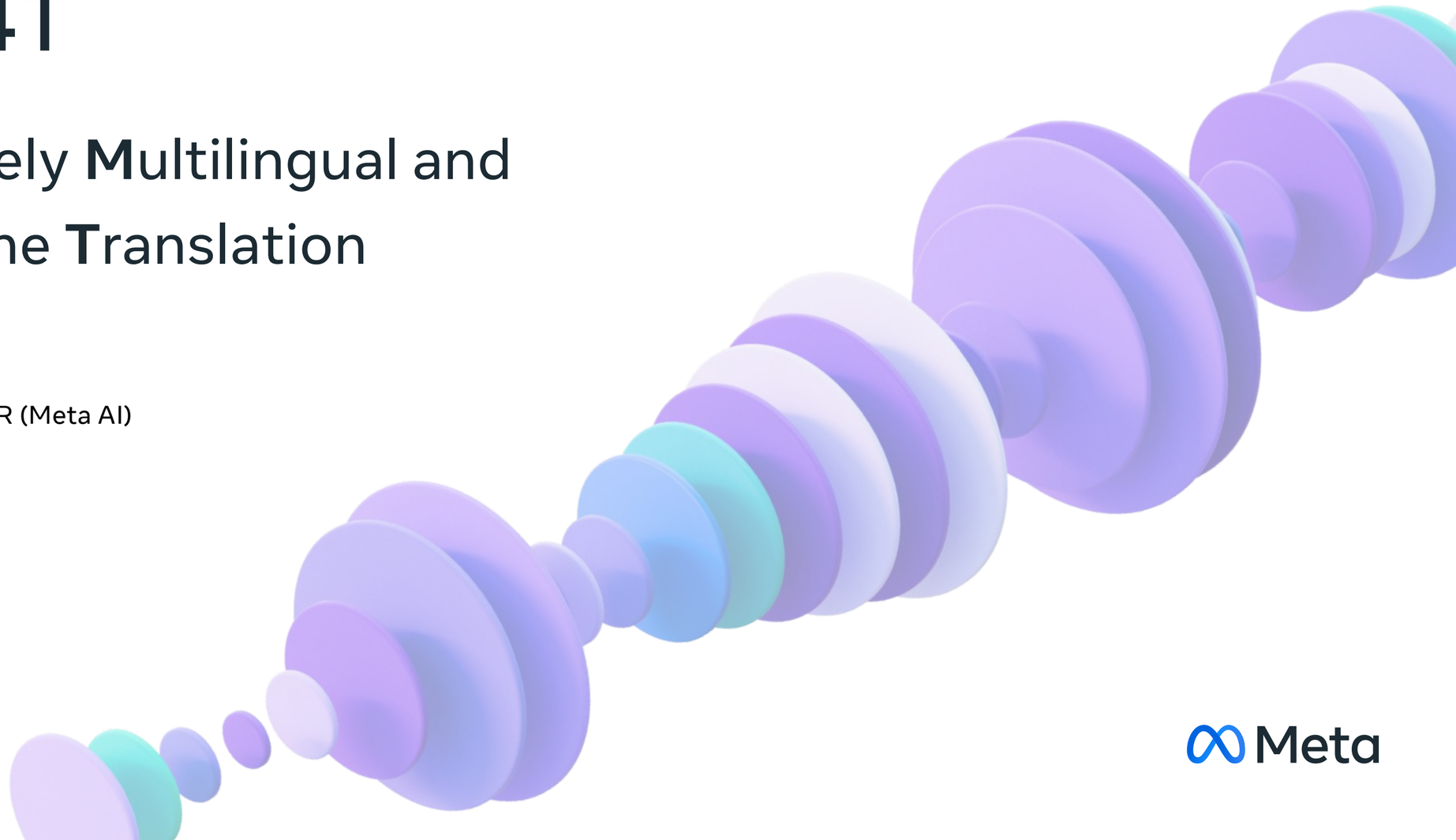
SeamlessM4T

All-in-One, **M**assively **M**ultilingual and **M**ultimodal **M**achine **T**ranslation

Maha Elbayad - Research scientist - FAIR (Meta AI)

09/22/2023



رياضيات المغرب
MATH MAROC

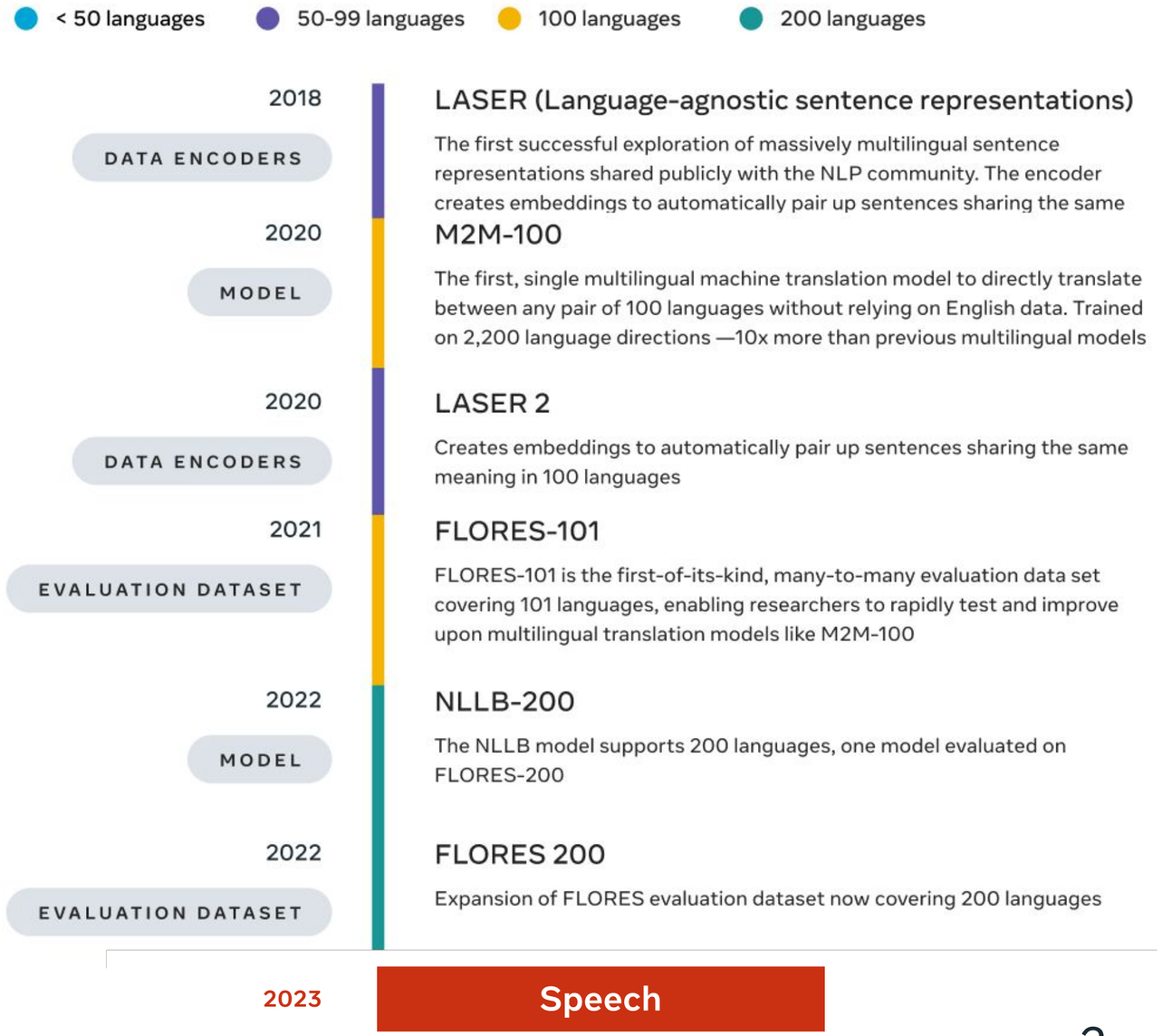
Meta

Our north star goal

Enable communication without language barriers.

