Fleurs-SLU: A Massively Multilingual Benchmark for Spoken Language Understanding

Fabian David Schmidt[1], Ivan Vulić[2], Goran Glavaš[1], David Ifeoluwa Adelani[3]
[1] Center For Artificial Intelligence and Data Science, University of Würzburg, Germany
[2] Language Technology Lab, University of Cambridge, United Kingdom
[3] Mila, McGill University and Canada CIFAR AI Chair

Abstract

Spoken language understanding (SLU) is indispensable for half of all living languages that lack a formal writing system. Unlike for high-resource languages, for these languages, we cannot offload semantic understanding of speech to the cascade of automatic speech recognition (ASR) and text-based large language models (LLMs). Even if low-resource languages possess a writing system, ASR for these languages remains unreliable due to limited bimodal speech and text training data. Nonetheless, the evaluation of multilingual SLU is limited to shallow tasks such as intent classification or language identification. This is why we present Fleurs-SLU, a multilingual SLU benchmark that encompasses (i) 692 hours of speech for topical utterance classification in 102 languages and (ii) multiple-choice question answering via listening comprehension spanning 944 hours of speech across 92 languages. We extensively evaluate end-to-end speech classification models, cascaded systems that combine speech-to-text transcription with subsequent LLM-based classification, and multimodal speech-LLMs on Fleurs-SLU. Our results show that cascaded systems are more robust in multilingual SLU, though well-pretrained speech encoders can perform competitively in topical speech classification. Closed-source speech-LLMs match or surpass the performance of cascaded systems. We observe a strong correlation between robust multilingual ASR, effective speech-to-text translation, and strong multilingual SLU, indicating mutual benefits between acoustic and semantic speech representations. [1]

1 Introduction

Only about half of the world's living languages possess a formal writing system, underscoring the need for massively multilingual speech technology (Ethnologue, 2017). Spoken language understanding (SLU) is hence a critical feature of inclusive technology for these languages, which cannot rely on combining automatic speech recognition (ASR) with language models to offload semantic speech understanding. However, the datasets currently available for evaluating SLU in truly relevant languages suffer from significant limitations. For instance, the Minds14 benchmark assesses SLU on 14 exclusively high-resource languages for intent classification, a task that often requires only shallow semantic processing, such as detecting specific keywords (Gerz et al., 2021). Likewise, while SpeechTaxi focuses on spoken topical classification of Bible verses in 28 diverse languages, SLU of Bible verse arguably does not reflect real-world usage (Keller & Glavaš, 2025).

Moreover, even when a low-resource language has a writing system, multilingual ASR often struggles to reliably transcribe that language due to the limited availability of bimodal speech and text data. Pre-trained on the self-supervised wav2vec-BERT objective (Baevski et al., 2020), these models embed acoustic features that capture phonetic rather than semantic information (Choi et al., 2024). Conversely, multilingual speech models that excel in semantic

[1]The datasets & per-language results are available on Huggingface: SIB-Fleurs; Belebele-Fleurs.

encoding tend to be more robust in ASR tasks by, e.g., leveraging context-related cues or cross-lingual similarities such as cognates or similar syntactic structures (cf. §5.1). This underscores that robust SLU is instrumental for genuinely inclusive speech technology.

To address the shortcomings of multilingual SLU evaluation of prior benchmarks, we compile Fleurs-SLU by realigning datasets derived from Flores (Team et al., 2022), a machine-translation benchmark with parallel sentences for over 200 languages. We first filter out silent instances from Fleurs (Conneau et al., 2022) and then map the remaining data back to Flores. The resulting dataset is then merged with SIB-200 (Adelani et al., 2024) and Belebele (Bandarkar et al., 2023), that both are based on Flores, to create a topical utterance classification benchmark for 102 languages, and dataset for multiple-choice question answering (QA) on spoken paragraphs for 92 languages, respectively.

Contributions. 1) To the best of our knowledge, we are the first to present a multilingual SLU benchmark that spans over 100 languages and enables end-to-end SLU evaluation of multilingual speech encoders (mSE). SIB-Fleurs supports utterance classification for 7 topics with 692 hours of speech spanning 102 languages, over 70 more additional languages than existing benchmarks. Belebele-Fleurs provides textual multiple-choice QA from spoken paragraphs, totaling 944 hours of speech across 92 languages and exceeding concurrent work by 18 languages (Costa-jussà et al., 2024). **2)** We extensively benchmark state-of-the-art speech models on SIB-Fleurs and Belebele-Fleurs. Like prior work, we observe that Cascaded Systems (CS) remain more robust than mSE, enabling SLU competitive with text-based language understanding (Keller & Glavaš, 2025). We additionally benchmark state-of-the-art speech-LLMs on multilingual SLU and find that often perform on par or better than CS. We are the first to show that mSE pre-trained on language understanding objectives can yield performance competitive to that of CS on speech classification. We further isolate utterance quality as an important factor in multilingual SLU and show that, in zero-shot cross-lingual transfer (ZS-XLT), worse utterance quality can substantially deteriorate transfer performance. **3)** We empirically demonstrate that strong multilingual SLU coincides with much more robust multilingual ASR and higher-quality speech-to-English-text translation (S2ETT). This suggests that the pre-training of multilingual speech models may take SLU into account to make multilingual ASR more robust.

2 Related Work

Multilingual Speech Representation Learning. Modern speech representation models are pre-trained with self-supervised objectives that are optionally followed-up by ASR and speech-to-text translation (S2TT) training. Among these models, mHubert employs the self-supervised wav2vec 2.0 objective on 90K hours of speech across 147 languages by predicting pseudo-labels derived from clustering raw speech features (Boito et al., 2024). MMS-1b is also pretrained with wav2vec 2.0 on a large corpus of 491K hours of speech spanning 1,406 languages (Pratap et al., 2024). Whisper-v3 is a Transformer encoder-decoder that has been multi-task pre-trained on multilingual ASR, speech-to-English-text-translation (S2ETT), spoken language identification, and Voice activity detection on 680k hours of audio (Radford et al., 2022). SeamlessM4Tv2-Large is a multilingual multimodal translation model (Seamless Communication et al., 2023). It combines a text-to-text translation (T2TT) model pre-trained on 105 languages and a Conformer speech encoder pre-trained with w2v-BERT 2.0 on 4.5M hours of audio. The model is then trained on T2TT, S2TT, ASR, and knowledge distillation (KD) objectives. The authors perform KD from the text encoder to the speech encoder by minimizing the KL-divergence between the decoder's token output distributions on bi-modal speech and text data.

