`.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE, 2020.

Javier Iranzo-Sánchez, Jorge Civera, and Alfons Juan. From simultaneous to streaming machine translation by leveraging streaming history. *arXiv preprint arXiv:2203.02459*, 2022.

Anastasia Iskhakova, Daniyar Wolf, and Roman Meshcheryakov. Automated destructive behavior state detection on the 1d cnn-based voice analysis. In *Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9,* *2020, Proceedings*, page 184–193, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-60275-8. doi: 10.1007/978-3-030-60276-5_19. URL https://doi.org/10.1007/978-3-030-60276-5_19.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184. IEEE, 2019a.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model. In *Proc. Interspeech 2019*, pages 1123–1127, 2019b. doi: 10.21437/Interspeech.2019-1951.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10120–10134. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/jia22b.html.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. CVSS corpus and massively multilingual speech-to-speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6691–6703, Marseille, France, June 2022b. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.720.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Youngjoo Jung and Bora Kim. Coexistence of multiple writing systems: Classifying digraphia in post-socialist countries. *Journal of Eurasian Studies*, page 18793665231188380, 2023.

Sameer Khurana, Antoine Laurent, and James Glass. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *arXiv preprint arXiv:2205.08180*, 2022.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478, 2021.

Tom Kocmi, Rachel Bawden, OndÅ™ej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal NovÃ¡k, Martin Popel, Maja PopoviÄ‡, and Mariya Shmatova. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 1–45, Abu Dhabi, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.1.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, September 13-15 2005. URL `https://aclanthology.org/2005.mtsummit-papers.11`.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14): 7684–7689, 2020.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

Robert Kraut, Jolene Galegher, Robert Fish, and Barbara Chalfonte. Task requirements and media choice in collaborative writing. *Human–Computer Interaction*, 7(4):375–407, 1992.

Taku Kudo and John Richardson. 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, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL `https://aclanthology.org/D18-2012`.

Amit Kumar and Nicholas Epley. It's surprisingly nice to hear you: Misunderstanding the impact of communication media can lead to suboptimal choices of how to connect with others. *Journal of Experimental Psychology: General*, 150(3):595, 2021.

Ravi Kumar Vuddagiri, Krishna Gurugubelli, Priyam Jain, Hari Krishna Vydana, and Anil Kumar Vuppala. IIITH-ILSC Speech Database for Indain Language Identification. In *Proc. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 56–60, 2018. doi: 10.21437/SLTU.2018-12.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021. doi: 10.1162/tacl_a_00430. URL `https://aclanthology.org/2021.tacl-1.79`.

Alon Lavie, Alexander H. Waibel, Lori S. Levin, Michael Finke, Donna Gates, Marsal Gavaldà, Torsten Zeppenfeld, and Puming Zhan. Janus-iii: speech-to-speech translation in multiple languages. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:99–102 vol.1, 1997.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3327–3339, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.235. URL `https://aclanthology.org/2022.acl-long.235`.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. Textless speech-to-speech translation on real data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 860–872, Seattle, United States, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.63. URL `https://aclanthology.org/2022.naacl-main.63`.

Sangmin-Michelle Lee. The effectiveness of machine translation in foreign language education: a systematic review and meta-analysis. *Computer Assisted Language Learning*, 36(1-2): 103–125, 2023.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.211. URL `https://aclanthology.org/2021.findings-emnlp.211`.

M. Paul Lewis, editor. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, sixteenth edition, 2009.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Multilingual speech translation from efficient finetuning of pretrained models. 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 827–838, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.68. URL `https://aclanthology.org/2021.acl-long.68`.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Multilingual speech translation from efficient finetuning of pretrained models. 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 827–838, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.68. URL `https://aclanthology.org/2021.acl-long.68`.

Alvin M Liberman. The relation of speech to reading and writing. In *Advances in psychology*, volume 94, pages 167–178. Elsevier, 1992.