Moving beyond text-based communication with **Seamless**:

1. communication through speech creates **stronger social bonds**
2. Speech is often the most **practical & accessible** communication channel
3. Text-based communication is further complicated by **script variance**



Speech Translation Today

Before SeamlessM4T:

Cascaded systems:
Speech-to-Text Translation = (1) ASR (automatic speech recognition)
→ (2) (text-to-text translation)

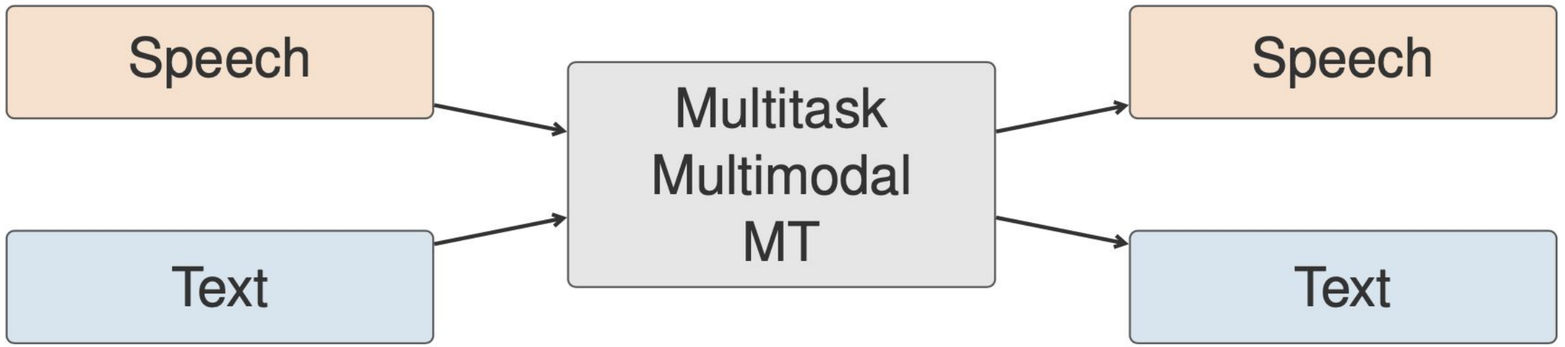
Speech-to-Speech Translation = (1) ASR (automatic speech recognition)
→ (2) (text-to-text translation)
→ (3) TTS (text-to-speech synthesis)

Limitations: Error propagation, high latency, domain mismatch between components

Direct systems: e.g.Whisper [Radford et al., 2022] , VALL-E X [Zhang et al. 2023], AudioPaLM [Rubenstein et al., 2023]

- Not massively multilingual (VALL-E)
- Do not outperform cascaded models
- Not evaluated on all supported tasks /languages (e.g. never evaluated on multilingual benchmarks out of English).

SeamlessM4T



A unified model supporting 5 multilingual tasks:

- ❏ Automatic Speech Recognition - ASR
- ❏ Speech-to-Speech Translation - S2ST
- ❏ Speech-to-Text Translation - S2TT
- ❏ Text-to-Speech Translation - T2ST
- ❏ Text-to-Text Translation - T2TT

With a large coverage of:

- ❏ 101+ languages for speech input
- ❏ 96 Languages for text input/output
- ❏ 35 languages for speech output.

Our efforts

(1) Data:
Automatically Creating Aligned Data for Speech

(2) Modeling:
Multilingual / Multimodal / Multitask MT
with multitask UnitY

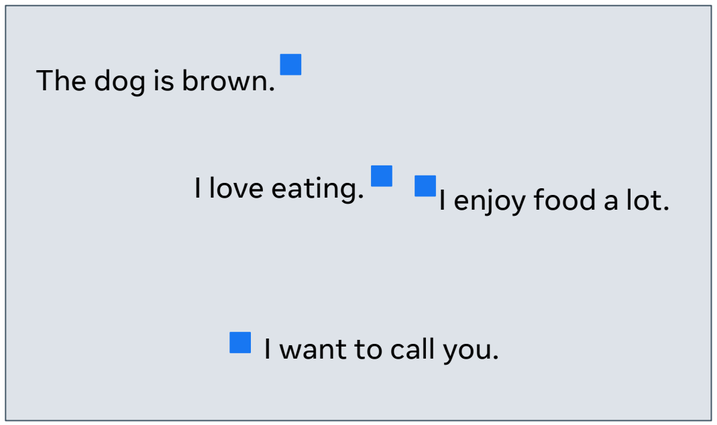
(3) Evaluation:
Automatic and human evaluation + Responsible AI

Data:
Automatically Creating Aligned Data for Speech
With T-modules

Methods

SeamlessM4T Data - SONAR

Mining for speech-text, text-text or speech-speech pairs

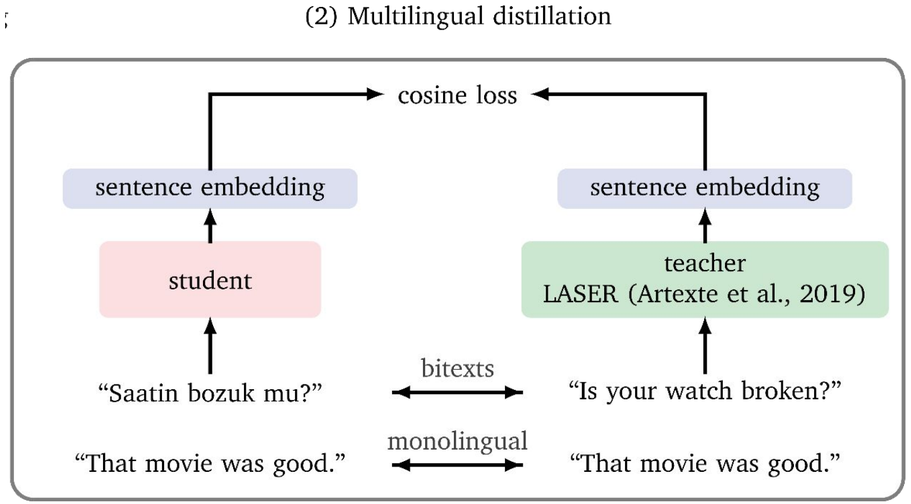


Sentences with similar meaning are ***close***.

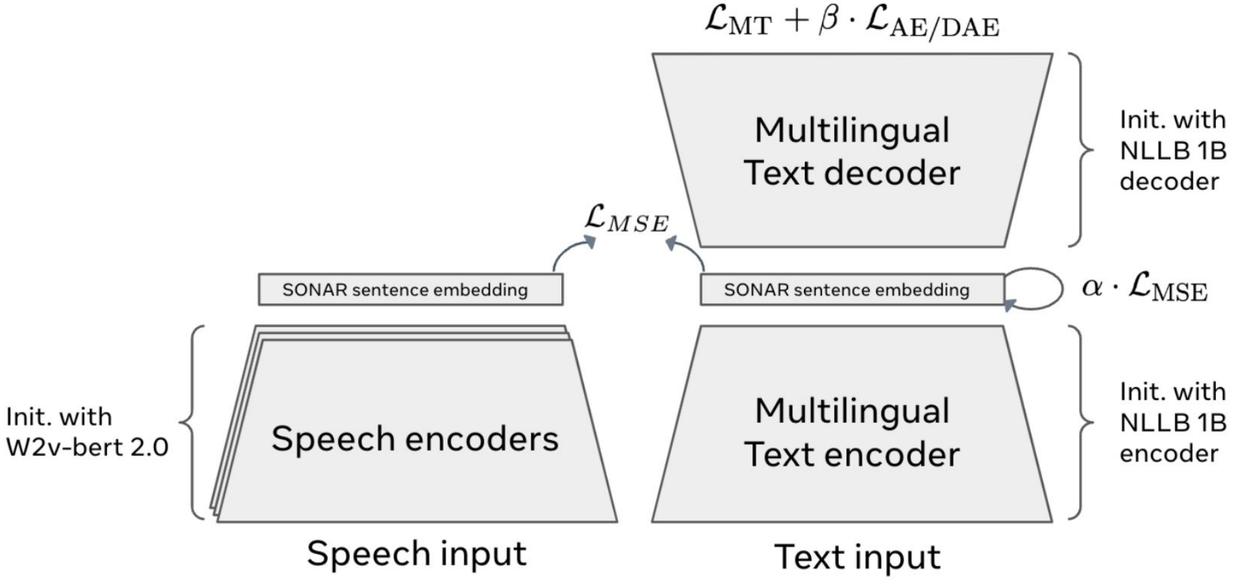
Sentences with similar meaning are ***close*** **independently of their language or their modality (text/speech)**

