Joint speech and text machine translation for up to 100 languages

https://doi.org/10.1038/s41586-024-08359-z

Received: 22 November 2023

Accepted: 6 November 2024

Published online: 15 January 2025

Open access

Check for updates

SEAMLESS Communication Team*

Creating the Babel Fish, a tool that helps individuals translate speech between any two languages, requires advanced technological innovation and linguistic expertise. Although conventional speech-to-speech translation systems composed of multiple subsystems performing translation in a cascaded fashion exist¹⁻³, scalable and high-performing unified systems^{4,5} remain underexplored. To address this gap, here we introduce SEAMLESSM4T-Massively Multilingual and Multimodal Machine Translation-a single model that supports speech-to-speech translation (101 to 36 languages), speech-to-text translation (from 101 to 96 languages), text-to-speech translation (from 96 to 36 languages), text-to-text translation (96 languages) and automatic speech recognition (96 languages). Built using a new multimodal corpus of automatically aligned speech translations and other publicly available data, SEAMLESSM4T is one of the first multilingual systems that can translate from and into English for both speech and text. Moreover, it outperforms the existing state-of-the-art cascaded systems, achieving up to 8% and 23% higher BLEU (Bilingual Evaluation Understudy) scores in speech-to-text and speech-to-speech tasks, respectively. Beyond quality, when tested for robustness, our system is, on average, approximately 50% more resilient against background noise and speaker variations in speech-to-text tasks than the previous state-of-the-art systems. We evaluated SEAMLESSM4T on added toxicity and gender bias to assess translation safety. For the former, we included two strategies for added toxicity mitigation working at either training or inference time. Finally, all contributions in this work are publicly available for noncommercial use to propel further research on inclusive speech translation technologies.

The Babel Fish from *The Hitchhiker's Guide to the Galaxy* is a fictional tool that translates between two languages. In the contemporary global landscape, characterized by increasing interconnectivity and mobile sociality, the social imperative to actualize these technologies and facilitate on-demand speech-to-speech translation (S2ST) both in the digital and in the physical worlds has never been greater. Despite the centrality of speech in everyday communication, machine translation (MT) systems today remain text-oriented. See Supplementary Information section I.1 for more details on why speech should be prioritized in MT. Although single, unimodal models such as No Language Left Behind (NLLB)⁶ pushed text-to-text translation (T2TT) coverage to more than 200 languages, unified S2ST models are far from achieving similar scope or performance. This disparity could be attributed to many causes, but audio data scarcity and modelling constraints remain key obstacles.

Existing S2ST systems have three main shortcomings. First, these systems tend to focus on high-resource languages, leaving many low-resource languages behind. Second, these systems mostly service translation from a source language into English (X–eng), not the reverse (eng–X). Third, most S2ST systems rely heavily on the cascading of several subsystems; for example, automatic speech recognition (ASR) + T2TT + text-to-speech (TTS). Although direct systems exist^{1,4,5}, they do not match the performance of their cascaded counterparts⁷.

See Supplementary Information section 1.2 for more details on the current technical landscape.

To address these limitations, we introduce SEAMLESSM4T (Massively Multilingual and Multimodal Machine Translation), a unified system that supports ASR, T2TT, speech-to-text translation (S2TT), text-to-speech translation (T2ST) and S2ST. To build this, we created a corpus of more than 470,000 h of automatically aligned speech translations (SEAMLESSALIGN) using a new sentence embedding space (Sentence-level Multimodal and Language-Agnostic Representations, or SONAR)⁸. We then combined a filtered subset of this corpus with human-labelled and pseudo-labelled data to develop the first multitasking system that performs S2ST from more than 100 languages into 36 languages, S2TT and ASR into 96 languages, zero-shot T2ST into 36 languages, as well as T2TT for 96 languages (see Table 1 for a comparative overview of language coverage and Supplementary Information section II for more details). Because of the unified architecture of SEAMLESSM4T (Fig. 1), the model can perform T2TT, S2TT or S2ST for non-English directions (X–X) in a zero-shot manner. It can also perform T2ST without being trained explicitly for this task. As a result of pretraining the speech encoder of SEAMLESSM4T on large amounts of unlabelled speech data (see section 'Unsupervised speech pretraining'), it can handle utterances mixing two or more languages.

^{*}A list of authors and their affiliations appears at the end of the paper.

Table 1 | State-of-the-art task and language coverage

Model	Task language coverage						
	S2TT	S2ST	ASR	T2TT	T2ST		
AudioPaLM-8B-S2ST ^{a,21}	113-eng	113-eng	98	-	-		
NLLB Team et al. ⁶	_	-	-	202-202	-		
WHISPER-LARGE-V220	96-eng	-	97	-	-		
MMS-L1107-CCLM-LSAH ²³	-	-	1107	-	-		
This work (SEAMLESSM4T)	101–96	101–36	96	96-96	96-36		

For each of our core tasks, we provide the language coverage of SEAMLESSM4T and existing state-of-the-art models.

^aAlthough other models in this table are open-sourced, AudioPaLM is a proprietary model.

To evaluate the quality of outputs of our model, we used several existing metrics spanning across tasks and modalities, as well as four main evaluation datasets. For example, we used chrF2++⁹ for T2TT, BLEU¹⁰ for S2TT, ASR-BLEU⁵ for S2ST and WER for ASR. See Supplementary Table 2 for details. We also tested our models for resilience against background noise or speaker variation, as well as other fronts for responsible deployment, such as gender bias, using the MULTILINGUAL HOLISTICBIAS datasets, or added toxicity, using new speech-based metrics (ASR-ETOX and MuTox). We mitigate added toxicity with a filtering strategy at training time and a beam filtering strategy at inference time¹¹.

Apart from building SEAMLESSM4T, we also discuss the social implications of our work and how it may contribute to greater degrees of world-readiness¹² in the long run (see section 'Social impact and conclusion'). To spur future research, we make the data tools, code and two sizes of SEAMLESSM4T models publicly available for non-commercial use.

In the subsequent sections, we trace the key results in developing our SEAMLESSM4T models. First, we outline our efforts to mine aligned speech and text data, starting with speech language identification to the mining of aligned speech and text segments using modality-agnostic encoders. Next, we report the main results of our direct translation systems trained in part with the aforementioned automatically aligned speech data. These results highlight the task versatility of SEAMLESSM4T models achieving multilingual state-of-the-art performance in ASR, T2TT, S2TT and S2ST. Then, we support the reported results with human evaluation analysis. Finally, we delineate our efforts to mitigate added toxicity and evaluate the robustness of our models to gender variations.

Data

Training speech translation systems requires labelled data; that is, speech-to-text and speech-to-speech aligned data. However, those resources are very limited for low-resource languages. We build on the multilingual and multimodal embedding space of SONAR⁸ and a large collection of raw speech and texts, as described in the Methods, to automatically mine aligned resources, complementing existing human-labelled and pseudo-labelled data.

Speech language identification

Processing raw speech from the web involves segmenting utterances into shorter chunks, followed by language identification. Building on an open-source model trained on VoxLingua107 with the ECAPA-TDNN architecture^{13,14}, we developed a new speech-based language identification (LID) model covering all 100 languages featured in this work (see Methods, 'Audio processing and speech LID' for more details).

To measure the precision and recall of LID models, we report F1 scores on the test data in Extended Data Table 1. The results are given for the 100 SEAMLESSM4T languages (Overall) and the 79 common languages between SEAMLESSM4T and VoxLingua107 (Intersection). Note that the macro-F1 on all languages for VL107HF is low because 21 languages are not covered by this model. We find that training on the additional languages slightly decreases the overall performance for the common set of languages, which is a direct consequence 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 accuracy (F1 difference greater than 5%) on 17 languages with an average gain of 14.6% without counting newly covered languages, while decreasing classification accuracy for 12 (with an average loss of 9.8%). We further filtered the data by applying a threshold on the LID score (likelihood). Language-specific thresholds have been tuned to maximize F1 score on the development data. By filtering out 8% of the data, we were able to further increase the F1 score by almost 3%.



Fig. 1 | **Schematic of the SEAMLESSM4T-V2 model.** The three main blocks of UNITY2 (S2ST fine-tuning) with its non-autoregressive (NAR) T2U are shown on the top left. Multitask-UNITY2 with its additional text encoder are shown on the

bottom left. Break down of the components of SEAMLESSM4T-V2 (a multitask-UNITY2 model) are shown on the right with the side panel showing the teacher T2U model used for pseudo-labelling (M4).

Table 2 | Comparison of similarity search error rates on all 200 FLORES languages and limited to the intersection of 98 languages on which each model has been trained

Model	Overall		Intersectio	n
	↓xsim	↓xsim++	↓xsim	↓xsim++
	(n=200)	(n=200)	(n=98)	(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

SONAR text embedding space

To mine automatically aligned translation data from language-identified segments, we rely on language- and modality-agnostic encoders. To this end, we build on the SONAR embedding space developed in ref. 8. Currently, we provide a single text encoder and decoder for 200 languages and speech encoders for 37 languages. The list of 200 languages is identical to the language list of the NLLB project⁶. In multilingual similarity search, xsim and xsim++¹⁵ are two well-known proxy metrics evaluating multilingual embedding spaces for the purpose of mining. As shown in Table 2, SONAR substantially outperforms other popular approaches such as LASER3¹⁶ or LaBSE¹⁷ with lower xsim and xsim++.

Furthermore, we evaluated the SONAR text encoders and decoders on T2TT tasks. The average performance over 200 languages is competitive compared with the medium-sized NLLB dense model, despite the replacement of encoder-decoder attention in SONAR with a bottleneck fixed-size embedding (see the T2TT columns of Extended Data Table 2). This result proves that the use of attention is not a requisite for reasonable translation accuracy. For more details on SONAR, see D2.