Multilingual SLU. The evaluation of multilingual SLU is constrained to a limited set of tasks. SLU has been significantly shaped by task-oriented dialogue (ToD), with datasets created focusing on ToD-specific tasks such as intent classification and slot filling. These tasks frequently require only basic semantic understanding, often reducing to merely detecting specific keywords. Additionally, commonly used utterance-level SLU tasks like language identification (LID) and sentiment classification do not assess content-based understanding but instead rely on phonetic or prosodic features of speech encodings. As a result, the majority of SLU datasets are predominantly in English. Multilingual exceptions are otherwise limited. The Minds14 dataset for intent classification only includes 14 high-resource languages (Gerz et al., 2021). SpeechTaxi offers spoken topical classification for Bible verses in 28 diverse languages, which however do not adequately represent real-world domains (Keller & Glavaš, 2025).

Concurrent Work. Costa-jussà et al. (2024) concurrently released the multilingual SLU benchmark 2M-BELEBELE for 74 languages. In this dataset, the authors first extend Fleurs by approx. 20% by incorporating human recordings for sentences that are part of Flores but were missing in Fleurs, as well as the questions and answers from Belebele. Costa-jussà et al. (2024) find that a CS-based approach on transcribing speech with Whisper-v3-Large and subsequently prompting Llama 3 70B with the transcription trails prompting the LLM on clean text by approx. 8 percentage points (pp).

3 Fleurs-SLU

We create Fleurs-SLU, a massively multilingual SLU benchmark for speech classification and multiple-choice QA from spoken paragraphs, from datasets that are based on Flores.

3.1 Core Datasets

Flores consists of professionally translated 3,001 sentences of English Wikipedia paragraphs to evaluate machine translation (Team et al., 2022).[2] **Fleurs** comprises 2.3 spoken utterances, on average, per sentence from the DEV and DEVTEST splits of 102 languages in Flores (Conneau et al., 2022).[3] Fleurs is used to evaluate multilingual ASR, LID, as well as speech-to-text and text-to-speech tanslation in all language directions. **SIB-200** refined the topical metadata annotations of sentences in the DEV and DEVTEST splits of Flores into 7 categories (Adelani et al., 2024).[4] The resulting SIB-200 is a topical classification benchmark for 205 language variants. **Belebele** is a multiple-choice reading comprehension benchmark for 122 languages (Bandarkar et al., 2024). The authors reconstruct paragraphs from Flores sentences. They generate 1-2 questions per English paragraph, which are professionally translated into 121 languages. Belebele comprises 900 questions that span across 488 passages.

3.2 Benchmark Creation

We compile Fleurs-SLU by carefully aligning data from the above benchmarks as follows.

1) Merging Fleurs & Flores. We begin by removing silent instances from Fleurs.[5] We first normalize the loudness to a target RMS level of -25 dB. We next apply voice activity detection using Silero-VAD (Team, 2024).[6] Samples are deemed silent if speech is detected in less than 5% of their duration.[7] Lastly, we verify our approach on 50 randomly sampled predicted silent instances. We find only one borderline misclassified example, which is noisy but comprehensible. We then conservatively merge Fleurs and Flores by matching instances first on exact string match and then by Levenshtein distance of 3 on normalized strings.[8]

2a) SIB-Fleurs. For each language, we pool instances from the training, validation, and test splits of SIB-200 (Adelani et al., 2024) and align the data with our merged Fleurs-Flores dataset with the same string alignment procedure as before. The data is then regrouped

[2] Flores comprises 3,001 sentences divided into DEV (997 sentences), DEVTEST (1,012 sentences), and TEST (992 sentences) sets. The authors did not release the TEST set.

[3] Almost all languages part of Fleurs however are missing a few hundred sentences from Flores.

[4] "science/technology", "travel", "politics", "sports", "health", "entertainment", "geography".

[5] Counts of removed examples are listed in Appendix A.6.

[6] https://github.com/snakers4/silero-vad

[7] The Silero-VAD pipeline also frequently removes inaudibly noisy samples as a side effect.

[8] Normalized strings remove characters that are Unicode punctuation codepoints.

into the training, validation, and test splits of the original Fleurs dataset. This segmentation ensures compatibility for speech models that may be trained on ASR using the training set of Fleurs prior to SIB-Fleurs evaluation. This also ensures that the speakers in the training set are different from those in the validation and test sets. Table 1 lists aggregated statistics on the instance- and the utterance-level for SIB-Fleurs. Appendix A.2 provides a list of samples by split per language.

Utterance Classification	
Languages	102
Classes (Topics)	7
Utterances per sample	2.2 (2.0)
Duration per utterance (s)	12.9 (11.6)
Total audio (hr)	692
Samples by Split	
Training	696 (728)
Validation	66 (70)
Test	163 (174)

Table 1: **Statistics of SIB-Fleurs.** Utterance-level metrics are aggregated by language and then pooled over languages. **Metrics:** either sums or averages (median).

Multiple-Choice QA Spoken Paragraphs	
Languages	92
Answer Choices	4
Questions per paragraph	1.8 (1.8)
Sentences per paragraph	3.6 (4.0)
Utterances per sentence	2.0 (2.0)
Duration per utterance (s)	12.6 (12.0)
Duration per paragraph (s)	47.0 (43.4)
Total audio (hr)	944
Samples (Paragraphs)	709 (771)

Table 2: **Statistics of Belebele-Fleurs.** Metrics are aggregated by language and then pooled over languages. **Metrics:** either sums or averages (median).

2b) Belebele-Fleurs. We merge our Fleurs-Flores sentences with Belebele paragraphs by intersecting the URLs of the texts. We discard all paragraphs that are not complete in Fleurs-Flores. We verify our reconstructed paragraphs against the original Belebele by ensuring that the Levenshtein distance for strings with removed punctuation is negligibly small (less than 3 characters). Table 2 provides a summary statistics on both the paragraph- and the sentence-level for Belebele-Fleurs, while Appendix A.3 lists the samples per language.

4 Experimental Setup

4.1 Tasks and Languages

SIB-Fleurs. We train mSE on the utterances and CS on the transcriptions of the English training set. For both mSE and CS, we feed the sequence-level representation pooled from token or speech frame embeddings into a classification head. Roberta-Large uses the `[CLS]` token as a sequence-level embedding. All other models average the token or speech frame output embeddings.