SeamlessM4T Data - SONAR

Teacher-student alignment



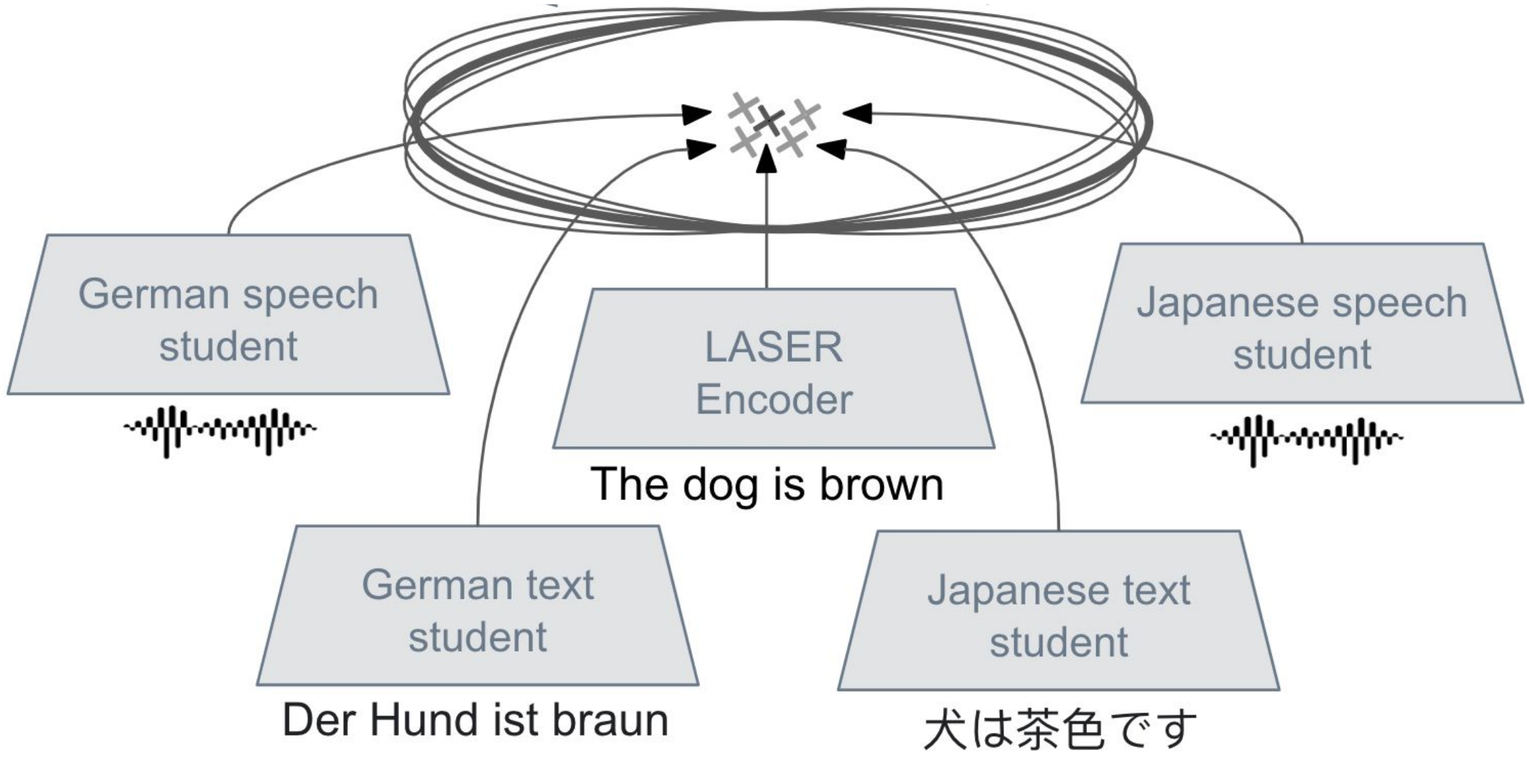
(1) Our previous work - LASER - on text with NLLB



(2) Extension to speech

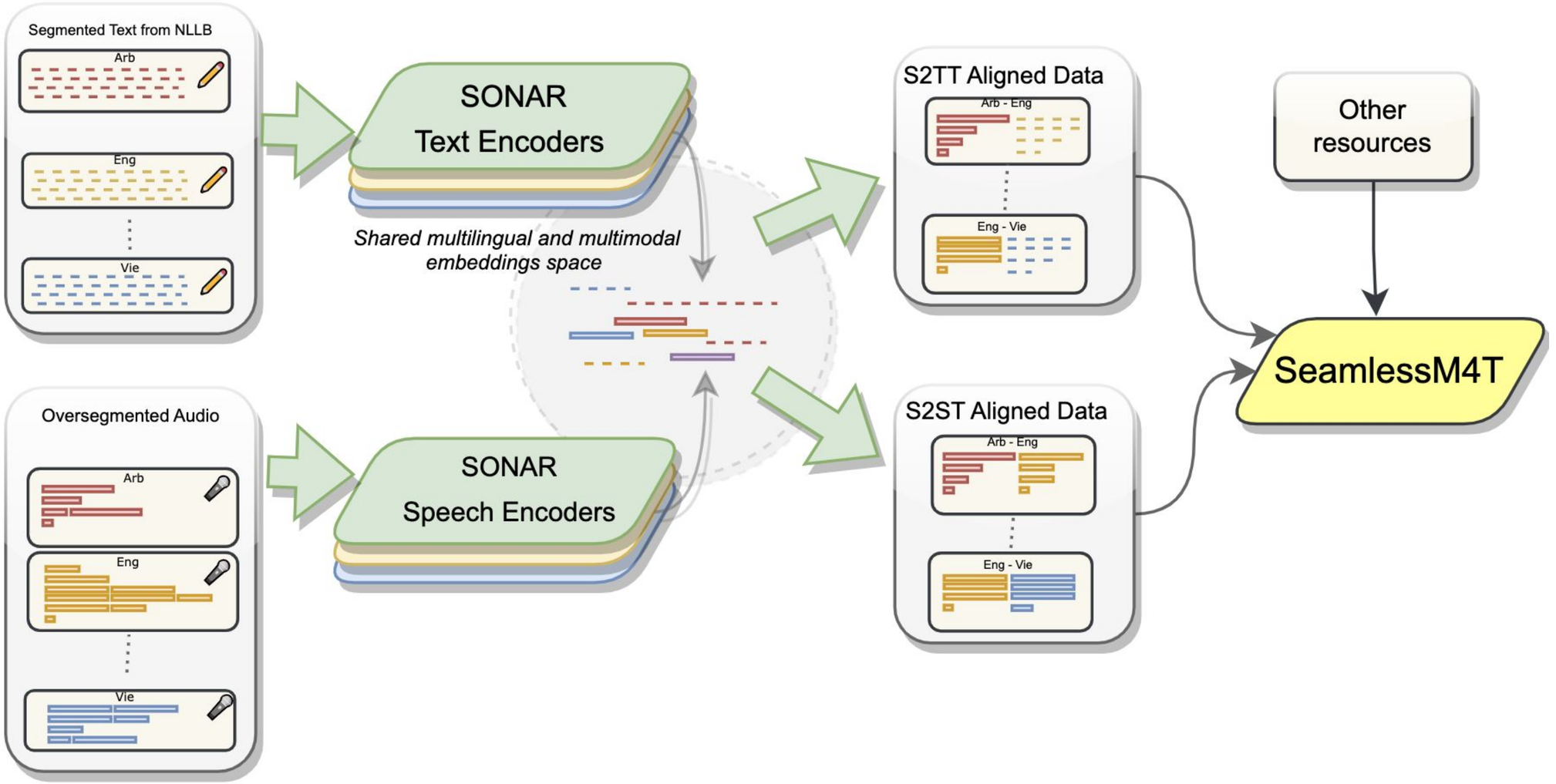
SeamlessM4T Data - SONAR

SONAR encoders



SeamlessM4T Data - Mining

Distance-based mining



Modeling:
Multilingual / Multimodal / Multitask MT
with multitask UnitY

Methods

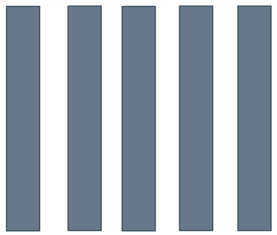
Discrete units for speech (101)

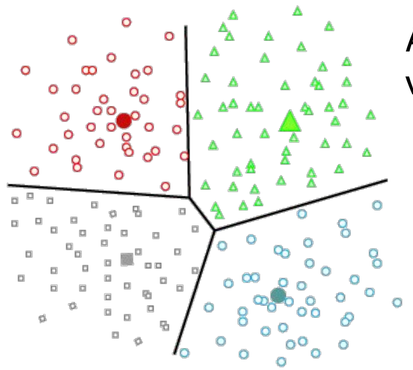
We need to map continuous representations from a speech encoder into discrete categories.

1) Randomly select and encode a set of unlabeled audio samples (a representation for every 20-ms) with a speech encoder (intermediate layer of HuBERT, XLSR, w2v-BERT, etc.)
2) Apply a k-means algorithm on these representations to estimate K cluster centroids.