Daniel Licht, Cynthia Gao, Janice Lam, Francisco Guzman, Mona Diab, and Philipp Koehn. Consistent human evaluation of machine translation across language pairs. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 309–321, Orlando, USA, September 2022. Association for Machine Translation in the Americas. URL `https://aclanthology.org/2022.amta-research.24`.

Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6162–6166. IEEE, 2022.

Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019.

Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno. Automatic language identification using deep neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5337–5341. IEEE, 2014.

Alicia Lozano-Diez, Ruben Zazo-Candil, Javier Gonzalez-Dominguez, Doroteo T Toledano, and Joaquin Gonzalez-Rodriguez. An end-to-end approach to language identification in short utterances using convolutional neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5302. URL `https://aclanthology.org/W19-5302`.

Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1851–1860, 2019.

Katerina Markelova. Illiteracy: "another form of slavery", Nov 2021. URL `https://en.unesco.org/courier/2021-5/illiteracy-another-form-slavery`.

Xiaoxiao Miao, Ian McLoughlin, and Yonghong Yan. A new time-frequency attention mechanism for tdnn and cnn-lstm-tdnn, with application to language identification. In *Interspeech*, pages 4080–4084, 2019.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL `https://doi.org/10.1145/3287560.3287596`.

Benjamin Muller, Belen Alastruey, Prangthip Hansanti, Elahe Kalbassi, Christophe Ropers, Eric Smith, Adina Williams, Luke Zettlemoyer, Pierre Andrews, and Marta R. Costa-jussà. Automatic pipeline for gender multilingual data characterisation at scale. *arXiv preprint arXiv:*, 2023.

S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, Jin-Song Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto. The atr multilingual speech-to-speech translation system. *Trans. Audio, Speech and Lang. Proc.*, 14(2):365–376, dec 2006. ISSN 1558-7916. doi: 10.1109/TSA.2005.860774. URL `https://doi.org/10.1109/TSA.2005.860774`.

Mikel K Ngueajio and Gloria Washington. Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review. In *International Conference on Human-Computer Interaction*, pages 421–440. Springer, 2022.

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. No language left behind: Scaling human-centered machine translation, 2022. URL `https://arxiv.org/abs/2207.04672`.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-4009. URL `https://aclanthology.org/N19-4009`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://aclanthology.org/P02-1040`.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL `https://aclanthology.org/2020.emnlp-main.617`.

Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. Self-Training for End-to-End Speech Translation. In *Proc. Interspeech 2020*, pages 1476–1480, 2020. doi: 10.21437/Interspeech.2020-2938.

Ingo Plag, Christiane Dalton-Puffer, and Harald Baayen. Morphological productivity across speech and writing. *English Language & Linguistics*, 3(2):209–228, 1999.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*, 2021.

Maja Popović. 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, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL `https://aclanthology.org/W15-3049`.

Bilal Porgali, Vítor Albiero, Jordan Ryda, Cristian Canton Ferrer, and Caner Hazirbas. The casual conversations v2 dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10–17, 2023.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W18-6319`.

Tomasz Potapczyk and Pawel Przybysz. Srpol's system for the IWSLT 2020 end-to-end speech translation task. In Marcello Federico, Alex Waibel, Kevin Knight, Satoshi Nakamura, Hermann Ney, Jan Niehues, Sebastian Stüker, Dekai Wu, Joseph Mariani, and François Yvon, editors, *Proceedings of the 17th International Conference on Spoken Language Translation, IWSLT 2020, Online, July 9 - 10, 2020*, pages 89–94. Association for Computational Linguistics, 2020a. doi: 10.18653/v1/2020.iwslt-1.9. URL `https://doi.org/10.18653/v1/2020.iwslt-1.9`.

Tomasz Potapczyk and Pawel Przybysz. SRPOL's system for the IWSLT 2020 end-to-end speech translation task. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 89–94, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.9. URL `https://aclanthology.org/2020.iwslt-1.9`.

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. Scaling speech technology to 1,000+ languages, 2023.

Marcelo Prates, Pedro Avelar, and Luís Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32, 05 2020. doi: 10.1007/s00521-019-04144-6.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.646. URL `https://aclanthology.org/2022.emnlp-main.646`.