Training speech encoders

The speech encoders were trained with a teacher–student approach on speech transcriptions only (see Methods, 'SONAR'). Evaluating iterations of each speech encoder in an end-to-end loop, that is, mining and training S2TT or S2ST translation systems, would be compute-intensive. Instead, we connected the speech encoder with the SONAR text decoder and evaluated this zero-shot S2TT system as a proxy for the quality of the encoder. As shown in the S2TT columns of Extended Data Table 2, the SONAR speech encoders compare favourably to a model such as WHISPER-LARGE-V2 on FLORES⁶ and FLEURS¹⁸ datasets, which was trained on massive amounts of supervised data. Gaps in accuracy can be observed in some high-resource languages such as German, Russian or Portuguese, but SONAR outperforms WHISPER-LARGE-V2 in several low-resource languages such as Swahili or Bengali (see Supplementary Table 8).

SEAMLESSALIGN

The SONAR text and speech encoders were used to mine for three types of aligned data: (1) English speech to non-English texts (Sen2Txx); (2) non-English speech to English texts (Sxx2Ten); and (3) non-English speech to English speech (Sxx2Sen). SEAMLESSALIGN provides 202,796 h of audio in Sen2Txx, 239,767 h of audio in Sxx2Ten and 29,161 h of audio in Sxx2En. These aligned data were mined from a total of 2.5M h of raw audio (of which English is nearly 40%). SONAR speech encoders were trained on 43,772 h of supervised ASR data. For statistics per language, see Supplementary Table 8.

For the text domain, we use the same data consolidated by the NLLB project⁶. The amount varies from 33 million or 55 million sentences for low-resource languages such as Maltese or Swahili, respectively, to 22,000 million English sentences.

Except for Maltese, for which we only had access to a small amount of raw audio, we were able to mine more than 100 h of speech alignments

with English speech for all languages. The alignments with English texts reached a thousand hours for most languages and exceeded 10,000 h for high-resource languages. Overall, SEAMLESSALIGN covers 37 languages for a total of 470,000 h.

Adding such large amounts of data to train a multilingual translation system is a substantial computational challenge. As described in the Methods, 'Modelling', not all of these data were used for modelling, but only a subset with the highest SONAR alignment scores. As our mined data can help support many different use cases, we open-sourced the meta-data needed to guide its recreation (up to a SONAR threshold of 1.15; see Methods, 'SpeechAlign') 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.

Modelling multitask translation systems

Combining modelling techniques outlined in the Methods, 'Modelling', with additional data from SEAMLESSALIGN (see Methods, 'Data'), we trained SEAMLESSM4T models in two sizes: large with 2.3B parameters and medium with 1.2B parameters. SEAMLESSM4T-MEDIUM is intended to be an accessible test bed to either fine-tune, improve on or engage in analysis with. We further trained an improved version of the large SEAM-LESSM4T, dubbed SEAMLESSM4T-V2, with a better speech encoder (see Methods, 'Unsupervised speech pretraining') and a more powerful unit decoder (see section 'S2ST fine-tuning'). All SEAMLESSM4T models support 96 source languages in the text modality and more than 100 source languages in the speech modality. On the target side, the models can output 96 languages in text form and 35 in speech form. The amount of supervised data per direction and per source (for example, M4 or SEAMLESSALIGN) is detailed in Supplementary Tables 12 and 13. This shows that, for some translation directions and given the lack of supervised data, our models will be evaluated zero-shot.

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 synthesis¹⁹). To generate text hypotheses, we decoded with beam-search. We scored with chrF2++ for T2TT and BLEU for S2TT. We measure BLEU scores with SacreBLEU and provide the signatures in Supplementary Table 2. For ASR, we scored with WER (word error rate) on normalized transcriptions and references following ref. 20.

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 a text-to-unit module (T2U; see section 'S2ST fine-tuning') to search for the best unit sequence hypothesis. We used a beam width of 5 for both searches. We evaluated S2ST and T2ST accuracy with ASR-BLEU using WHISPER models. We set the decoding temperature of WHISPER at zero and used greedy decoding to ensure a deterministic behaviour of the ASR model. The transcribed hypotheses, as well as the references, are normalized following ref. 20 before computing BLEU scores.

Comparison with cascaded approaches for speech translation

On the set of languages supported by both SEAMLESSM4T and WHIS-PER, we compare in Table 3 (S2TT columns) the performance of our direct S2TT model to that of cascaded models, namely, combinations of WHISPER ASR models and NLLB T2TT models. SEAMLESSM4T-V2 surpasses the cascaded models with less than 3B parameters in X-eng directions by 4.6 BLEU points (from 22.0 to 26.6) and in eng-X directions by 1 BLEU point (from 21.1 to 22.2). We also added to the comparison in Table 3 cascaded models with the large NLLB-3.3B T2TT model. These models exceed 4B parameters and are largely surpassed by SEAMLESSM4T-V2 in X-eng (+3.9); they only marginally outperform SEAMLESSM4T-V2 in eng-X directions by 0.2 BLEU points.

Table 3 | State-of-the-art S2TT/S2ST models

Model			S2TT FLEURS (↑BLEU)		eurs Bleu)	S2ST CVSS (↑ASR-BLEU)	
	Size	X-eng (n=81)	eng-X (n=88)	X-eng (n=81)	eng-X (n=26)	X-eng (n=21)	
WL-V2 (S2TT)	1.5B	17.9	-	17.8	-	29.6	
WL-V3 (S2TT)	1.5B	16.9ª	-				
A8B (S2TT)	8B	19.7	-				
WM (ASR)+NLLB-1.3B	2B	19.7	20.7	20.7	21.5		
WM (ASR)+NLLB-3.3B	4B	20.4	22.0	21.4	22.4		
WL-V2 (ASR)+NLLB-1.3B	2.8B	22.0	21.2	22.9	21.8		
WL-V2 (ASR)+NLLB-3.3B	4.8B	22.7	22.4	23.7	22.7		
SEAMLESSM4T-MEDIUM	1.2B	20.9	19.4	20.2	15.8	30.6	
SEAMLESSM4T-LARGE	2.3B	24.1	21.8	25.8	20.9	35.7	
SEAMLESSM4T-V2	2.3B	26.6	22.2	29.7	26.1	39.2	

Comparison with cascaded ASR+T2TT models on FLEURS S2TT, and with two-stage and three-stage cascaded models on FLEURS and CVSS S2ST X-eng. S2ST cascaded systems rely on a TTS model as the last subsystem, for this we use YOURTTS²² to synthesize English speech and MMS²³ to synthesize non-English speech (these models are not factored into the system size and are omitted from the nomenclature of the models). We abbreviate WHISPER-LARGE as WL, WHISPER-MEDIUM as WM and AUDIOPALM-2-8B-AST as A8B. The results of the cascaded models are shown in italic and the best score for each task is shown in bold.

^aWe evaluated WHISPER-LARGE-V3 on S2TT FLEURS X-eng using https://github.com/openai/ whisper/. For WHISPER-LARGE-V2, we used the results from ref. 20.

Compared with previous direct S2TT SOTA models that lagged behind cascaded systems (for example, AUDIOPALM-2-8B-AST. ref. 21), SEAMLESSM4T-V2 improves on FLEURS X-eng S2TT BLEU score by 6.9 points (from 19.7 to 26.6; that is, an improvement of 35%).

Table 3 (S2ST columns) also compares S2ST between SEAMLESSM4T models and cascaded models. For S2ST, we explore two options for cascading: (1) three-stage with ASR, T2TT and TTS and (2) two-stage with S2TT and TTS. Both types of cascaded systems rely on a TTS model to synthesize translated speech, and for this we use YOURTTS²² when synthesizing English speech and MMS²³ when synthesizing speech in the 26 non-English languages of comparison (overlap between the support of SEAMLESSM4T and the support of TTS systems of MMS). Our SEAMLESSM4T-LARGE outperforms the two-stage cascaded models on FLEURS X-eng directions by 8 ASR-BLEU points (17.8–25.8). It also outperforms stronger three-stage cascaded models (WHISPER-LARGE-V2 + NLLB-3.3B + YOURTTS) by 2.1 ASR-BLEU points (23.7-25.8). The improved SEAMLESSM4T-V2 further strengthen this lead on S2ST FLEURS X-eng with an additional +3.9 ASR-BLEU points (25.8-29.7). On CVSS, SEAMLESSM4T-V2 outperforms the two-stage cascaded model (WHISPER-LARGE-V2 + YOURTTS) by a large margin of 9.6 ASR-BLEU points (29.6–39.2). On FLEURS S2ST eng-X directions, we reduce the evaluation set to the 26 languages supported by both TTS of MMS and SEAMLESSM4T. The medium-size model (SEAMLESSM4T-MEDIUM) scores an average ASR-BLEU of 15.8. SEAMLESSM4T-LARGE achieves an average ASR-BLEU of 20.9 and with its improved speech encoder and non-autoregressive T2U model, SEAMLESSM4T-V2 further gains +5.2 ASR-BLEU points (20.9-26.1). By contrast, the best three-stage cascaded system with MMS (WHISPER-LARGE-V2 + NLLB-3.3B + MMS) scores an average 22.7 ASR-BLEU, that is, SEAMLESSM4T-V2 surpasses state-of-the-art cascaded models by 15% (22.7-26.1).

We share in Supplementary Information section IV.1 evaluation results for the tasks of S2TT and S2ST with additional metrics, including our modality-agnostic BLASER 2.0.