Belebele-Fleurs. We train and validate CS models on the English training and dev sets of Belebele (Bandarkar et al., 2024), respectively. We jointly embed the paragraph, question, and choices with text encoders. We then average the token encodings of each choice $c_i \in C$ and project the choice embedding via head $H^{D\times 1}$ to a logit $\mathbf{l}_{\mathbf{c_i}}$. We minimize the cross-entropy of the concatenated choice logits $\{\mathbf{l}_{\mathbf{c_i}}\}_{i=1}^{|C|}$ to the label choice.[9]

4.2 Cross-Lingual Transfer Setups

We experiment on two commonly used cross-lingual transfer (XLT) paradigms. Zero-shot cross-lingual transfer (ZS-XLT) and Translate-Test (TTEST) allow us to evaluate XLT without requiring additional annotation for any target language. In ZS-XLT, we first train a multilingual model on the English source-language data (cf. §4.1) and then directly run inference on the target-language test instances. In TTEST, the model is also first fine-tuned on labeled English source-language data. At test time, the target-language examples are translated

[9]We do not evaluate speech LLMs as existing models are based on Whisper-v3 which is limited to 30 seconds of audio input. We leave such evaluation to future work.

to the source language prior to inference, which enables XLT with monolingual LLMs. We additionally evaluate zero-shot prompting of speech-LLMs. Here, we provide the context of the task in English and all relevant input in-language in the respective modality, i.e. text or utterances, to the speech-LLMs (cf. §A.1).

Speech Classification. We evaluate state-of-the-art mSE for speech classification (cf. §2). We include MMS-1B without fine-tuning ('MMS-1B'), with ASR fine-tuning on Fleurs ('MMS-1B-Fleurs'), and with ASR fine-tuning on multilingual datasets ('MMS-1B-all'), allowing us to analyze the impact of ASR fine-tuning on cross-lingual SLU. For ASR fine-tuning, Pratap et al. (2024) train language adapters and language-specific decoding heads while keeping all other parameters frozen. For task fine-tuning, we freeze the adapters to facilitate ZS-XLT. We further evaluate mHubert and the speech encoders of both Whisper-v3-Large and SeamlessM4Tv2-Large (cf. §2).

Cascading & Text. Cascaded systems (CS) perform XLT in two steps: (i) an ASR model transcribes speech into text, and (ii) a text encoder processes the transcription via a classification head. We consider two transcription targets: the target language (in-language) and English (speech-to-English translation, S2ETT). S2ETT corresponds to TTEST, enabling XLT with monolingual LLMs. We use SeamlessM4Tv2-Large and Whisper-v3-Large as ASR backends. For languages not supported by the model, we manually select the closest available language for in-language transcription. All decoding is greedy, as more complex decoding strategies do not improve XLT performance (Ebing & Glavaš, 2024). We additionally evaluate on the original text (TEXT) of SIB-200 and Belebele. For TTEST, we translate this text to English using SeamlessM4Tv2-Large.[10] For classification, we evaluate three text encoders: Roberta-Large (Liu et al., 2019), LLM2Vec (BehnamGhader et al., 2024), and NLLB-LLM2Vec (Schmidt et al., 2024). LLM2Vec is a sequence encoder based on Llama 3 8B (AI@Meta, 2024), trained with bidirectional attention on masked next-token prediction and SimCSE (Gao et al., 2021). NLLB-LLM2Vec extends LLM2Vec with the encoder of the NLLB translation model, covering 200+ languages for robust multilingual NLU (Schmidt et al., 2024).

Speech-LLMs. We further evaluate two multimodal speech-LLMs, Qwen 2.5 7B-Omni (Xu et al., 2025) and Gemini 2.0 Flash[11], in a zero-shot prompting setup.[12] For each task, we provide an English task description and supply the task-specific input (e.g., question, paragraph, or utterance) in the target language, using the appropriate modality (text or speech). To ensure consistent audio quality across languages, we normalize all utterances to a root mean square (RMS) level of 0.07. For Belebele-Fleurs, where inputs consist of multiple utterances, we normalize the concatenated sequence. Appendix §A.1 details the prompt formats and input preprocessing used for each task and modality.

4.3 Further Details

Hyperparameters. We train all mSE and CS models with AdamW (Loshchilov & Hutter, 2019), weight decay of 0.01, and with 10% linear warm-up and followed by linear decay on an effective batch size of 32. For SIB-Fleurs, we run a grid search over the learning rates $\{1, 2, \ldots, 9, 10\}e^{-5}$, since suitable hyperparameters have not yet been extensively studied for mSE on such a downstream task. For Belebele-Fleurs, we fine-tune the LLMs on the learning rates $\{1, 2, 3\}e^{-5}$ with LoRAs of rank r=16 and alpha α=32 attached onto all linear layers, as full fine-tuning is prohibitively expensive. We train models for 20 and 3 epochs for SIB-Fleurs and Belebele-Fleurs, respectively, and validate at every 10% of training steps. Experimental results are averaged across 3 random seeds. We report results for runs on the learning rate that performs best, on average, on the English validation sets. For experiments with Qwen 2.5 7B-Omni and Gemini 2-Flash, we perform greedy decoding. We report further experimental details of our prompting experiments in §A.1.

[10]We use the gold Q&A in English for TTEST. Otherwise, MT models would need to be combined, e.g., S2ETT of paragraphs with Whisper, and translation of textual Q&A with SeamlessM4Tv2.

[11]`https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash`

[12]Note that the audio encoder of Qwen 2.5 7B-Omni is initialized with Whisper-v3-Large prior to further cross-modal instruction tuning.

Text vs. Utterance Quality. We evaluate models trained on the original text (TEXT) as well as on speech data of varying quality. Each sentence in Fleurs is associated with one or more utterances. To quantify speech quality, we compute the character error rate (CER) between the reference Flores sentence and its transcriptions produced by Whisper-v3-Large and SeamlessM4Tv2-Large. Based on these CER scores, we construct two utterance subsets: one comprising the lowest-CER utterances (best quality), and one with the highest-CER utterances (worst quality). We fine-tune the mSE and CS models and evaluate all models (incl. speech-LLMs) separately on each subset to isolate the effect of utterance quality on downstream performance. This setup also balances the training data to enable a fair comparison between mSE, CS, and TEXT.