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

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapr. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *TACL*, 10: 145–162, 2022.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL `https://aclanthology.org/D19-1410`.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.365. URL `https://aclanthology.org/2020.emnlp-main.365`.

Adithya Renduchintala and Adina Williams. Investigating failures of automatic translationin the case of unambiguous gender. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3454–3469, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.243. URL `https://aclanthology.org/2022.acl-long.243`.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.15. URL `https://aclanthology.org/2021.acl-short.15`.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Ryan Padfield, James Qin, Daniel Rozenberg, Tara N. Sainath, Johan Schalkwyk, Matthew Sharifi, Michelle D. Tadmor, Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirovi'c, Damien Vincent, Jiahui Yu, Yongqiang Wang, Victoria Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukás Zilka, and Christian Havnø Frank. Audiopalm: A large language model that can speak and listen. *ArXiv*, abs/2306.12925, 2023. URL `https://api.semanticscholar.org/CorpusID:259224345`.

Oleg Rybakov, Fadi Biadsy, Xia Zhang, Liyang Jiang, Phoenix Meadowlark, and Shivani Agrawal. Streaming parrotron for on-device speech-to-speech conversion. *arXiv preprint arXiv:2210.13761*, 2022.

Tara N. Sainath, Ruoming Pang, David Rybach, Yanzhang He, Rohit Prabhavalkar, Wei Li, Mirkó Visontai, Qiao Liang, Trevor Strohman, Yonghui Wu, Ian McGraw, and Chung-Cheng Chiu. Two-Pass End-to-End Speech Recognition. In *Proc. Interspeech 2019*, pages 2773–2777, 2019. doi: 10.21437/Interspeech.2019-1341.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. The Multilingual TEDx Corpus for Speech Recognition and Translation. In *Proc. Interspeech 2021*, pages 3655–3659, 2021. doi: 10.21437/Interspeech.2021-11.

Juliana Schroeder, Michael Kardas, and Nicholas Epley. The humanizing voice: Speech reveals, and text conceals, a more thoughtful mind in the midst of disagreement. *Psychological science*, 28(12):1745–1762, 2017.

Holger Schwenk. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2037. URL `https://aclanthology.org/P18-2037`.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. CCMatrix: Mining billions of high-quality parallel sentences on the web. 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 6490–6500, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.507. URL `https://aclanthology.org/2021.acl-long.507`.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL `https://aclanthology.org/P16-1162`.

Khetam Al Sharou and Lucia Specia. A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–180, Ghent, Belgium, June 2022. European Association for Machine Translation. URL `https://aclanthology.org/2022.eamt-1.20`.

Peng Shen, Xugang Lu, Sheng Li, and Hisashi Kawai. Feature representation of short utterances based on knowledge distillation for spoken language identification. In *Interspeech*, pages 1813–1817, 2018.

Peng Shen, Xugang Lu, Sheng Li, and Hisashi Kawai. Interactive learning of teacher-student model for short utterance spoken language identification. In *ICASSP 2019-2019 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5981–5985. IEEE, 2019.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *CoRR*, abs/2201.03110, 2022. URL `https://arxiv.org/abs/2201.03110`.

Silero. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`, 2021.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.emnlp-main.625`.

David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. Spoken language recognition using x-vectors. In *Odyssey*, volume 2018, pages 105–111, 2018.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wmt-1.71`.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1164. URL `https://aclanthology.org/P19-1164`.

Tzu-Wei Sung, Jun-You Liu, Hung-yi Lee, and Lin-shan Lee. Towards end-to-end speech-to-text translation with two-pass decoding. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7175–7179, 2019. doi: 10.1109/ICASSP.2019.8682801.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. Improving speech translation by understanding and learning from the auxiliary text translation task. 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 4252–4261, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.328. URL https://aclanthology.org/2021.acl-long.328.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Speech-to-speech translation between untranscribed unknown languages. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 593–600. IEEE, 2019.