Multitasking results

We report in Table 4 results on the FLEURS benchmark for the tasks of ASR and zero-shot T2ST (X-eng and eng-X), and the related FLORES

Model		ASR T2TT (↓WER) (↑chrF2++)		+)	T2ST (↑ASR-BL	
	Size	FLEURS (n=77)	FLORES X-eng (n=95)	FLORES eng–X (n=95)	FLEURS X-eng (n=88)	FLEURS eng–X (n=26)
NLLB-3.3B	3.3B	-	60.7	49.6		
NLLB-3.3B + YOURTTS/MMS	3.4B	-	-	-	36.4	23.7
WHISPER-LARGE-V2	1.5B	41.7				
SEAMLESSM4T- MEDIUM	1.2B	21.9	55.4	48.4	26.3	18.4
SEAMLESSM4T- LARGE	2.3B	22.6	60.8	50.9	34.1	21.8
SEAMLESSM4T- LARGE V2	2.3B	18.5	59.2	49.3	35.9	27.6

Performance of SEAMLESSM4T-LARGE on the auxiliary tasks of ASR and T2TT and the zero-shot task of T2ST compared with SOTA single-task models. The results of cascaded models are shown in italics and the best result in each task is shown in bold. 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 ref. 20.

benchmark for T2TT (X-eng and eng-X). In ASR, SEAMLESSM4T-LARGE outperforms WHISPER-LARGE-V2²⁰ on the overlapping 77 supported languages with a WER reduction of 46% (from 41.7 to 22.6), whereas SEAMLESSM4T-V2 improves over WHISPER-LARGE-V2 by 56% (from 41.7 to 18.5). We also compared in Supplementary Table 9 against MMS²³ on FLEURS-54, a subset of FLEURS languages that MMS and WHISPER both support. SEAMLESSM4T-V2 outperforms the MMS variants evaluated with CTC by more than 38% WER (from 31.0 to 19.1), but it is surpassed by the variants that leverage monolingual n-gram language models (5% better WER with 18.6).

In the T2TT support task, results in Table 4 show that our SEAM-LESSM4T models are on par with NLLB-3.3B (ref. 6) in both X-eng and eng-X directions.

We next evaluated SEAMLESSM4T models on the task of T2ST in a zero-shot way. Given that FLEURS collected three recordings by three different native speakers for each sample, we randomly selected one for the task of T2ST (the input being text). We report in Table 4 (the T2ST columns) a comparison between SEAMLESSM4T models and cascaded models with NLLB and either YOURTTS (English TTS) or MMS (non-English TTS) for synthesizing translated text. We averaged ASR-BLEU scores over 88 X-eng directions (the overlap between FLEURS and the languages supported by SEAMLESSM4T). We also averaged ASR-BLEU over 26 eng-X directions (overlap of SEAM-LESSM4T with TTS models of MMS). Compared with cascaded models, the zero-shot capability of SEAMLESSM4T-LARGE V2 is on par with NLLB-3.3B + YOURTTS in X-eng and outperforms NLLB-3.3B + MMS by more than +3.9 ASR-BLEU points in eng-X (from 23.7 to 27.6). This result demonstrates that (1) the quality of SEAMLESSM4T on zero-shot T2ST is on par with the supervised tasks and (2) non-English speech source is the most challenging input to translate with our model.

To further understand where the improvements in FLEURS S2TT Xeng directions were coming from, we bucketed languages by resource level (see the exact list of languages in Supplementary Table 12) and report average BLEU scores per resource level in Table 5. The results show that SEAMLESSM4T strongly improves the quality of translating from low-resource languages with an improvement of +10.2 BLEU (from 18.0 to 28.2, that is, 57% 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 +7.8 BLEU in that subset of directions suggests that this improvement goes beyond sheer

Table 5 | FLEURS S2TT X-eng by resource level

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	27.0	25.3	25.6	27.1		
SEAMLESSM4T-LARGE V2	28.8	28.3	28.2	29.8		

In each resource level (that is, high, medium and low), we averaged over languages that are covered by all three models. In low, we excluded low-resource languages that are evaluated as zero-shot by AUDIOPALM-2-8B-AST. In each subset, the highest BLEU score is shown in bold.

supervision but instead should be attributed to the quality of supervised data and the training recipes.

Automatic and human evaluation

Semantic accuracy in speech translation is generally evaluated with the automatic metric BLEU¹⁰ for S2TT or its extension ASR-BLEU for S2ST. Moreover, we use BLASER 2.0 (ref. 24), an extension of BLASER²⁵, which now enables modality-agnostic evaluation and quality estimation for both speech and text.

To complement the utility of automatic metrics, we also relied on extensive human evaluation of our models. In the following, we provide human evaluation for S2TT and S2ST tasks with the XSTS (cross-lingual semantic textual similarity) protocol²⁶ and MOS (mean opinion score) protocol for speech outputs (see Methods, 'Human evaluation') on the FLEURS test set. However, we cover a limited number of models (SEAMLESSM4T-LARGE, SEAMLESSM4T-LARGE V2 and a cascaded baseline composed of WHISPER-LARGE-V2 for ASR, NLLB 3.3B for translation and YOURTTS or MMS for TTS for the S2ST task) and translation directions (23 languages from and into English, 10 languages X-eng for MOS) because of resource restrictions.

XSTS scores show that SEAMLESSM4T-LARGE V2 outperforms both the cascaded baseline systems and SEAMLESSM4T-LARGE in terms of both average language-level XSTS score and win rate (the fraction of evaluated languages for which XSTS performance is superior), for all tasks and language directions with high confidence. For the S2ST task, where relative performance of SEAMLESSM4T-V2 was the strongest, win rate of SEAMLESSM4T-V2 approaches 100% compared with both cascaded baseline and SEAMLESSM4T-LARGE for both X-eng and eng-X directions, with average language XSTS scores about 0.5 points higher than the cascaded baseline and 0.36–0.51 points higher compared with SEAMLESSM4T-LARGE for eng-X and X–eng, respectively. See Supplementary Tables 22 and 24 for full language-level and summarized XSTS results, respectively.

We also measure the quality of speech output for the S2ST using a Mean Opinion Score protocol that assesses (1) sound quality, (2) clarity of speech and (3) naturalness. We find that generally, across all MOS aspects, SEAMLESSM4T-LARGE V2 tends to be preferred to SEAM-LESSM4T-LARGE, which tends to be preferred to the cascaded model baselines, with the exception of X-eng, for which SEAMLESSM4T-LARGE generations are strongly preferred (+1 point average difference between SEAMLESSM4T-LARGE and SEAMLESSM4T-LARGE V2 generations), an unexpected result that may be a consequence of differences in model architectures that otherwise have improved generation quality. See Supplementary Tables 23 and 24 for full language-level and summarized MOS results, respectively.

Using the XSTS evaluations of the S2ST task, BLASER 2.0 (averaged over all evaluation items in a given language direction) achieves superior Spearman correlations with calibrated language-level XSTS scores in both X-eng direction (0.845 for BLASER 2.0 compared with 0.74 for

Table 6 | Results for S2TT and S2ST averaged across 28 directions that add toxicity

Model	FLEURS	X-eng	FLEURS	eng-X	HOLISTI	CBIAS
	ETOX% (↓)	MuTox (↓)	ETOX% (↓)	Mutox (↓)	ETOX% (↓)	MuTox (↓)
S2TT						
Baseline	0.21	0.05	0.23	0.08	0.32	0.39
SEAMLESSM4T- LARGE	0.20	0.02	0.24	0.07	0.32	0.37
SEAMLESSM4T-V2	0.22	0.01	0.16	0.08	0.15	0.39
SEAMLESSM4T-V2 + MinTox	0.22	0.01	0.07	0.01	0.03	0.37
S2ST						
Baseline	0.05	0.05	0.30	0.02	0.32	0.32
SEAMLESSM4T- LARGE	0.05	0.01	0.15	0.04	0.26	0.29
SEAMLESSM4T-V2	0.04	0.01	0.11	0.02	0.15	0.26
SEAMLESSM4T-V2 + MinTox	0.04	0.01	0.05	0.02	0.03	0.25

ETOX is ASR-ETOX in the case of speech outputs. The baseline corresponds to WHISPER-LARGE-V2 for S2TT X-eng; WHISPER-LARGE-V2+NLLB-3.3B for S2TT X-eng; WHISPER-LARGE-V2+YOURTTS for S2ST X-eng.

ASR-BLEU) and in particular for the eng–X direction (0.81 for BLASER 2.0 compared with 0.246 for ASR-BLEU). Similar results hold for the S2TT task (see Supplementary Table 21 for full results).

Finally, we tested our models for robustness in terms of noise and speaker variations by creating open robustness benchmarks based on FLEURS (see Methods, 'Robustness'). To that end, we find that SEAM-LESSM4T-V2 is, on average, approximately 42% and 66% more resilient against background noise and speaker variation, respectively, when compared with WHISPER-LARGE-V2 (see full results in Supplementary Information section V.2).

Responsible AI

Toxicity

Toxicity can be defined as instances of profanity or language that may incite hate, violence or abuse against an individual or a group (such as a religion, race or gender). When it comes to massively multilingual toxicity classifiers for text, the ETOX toolkit seems to be the only openly accessible option with the largest language coverage²⁷. In the context of speech translation, we were primarily worried about added toxicity– the introduction in translations of toxic elements not present in source utterances. Speech toxicity has been evaluated for English in ref. 28 and, recently, for tens of languages with MuTox²⁹.

Therefore, for speech and text multilingual toxicity detection, we used ETOX (or ASR-ETOX for S2ST) and MuTox (both for speech and text) as metrics to detect and evaluate added toxicity. For toxicity mitigation, we implemented two techniques to deal with added toxicity. Before training, we filtered out training pairs with imbalanced toxicity. Moreover, we used Mintox¹¹ at inference time (see Methods, 'Toxicity detection').