5 Results

Table 3 summarizes the main results for both tasks by approach across our XLT setups. We dissect the results along several axes of analysis.

English. The first column presents the in-language English performance by task, modality (TEXT vs. mSE, CS, and speech-LLM), and utterance quality by model.

Text, CS, and Speech-LLMs. English ASR performance is strong for both transcription models (cf. Figure 1), and CS effectively leverages the NLU capabilities of LLMs, thanks to their extensive pre-training on both tasks. CS are even on par with and sometimes outperform models trained on gold text, underscoring the quality of ASR on English. The slight outperformance of CS over TEXT in SIB-Fleurs results from model selection on comparatively small validation splits (cf. Table 1). As a result, CS outperforms all mSE models on SIB-Fleurs. On Belebele-Fleurs, both CS models backed by LLM2Vec and NLLB-LLM2Vec perform comparably. In sum, the transcription model has no clear impact on English performance. The speech-LLMs perform on par or better on Belebele-Fleurs while slightly trailing the best fine-tuned models on both CS and TEXT on SIB-Fleurs. Gemini 2.0 Flash consistently performs better than Qwen 2.5B 7B-Omni. Notably, their cross-modal gap (TEXT vs. speech-LLM) is small (approx. -1.5%).

mSE. The English results on SIB-Fleurs suggest that the SLU capabilities of mSE models are strongly shaped by their pre-training curriculum. mHubert (41.1%) and MMS-1B (44.6%), which are pre-trained solely in a self-supervised manner, underperform other mSE on SIB-Fleurs. Fine-tuning MMS-1B on ASR, either on Fleurs (MMS-1B-Fleurs) or Fleurs combined with additional data (MMS-1B-All), results in significant improvements on SIB-Fleurs of +10-20 percentage points (pp). Notably, MMS-1B-Fleurs outperforms MMS-1B-All (+9.6 pp), suggesting that the broader domain mixture in the MMS-1B-All training set negatively affects this task. Whisper-v3-Large surpasses all MMS-1B variants (78.3%). The large-scale pre-training of Whisper-v3 on both multilingual ASR and S2ETT enhances its SLU performance. SeamlessM4Tv2-Large is the best performing mSE on the English test set of SIB-Fleurs (87.4%), with only a slight performance gap compared to CS (approx. -3 pp). The joint pre-training on multilingual ASR as well as multilingual and cross-modal MT with text-to-speech knowledge distillation (KD) significantly enriches the semantics in speech representations of SeamlessM4Tv2-Large. The results particularly highlight that (i) multilingual ASR, (ii) S2ETT pre-training, and (iii) text-to-speech KD are crucial training objectives for enabling speech encoders to acquire strong SLU capabilities.

Utterance Quality. The quality of English utterances does not affect performance of the models across both tasks. This can be attributed to the large-scale training on English ASR data of all models. This is why CS are also on par with TEXT.

ZS-XLT. The right-hand side of Table 3 reports ZS-XLT performance. For each task, we group the languages into (i) languages supported by Whisper-v3, (ii) languages supported by SeamlessM4Tv2, and (iii) languages unsupported by either model. Whisper-v3 and SeamlessM4Tv2 overlap in their support for 81 Fleurs languages.[13]

[13]For unsupported languages, we hand-select the closest supported language to transcribe into.

			English		Non-English							
Utterance Quality (**B**est; **W**orst)			**B**	**W**	**B**	**W**	**B**	**W**	**B**	**W**	**B**	**W**
Language Groups (Size)		**Setup**	**EN**		**Whisper**		**S4T**		**Unsup.**		**Non-EN**	
SIB-Fleurs			**(1)**		**(85)**		**(90)**		**(7)**		**(101)**	
mSE	mHuBERT	X	41.1	39.7	27.3	19.1	27.2	19.1	26.5	16.3	26.9	18.6
	MMS-1B	X	44.6	46.9	19.9	13.6	19.8	13.5	20.8	13.3	19.8	13.4
	MMS-1B-Fleurs	X	64.8	64.2	25.2	20.8	24.7	20.3	27.1	22.1	25.0	20.6
	MMS-1B-All	X	55.2	53.1	21.7	18.1	21.4	17.9	24.9	21.0	21.7	18.2
	Whisper-v3-L	X	78.3	76.3	46.0	42.0	45.5	41.3	39.4	38.4	44.7	40.8
	SeamlessM4Tv2-L	X	87.4	88.5	82.3	79.3	82.2	79.1	55.7	52.9	79.0	75.8
CS	RoBERTa$_{\text{Large}}$-Wh-EN	T	90.4	91.7	75.8	72.7	73.6	70.8	52.0	48.7	71.5	68.4
	RoBERTa$_{\text{Large}}$-S4T-EN	T	91.7	90.8	86.2	84.5	85.9	84.2	56.3	55.2	82.3	80.6
	LLM2Vec-Wh-EN	T	91.1	92.3	76.6	75.3	74.5	73.4	53.0	53.2	72.4	71.1
	LLM2Vec-S4T-EN	T	93.0	91.3	**87.1**	**85.2**	**86.9**	**84.9**	58.1	56.6	**83.4**	81.4
	NLLB-LLM2Vec-Wh	X	92.5	91.0	77.5	73.9	75.4	72.2	54.4	53.2	73.7	70.4
	NLLB-LLM2Vec-S4T	P	**94.0**	**92.7**	85.1	84.9	85.3	**84.9**	**62.3**	**60.7**	82.3	**81.9**
Speech-LLM	Qwen 2.5 7B-Omni	P	87.0	87.6	59.5	58.4	58.3	57.3	45.6	45.6	56.9	55.9
	Gemini 2.0 Flash	P	87.6	88.7	82.1	81.4	80.8	80.0	59.4	59.3	78.6	77.8
Text	RoBERTa$_{\text{Large}}$-S4T-EN	T	**92.3**		87.4		**89.1**		55.3		85.2	
	LLM2Vec-S4T-EN	T	91.0		86.4		87.9		55.6		84.1	
	NLLB-LLM2Vec	X	**92.3**		**88.1**		87.9		**80.7**		**87.2**	
	Qwen 2.5 7B-Omni	P	88.7		76.4		75.1		53.9		73.0	
	Gemini 2.0 Flash	P	90.4		87.0		86.8		77.2		86.1	
Belebele-Fleurs			**(1)**		**(79)**		**(84)**		**(4)**		**(91)**	
CS	RoBERTa$_{\text{Large}}$-Wh-EN	T	82.6	81.3	57.5	56.2	56.4	55.2	46.2	44.7	55.5	54.3
	RoBERTa$_{\text{Large}}$-S4T-EN	T	82.0	79.4	67.5	66.4	67.0	65.8	47.2	46.2	65.2	64.1
	LLM2Vec-Wh-EN	T	94.9	**94.4**	74.6	73.5	73.6	72.6	63.5	62.9	72.7	71.8
	LLM2Vec-S4T-EN	T	**95.5**	**94.4**	84.0	83.1	**83.5**	**82.5**	**64.7**	**63.6**	**81.8**	**80.9**
	NLLB-LLM2Vec-Wh	X	94.8	93.5	58.0	56.2	56.8	55.0	40.9	41.0	55.5	53.8
	NLLB-LLM2Vec-S4T	X	94.9	93.5	61.3	60.2	60.8	59.6	43.9	41.1	59.2	57.8
Speech-LLM	Qwen 2.5 7B-Omni	P	91.6	91.0	51.4	50.8	50.4	49.7	34.4	34.4	49.0	48.4
	Gemini 2.0 Flash	P	94.1	93.4	**84.4**	**83.5**	82.9	82.0	59.7	59.4	81.1	80.2
Text	RoBERTa$_{\text{Large}}$-S4T-EN	T	83.3		72.1		72.6		48.3		70.5	
	LLM2Vec-S4T-EN	T	95.3		87.8		88.3		66.0		86.3	
	NLLB-LLM2Vec	X	95.1		65.6		64.6		55.2		63.8	
	Qwen 2.5 7B-Omni	P	94.1		69.1		67.4		36.6		64.9	
	Gemini 2.0 Flash	P	**95.9**		**90.8**		**89.7**		**82.0**		**89.1**	