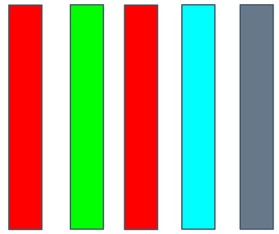
Converting to units:

Intermediate features
$X_1, x_2, \ldots, x_T$





Assign each feature vector x_t to a cluster



Represent the audio waveform with a sequence of cluster indices

Speech Resynthesis from Discrete Disentangled Self-Supervised Representations

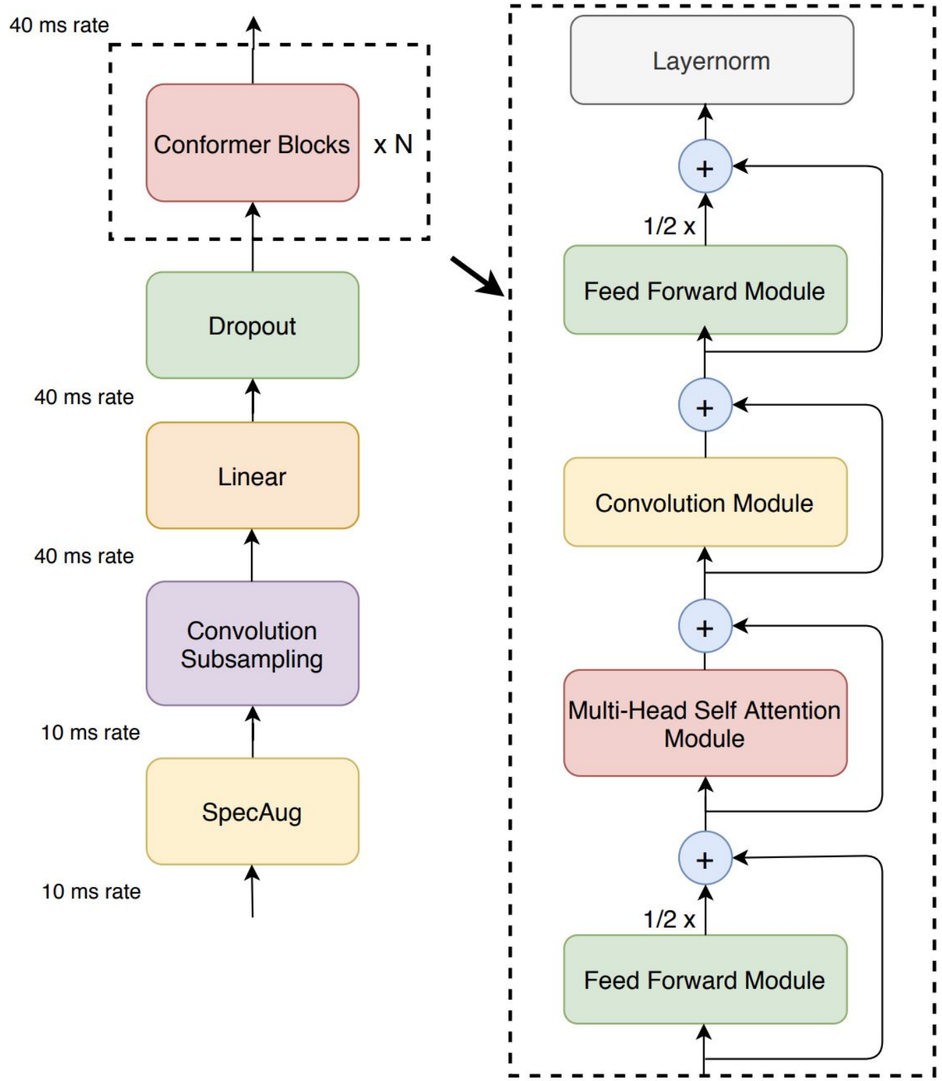
Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux

https://arxiv.org/abs/2104.00355

Conformer (101)

Similar to Transformer but with:

- Two macaron-like FFN layers
- An additional convolutional modules after self-attention



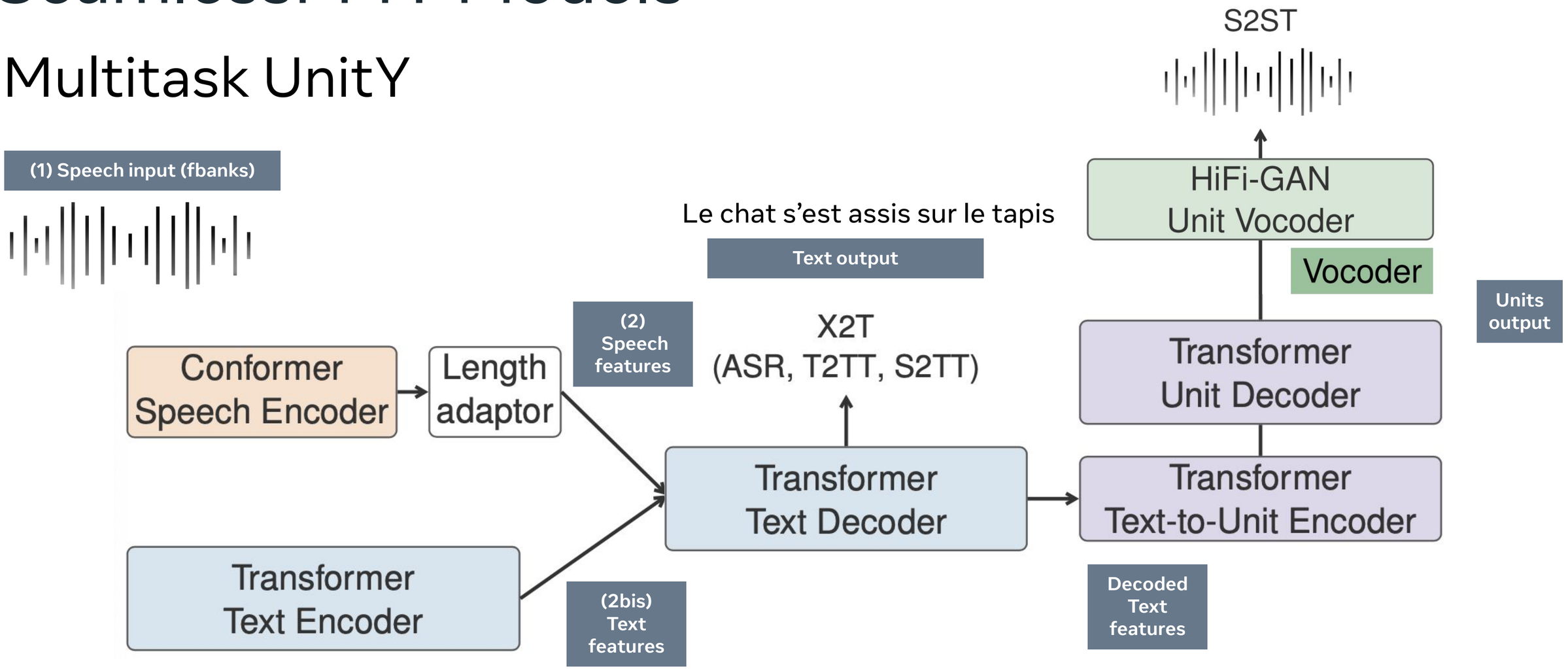
Conformer: Convolution-augmented Transformer for Speech Recognition

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang

https://arxiv.org/abs/2005.08100

SeamlessM4T Models

Multitask UnitY



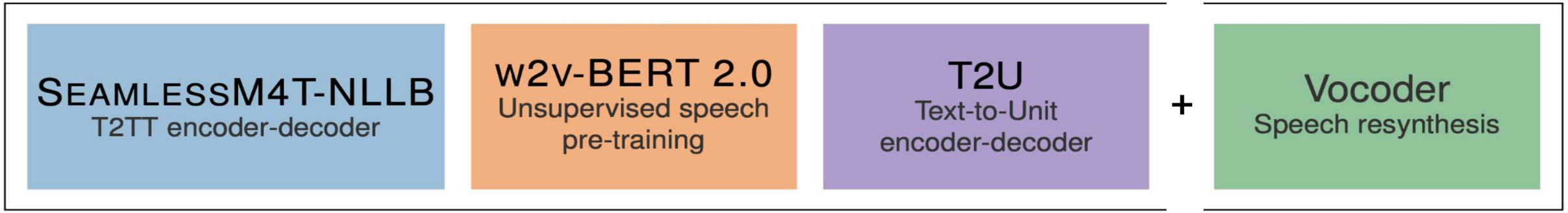
S2ST = X2T (direct) + T2U + Vocooder

vocoder: a separate module not jointly finetuned with multitask-unitY

SeamlessM4T Models

How is Multitask UnitY trained?

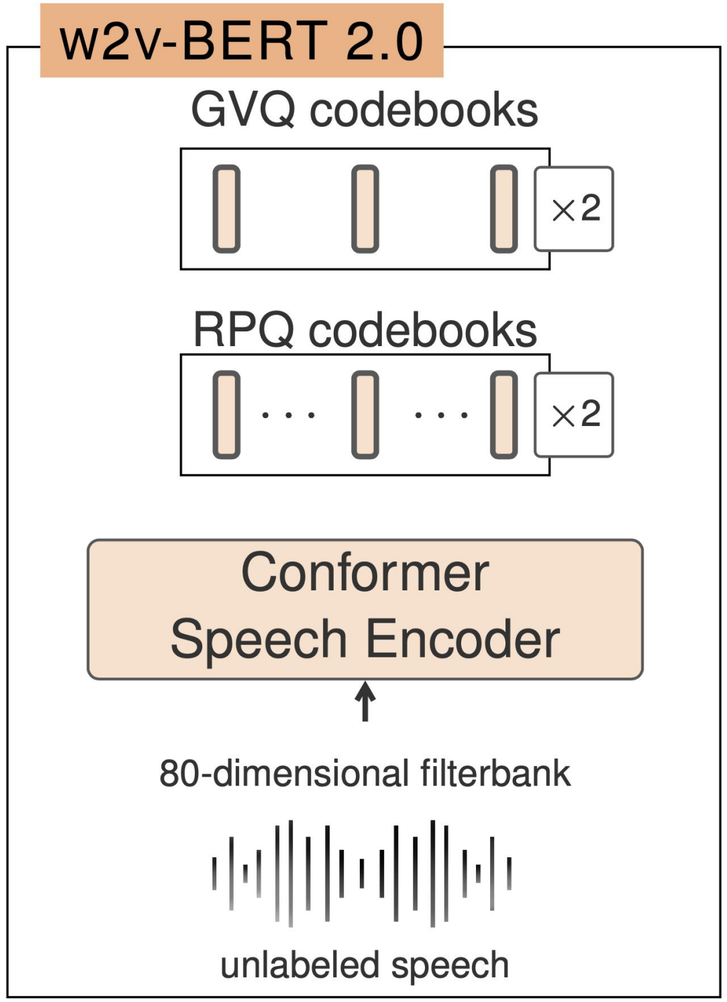
All components of multitask UnitY are independently pre-trained.