We computed added toxicity in two datasets (FLEURS and HOLIS-TICBIAS²⁷) across 24 translation directions with English (arb, ben, cat, ces, dan, deu, est, fin, fra, hin, ind, ita, nld, pes, pol, por, rus, slk, spa, swh, tgl, tur, urd, vie), using the languages at the intersection of the coverage of systems and MuTox having been benchmarked (note that MuTox has wider language coverage, similar to SONAR, but it has only been benchmarked in 30 languages²⁹). See results with more translation directions evaluated with ETOX in Supplementary Information section VI.1. Table 6 shows that although levels and types of added toxicity vary significantly as a function of language and dataset, the added

toxicity in our systems has a relatively low prevalence consistent across our two toxicity detection metrics (<0.4%). Table 6 shows that MinTox is capable of mitigating added toxicity consistently. The lowest toxicity for all modalities, directions and datasets is consistently obtained with SEAMLESSM4T-V2 + MinTox, achieving reductions of toxicity up to 5% in terms of MuTox (and up to 80% in terms of ETOX) when comparing with the same model without using MinTox, and up to 20% in terms of MuTox (up to 90% in terms of ETOX) when comparing with the baseline (see complete results in Supplementary Information section VI.1).

Gender bias

Gender bias in the context of MT can be defined as errors in grammatical gender determination. This bias may manifest explicitly as an overgeneralization to one gender when translating non-gendered to gendered forms (for example, outputs favouring masculine representations) or as a lack of robustness when varying the quality of the translation for sentences that differ only in gender inflection.

Previous work on this matter is mostly in the text modality³⁰⁻³² and tends to be English-centric, with few demographic axes and multilingual references. Similar efforts for the speech modality remain sparse^{33,34}.

We used MULTILINGUAL HOLISTICBIAS³⁵ and its speech extension to compare the performance of S2TT and S2ST. The eng–X direction enables 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 English–Spanish language pair 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, whereas in the other direction, we can evaluate the robustness of the translation to gender inflection (see Methods, 'Speech extension of MULTILINGUAL HOLISTICBIAS').

We conducted a set of comprehensive evaluations on translation biases for S2TT and S2ST (see Extended Data Table 3 for average results and Supplementary Information section VI.2 for detailed results). SEAM-LESSM4T-V2 consistently improves robustness in gender variations across metrics and tasks. When compared with the previous model, SEAMLESSM4T-V2 improves over SEAMLESSM4T-LARGE by 0.4% in S2TT and by 0.1% BLASER 2.0 in S2ST, and it beats the external baseline system of WHISPER-LARGE-V2 (+YOURTTS) by 0.1% in S2TT and by 0.9% BLASER 2.0 for S2ST. However, SEAMLESSM4T-V2 is not able to consistently improve in terms of gender overgeneralization compared with the previous model. SEAMLESSM4T-V2 is comparable in terms of BLASER 2.0 to SEAMLESSM4T-LARGE, but it lags far behind in terms of ASRchrf (by 2.2%), and overgeneralization is increased by 0.2% when it comes to S2TT. Although we can increase bias robustness by improving the overall quality of the model, it seems that we need specific techniques to counteract the overgeneralization of the model towards a specific gender.

Social impact and conclusion

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 before³⁶. The current social order places a demand on the world-readiness of a person¹², 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 our mother tongue for both instrumental (that is, employment or schooling) and cultural reasons (that is, to become a global citizen). That said, although we believe that language acquisition should remain a key mechanism for boosting our world-readiness, we acknowledge that doing so requires 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. SEAMLESSM4T-supported applications could act as a co-piloting mechanism that supports users in multilingual conversations and boosts their confidence in speech-heavy interactions. As speech-based interfaces (for example, audio assistants, voice memos and live transcriptions) and auditory content (for example, podcasts, audiobooks and short-form videos) become ever more present, SEAMLESSM4T-enabled downstream applications could unlock a greater variety of multilingual experiences.

From an inclusion standpoint, the focus of SEAMLESSM4T on multimodality could make a meaningful difference in augmenting the world-readiness of those with accessibility needs and those whose languages can be transcribed with multiple writing systems. For many who lack reading or writing skills, or cannot rely on sight (that is, people who are blind or with visual impairment), voice-assisted technologies are essential to how they communicate and stay connected³⁷. The ability to translate speech gives these groups more comprehensive access to information not only beyond their native languages but also in a manner that is better suited for their communicative needs.

As with most technologies, the distribution of benefits varies based on user demographics and social situation³⁸. Although 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, similar to many other speech technologies, the ASR performance of SEAMLESSM4T may vary based on gender, race, accent or language^{39,40}. Moreover, the performance of our system in translating slang or proper nouns may also be inconsistent across high and low-resource languages.

Another challenge for S2ST is that speech tends to hinge on immediate reception and feedback more than written language does. In other words, speakers are 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 risk when it comes to mistranslations or toxicity. We urge researchers and developers who fine-tune or build artefacts using SEAMLESSM4T to think critically about design features that could help users circumvent these potential obstacles. Importantly, 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 (for example, rhythm, stress, intonation or tone) and emotional components that deserve further research⁴¹. To create S2ST systems that feel organic and natural, more research should be directed at output generation that preserves expressivity⁴². Moreover, the consummate realization of the Babel Fish requires deeper investments into research on low-latency speech translation. Developing systems that enable streaming (that is, incrementally translating an input sentence as it is being presented) may increase the adoption of these systems across institutional contexts^{43,44}. We hope that SEAMLESSM4T opens up new possibilities for both these research areas.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-08359-z.

2. Wahlster, W. (ed.) Verbmobil: Foundations of Speech-to-Speech Translation (Springer, 2000).

Lavie, A. et al. Janus-III: speech-to-speech translation in multiple languages. In Proc. 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing Vol. 1, 99–102 (IEEE, 1997).

- Nakamura, S. et al. The atr multilingual speech-to-speech translation system. IEEE Trans. Audio Speech Lang. Process. 14, 365–376 (2006).
- Jia, Y. et al. Direct speech-to-speech translation with a sequence-to-sequence model. In Proc. Interspeech 2019, 1123–1127 (ISCA, 2019).
- Lee, A. et al. Direct speech-to-speech translation with discrete units. In Proc. 60th Annual Meeting of the Association for Computational Linguistics (eds Muresan, S. et al.) Vol. 1, 3327–3339 (Association for Computational Linguistics, 2022).
- NLLB Team Scaling neural machine translation to 200 languages. Nature 630, 841–846 (2024).
 Agarwal, M. et al. Findings of the IWSLT 2023 evaluation campaign. In Proc. 20th International Conference on Spoken Language Translation (IWSLT 2023) (eds Salesky, E. et al.) 1–61 (Association for Computational Linguistics, 2023).
- Duquenne, P.-A., Schwenk, H. & Sagot, B. SONAR: sentence-level multimodal and languageagnostic representations. Preprint at https://arxiv.org/abs/2308.11466 (2023).
- Popović, M. chrF: character n-gram F-score for automatic MT evaluation. In Proc. Tenth Workshop on Statistical Machine Translation (eds Bojar, O. et al.) 392–395 (Association for Computational Linguistics, 2015).
- Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In Proc. 40th Annual Meeting of the Association for Computational Linguistics (eds Isabelle, P. et al.) 311–318 (Association for Computational Linguistics, 2002).
- Costa-jussà, M. R., Dale, D., Elbayad, M. & Yu, B. Added toxicity mitigation at inference time for multimodal and massively multilingual translation. In Proc. 25th Annual Conference of the European Association for Machine Translation (eds Scarton, C. et al.) Vol. 1, 360–372 (European Association for Machine Translation, 2024).
- ACTFL. World-readiness standards for learning languages. https://www.actfl.org/educatorresources/world-readiness-standards-for-learning-languages (2023).
- Desplanques, B., Thienpondt, J. & Demuynck, K. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In Proc. Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event (eds Meng, H. et al.) 3830–3834 (ISCA, 2020).
- Valk, J. & Alumäe, T. VOXLINGUA107: a dataset for spoken language recognition. In Proc. 2021 IEEE Spoken Language Technology Workshop, 652–658 (IEEE, 2021).
- Chen, M., Heffernan, K., Çelebi, O., Mourachko, A. & Schwenk, H. xSIM++: An improved proxy to bitext mining performance for low-resource languages. In Proc. 61st Annual Meeting of the Association for Computational Linguistics (eds Rogers, A. et al.) Vol. 2, 101–109 (Association for Computational Linguistics, 2023).
- Heffernan, K., Çelebi, O. & Schwenk, H. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2101–2112 (eds Goldberg, Y. et al.) (Association for Computational Linguistics, 2022).
- Feng, F., Yang, Y., Cer, D., Arivazhagan, N. & Wang, W. Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (eds Muresan, S. et al.) Vol. 1, 878–891 (Association for Computational Linguistics, 2022).
- Conneau, A. et al. FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech. In 2022 IEEE Spoken Language Technology Workshop (SLT), 798–805 (IEEE, 2023).
- 19. Zhang, Z.-H. et al. Speak foreign languages with your own voice: cross-lingual neural codec language modeling. Preprint at https://arxiv.org/abs/2303.03926 (2023).
- Radford, A. et al. Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning, article no. 1182, 28492–28518 (ACM, 2023).
- Rubenstein, P. K. et al. AudioPaLM: a large language model that can speak and listen. Preprint at https://arxiv.org/abs/2306.12925 (2023).
- Casanova, E. et al. YourTTS: towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In Proc. 39th International Conference on Machine Learning, Vol. 162, 2709–2720 (PMLR, 2022).
- Pratap, V. et al. Scaling speech technology to 1,000+ languages. J. Mach. Learn. Res. 25, 1–52 (2024).
- Dale, D. & Costa-jussà, M. R. BLASER 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (eds Al-Onaizan, Y. et al.) 16075–16085 (Association for Computational Linguistics, 2024).
- Chen, P.-J. et al. Speech-to-speech translation for a real-world unwritten language. In Findings of the Association for Computational Linguistics: ACL 2023 (eds Rogers, A. et al.) 4969–4983 (Association for Computational Linguistics, 2023).
- Licht, D. et al. Consistent human evaluation of machine translation across language pairs. In Proc. 15th Biennial Conference of the Association for Machine Translation in the Americas (eds Duh, K. & Guzmán, F.) Vol. 1, 309–321 (Association for Machine Translation in the Americas, 2022).
- Costa-jussà, M. et al. Toxicity in multilingual machine translation at scale. In Findings of the Association for Computational Linguistics: EMNLP 2023, 9570–9586 (Association for Computational Linguistics, 2023).
- Ghosh, S., Lepcha, S., Sakshi, S., Shah, R. R. & Umesh, S. DeToxy: a large-scale multimodal dataset for toxicity classification in spoken utterances. In *Proc. Interspeech* 2022, 5185–5189 (2022); https://api.semanticscholar.org/CorpusID:247940269.
- Costa-jussà, M. R. et al. MuTox: universal multilingual audio-based toxicity dataset and zero-shot detector. In Findings of the Association for Computational Linguistics: ACL 2024 (eds Ku, L.-W. et al.) 5725–5734 (Association for Computational Linguistics, 2024).
- Stanovsky, G., Smith, N. A. & Zettlemoyer, L. Evaluating gender bias in machine translation. In Proc. 57th Annual Meeting of the Association for Computational Linguistics (eds Korhonen, A. et al.) 1679–1684 (Association for Computational Linguistics, 2019).
- Levy, S., Lazar, K. & Stanovsky, G. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (eds Moens, M. F. et al.) 2470–2480 (Association for Computational Linguistics, 2021).