Alexandre Trilla and Francesc Alias. Sentence-based sentiment analysis for expressive text-to-speech. *IEEE transactions on audio, speech, and language processing*, 21(2):223–233, 2012.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. SHAS: Approaching optimal Segmentation for End-to-End Speech Translation. In *Proc. Interspeech 2022*, pages 106–110, 2022. doi: 10.21437/Interspeech.2022-59.

Jörgen Valk and Tanel Alumäe. VoxLingua107: a dataset for spoken language recognition. In *Proc. IEEE SLT Workshop*, 2021.

Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3614–3633, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wolfgang Wahlster. Verbmobil: Foundations of speech-to-speech translation. In *Artificial Intelligence*, 2000. URL https://api.semanticscholar.org/CorpusID:30807920.

Li Wan, Prashant Sridhar, Yang Yu, Quan Wang, and Ignacio Lopez Moreno. Tuplemax loss for language identification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5976–5980. IEEE, 2019.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.517.

Changhan Wang, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Ann Lee, Peng-Jen Chen, Jiatao Gu, and Juan Pino. fairseq sˆ2: A scalable and integrable speech synthesis toolkit. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 143–152, Online and Punta Cana, Dominican Republic, November 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-demo.17. URL https://aclanthology.org/2021.emnlp-demo.17.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 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, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.80. URL https://aclanthology.org/2021.acl-long.80.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. CoVoST 2 and Massively Multilingual Speech Translation. In *Proc. Interspeech 2021*, pages 2247–2251, 2021c. doi: 10.21437/Interspeech.2021-2027.

Chengyi Wang, Sanyuan Chen, Yu Wu, Zi-Hua Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *ArXiv*, abs/2301.02111, 2023a. URL https://api.semanticscholar.org/CorpusID:255440307.

Skyler Wang, Ned Cooper, Margaret Eby, and Eun Seo Jo. From human-centered to social-centered artificial intelligence: Assessing chatgpt's impact through disruptive events. *arXiv preprint arXiv:2306.00227*, 2023b.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. VioLA: Unified codec language models for speech recognition, synthesis, and translation. May 2023c.

Yiming Wang, Jinyu Li, Heming Wang, Yao Qian, Chengyi Wang, and Yu Wu. Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7097–7101. IEEE, 2022.

Barack Wanjawa, Lilian Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. Kencorpus: A kenyan language corpus of swahili, dholuo and luhya for natural language processing tasks. *arXiv preprint arXiv:2208.12081*, 2022.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Z. Chen. Sequence-to-sequence models can directly translate foreign speech. In *Interspeech*, 2017a.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly translate foreign speech. In *Interspeech*, 2017b.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Deliberation networks: Sequence generation beyond one-pass decoding. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/c6036a69be21cb660499b75718a3ef24-Paper.pdf.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5370–5378. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/746. URL https://doi.org/10.24963/ijcai.2019/746.

Midia Yousefi and Dimitra Emmanouilidou. Audio-based toxic language classification using self-attentive convolutional neural network. In *29th European Signal Processing Conference, EUSIPCO 2021, Dublin, Ireland, August 23-27, 2021*, pages 11–15. IEEE, 2021. doi: 10.23919/EUSIPCO54536.2021.9616001. URL https://doi.org/10.23919/EUSIPCO54536.2021.9616001.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022. URL https://api.semanticscholar.org/CorpusID:236149944.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. Google usm: Scaling automatic speech recognition beyond 100 languages, 2023a.

Zi-Hua Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *ArXiv*, abs/2303.03926, 2023b. URL https://api.semanticscholar.org/CorpusID:257378493.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1663–1676, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.108.

Ding Zhao, Tara N. Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang. Shallow-Fusion End-to-End Contextual Biasing. In *Proc. Interspeech 2019*, pages 1418–1422, 2019. doi: 10.21437/Interspeech.2019-1209.