SeamlessM4T Models

w2v-BERT 2.0

w2v-BERT 2.0
Unsupervised speech pre-training



Model	Languages	Hours	Model type	Open model
XLS-R-2B-S2T	128	0.4M	wav2vec 2.0	✓
USM	over 300†	12M	BEST-RQ	
MMS	1406	0.5M	wav2vec 2.0	✓
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.

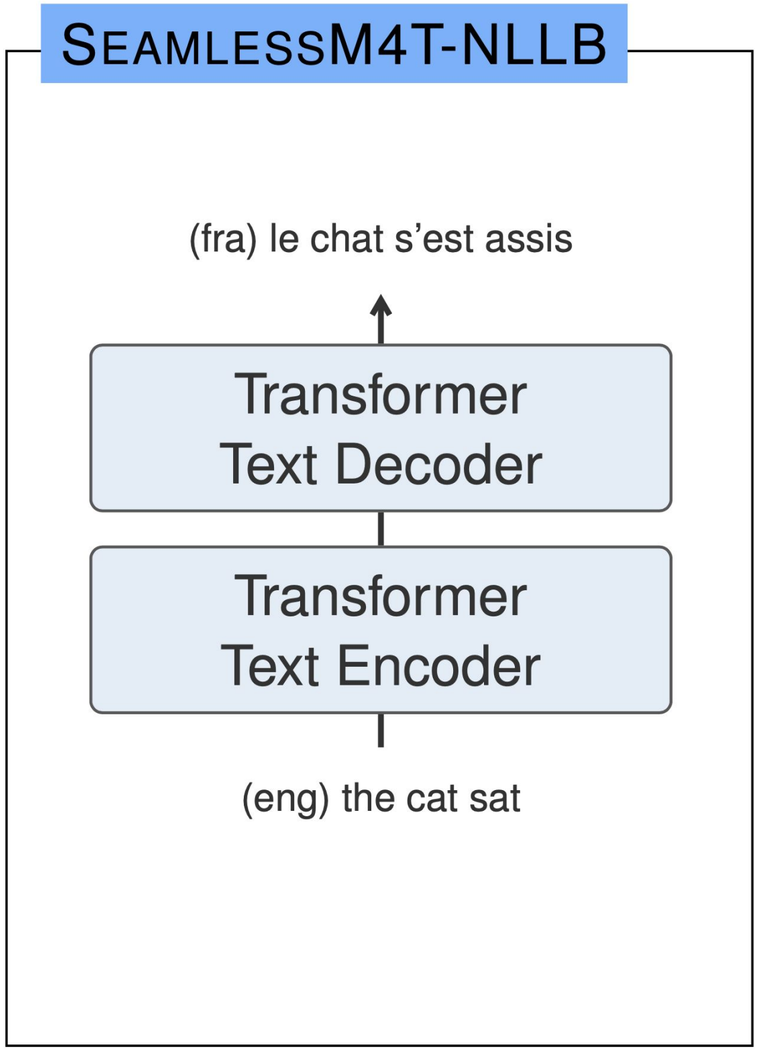
Trained on **1M Hours** covering **143+ languages**

w2v-BERT 2.0
= w2v-BERT (contrastive learning + masked prediction learning)
+ additional codebooks in both objectives.

SeamlessM4T Models

SeamlessM4T-NLLB (NLLB v2)

SEAMLESSM4T-NLLB
T2TT encoder-decoder

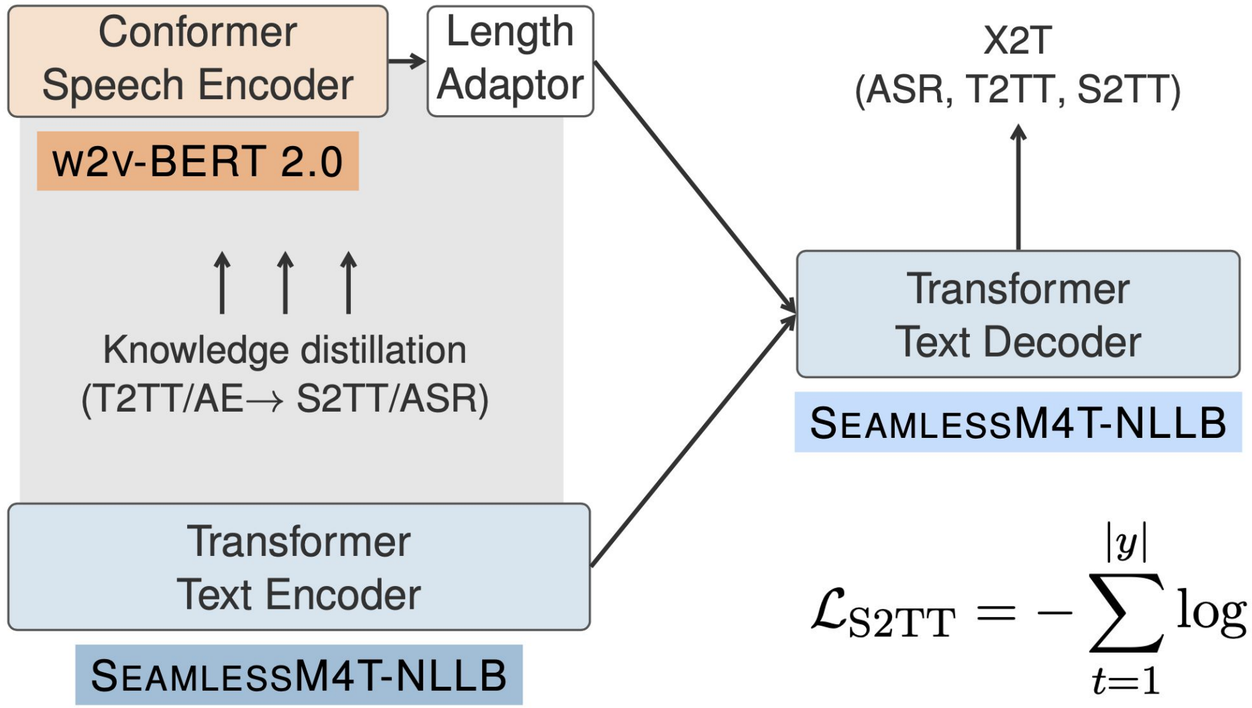


Re-trained a sentencepiece tokenizer (256K tokens) while forcing the addition of Simplified Chinese characters, Traditional Chinese characters, and Japanese kanji character

Trained a dense 1.3B model on NLLB MT bitexts (primary, mined and BT) covering **96 languages**
(Chinese in two scripts and two written forms for Norwegian)

SeamlessM4T Models

X2T: First finetuning stage



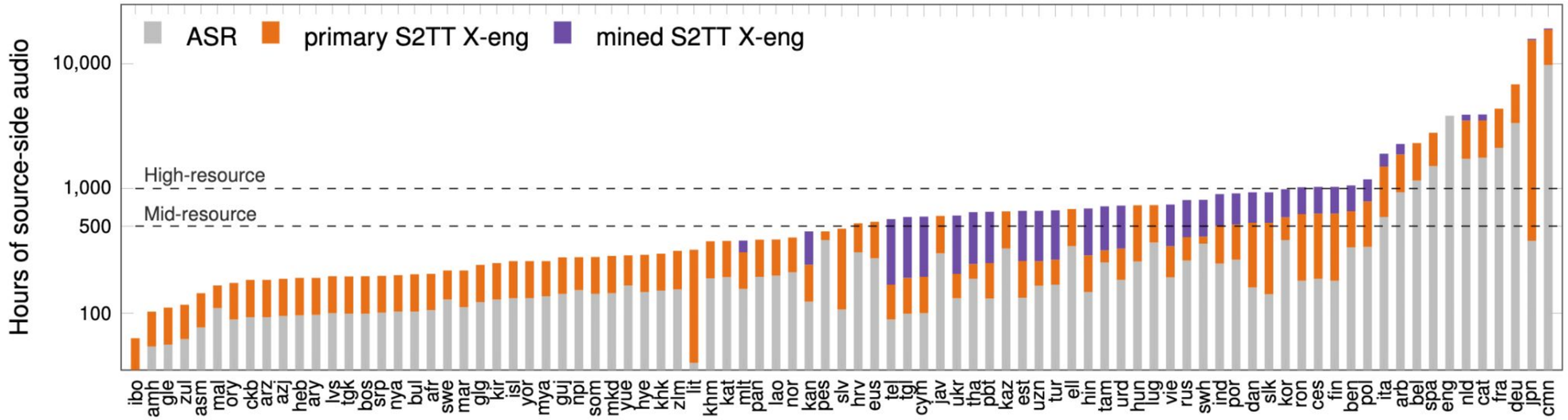
Training on triplets of source audio, source text and target text:

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

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

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

SeamlessM4T X2T Data



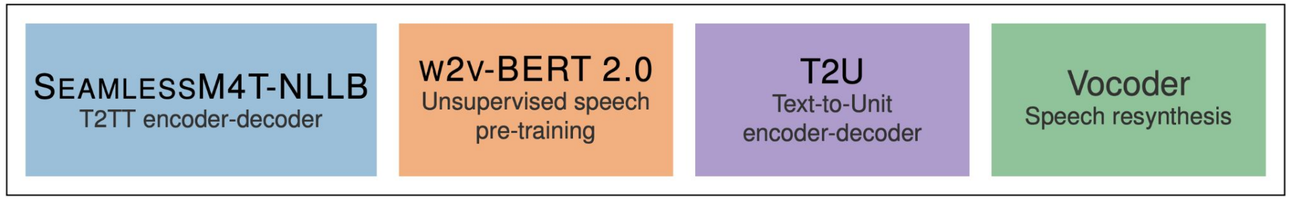
Primary S2TT data mostly comes from **pseudo-labeling** ASR data.

Approximately 2000 hours in each eng-X direction.

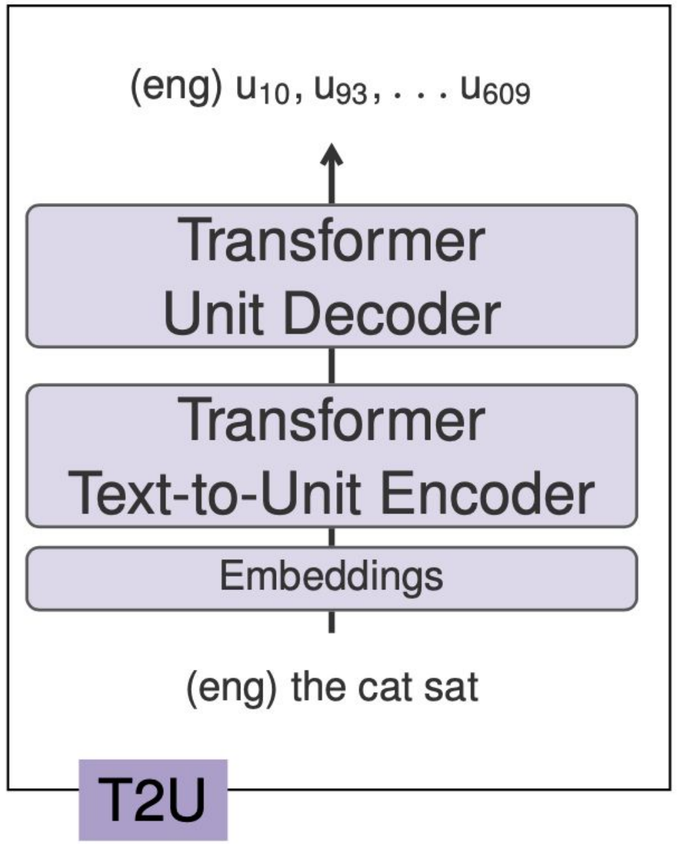
Imbalance data requires temperature samplins to improve synergy between languages (T=2).

A total of **286K hours** of audio.

SeamlessM4T Models



Pre-trained T2U



We trained T2U models for two purposes:

(1) **Pseudo-labeling** of S2TT data instead of using TTS.

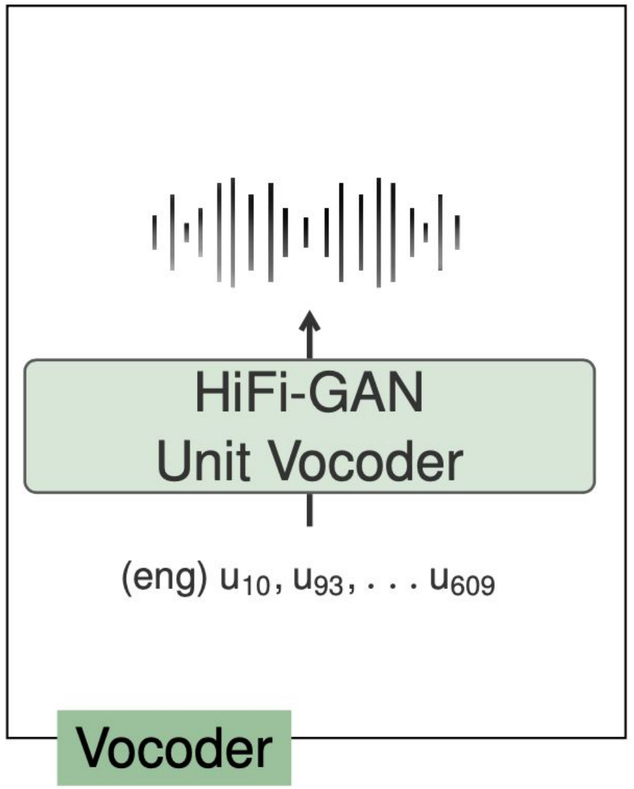
Architecture: 12 encoder-decoder

(2) **initializing the model's T2U**.

Architecture: A smaller 6 encoder-decoder

SeamlessM4T Models

Multilingual HiFi-GAN vocoder



Units vocabulary.

1) Features from the 35th layer of XLSR-1B

2) Encoded 10K audio samples from each of the 35 supported target languages

3) Applied a k-means on these representations to estimate K=10,000 cluster centroids

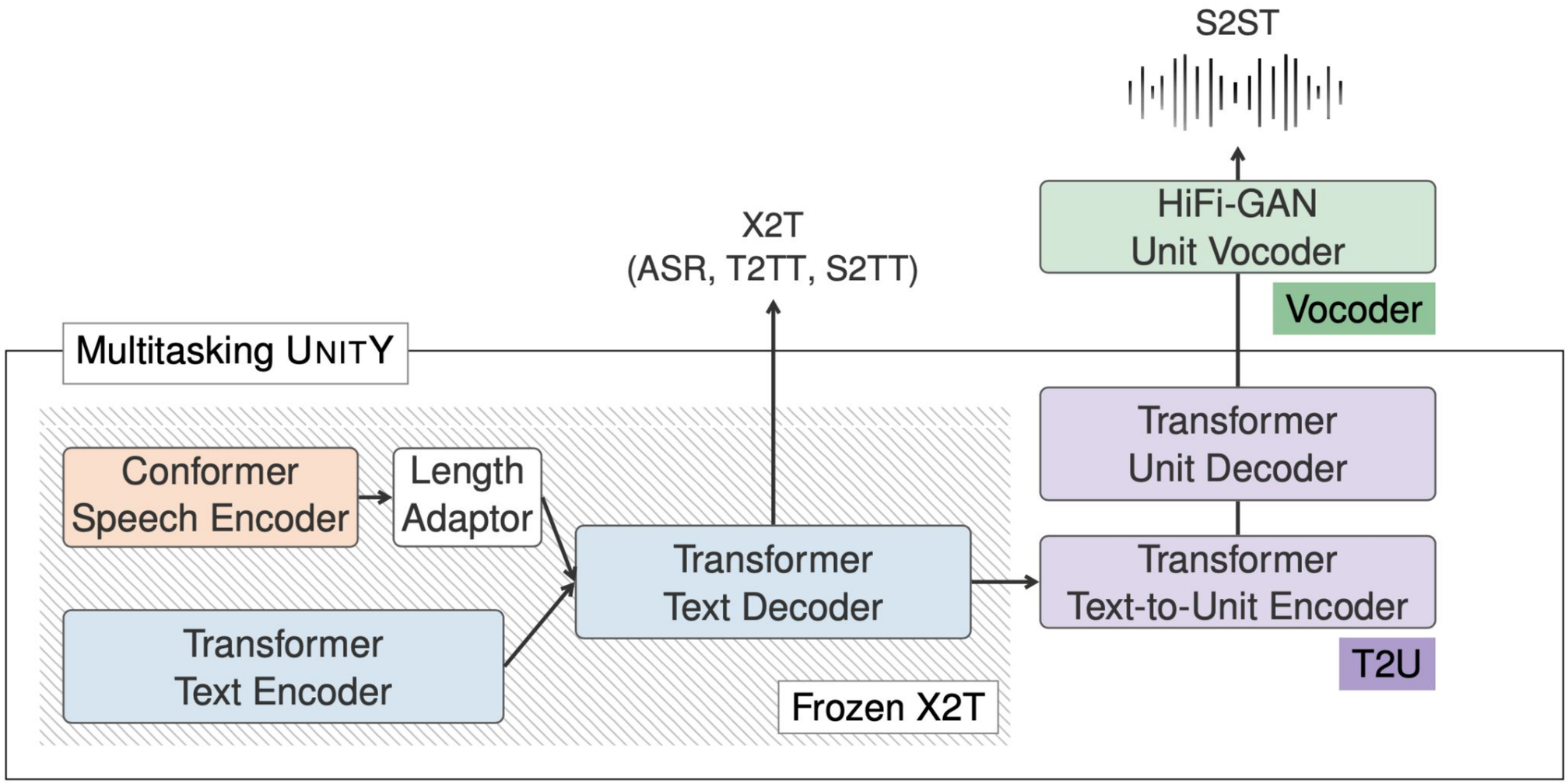
Vocoder

To mitigate **cross-lingual interference,** language identification is used as an auxiliary loss in multilingual training.