- Costa-jussà, M. R. et al. Interpreting gender bias in neural machine translation: Multilingual architecture matters. Proc. AAAI Conf. Artif. Intell. 36, 11855–11863 (2022).
- Costa-jussà, M. R., Basta, C. & Gállego, G. I. Evaluating gender bias in speech translation. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (eds Calzolari, N. et al.) 2141–2147 (European Language Resources Association, 2022).
- Bentivogli, L. et al. Gender in danger? Evaluating speech translation technology on the MuST-SHE corpus. In Proc. 58th Annual Meeting of the Association for Computational Linguistics (eds Jurafsky, D. et al.) 6923–6933 (Association for Computational Linguistics, 2020).
- 35. Costa-jussà, M. et al. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 14141–14156 (Association for Computational Linguistics, 2023).
- Zuckerman, E. The polyglot internet. https://ethanzuckerman.com/the-polyglot-internet/ (2008).
- Belekar, A., Sunka, S., Bhawar, N. & Bagade, S. Voice based e-mail for the visually impaired. Int. J. Comput. Appl. 175, 8–12 (2020).
- Wang, S., Cooper, N. & Eby, M. From human-centered to social-centered artificial intelligence: Assessing ChatGPT's impact through disruptive events. *Big Data Soc.* https://doi.org/10.1177/20539517241290220 (2024).
- Koenecke, A. et al. Racial disparities in automated speech recognition. Proc. Natl Acad. Sci. USA 117, 7684–7689 (2020).
- Ngueajio, M. K. & Washington, G. Hey ASR system! Why aren't you more inclusive? Automatic speech recognition systems' bias and proposed bias mitigation techniques. A literature review. In Proc. International Conference on Human-Computer Interaction (eds Chen, J. Y. C. et al.) Vol. 13518, 421–440 (Springer, 2022).
- Elbow, P. The shifting relationships between speech and writing. Coll. Compos. Commun. 36, 283–303 (1985).
- Trilla, A. & Alias, F. Sentence-based sentiment analysis for expressive text-to-speech. IEEE Trans. Audio Speech Lang. Process. 21, 223–233 (2013).
- Iranzo-Sánchez, J., Civera, J. & Juan, A. From simultaneous to streaming machine translation by leveraging streaming history. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 6972–6985 (Association for Computational Linguistics, 2022).
- 44. Rybakov, O. et al. Streaming Parrotron for on-device speech-to-speech conversion. Preprint at https://arxiv.org/abs/2210.13761 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Composed on the article's Creative Commons Licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permitssion directly from the copyright holder. To view a copy of this licence, visit http:// creativecommons.org/licenses/by-nc-nd/4.0/.

© Meta 2025

SEAMLESS Communication Team

Loïc Barrault¹⁴, Yu-An Chung¹⁴, Mariano Coria Meglioli¹⁴, David Dale¹⁴, Ning Dong¹⁴, Paul-Ambroise Duquenne¹²⁴, Hady Elsahar^{1,4}, Hongyu Gong¹⁴, Kevin Heffernan¹⁴, John Hoffman¹⁴, Christopher Klaiber¹⁴, Pengwei Li¹⁴, Daniel Licht¹⁴, Jean Maillard¹⁴, Alice Rakotarison^{1,4}, Kaushik Ram Sadagopan¹⁴, Guillaume Wenzek¹⁴, Ethan Ye¹⁴, Bapi Akula¹, Peng-Jen Chen¹, Naji El Hachem¹, Brian Ellis¹, Gabriel Mejia Gonzalez¹, Justin Haaheim¹, Prangthip Hansanti¹, Russ Howes¹, Bernie Huang¹, Min-Jae Hwang¹, Hirofumi Inaguma¹, Somya Jain¹, Elahe Kalbassi¹, Amanda Kallet¹, Ilia Kulikov¹, Janice Lam¹, Daniel Li¹, Xutai Ma¹, Ruslan Mavlyutov¹, Benjamin Peloquin¹, Mohamed Ramadan¹, Abinesh Ramakrishnan¹, Anna Sun¹, Kevin Tran¹, Tuan Tran¹, Igor Tufanov¹, Vish Vogeti¹, Carleigh Wood¹, Yilin Yang¹, Bokai Yu¹, Pierre Andrews¹⁵, Can Balioglu¹⁵, Marta R. Costa-jussà^{15^[5]}, Onur Çelebi¹⁵, Maha Elbayad¹⁵, Cynthia Gao¹⁵, Francisco Guzmán¹⁵, Justine Kao¹⁵, Ann Lee¹⁵, Alexandre Mourachko¹⁵, Juan Pino¹⁵, Sravya Popuri¹⁵, Christophe Ropers¹⁵, Safiyyah Saleem¹⁵, Holger Schwenk¹⁵, Paden Tomasello¹⁵, Changhan Wang¹⁵, Jeff Wang¹⁵ & Skyler Wang^{13,5}

¹Foundational AI Research, (FAIR) Meta, Menlo Park CA, USA. ²INRIA, Meta, Paris, France. ³UC Berkeley, Berkeley, CA, USA. ⁴These authors contributed equally: Loic Barrault, Yu-An Chung, Mariano Coria 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. ⁵These authors jointly supervised this work: Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Çelebi, 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. ⁵⁵e-mail: costajussa@meta.com

Article Methods

Data

Developing an effective multilingual and multimodal translation system such as SEAMLESSM4T requires sizeable resources across languages and modalities. Some human-labelled resources for translation are freely available, but often limited to a small set of languages or in very specific domains. Well-known examples are parallel text collections such as Europarl⁴⁵ and the United Nations Corpus⁴⁶. A few human-created collections also involve the speech modality, such as CoVoST^{47,48} and mTEDx⁴⁹. Yet no open dataset currently matches the size of those used in initiatives such as WHISPER²⁰ or USM⁵⁰, which proved to unlock unprecedented performance.

Parallel data mining emerges as an alternative to using closed data, in terms of both 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 find parallel instances based on a similarity metric. Mining is then performed by pairwise comparison over massive monolingual corpora, in which sentences with similarity above a certain threshold are considered mutual translations^{51,52}. This approach was first introduced using the multilingual LASER space⁵³. Teacher–student training was then used to scale this approach to 200 languages^{6,16} and, subsequently, the speech modality^{54,55}.

Audio processing and speech LID. We started with 4 million hours of raw audio originating from a publicly available repository of crawled web data on which we applied several cleaning and filtering operations. To maximize the recall of mining, it is important that all segments have a similar granularity. For the text domain, a sentence is generally well-defined. This is less obvious for raw speech because pauses are not necessarily used at sentence boundaries. First, we used an open Voice Activity Detection model⁵⁶ to split audio files into shorter segments. Second, a newly developed speech LID model was applied to each segment. Our model follows the ECAPA-TDNN architecture¹³ and extends the open-source model trained on VoxLingua107¹⁴ by 15 new languages. Finally, we applied an over-segmentation approach that simultaneously proposed multiple, potentially overlapping speech segmentations. We relied on the mining approach to align the most likely ones. Supplementary Fig. 1 shows this pipeline.