Jinming Zhao, Hao Yang, Gholamreza Haffari, and Ehsan Shareghi. M-Adapter: Modality Adaptation for End-to-End Speech-to-Text Translation. In *Proc. Interspeech 2022*, pages 111–115, 2022. doi: 10.21437/Interspeech.2022-592.

Qiu-Shi Zhu, Jie Zhang, Zi-Qiang Zhang, Ming-Hui Wu, Xin Fang, and Li-Rong Dai. A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3174–3178. IEEE, 2022.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1561.

Ethan Zuckerman. The polyglot internet, October 2008. URL https://ethanzuckerman.com/the-polyglot-internet/.

A. FAIRSEQ2

FAIRSEQ2 is an open-source library of sequence modeling components that provides researchers and developers with building blocks for machine translation, language modeling, and other sequence generation tasks specifically around text and audio data format. FAIRSEQ2 is distributed with a MIT license and is available on GitHub at https://github.com/pytorch/fairseq2.

FAIRSEQ2 features: (i) state of-the-art implementations of transformers and their components (transformer layers, embedding layers, layernorms, attention blocks, etc.); (ii) `fairseq2.data` – a scalable pipeline API that enables text and audio data pre-processing, transformation, shuffling and batching in a streaming manner, allowing training over multi-terabyte datasets without explicit data preparation steps or data loading timeouts; (iii) core building components for efficient model training (optimizers, LR schedulers, loss implementations); (iv) sequence generators for optimized inference with incremental beam search.

Following the spirit of its predecessor FAIRSEQ [Ott et al., 2019], FARSEQ2 was built with the idea of extensibility in mind. The library-like structure of the code enables effortless component drop-ins, including those that were initially written in FAIRSEQ. We expect a continuous population of the library with new components by us and by the open source community in the next years.

Another guiding principle for FAIRSEQ2 is a clear separation of core and experimental code. The original FAIRSEQ has become a hub for numerous research ideas. Often they were added in the form of if-else statements mixed with the core functionality. Over time, the number of such if-else statements and associated command line options has grown, with each option poorly supported and often subtly incompatible with other options. To prevent this scenario In FAIRSEQ2 all basic components are designed with the "dependency inversion" principle making it possible to compose them easily. Existing model architectures can be modified with just a few lines of code without requiring copy/pasting large amounts of code, All plug-ins and modifications exist as separate components, not interfering with the parent blocks and not hindering the access to them for other users. Larger efforts (like UnitY or Sonar described in this paper) are moved into separate repositories and use FAIRSEQ2 as a dependency.

We acknowledge the wide range of training and execution environments for Deep Learning models that exist today (from a single-container training via on-demand Cloud Computing Services to huge LLMs training jobs running on exaFLOPS supercomputers with tens of thousands GPUs; from a very limited inference capabilities of edge devices to the power of accelerated inference on ASICs). To meet the diverse expectations of these environments, FAIRSEQ2 has shifted from the idea of a self-contained single-stop for all training, evaluation and inference pipelines towards a set of independent components that can be used and extended outside of FAIRSEQ2. We put an emphasis on compatibility with the existing alternatives in PyTorch and other Deep Learning frameworks, following common API conventions and inheriting from the same base classes. That guarantees effortless drop-in replacement of components from different origin. The user is offered with a wide range of usage scenarios: from implementing a complete pipeline using FAIRSEQ2, to fusing multiple Deep Learning frameworks in their project, or even picking a single block like the efficient implementation of an optimizer.

B. Data Statistics

We provide in Table 35 statistics of ASR and S2TT data (in hours of speech audio) used to train the X2T models of SeamlessM4T. Similarly, we provide in Table 36 statistics of S2ST training data.