Table 3: **zs-xlt with mSE, CS, and Text and zero-shot prompting with speech-LLMs (cf. §4).** We report accuracy averaged over 3 seeds on checkpoints that maximize perf. on English validation sets. CS: The suffixes Wh and S4T denote Transcription & Translation (incl. -En: {Text,Speech}-to-English-Text-Translation), respectively. Numbers in parentheses denominate size of group (e.g., Whisper-v3 supports 84 languages of SIB-Fleurs). **Setup**: X=Zero-shot Cross-Lingual Transfer;T=Translate-Test;P=Zero-Shot Prompting (cf. §4). **Abbreviations**: EN=English;Whisper=Whisper-v3-Large;S4T=SeamlessM4Tv2-Large; Unsup.=Unsupported. The (second-)best model in each column in **bold** (underline).

Text, CS, and Speech-LLMs. Across both tasks and all language groups, models trained on transcriptions of SeamlessM4Tv2-Large consistently outperform models fine-tuned on transcriptions of Whisper-v3-Large. The performance gap grows as Whisper's support decreases or becomes unavailable for the target languages (cf. 'S4T' and 'Unsup.' language groups in Table 3). Nevertheless, while CS are mostly competitive on SIB-Fleurs, they trail models evaluated on ground-truth paragraphs more significantly on Belebele-Fleurs (approx. -5 pp). We presume that transcription and S2ETT errors propagate more severely in multiple-choice QA. Moreover, the best speech-LLM, Gemini 2.0 Flash, performs highly competitively across tasks, boasting on par or better performance than other models.

For SIB-Fleurs, all CS with SeamlessM4Tv2-Large deteriorate only slightly in xlt performance to all 101 target language, on average, relative to English (approx. -10 pp). Most of this xlt gap comes from the 7 target languages that SeamlessM4Tv2 does not support (approx. -35 pp on average). In contrast, pairing LMs with Whisper-v3-Large causes more pronounced drops on languages that Whisper supports (approx. -13 pp) and again a larger deficit on languages neither model supports (approx. -32 pp). CS are highly competitive to Text (approx. -0.7 pp). Only NLLB-LLM2Vec evaluated on gold text tremendously outperforms all other models on unsupported languages, since they are supported by NLLB (Team et al., 2022).

For Belebele-Fleurs, S2ETT paired with LLM2Vec (i.e., TTEST) outperforms ZS-XLT cascading on in-language transcriptions and NLLB-LLM2Vec, except for unsupported languages. This suggests that S2ETT of both Whisper-v3 and SeamlessM4Tv2 sufficiently translates the target languages into English for successful NLU. On unsupported languages, however, NLLB-LLM2Vec performs better for likely two reasons. First, for languages with low S2ETT quality, in-language transcriptions to closely related languages likely better preserve the core meanings of input sequences. Second, in addition to translation, NLLB was pre-trained with denoising autoencoding, making NLLB-LLM2Vec more resistant to noisy inputs than LLM2Vec. Overall, the findings suggest that SeamlessM4Tv2-Large is more robust for both in-language ASR as well as S2ETT (cf. Figure 1).

The performance of speech-LLMs varies substantially across models and language groups. Gemini 2.0 Flash strongly outperforms Qwen 2.5 7B-Omni on both tasks in non-English settings. Qwen exhibits a marked performance degradation when transitioning from English to non-English inputs, dropping by approx. 30 pp in TEXT, and even more steeply in the speech modality, where the cross-modal gap widens from approx. 3% (English) to 18% (non-English). In contrast, Gemini maintains strong performance on non-English TEXT inputs and shows only a moderate cross-modal drop of approx. 9%. This however exceeds the gap for the best fine-tuned CS models vs. their TEXT counterparts (approx. 3%).
Qwen relies on Whisper-v3-Large as its audio encoder, making it sensitive to Whisper's transcription limitations. Moreover, its instruction tuning possibly is focused on a small subset of high-resource languages, potentially limiting its generalization to languages not seen during alignment or fine-tuning (Xu et al., 2025). This is consistent with its sharp drop in performance on unsupported languages. By contrast, Gemini's closed-source nature precludes detailed analysis, but its robustness across modalities and languages suggests a more extensive and diverse instruction tuning setup. It may also benefit from large-scale multilingual pretraining, enhanced alignment objectives, or broader training coverage.

mSE. The ZS-XLT performance of mSE on SIB-Fleurs mirrors the trends we observed for English performance. When trained solely with the wav2vec 2.0 objective, MMS-1B (19%) only slightly surpasses random performance (14%) in ZS-XLT for the highest-quality utterances, regardless of the target language group. In contrast to English, post-hoc modular ASR fine-tuning of MMS-1B with language adapters and language-specific decoding heads, whether on Fleurs alone (MMS-1B-Fleurs) or on Fleurs with additional data (MMS-1B-All), yields much smaller gains in ZS-XLT (approx. +1-3 pp). Furthermore, despite employing fully parameter-shared multilingual ASR and S2ETT pre-training, Whisper-v3-Large fails to transfer strong English in-language performance to other languages effectively, with the XLT gap ranging from -32.3 pp for supported languages to -38.9 pp for unseen languages. In contrast, only SeamlessM4Tv2-Large achieves performance comparable to English on supported languages (-5.2 pp), though also deflates in performance on unsupported languages (55.7%). Overall, these findings nevertheless indicate that multilingual cross-modal translation and multilingual text-to-speech distillation align and semantically enrich multilingual speech representations to enhance cross-lingual SLU.