SONAR. The SONAR text and speech encoders were developed in ref. 8 using a two-step approach (Supplementary Fig. 2). First, a massively multilingual representation was learnt for the text modality only. Then, teacher-student training was used to extend the embedding space to the speech modality. The text embedding space was trained with an encoder-decoder approach using a combination of multiple objectives: translation, denoising auto-encoding and mean squared error (MSE) loss objective in the sentence embedding space. The training data were identical to those used to train the NLLB model⁶, that is, parallel data to translation from 200 to 200 languages. Speech encoders were trained only on ASR data and by grouping languages into linguistic genealogical groups following ref. 16, for example, Italic, Common Turkic or Indo-Iranian languages. To obtain optimal performance, we determined the optimal convergence separately for each language (that is, when to stop training). This yielded a separate speech encoder for each language. The amount of available ASR data for each language is provided in Supplementary Table 8. The speech encoders were initialized with w3v-best 2.0 speech front end. Preceding work performed max pooling or mean pooling of the output states of the speech front end to obtain a fixed-size embedding of the speech signal^{54,57}. An ablation study has shown that better results can be obtained by using a three-layer transformer decoder⁸. Teacher-student training consisted of minimizing the MSE loss with respect to the embedding of the ASR text transcriptions. These embeddings were obtained by the SONAR text encoder, which was kept constant. No translations (into English) were used.

SpeechAlign. We first calculated the embeddings of all over-segmented speech segments. For the text domain, we used exactly the same texts as the NLLB project⁶ and embedded them with the SONAR encoder. Exhaustive pairwise comparison can be efficiently performed with the FAISS toolkit⁵⁸. Similarity is measured with a margin criterium as first introduced in ref. 52:

$$score(x, y) = margin$$

$$\left(\cos(x,y),\sum_{z\in NN_k(x)}\frac{\cos(x,z)}{2k}+\sum_{v\in NN_k(y)}\frac{\cos(y,v)}{2k}\right) \quad (1)$$

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

As an example, this amounts to comparing a hundred thousand hours of speech with more than 20,000 million English sentences, which yielded about eight thousand hours of aligned Arabic speech.

Modelling

The SEAMLESSM4T models rely on our multitask UNITY architecture. Our proposed unified translation model builds on vanilla UNITY⁵⁹, a two-pass decoding framework that first generates text and subsequently generates speech by predicting discrete acoustic units (see section 'Multilingual discrete acoustic units'). Compared with the vanilla UNITY model⁵⁹, (1) the core S2TT model, initialized from scratch in UNITY, is replaced with an X2T model that supports text as input and is pretrained to jointly optimize the tasks of ASR, S2TT and T2TT (see section 'X2T fine-tuning'), and (2) the shallow T2U model (referred to as T2U unit encoder and second-pass unit decoder in ref. 59) is replaced with a deeper transformer-based encoder-decoder model with six transformer layers that are pretrained on ASR data (see section 'S2ST fine-tuning'). An improved version of UNITY, dubbed UNITY2, replaces the autoregressive T2U with a new non-autoregressive (NAR) T2U decoder. This NAR T2U model delivers stronger accuracy because of its hierarchical upsampling from subwords to characters and then to units.

The pretraining of X2T yielded a stronger speech encoder and a higher quality first-pass text decoder, whereas the scaling and pretraining of the T2U model allowed us to better handle multilingual unit generation without interference. Furthermore, the switch to non-autoregressive T2U decoding improved S2ST inference speed by three times.

Multilingual discrete acoustic units. Recent works have achieved state-of-the-art translation performance by using self-supervised discrete acoustic units as targets for building direct speech translation models^{5,60}. This consists of decomposing the S2ST problem into a speech-to-unit translation step and a unit-to-speech conversion step. We extracted continuous speech representations using XLS-R⁶¹ and mapped these representations to discrete tokens. The set of discrete tokens (also referred to as unit vocabulary) is learnt by applying a *k*-means algorithm to a set of multilingual audio samples. The *k*-means 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. We used a unit vocabulary size K = 10,000 with features from the 35th layer of XLS-R-1B to represent the 35 supported target languages.

For the unit-to-speech conversion step, we followed ref. 62 and built a multilingual vocoder for speech synthesis from the learnt multilingual units. This model is responsible for synthesizing audios from a sequence of units that SEAMLESSM4T models will predict.

Unsupervised speech pretraining. Self-supervised pretraining with unlabelled speech audio data is a practical approach for leveraging

unlabelled data. With pretraining, we can bootstrap the quality of translation models and make the most out of our supervised paired data. We pretrained a speech encoder following our improved W2V-BERT 2.0. It follows w2v-BERT⁶³ in combining contrastive learning with masked prediction learning. W2V-BERT 2.0 uses more codebooks and an additional masked prediction task using random projection quantizers⁶⁴ (RPQ). Our W2V-BERT 2.0 model is first trained on 1 million hours of open speech audio data that covers over 143 languages. It follows the w2v-BERT XL architecture⁶³, which has 24 Conformer layers⁶⁵ and approximately 600 million model parameters. For the v2 version, we scaled up the amount of unlabelled data from 1 million to 4.5 million hours of audio. The most recent and publicly available multilingual speech pretrained model is MMS²³. It is trained on only 0.5 million hours, spanning over 1.400 languages. The largest model in scale is USM⁵⁰. It is a proprietary multilingual speech pretrained model with 12 million hours of data and more than 300 languages in coverage.

Text-to-text translation models. The text processing components of our SEAMLESSM4T models were pretrained on the task of text-to-text translation, a much more resourced task than speech translation. Consider, for instance, the English–Italian direction, one of the highly resourced pairs in T2TT with more than 128 million parallel sentences– only 2 million pairs of English text paired with Italian audio are available for S2TT.

A key step in training multilingual text-to-text translation models is learning a shared vocabulary with a text tokenizer. Following ref. 6, we used SentencePiece⁶⁶ with the BPE algorithm⁶⁷ for this purpose. The tokenizer used in NLLB-200⁶ suffers from missing key Chinese characters because of artefacts of sampling. This sampling does not favour logo-graphic writing systems with a large number of unique symbols. To fix this issue, we forced the inclusion of these characters. Our new tokenizer improves the coverage of the MTSU top 5K Chinese characters from 54% to 84%.

To train our multilingual text-to-text model, we followed the same data preparation and training pipelines in ref. 6 using STOPES⁶⁸. Having a smaller language coverage allowed us to significantly decrease the size of the model to 1.3B parameters and only use NLLB-200 training data in the 95 SEAMLESSM4T languages.

Data augmentation with pseudo-labelling. As with any sequenceto-sequence task, speech translation performance is dependent on the availability of high-quality training data. However, the amount of human-labelled data is scarce compared with its T2TT or ASR counterparts. To address this shortage of labelled data, we resort to pseudo-labelling^{69,70} the ASR data with a multilingual T2TT model (for example, NLLB models) to generate pseudo-labelled S2TT data.

To augment S2ST data, it is common practice to use TTS models to convert text from speech-to-text data sets into synthetic speech^{4,5}. This synthetic speech is, in turn, converted into discrete units for training. This two-step unit extraction process is a slow process and is harder to scale given the dependencies on TTS models. We circumvented the need for synthesizing speech and trained multilingual text-to-unit (T2U) models on all 36 target speech languages. These models can directly convert the text into target discrete units and can be trained on ASR data sets that are readily available.

X2T fine-tuning. The first key part of our multitask UNITY framework is the X2T model, a multi-encoder sequence-to-sequence model with a conformer-based encoder⁶⁵ for speech input and another transformer-based encoder⁷¹ for text input. Both encoders were joined with the same text decoder and fine-tuned jointly to optimize the tasks of ASR, S2TT and T2TT.

Our X2T model consists of joining the speech encoder, W2V-BERT 2.0 from M2, post-fixed with a length adapter to downsample long audio sequences, with the text encoder-decoder from M3 (Supplementary

Fig. 4). For the length adapter, we used a modified version of M-adapter⁷², in which we replaced the three independent pooling modules for Q, K and V with a shared pooling module to improve efficiency.

X2T was fine-tuned on S2TT data triplets with speech audio (x^{speech}) in a source language $\langle \ell_s \rangle$, paired with its transcription (x^{text}) and text translation (y^{text}) in a target language $\langle \ell_t \rangle$. To enable meaning transfer across modalities, X2T model was fine-tuned to jointly optimize the following objective functions:

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

$$\mathcal{L}_{T2TT} = -\sum_{t=1}^{|y|} \log p(y_t^{\text{text}} | y_{< t}^{\text{text}}, x^{\text{text}}).$$
(2)

We additionally optimized an auxiliary objective function in the form of token-level knowledge distillation (\mathcal{L}_{KD}) to further transfer knowledge from the strong MT model to the student speech translation task (S2TT).

$$\mathcal{L}_{\text{KD}} = \sum_{t=1}^{|y|} D_{\text{KL}}[p(.|y_{t}^{\text{text}}, x^{\text{text}}) \parallel p(.|y_{t}^{\text{text}}, x^{\text{speech}})].$$
(3)

The final loss is a weighted sum of all three losses: $\mathcal{L} = \alpha \mathcal{L}_{S2TT} + \beta \mathcal{L}_{T2TT} + \gamma \mathcal{L}_{KD}$, where α , β and γ are scalar hyper-parameters tuned on the development data.

S2ST fine-tuning. In the last stage of fine-tuning multitask UNITY, we initialized the model with the pretrained X2T model (see section 'X2T fine-tuning') and a pretrained T2U model, similar to the one used for pseudo-labelling S2ST data in M4. The T2U model used for pseudo-labelling is referred to as the teacher T2U model with 12 transformer layers encoder-decoder. For initialization, we used a smaller student T2U model with only six layers to optimize inference and distill the labels of the stronger T2U. In the second version of SEAMLESSM4T, UNITY2 replaces the second-pass autoregressive unit decoder in UNITY with a NAR unit decoder. We adopted the decoder architecture of Fast-Speech2⁷³ and extended it to discrete unit generation. UNITY2 starts with hierarchically upsampling the T2U encoder output from subword length to character length and then to unit length. The unit duration predictor, the key to the hierarchical upsampling, is supervised during training by a multilingual aligner based on RAD-TTS⁷⁴. The architecture is shown in detail in Supplementary Information section IV.8.