language code	ASR	S2TTX–eng		Resource	S2TTeng–X	
		P	M		P	M
Total	40,012	50,596	12,682		17,6827	5,701
afr	106	101		low	2069	
amh	54	49		low	1921	
arb	934	942	400	high	1959	200
ary	97	95		low	1776	
arz	93	92		low	2014	
asm	77	68		low	1698	
ast	0	0		zero-shot	0	
azj	95	94		low	1901	
bel	1160	1157		high	1641	
ben	338	320	400	high	1987	200
bos	99	99		low	2113	
bul	103	102		low	1881	
cat	1767	1758	400	high	1781	200
ceb	0	0		zero-shot	2020	
ces	189	442	400	high	2066	200
ckb	93	92		low	2001	
cmn	9784	9027	400	high	1947	
cym	100	96	400	medium	1676	
dan	161	371	400	medium	1954	200
deu	3354	3490		high	2043	200
ell	345	339		medium	1725	
eng	3845	0		high	0	
est	133	130	400	medium	1803	200
eus	276	265		medium	1998	
fin	182	449	400	high	1933	200
fra	2123	2247		high	2304	200
fuv	0	0		zero-shot	0	
gaz	0	0		zero-shot	1766	
gle	56	55		low	1973	
glg	123	121		low	2116	
guj	143	138		low	1990	
hau	0	0		zero-shot	0	
heb	96	96		low	2092	
hin	148	143	400	medium	2066	200
hrv	308	219		medium	2119	
hun	260	474		medium	1900	
hye	148	146		low	1696	
ibo	35	28		low	1738	
ind	250	254	400	medium	1818	200
isl	132	130		low	2059	
ita	591	910	400	high	2278	200
jav	302	301		medium	2122	
jpn	381	15141	400	high	1798	200
kam	0	0		zero-shot	0	
kan	124	121	208	low	1954	
kat	195	185		low	1639	
kaz	330	327		medium	1895	
kea	0	0		zero-shot	0	
khk	152	148		low	1657	
khm	191	187		low	1661	
kir	129	123		low	1839	
kor	387	201	400	medium	2125	
lao	200	190		low	1959	
lin	0	0		zero-shot	0	
lit	40	283		low	1920	
ltz	0	0		zero-shot	0	
lug	369	368		medium	1890	
luo	0	0		zero-shot	1975	
lvs	100	98		low	1779	
mai	0	0		zero-shot	2004	
mal	110	57		low	1754	
mar	112	108		low	1848	
mkd	145	143		low	1918	
mlt	157	151	74	low	1699	200
mni	0	0		zero-shot	1257	
mri	0	0		zero-shot	0	
mya	137	125		low	1860	
nld	1734	1780	400	high	2249	200
nor	214	193		low	2134	
npi	153	129		low	1714	
nso	0	0		zero-shot	0	
nya	103	99		low	2058	
oci	0	0		zero-shot	0	
ory	89	86		low	1721	
pan	196	193		low	1641	
pbt	131	121	400	medium	1847	
pes	386	68		low	1980	
pol	341	446	400	high	1914	200
por	269	246	400	medium	2250	200
ron	182	443	400	high	2131	200
rus	264	144	400	medium	2161	200
sat	0	0		zero-shot	0	
slk	142	390	400	medium	1931	200
slv	107	370		low	1800	
sna	0	0		zero-shot	2067	
snd	0	0		zero-shot	1958	
som	143	140		low	1851	
spa	1514	1285		high	2505	200
srp	101	98		low	1910	
swe	129	91		low	1810	200
swh	361	50	400	medium	1930	200
tam	256	64	400	medium	1569	
tel	89	80	400	medium	1934	
tgk	99	98		low	1820	
tgl	99	93	400	medium	2015	
tha	189	59	400	medium	1941	101
tur	169	100	400	medium	2135	200
umb	0	0		zero-shot	0	
ukr	132	75	400	medium	2052	200
urd	185	145	400	medium	1844	200
uzn	166	96	400	medium	1801	200
vie	194	151	400	medium	2396	200
wol				zero-shot		
xho	0	0		zero-shot	0	
yor	132	130		low	1384	
yue	167	124		low	1931	
zlm	155	161		low	0	
zul	62	55		low	2063	

Table 35: Statistics of ASR and S2TT data used to train our SeamlessM4T model. We list the data size in hours of speech between primary (P) i.e., open-source S2TT and pseudo-labeled ASR data, and mined (M). For each language we distinguish between eng–X for translating from English into that language, and X–eng for translating into English. We qualify as high-resource, languages with more than 1000 hours of supervision. Languages with between 500 and 1000 hours are dubbed medium-resource, and languages with less than 500 hours are low-resource. If a language is not supervised during the 1+2 stages of finetuning then it is evaluated as zero-shot.