Utterance Quality. The utterance quality consistently affects performance on both tasks for all model configurations evaluated on non-English languages. Notably, CS and speech-LLMs seem to be less affected than mSE. We attribute this to the much more sizable pre-training on text of diverse quality of LLMs compared to speech-only models. For SIB-Fleurs, the magnitude of the performance gaps between best and worst quality utterances in ZS-XLT mimics how well the speech model backbones perform across both tasks in both CS and end-to-end speech classification. SeamlessM4Tv2-Large is more robust to noisy utterances than Whisper-v3-Large, whereas MMS variants suffer from the largest drops (approx. 5 pp).

5.1 Further Analyses and Discussion

ASR & Translation Performance. To further understand the underlying factors behind our main results (cf. Table 3), we benchmark Whisper-v3-Large and SeamlessM4Tv2-Large on ASR and Speech-to-English-text-translation (S2ETT) on all 102 Fleurs languages in Figure 1. We first compute CER and sacreBLEU between the Flores sentences and the in-language

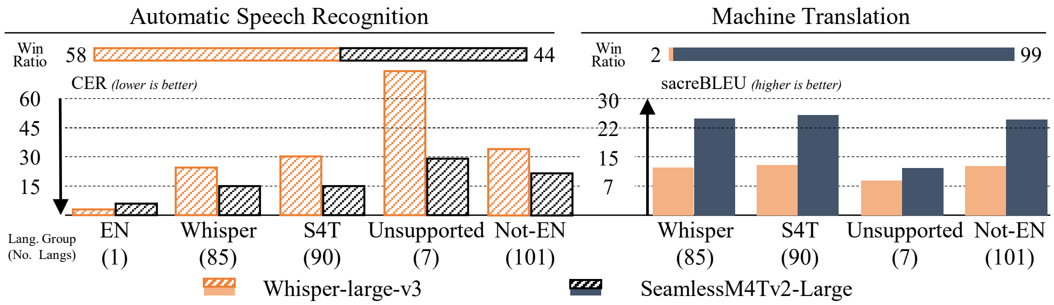

Figure 1: **ASR & Speech-to-English-Text Translation (S2ETT).** CER and sacreBLEU scores for ASR and S2ETT outputs on Fleurs utterances, evaluated against original Flores sentences and pooled across all splits. Numbers in parentheses indicate the number of supported languages per model; e.g., Whisper-v3 (SeamlessM4Tv2) supports 85 (90) of the 101 non-English Fleurs languages. The two models combined do not support 7 languages. 'Win Ratio' denotes the number of languages on which a model outperforms the other. Abbreviations: Whisper = Whisper-v3-Large; S4T = SeamlessM4Tv2-Large.

transcriptions and S2ETT transcriptions of utterances in Fleurs pooled across all splits, respectively.[14] We 'macro-average' the metrics over all languages.

ASR. Whisper-v3-Large outperforms SeamlessM4Tv2-Large for ASR on 58 out of 102 Fleurs languages (cf. 'win ratio'). Nevertheless, SeamlessM4Tv2-Large is the overall more robust transcription model: in transcription for languages *other than English*, the CER of Whisper-v3-large is more than twice as high as the CER of SeamlessM4Tv2-Large, on average. The full per-language results (cf. Appendix A.5)) confirm that SeamlessM4Tv2-Large exhibits much better ASR support for low(er)-resource languages, while being competitive in transcription quality for higher-resource languages.

S2ETT. SeamlessM4Tv2-Large outperforms Whisper-v3-large across the board in speech-to-English-text-translation and is favored for all but 2 languages. For both groups of languages supported by Whisper-v3-Large and SeamlessM4Tv2-Large, respectively, SeamlessM4Tv2-Large achieves on average about 8 higher sacreBLEU than Whisper-v3-Large. The gap from SeamlessM4T-Large to Whisper-v3-Large reduces to about 5 sacreBLEU for unsupported languages. This supports the notion that SeamlessM4Tv2-Large is a more robust multilingual speech encoder.

The more robust ASR and much stronger translation performance of SeamlessM4Tv2-Large results stem from its pre-training. SeamlessM4Tv2 first initializes a text encoder and decoder with weights from a pre-trained NLLB translation model (Team et al., 2022) and a speech encoder pre-trained on 4.5M hours of self-supervised training on w2v-BERT 2.0 objective. The model is then trained on translation objectives from text and speech to text between any two languages in both translation directions. The text encoder-decoder backbone is used to train the speech encoder with token-level knowledge distillation objectives on decoder output representations. On the contrary, Whisper-v3 trains models from scratch on, among others, in-language ASR and S2ETT (cf. §4). Consequently, the mixture of strong initialization from existing MT backbones, text-to-speech knowledge distillation, and multimodal translation objectives result in much stronger translation performance.

The main results (cf. Table 3), together with our ASR and S2ETT analyses, underscore that sizable text and pre-training coupled with cross-modal and multilingual translation and text-to-speech knowledge distillation infuses rich semantic knowledge into mSE, as witnessed by the SLU performance of SeamlessM4Tv2-based models on Fleurs-SLU.

[14] We verify that there is no leakage of dev and test splits of Fleurs in SeamlessM4Tv2-Large and Whisper-v3-Large. We observe that the averages and the standard deviations CER and sacreBLEU by task and language for either model are highly comparable across splits. We trust that the models have not been trained on the dev and test set of Fleurs.