We fine-tuned the S2ST task with a combination of X-eng and eng-X S2ST translation data totalling 121,000 h. 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.

Automatic and human evaluation

BLASER 2.0. BLASER 2.0 (ref. 24) is the new version of BLASER⁷⁵, which works with both speech and text modalities, and 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 Supplementary Information section III.3.1, supports 57 languages in speech and 202 in text (coverage of languages by SONAR at the moment of submission of this paper) and is extendable to future encoders for new languages or modalities that share the same embedding spaces. For the purposes of evaluating speech outputs (and unlike ASR-based metrics), BLASER 2.0 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. 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.

Human evaluation. Apart from automatic metrics such as (ASR) BLEU and BLASER 2.0, we used human metrics such as XSTS²⁶, which measures semantic similarity between a source and target translation, and a standard Mean Opinion Score (as standardized in Recommendation ITU-T P.800, henceforth MOS), which measures (1) naturalness, (2) sound quality and (3) clarity of audio generations to evaluate our models. To obtain more robust language-level scores, we also incorporate a calibration set and calibration methodology, the same used to evaluate the NLLB models⁶. Apart from XSTS, we also obtained MOS evaluations to understand other aspects of audio quality in the target speech. For additional information about human evaluation protocols and analysis, see Supplementary Information section V.1.

Robustness. We built a replicable noise-robustness evaluation benchmark based on FLEURS (noisy FLEURS), which covers 102 languages, two speech tasks (S2TT and ASR), and various noise types (natural noises and music). To create simulated noisy audios, we sampled audio clips from MUSAN⁷⁶ 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 compared 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 (that is, when SNR decreases). For low-resource languages, the clean speech setup is already challenging, let alone a noisy one. Thus, we focused on four high-resource languages (French, Spanish, Modern Standard Arabic and Russian) belonging to three different language families for our noise-robustness analysis.

We followed ref. 47 to evaluate model robustness against speaker variations by calculating the average by-group mean score and by-group coefficient of variation using 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 CoefVar_{MS}, defined as follows:

$$\operatorname{chr}_{\mathsf{MS}} = \frac{1}{|G|} \sum_{g \in G} \operatorname{Mean}(g)$$
$$\operatorname{CoefVar}_{\mathsf{MS}} = \frac{1}{|G'|} \sum_{g' \in G'} \frac{\operatorname{Standard} \operatorname{deviation}(g)}{\operatorname{Mean}(g)}$$

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

Responsible AI

Toxicity detection. 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 S2ST ability of SEAMLESSM4T. Essentially, this metric follows a cascaded framework by first deploying a standard ASR module (that is, the same that is used for ASR-BLEU as defined in Supplementary Table 2), then the toxicity detection module, ETOX²⁷, which uses the Toxicity-200 word lists⁶. For S2TT, the translated output

can be directly evaluated with ETOX. In both cases (S2ST and S2TT), we measured added toxicity at the utterance or sentence level. We first computed toxicity detection for each input in the evaluation dataset and the corresponding output. Then, we compared them and counted a case as containing added toxicity only when the output value exceeds that of the input. Moreover, we used the recently proposed MuTox metric that can be applied to text or speech output with no need for ASR. This classifier has been trained on both speech and text toxicity labelled data for 30 languages. As MuTox relies on SONAR embeddings²⁹, MuTox the same number of languages by the zero-shot property. However, accounting for validated quality, we report MuTox only the languages that have been benchmarked²⁹. Again in both cases (S2ST and S2TT), we measured added toxicity at the utterance or sentence level. In this case, a sentence contains added toxicity if MuTox scores is >0.9 in the output and <0.5 in the input. We have experimentally validated these thresholds for several languages with human bilingual speakers for several pairs of languages. For S2TT, we computed MuTox in transcribed speech and target text. For S2ST, we computed MuTox in source and target speech.

For toxicity mitigation, we implement two techniques for the mitigation of added toxicity. Before training, we filter out training pairs with imbalanced toxicity. Furthermore, we use Mintox¹¹ at inference time. In particular, the main workflow generates a translation hypothesis with an unconstrained search. Then, the toxicity classifier is run on this hypothesis. If no toxicity is detected, we provide the translation hypothesis as it is. However, if toxicity is detected in the output, we run the classifier on the input. If the toxicity is unbalanced (that is, no toxicity is detected in the input), we re-run the translation with mitigation, which is the BEAMFILTERING step. This BEAMFILTERING consists of taking as input the multi-token expressions that should not appear in the output and excluding them from the beam-search hypotheses. Note that we do not apply mitigation in cases in which there is toxicity in the input (in other words, we do not deal with cases in which there is toxicity in the input but more toxicity in the output).

We used two datasets to analyse added toxicity. First, we deployed FLEURS to better align with our human evaluation effort and other evaluative components of this work. Furthermore, we used the English-only HOLISTICBIAS framework⁷⁷, which has been shown to trigger true added toxicity in previous studies²⁷. In this work, we extend HOLISTICBIAS to speech by applying the default English TTS model from MMS²³.

Speech extension of MULTILINGUAL HOLISTICBIAS. To compare the performances across modalities (S2ST and S2TT), we extended MULTILINGUAL HOLISTICBIAS to speech²³ (https://github.com/facebookresearch/fairseq/tree/main/examples/mms#tts-1). We used 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 is 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 Supplementary Table 26.

In terms of evaluation metrics for S2TT, we used chrF. For S2ST, we used ASRchrf (the transcription is done by WHISPER-LARGE and WHISPER-MEDIUM²⁰ for eng–X and X–eng, respectively, and chrF has been calculated the same way as S2TT except that in S2ST, the text from both prediction and reference were normalized) and BLASER 2.0. It is worth noting that when evaluating on BLASER 2.0, we only included 14 languages (arb, cat, deu, eng, fra, nld, por, ron, rus, spa, swe, tha, ukr and urd) for the eng–X direction (overlapping languages from the generated TTS data and the languages available in our S2ST model).

Data availability

To make our work available to the community, we provide open source of the following data sets, models and code at GitHub (https://github. com/facebookresearch/seamless_communication): (1) 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 modelling toolkit FAIRSEQ2 (https://github.com/facebookresearch/ fairseq2); (2) tools for creating aligned speech data, including metadata to recreate the unfiltered 470,000 h of SEAMLESSALIGN, STOPES-based pipelines (https://github.com/facebookresearch/stopes) to create alignments similar to SEAMLESSALIGN and SONAR for speech encoders in 37 languages and text encoders in 200 languages (https://github. com/facebookresearch/SONAR) and (3) a text-free S2ST automatic evaluation model, BLASER 2.0, inclusive of model weights and inference scripts.

- Koehn, P. Europarl: A parallel corpus for statistical machine translation. In Proceedings of Machine Translation Summit X: Papers, 79–86 (2005).
- Ziemski, M., Junczys-Dowmunt, M. & Pouliquen, B. The United Nations Parallel Corpus v1.0. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (eds Calzolari, N. et al.) 3530–3534 (European Language Resources Association, 2016).
- Wang, C., Pino, J., Wu, A. & Gu, J. CoVoST: a diverse multilingual speech-to-text translation corpus. In Proceedings of the Twelfth Language Resources and Evaluation Conference (eds Calzolari, N. et al.) 4197–4203 (European Language Resources Association, 2020).
- Wang, C., Wu, A., Gu, J. & Pino, J. CoVoST 2 and Massively Multilingual Speech Translation. In Proc. Interspeech 2021, 2247–2251 (ISCA, 2021).
- Salesky, E. et al. The multilingual TEDx corpus for speech recognition and translation. In Proc. Interspeech 2021, 3655–3659 (ISCA, 2021).
- Zhang, Y. et al. Google USM: scaling automatic speech recognition beyond 100 languages. Preprint at https://arxiv.org/abs/2303.01037 (2023).
- Schwenk, H. Filtering and mining parallel data in a joint multilingual space. In Proc. 56th Annual Meeting of the Association for Computational Linguistics (eds Gurevych, I. & Miyao, Y.) Vol. 2, 228–234 (Association for Computational Linguistics, 2018).
- Artetxe, M. & Schwenk, H. Margin-based parallel corpus mining with multilingual sentence embeddings. In Proc. 57th Annual Meeting of the Association for Computational Linguistics (eds Korhonen, A. et al.) 3197–3203 (Association for Computational Linguistics, 2019).
- Artetxe, M. & Schwenk, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguist.* 7, 597–610 (2019).
- Duquenne, P.-A., Gong, H. & Schwenk, H. Multimodal and multilingual embeddings for large-scale speech mining. In Advances in Neural Information Processing Systems (eds Ranzato, M. et al.) Vol. 34, 15748–15761 (Curran Associates, 2021).
- Duquenne, P.-A. et al. SpeechMatrix: a large-scale mined corpus of multilingual speechto-speech translations. In Proc. 61st Annual Meeting of the Association for Computational Linguistics (eds Rogers, A. et al.) Vol. 1, 16251–16269 (2023).
- Silero. Silero VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier. *GitHub* https://github.com/snakers4/silero-vad (2021).
- Khurana, S., Laurent, A. & Glass, J. SAMU-XLSR: Semantically-Aligned Multimodal Utterance-Level Cross-Lingual Speech Representation. *IEEE J. Sel. Top. Signal Process.* 16, 1493–1504 (2022).
- 58. Douze, M. et al. The Faiss library. Preprint at https://arxiv.org/abs/2401.08281 (2024).
- Inaguma, H. et al. UnitY: two-pass direct speech-to-speech translation with discrete units. In Proc. 61st Annual Meeting of the Association for Computational Linguistics (eds Rogers, A. et al.) Vol. 1, 15655–15680 (Association for Computational Linguistics, 2023).
- Tjandra, A., Sakti, S. & Nakamura, S. Speech-to-speech translation between untranscribed unknown languages. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 593–600 (IEEE, 2019).
- Babu, A. et al. XLS-R: self-supervised Cross-lingual Speech Representation Learning at Scale. In Proc. Interspeech 2022, 2278–2282 (ISCA, 2022).
- Gong, H. et al. Multilingual speech-to-speech translation into multiple target languages. Preprint at https://arxiv.org/abs/2307.08655 (2023).
- Chung, Y.-A. et al. 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), 244–250 (IEEE, 2021).
- Chiu, C.-C., Qin, J., Zhang, Y., Yu, J. & Wu, Y. Self-supervised learning with random-projection quantizer for speech recognition. In Proc. 39th International Conference on Machine Learning, 3915–3924 (PMLR, 2022).
- 65. Gulati, A. et al. Conformer: convolution-augmented transformer for speech recognition. In Proc. Interspeech 2020, 5036–5040 (ISCA, 2020).