	S2ST					S2ST			
	X–eng		eng–X			X–eng		eng–X	
Language	Primary	Mined	Primary	Mined	**Language**	Primary	Mined	Primary	Mined
Total	26,254	23,171	49,425	21,983					
afr	100	0	0	0	lin	52	0	0	0
amh	46	0	0	0	lit	279	0	0	0
arb	898	736	895	681	ltz	2	0	0	0
ary	94	0	0	0	lug	362	0	0	0
arz	91	0	0	0	lvs	95	0	0	0
asm	62	0	0	0	mal	103	0	0	0
ast	0	0	0	0	mar	106	0	0	0
azj	92	0	0	0	mkd	141	0	0	0
bel	285	0	0	0	mlt	149	46	688	39
ben	292	246	652	221	mya	123	0	0	0
bos	99	0	0	0	nld	1,777	1,061	1,003	962
bul	101	0	0	0	nor	189	0	0	0
cat	276	278	692	293	npi	114	0	0	0
ces	437	522	832	528	nya	99	0	0	0
ckb	89	0	0	0	oci	0	0	0	0
cmn	350	1,318	857	1,388	ory	84	0	0	0
cym	93	197	700	185	pan	188	0	0	0
dan	368	420	684	450	pbt	114	0	0	0
deu	2,570	1,661	962	1,618	pes	366	0	881	0
ell	330	0	0	0	pol	591	667	726	657
est	128	502	691	477	por	355	606	983	508
eus	263	0	0	0	ron	469	588	951	521
fin	446	442	684	414	rus	290	1,093	959	1,075
fra	2,255	2,438	937	2,303	slk	402	427	686	426
gle	55	0	0	0	slv	377	0	0	0
glg	120	0	0	0	som	138	0	0	0
guj	135	0	0	0	spa	1,694	2,335	1,035	2,209
hau	78	0	0	0	srp	99	0	0	0
heb	96	0	0	0	swe	124	0	688	0
hin	138	466	656	430	swh	342	411	682	392
hrv	218	0	0	0	tam	241	664	654	685
hun	468	0	0	0	tel	76	426	655	403
hye	141	0	0	0	tgk	98	0	0	0
ibo	24	0	0	0	tgl	82	213	661	169
ind	248	443	684	375	tha	183	462	641	408
isl	127	0	0	0	tur	156	375	998	411
ita	930	716	1,020	636	ukr	129	349	662	329
jav	291	0	0	0	urd	179	555	682	502
jpn	624	993	681	779	uzn	162	139	695	147
kan	119	170	703	135	vie	176	666	954	684
kat	180	0	0	0	wol	13		0	0
kaz	319	0	0	0	xho	6	0	0	0
khk	143	0	0	0	yor	128	0	0	0
khm	184	0	0	0	yue	136	0	0	0
kir	120	0	0	0	zlm	157	0	0	0
kor	350	541	666	541	zul	48	0	0	0
lao	183	0	0	0					

Table 36: Statistics of S2ST data used to train our SeamlessM4T model. We list the data size in hours of speech. For each language we distinguish between Eng-X for translating from English into that language, and X-Eng for translating into English.