				ZS-XLT					
		EN (1)		S4T (89)		N/A (12)		AVG (101)	
S4Tv2-L	Size	B	W	B	W	B	W	B	W
INCL. LA	635M	87.4	88.5	82.2	79.1	55.7	52.9	79.0	75.8
EXCL. LA	588M	88.5	86.4	80.9	77.5	55.6	52.0	77.8	74.4
Δ	47M	+1.1	-2.1	-1.3	-2.6	-0.1	-0.9	-1.2	-1.4

Table 4: **Ablation of Length Adaptor.** We benchmark SeamlessM4Tv2-Large with and without the Length Adaptor on SIB-Fleurs. See Table 3 for further details.

Length Adaptor. The sequence length of encoded speech typically far exceeds the length of embedded tokenized text for the same input. The mSE of SeamlessM4Tv2-Large thus appends a temporal convolution, a 'length adaptor', as its final layer to reduce the resolution of speech frames by a factor of 8 and to better align the modalities (Seamless Communication et al., 2023).[15] This begs the question whether a length adaptor is essential for pooling semantics from speech output tokens. We hence compare SeamlessM4Tv2-Large's performance on SIB-Fleurs with and without the length adaptor.

Table 4 presents the inconclusive results for both variants. For English, performance improves on high-quality utterances but decreases on lower-quality ones. In contrast, performance most pronouncedly declines for low-quality utterances in ZS-XLT to supported languages (-2.6%). The ZS-XLT unsupported languages is not affected for high-quality utterances (-0.1%) and only slightly for low-quality ones (-0.9%). Three factors may explain this modest gap. First, removing the length adaptor reduces model size by 47M parameters (-7.4 pp), which were part of substantial pre-training. Second, the length adaptor, as the final layer of the mSE, is most explicitly trained to embed speech into a shared multilingual space, attended to by the text decoder at every layer. Finally, the length adaptor might filter noisy frames through temporal downsampling to improve the robustness of speech encodings. While a modest gap persists across setups, we cannot decisively infer whether appending a length adaptor is, as opposed to replacing model capacity with other layers of equal parameter size during pre-training, crucial for improved multilingual SLU. We conclude that the pre-training regime is more vital for multilingual SLU capabilities of mSE than nuanced architectural design choices. We leave a more detailed investigation into the utility of length adaptors in SLU to future work.

6 Conclusion

We introduce Fleurs-SLU, a multilingual SLU benchmark for semantic speech classification across 102 languages and multiple-choice question answering from spoken paragraphs in 92 languages. Using Fleurs-SLU, we evaluate massively multilingual speech models in both end-to-end speech classification and a cascaded approach that combines initial speech-to-text transcription and subsequent text-based classification with LLMs. Our findings indicate that, while cascaded systems remain the most robust option, multilingual speech encoders can achieve competitive performance when adequately pre-trained. Moreover, speech-LLMs can yield state-of-the-art performance on par with the best targeted SLU pipelines when appropriately aligned for multilingual spoken language instruction following. Moreover, speech-LLMs can achieve state-of-the-art performance on par or better than the best targeted SLU pipelines when properly aligned for multilingual spoken language instruction following. Furthermore, we observe a strong correlation between strong multilingual SLU and both the robustness of multilingual ASR and the effectiveness of cross-modal speech translation to English text. This suggests that multilingual SLU and multilingual ASR can be mutually beneficial. We hope that our findings inspire future work towards developing more efficient multilingual speech encoders that are jointly pre-trained for both multilingual ASR and SLU to close the performance gap between end-to-end speech classification and cascaded approaches.

[15] For architectural details, we refer to the original paper (Communication et al., 2023).

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 226–245, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.14/.

AI@Meta. Introducing meta llama 3, 2024. URL https://ai.meta.com/blog/meta-llama-3/.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*, 2023.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=IW1PR7vEBf.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. mHuBERT-147: A Compact Multilingual HuBERT Model. In *Interspeech 2024*, 2024.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. Self-supervised speech representations are more phonetic than semantic, 2024. URL https://arxiv.org/abs/2406.08619.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. Seamlessm4t: Massively multilingual & multimodal machine translation, 2023. URL https://arxiv.org/abs/2308.11596.

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

Marta R. Costa-jussà, Bokai Yu, Pierre Andrews, Belen Alastruey, Necati Cihan Camgoz, Joe Chuang, Jean Maillard, Christophe Ropers, Arina Turkantenko, and Carleigh Wood. 2m-belebele: Highly multilingual speech and american sign language comprehension dataset, 2024. URL https://arxiv.org/abs/2412.08274.

Benedikt Ebing and Goran Glavaš. To translate or not to translate: A systematic investigation of translation-based cross-lingual transfer to low-resource languages. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5325–5344, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.298. URL https://aclanthology.org/2024.naacl-long.298/.

Ethnologue. How many languages in the world are unwritten? — web.archive.org. https://web.archive.org/web/20170227130129/https://www.ethnologue.com/enterprise-faq/how-many-languages-world-are-unwritten-0, 2017.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL https://aclanthology.org/2021.emnlp-main.552/.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. Multilingual and cross-lingual intent detection from spoken data, 2021. URL https://arxiv.org/abs/2104.08524.

Lennart Keller and Goran Glavaš. Speechtaxi: On multilingual semantic speech classification. In *Proceedings of the 2025 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. to appear, 2025.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024. URL http://jmlr.org/papers/v25/23-1318.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Fabian David Schmidt, Philipp Borchert, Ivan Vulić, and Goran Glavaš. Self-distillation for model stacking unlocks cross-lingual NLU in 200+ languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6724–6743, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.394. URL https://aclanthology.org/2024.findings-emnlp.394/.

Loïc Barrault Seamless Communication, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. Seamless: Multilingual expressive and streaming speech translation. 2023.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.

Silero Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

A Appendix

A.1 Further Experimental Details

Compute infrastructure. All experiments run on a single Nvidia L40S 48GB or A100 80GB, respectively. We estimate that the total compute budget accumulates to about 2K GPU hours. Training runs on SIB-Fleurs require about 30-45 minutes, while fine-tuning LLMs for Belebele require roughly 9 hours per run. Evaluation, per checkpoint, across all languages supported by the corresponding benchmark, requires about 1 hour due to the comprehensive setups (e.g., original text and two types of transcriptions for speech recordings).

Prompting Qwen 2.5 7B-Omni & Gemini 2-Flash.