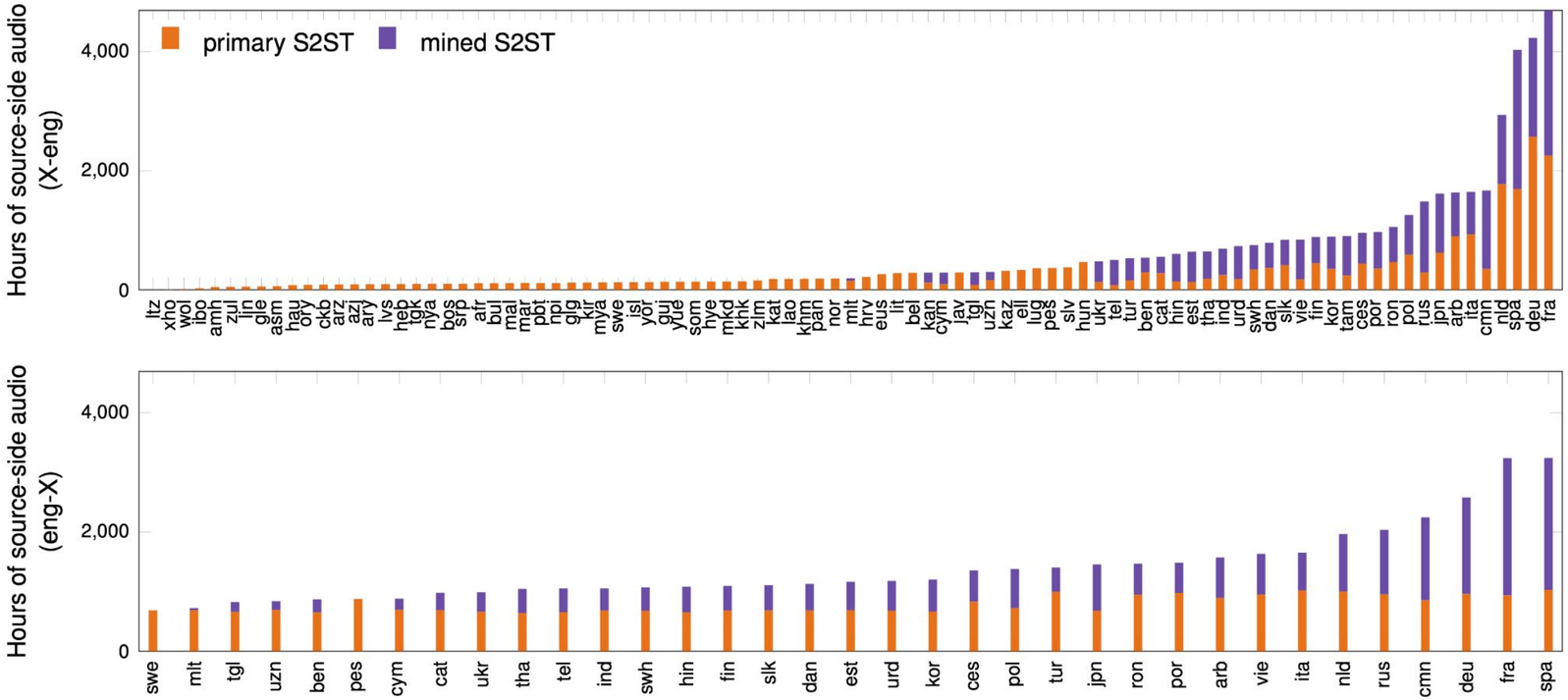
For more see this work.

SeamlessM4T Models

S2ST Last finetuning stage



SeamlessM4T S2ST Data



Primary S2ST data comes from pseudo-labeling S2TT primary data.

A total of **120K hours** of audio

Results

SeamlessM4T - Results

Comparison against direct & cascaded models (S2TT)

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.

X-eng

+34% from whisper

+22% from AudioPaLM

+6% from best cascade

eng-X

+2% from ~cascade

-3% from best cascade

SeamlessM4T - Results

Comparison against cascaded models (S2ST)

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		4.1B	20.6	
+WHISPER-LARGE-V2 (ASR)+ NLLB-1.3B		2.9B	22.1	
+WHISPER-LARGE-V2 (ASR)+ NLLB-3.3B		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.

Fleurs X-eng

+11% from best cascade

SeamlessM4T - Results

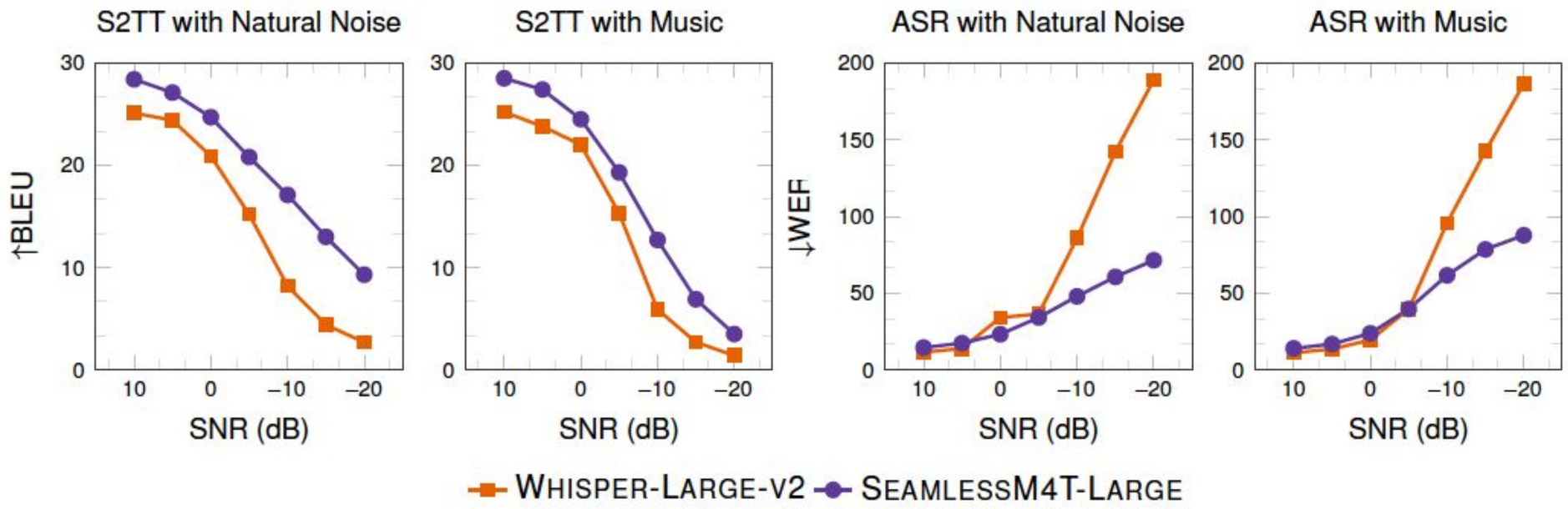
Multitasking results

Model	size	ASR (↓WER) FLEURS (n=77)	ASR (↓WER) FLEURS-54 (n=54)	T2TT (↑chrF++) FLORES X–eng (n=95)	T2TT (↑chrF++) 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**

Model	FLEURS T2ST (↑ASR-BLEU) X–eng (n=88)	FLEURS T2ST (↑ASR-BLEU) eng–X (n=35)	FLEURS T2ST (↑ASR-BLEU) eng–X (n=32) zero-shot
SEAMLESSM4T-LARGE	34.9	20.7	22.5

SeamlessM4T - Results beyond BLEU scores

Robustness to noise (↑38%)



Methodology:

- Sampled audio clips from MUSAN [Snyder et al., 2015] on the “noise” and “music” categories, and mixed them with Fleurs speech audios under different SNR.
- 4 high-resource languages (French, Spanish, Modern Standard Arabic, and Russian), X-Eng

SeamlessM4T - Results beyond BLEU scores

Robustness to speaker variations (↑49%)

Languages	Average #	Whisper-Large-v2		SeamlessM4T-Large	
($\geq$ 40 content groups)	cont. groups	$\text{chrF}_{MS}\uparrow$	$\text{CoefVar}_{MS}\downarrow$	$\text{chrF}_{MS}\uparrow$	$\text{CoefVar}_{MS}\downarrow$
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**

Methodology:

- Calculating average by-group mean score and by-group coefficient of variation of an utterance-level quality metric (chrF).
- Fleurs languages that have at least 40 content groups in the test sets.