C. Model Card - SeamlessM4T

Model Details[a]

- Person or organization developing model: *Developed by Meta AI Research*
- Model date: *August 22nd, 2023*
- Model version: SeamlessM4T-Large and SeamlessM4T-Medium
- Model type: *Multitasking* UnitY *with (a) Conformer speech encoder, (b) Transformer text encoder-decoder and (c) Transformer encoder-decoder for T2U.*
  - *The exact training algorithm and data used to train* SeamlessM4T-Large *and* SeamlessM4T-Medium *are described in the paper: Seamless Communication et al, SeamlessM4T—Massively Multilingual & Multimodal Machine Translation, Arxiv, 2023*
  - *License: CC-BY-NC 4.0* [b]
  - *Where to send questions or comments about the model:* *https://github.com/facebookresearch/seamless_communication/issues*

Intended Use

- Primary intended uses: SeamlessM4T-Large *and* SeamlessM4T-Medium *are multilingual and multimodal translation models primarily intended for research in speech and text translation. It allows for:*
  - *ASR: Automatic speech recognition for 96 languages.*
  - *S2ST: Speech-to-Speech translation from 100 source speech languages into 35 target speech languages.*
  - *S2TT: Speech-to-text translation from 100 source speech languages into 95 target text languages.*
  - *T2ST: Text-to-Speech translation from 95 source text languages into 35 target speech languages.*
  - *T2TT: Text-to-text translation (MT) from 95 source text languages into 95 target text languages.*
  - *TTS: Text-to-speech synthesis for 36 languages.*

  Information on how to use the model can be found in seamless_communication repository along with recipes for finetuning.
- Primary intended users: *Primary users are researchers and machine translation (speech and text) research community.*
- Out-of-scope use cases: SeamlessM4T *is a research model and is not released for production deployment.* SeamlessM4T *is trained on general domain data and is not intended to be used with domain specific inputs, such as medical domain or legal domain. The model is not intended to be used for long-form translation. The model*

was trained on short text and speech inputs, therefore translating longer sequences might result in quality degradation. SeamlessM4T *translations can not be used as certified translations.*

Metrics

- Model performance measures: *For the S2TT task,* SeamlessM4T *models were evaluated using the* BLEU *metric adopted by SOTA models in speech-to-text translation. The models were additionally evaluated with* spBLEU *and* Blaser 2.0 *on S2TT. For S2ST, the models are evaluated with* ASR-BLEU *and* Blaser 2.0. *For the T2TT taks, we report quality in terms of chrF++. For ASR, we report the widely adopted metric of* WER *with the text normalized following the normalization in Radford et al. [2022]. Additionally, we performed human evaluation with the XSTS protocol and measured added toxicity, robustness and bias of* SeamlessM4T-Large. *Please refer to Table 4 of the* SeamlessM4T *paper for an exhaustive list of metrics.*

Evaluation Data

- Datasets: Fleurs, Flores, CoVoST *2 and* CVSS, HolisticBias *and* Multilingual HolisticBias *described in Sections 2.2 and 6 of the* SeamlessM4T *paper.*
- Motivation: *We used* Fleurs *as it provides an n-way parallel speech and text dataset in 102 languages, on which we can evaluate* SeamlessM4T *models on multiple tasks.*

Training Data

- *We used parallel multilingual data from a variety of sources to train the model.*

Ethical Considerations

- *In this work, we took a reflexive approach in technological development to ensure that we prioritize human users and minimize risks that could be transferred to them. While we reflect on our ethical considerations throughout the article, here are some additional points to highlight. For one, many languages chosen for this study are low-resource languages. While quality translation could improve education and information access in many in these communities, such an access could also make groups with lower levels of digital literacy more vulnerable to misinformation or online scams. The latter scenarios could arise if bad actors misappropriate our work for nefarious activities, which we conceive as an example of unintended use. Regarding data acquisition, the training data used for model development were mined from various publicly available sources on the web. Although we invested heavily in data cleaning, personally identifiable information may not be entirely eliminated. Finally, although we did our best to optimize for translation quality, mistranslations produced by the model could remain. Although the odds are low, this could have adverse impact on those who rely on these translations to make important decisions (particularly when related to health and safety).*

Caveats and Recommendations

- Limitations: *Researchers should consider implementing additional integrity mitigations for "added toxicity" when using the model in a research application.*

---

a. For this card, we use the template from Mitchell et al. [2019].

b. `https://creativecommons.org/licenses/by-nc/4.0/legalcode`