- Kudo, T. & Richardson, J. SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. In Proc. 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (eds Blanco, E. & Lu, W.) 66–71 (Association for Computational Linguistics, 2018).
- Sennrich, R., Haddow, B. & Birch, A. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (eds Erk, K. & Smith, N. A.) Vol. 1, 1715–1725 (Association for Computational Linguistics, 2016).
- Andrews, P. et al. stopes modular machine translation pipelines. In Proc. 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (eds Che, W. & Shutova, E.) 258–265 (Association for Computational Linguistics, 2022).
- Jia, Y. et al. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7180–7184 (IEEE, 2019).
- Pino, J., Xu, Q., Ma, X., Dousti, M. J. & Tang, Y. Self-training for end-to-end speech translation. In Proc. Interspeech 2020, 1476–1480 (2020).
- Vaswani, A. et al. Attention is all you need. In Advances in Neural Information Processing Systems (eds Guyon, I. et al.) Vol. 30 (Curran Associates, 2017).
- Zhao, J., Yang, H., Haffari, G. & Shareghi, E. M-Adapter: modality adaptation for end-to-end speech-to-text translation. In Proc. Interspeech 2022, 111–115 (ISCA, 2022).
- 73. Ren, Y. et al. FastSpeech 2: fast and high-quality end-to-end text to speech. In Proc. 2021 International Conference on Learning Representations (2021).
- Shih, K. J. et al. RAD-TTS: parallel flow-based tts with robust alignment learning and diverse synthesis. In ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models (2021).
- Chen, M. et al. BLASER: A text-free speech-to-speech translation evaluation metric. In Proc. 61st Annual Meeting of the Association for Computational Linguistics (eds Rogers, A. et al.) Vol. 1, 9064–9079 (Association for Computational Linguistics, 2023).
- Snyder, D., Chen, G. & Povey, D. MUSAN: a music, speech, and noise corpus. Preprint at https://arxiv.org/abs/1510.08484 (2015).
- Smith, E. M., Hall, M., Kambadur, M., Presani, E. & Williams, A. "I'm sorry to hear that": finding new biases in language models with a holistic descriptor dataset. In Proc. 2022 Conference on Empirical Methods in Natural Language Processing (eds Goldberg, Y. et al.) 9180–9211 (Association for Computational Linguistics, 2022).

Acknowledgements We want to extend our gratitude to those who made this work possible. We thank S. Edunov and A. Fan for helping shape the earlier stages of the project: S. Bhosale. V. Goswami, F. Hernandez and Y. Tang for their help in building stronger models: M. Chen for his contributions to BLASER 1.0; K. Klyushkin for his help in building better experiences A. Kozhevnikov for his contributions to FAIRSEQ2 and SONAR inference; Z. Ni and X. Zhang for benchmarking audio denoising models; N. Seejoor and M. Duppenthaler for their help in setting up the demo; V. Goswami, S. Hsia, B. Acun-Uyan, and C.-J. Wu for helping with efficiency optimizations; B. Alastruey, M. Anwar, H.-J. Chang, H. Han, C.-W. Huang, H. Lu, S. Ouyang, Y. Peng, P. Rust, J. Shi, N. Verma, S.-L. Yeh and all of our interns and residents for the generative discussions they brought to the team; M. Clark, L. Cohen, J. Pak and H. Rudolph for their advice and guidance; E. Astbury, L. Baillergeau, D. Beaty, J. Bennett, J. Carvill, A. Davidson, A. Farooq, A. Gabriel, G. Jhala, C. Johnson, S. Miles, A. P. K. Mofarrej, R. Nayani, A. Newcomb, T. Piksa, M. Restrepo, N. Rizk and A. Tharinger for helping our research reach new audiences; G. Chauhan, A. Gunapal, C. Ho, D. Kannappan, A. Kokolis, T. Li, M. Reso, S. Sengupta, H. Shojanazeri and X. Zhang for assisting us with compute resources and infrastructure; E. Dupoux and E. M. Smith for their feedback on the paper; C. Moghbel, M. Paluri, J. Pineau, L. van der Maaten and M. Williamson for their continued support of the project

Author contributions Core contributors to SEAMLESSM4T and their areas of contribution are specified as follows: L.B., P.-A.D., H.E., K.H., G.W., P.A., O.Ç., C.G., J.K. and H.S. contributed to the data workstream of the project, which includes data acquisition and developing tools to facilitate data mining, cleaning and consolidation. P.A., M.R.C., D.D., J. Hoffman and D.L. contributed to the evaluation workstream, which assessed automatic and human quality. P.A., N.D., M.E., C.K., A. Rakotoarison, E.Y., C.B. and J.W. contributed to externalization, including fairseq2 development and demo experiences. P.L., K.R.S., M.E., A.L., S.P., PT., C. Wang, Y.-A.C., N.D., H.G. and J.M. contributed to the modelling workstream. P.A., M.R.C., M.C.M., H.E., M.E., H.G., C.R. and S.W. contributed to Responsible AI, which includes toxicity, bias, linguistics and social impact evaluations. F.G., J.K., A.M., J.P., H.S., S.S. and P.T. provided leadership for the project. Contributors to SEAMLESSM4T are B.A., P.-J.C., N.E.H., B.E., G.M.G., J. Haaheim, P.H., R.H., B.H., M.-J.H., H.J., S.J., E.K., A.K., I.K., J.L., D.L., X.M., R.M., B.P., M.R., A.S., K.T., T.T., I.T., V.V., C.W., YY. and B.Y.; M.R.C., C.R., M.E. and S.W. constitute the editorial team and prepared the paper for publication.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-08359-z.

Correspondence and requests for materials should be addressed to Marta R. Costa-jussà. **Peer review information** *Nature* thanks Tanel Alumäe, Allison Koenecke and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.

Extended Data Table 1 | F1 micro and macro averages over all SEAMLESSM4T languages and the intersection of supported languages across models

	Ove	erall	Intersection		
	\uparrow F1-micro (n=100)	\uparrow F1-macro (n =100)	$\overline{\uparrow \mathrm{F1} ext{-micro}}_{(n=79)}$	\uparrow F1-macro $(n=79)$	
VL107 HF LID100	$82.3\%\ 86.0\%$	$69.1\%\ 81.9\%$	$94.1\%\ 92.9\%$	$92.6\%\ 91.1\%$	

Extended Data Table 2 Average T2TT and S2TT performance on FLORES devtest and FLE ${\sf FLE}$						
	T2TT I ↑spB		S2TT FLEURS †BLEU			
Model	$ ext{X-eng}\ (n{=}200)$	${ m eng-X} \ (n{=}200)$	X-eng $(n=37)$			
NLLB-1.3B Whisper-Large-v2	35.2	24.9	$_{22.5}^{-}$			
SONAR	32.7	21.6	23.3			

Extended Data Table 3 | The averaged points across modalities and genders for assessing the overgeneralization (eng-X) and the robustness (X-eng)

		$\mathbf{eng}-\mathbf{X}$			\mathbf{X} –eng
S2TT	chrF_f	chrF_m	$\Delta\downarrow\%$	chrF_f	chrF_m
Baseline SeamlessM4T-Large SeamlessM4T-v2	47.4 45.0 45.2	52.7 49.9 50.2	11.2 10.9 11.1	50.4 52.4 54.2	52.1 54.3 56.0
S2ST	$\mathrm{ASRchrf}_f$	$\mathrm{ASRchrf}_m$	$\Delta\downarrow\%$	$\mathrm{ASRchrf}_f$	$\mathrm{ASRchrf}_m$
Baseline SeamlessM4T-Large SeamlessM4T-v2	38.4 45.6	41.6 50.4	- 8.3 10.5	52.1 52.7 56.1	53.9 54.5 58.0
S2ST	Blaser 2.0_f	Blaser 2.0_m	$\Delta\downarrow\%$	Blaser 2.0_f	Blaser 2.0
Baseline SeamlessM4T-Large SeamlessM4T-v2	- 3.5 3.7	3.5 3.7	- 0.0 0.0	2.8 3.6 3.7	2.9 3.7 3.8

 Δ represents the relative difference between masculine and feminine ($\Delta = \omega(M - F)/\omega(min(M, F)), \omega \in \{chrF, ASRchrf, BLASER 2.0\}$). Baseline corresponds to WHISPER-LARGE-V2 for S2TT X–eng; WHISPER-LARGE-V2 + NLLB-3.3B for S2TT X–eng; WHISPER-LARGE-V2 + YOURTTS for S2ST X–eng.