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

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

SeamlessM4T - Results beyond BLEU scores

Blaser 2.0: Modality-Agnostic Model-based

New Capabilities

- Modality-agnostic
- Quality estimation
- 83 speech and 200 text languages

Results - Correlation with XSTS

- Supervised Blaser 2.0 comparable to supervised Blaser (0.58)
- Unsupervised Blaser 2.0 improves ~18% unsupervised Blaser (0.58 vs 0.49)
- Blaser-QE 2.0 achieves 93% of this performance (0.54)

SeamlessM4T - Artifacts

(1) SeamlessM4T models:

Inference code and fine-tuning recipes powered by our new modeling toolkit Fairseq2

(2) Tools for creating aligned speech data, including metadata to recreate SeamlessAlign

Paper: https://arxiv.org/abs/2308.11596

Blog post: https://ai.meta.com/blog/seamless-m4t/

Demo: https://seamless.metademolab.com/

Github repo: https://github.com/facebookresearch/seamless_communication

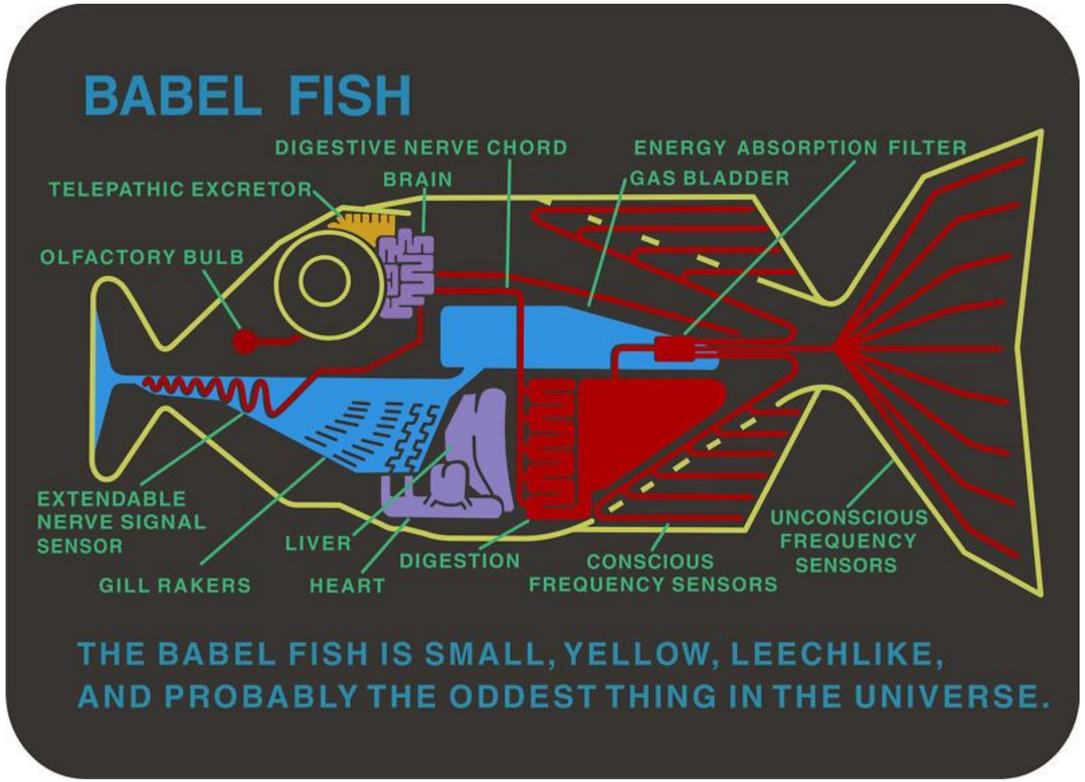
Fairseq2: https://github.com/facebookresearch/fairseq2

What’s next?

Low-latency streaming. Feeling like a normal in-person conversation with high quality; latency on par or better than that of human simultaneous interpreter.

Expressivity preserving models. Enable users to have control over the voice and create human-like, emotive speech translations (global and local prosody).

Putting everything together! Massively multilingual, low-latency and expressive.




Meta

Robustness to speaker variations (↑49%)

Languages	Average #	WHISPER-LARGE-V2		SEAMLESSM4T-LARGE	
(≥ 40 content groups)	cont. groups	$\text{chrF}_{MS}\uparrow$	$\text{CoefVar}_{MS}\downarrow$	$\text{chrF}_{MS}\uparrow$	$\text{CoefVar}_{MS}\downarrow$
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**

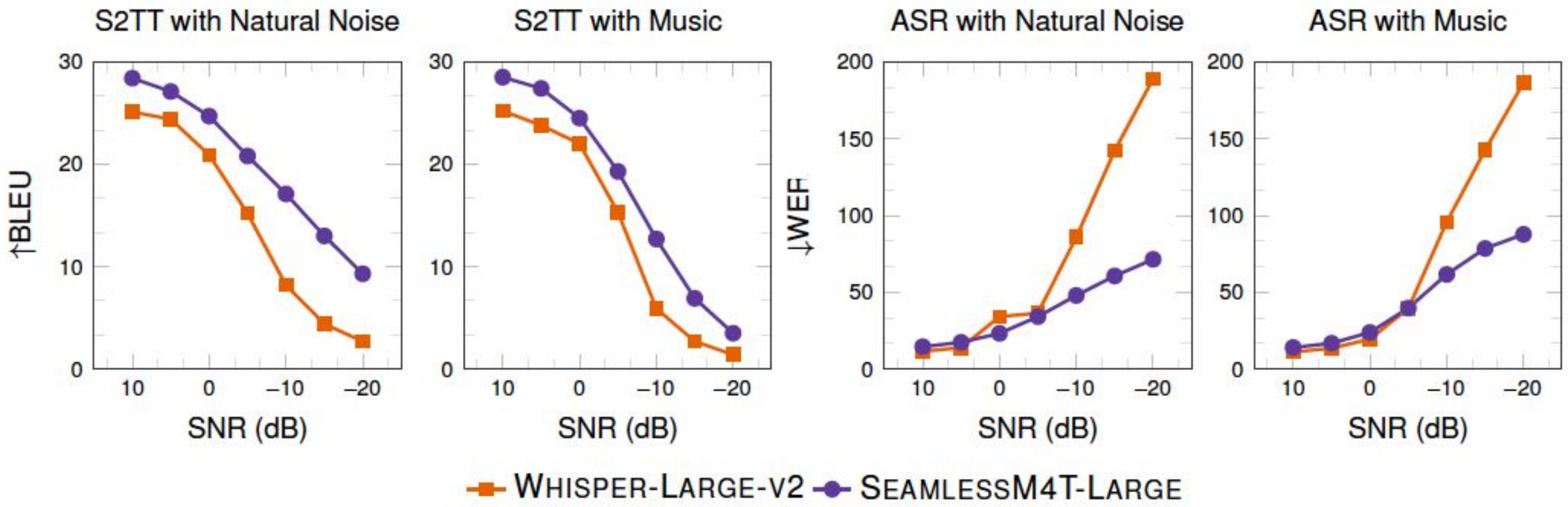
Methodology:

- Calculating average by-group mean score and by-group coefficient of variation of an utterance-quality metric (chrF).
- Fleurs languages that have at least 40 content gr in the test sets.

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

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

Robustness to noise (↑38%)



Methodology:

- Sampled audio clips from MUSAN [Snyder et al., 2015] on the “noise” and “music” categories, and mixed them with Fleurs speech audios under different SNR.
- 4 high-resource languages (French, Spanish, Modern Standard Arabic, and Russian), X-Eng

Human Evaluation: XSTS applied to S2TT

Direction	System	Avg. XSTS (S2TT)	% 3+	% 4+
eng–X	Human reference	4.69	95.98	78.66
	SeamlessM4T-Large	4.53	87.69	73.28
X–eng	Human reference	4.67	95.23	76.86
	Whisper-Large-v2	4.05	70.11	58.00
	SeamlessM4T-Large	**4.16**	**72.51**	**59.86**

Table 29: For S2TT task: overall average XSTS human evaluation results into and out of English, over all 24 evaluated languages. Results were computed for each language direction (see Table 31 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.

SeamlessM4T

Collection of models

	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 .

Two models part of the seamlessM4T release.

+ 2 smaller on-device models

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.

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]