For the multimodal speech-LLMs Qwen 2.5 7B-Omni and Gemini 2-Flash, we apply the following utterance pre-processing. Each audio segment is normalized to a target root mean square (RMS) level of 0.07 to ensure consistent loudness across samples and reduce variability due to recording conditions. Specifically, we first normalize the loudness of individual utterances; for Belebele-Fleurs, we additionally normalize the loudness after concatenation. The models are then prompted using the task- and modality-specific templates detailed below.

SIB-Fleurs

Speech.
The utterance belongs to one of the following topics.
Utterance: `[IN-LANGUAGE UTTERANCE]`

Text.
The sentence belongs to one of the following topics.
Sentence: `[IN-LANGUAGE SENTENCE]`

Topics:

1. Entertainment
2. Geography
3. Health
4. Politics
5. Science and Technology
6. Sports
7. Travel

Please respond with only the number of the correct answer (1, 2, 3, 4, 5, 6, or 7).

Belebele-Fleurs

Speech.
Listen to the audio passage. Based on the audio, answer the following multiple-choice question.
Passage: `[IN-LANGUAGE CONCATENATED UTTERANCES]`

Text.
Given the paragraph, answer the following multiple-choice question.
Paragraph: `[IN-LANGUAGE PARAGRAPH]`

Question: `[IN-LANGUAGE QUESTION]`

Options:

1. `[IN-LANGUAGE MULTIPLE-CHOICE-ANSWER 1]`
2. `[IN-LANGUAGE MULTIPLE-CHOICE-ANSWER 2]`
3. `[IN-LANGUAGE MULTIPLE-CHOICE-ANSWER 3]`
4. `[IN-LANGUAGE MULTIPLE-CHOICE-ANSWER 4]`

Please respond with only the number of the correct answer (1, 2, 3, or 4).

A.2 SIB-Fleurs

Language Code	Language Name	Train	Validation	Test
afr_Latn	Afrikaans	406	86	95
amh_Ethi	Amharic	752	54	149
arb_Arab	Modern Standard	579	64	133
asm_Beng	Assamese	730	71	176
ast_Latn	Asturian	701	69	177
azj_Latn	North Azerbaijani	712	71	174
bel_Cyrl	Belarusian	690	71	177
ben_Beng	Bengali	742	71	176
bos_Latn	Bosnian	746	71	177
bul_Cyrl	Bulgarian	749	70	176
cat_Latn	Catalan	683	71	177
ceb_Latn	Cebuano	741	61	149
ces_Latn	Czech	732	68	172
ckb_Arab	Central Kurdish	738	70	176
cym_Latn	Welsh	739	71	177
dan_Latn	Danish	696	70	177
deu_Latn	German	736	69	175
ell_Grek	Greek	750	67	168
eng_Latn	English	738	71	177
est_Latn	Estonian	700	71	176
fin_Latn	Finnish	735	71	175
fra_Latn	French	753	65	164
fuv_Latn	Nigerian Fulfulde	752	68	166
gaz_Latn	West Central Oromo	574	6	17
gle_Latn	Irish	731	71	176
glg_Latn	Galician	660	71	174
guj_Gujr	Gujarati	752	71	177
hau_Latn	Hausa	753	70	166
heb_Hebr	Hebrew	754	70	175
hin_Deva	Hindi	653	60	132
hrv_Latn	Croatian	756	71	176
hun_Latn	Hungarian	750	71	177
hye_Armn	Armenian	741	71	177
ibo_Latn	Igbo	737	71	177
ind_Latn	Indonesian	728	69	167
isl_Latn	Icelandic	381	18	23
ita_Latn	Italian	743	69	175
jav_Latn	Javanese	740	67	171
jpn_Jpan	Japanese	662	62	164
kam_Latn	Kamba	752	69	179
kan_Knda	Kannada	660	70	174
kat_Geor	Georgian	557	69	177
kaz_Cyrl	Kazakh	749	70	176
kea_Latn	Kabuverdianu	725	71	175
khk_Cyrl	Halh	743	71	177
khm_Khmr	Khmer	588	69	168
kir_Cyrl	Kyrgyz	729	71	177
kor_Hang	Korean	669	61	141
lao_Laoo	Lao	591	54	132
lin_Latn	Lingala	755	59	139
lit_Latn	Lithuanian	730	71	178
ltz_Latn	Luxembourgish	703	71	176
lug_Latn	Ganda	691	70	173
luo_Latn	Luo	698	39	98
lvs_Latn	Standard Latvian	634	69	174
mal_Mlym	Malayalam	723	68	174
mar_Deva	Marathi	749	71	177
mkd_Cyrl	Macedonian	680	71	177
mlt_Latn	Maltese	731	71	176
mri_Latn	Maori	749	71	176
mya_Mymr	Burmese	746	71	175
nld_Latn	Dutch	729	58	123
nob_Latn	Norwegian Bokmål	723	51	127
npi_Deva	Nepali	754	70	175
nso_Latn	Northern Sotho	633	70	169
nya_Latn	Nyanja	720	68	169
oci_Latn	Occitan	756	71	177
ory_Orya	Odia	442	71	168
pan_Guru	Eastern Panjabi	580	56	143
pbt_Arab	Southern Pashto	701	55	144
pes_Arab	Western Persian	692	66	165
pol_Latn	Polish	723	68	165
por_Latn	Portuguese	728	70	177
ron_Latn	Romanian	734	69	177
rus_Cyrl	Russian	733	71	173
slk_Latn	Slovak	628	71	169
slv_Latn	Slovenian	704	71	174
sna_Latn	Shona	689	71	176
snd_Arab	Sindhi	749	71	177
som_Latn	Somali	746	70	177
spa_Latn	Spanish	676	71	177
srp_Cyrl	Serbian	730	63	164
swe_Latn	Swedish	686	71	168
swh_Latn	Swahili	745	65	154
tam_Taml	Tamil	693	71	169
tel_Telu	Telugu	658	66	153
tgk_Cyrl	Tajik	680	69	163
tgl_Latn	Tagalog	604	71	176
tha_Thai	Thai	710	71	176
tur_Latn	Turkish	692	67	164
ukr_Cyrl	Ukrainian	732	67	164
umb_Latn	Umbundu	473	39	108
urd_Arab	Urdu	636	65	120
uzn_Latn	Northern Uzbek	734	69	175
vie_Latn	Vietnamese	737	70	176
wol_Latn	Wolof	643	52	123
xho_Latn	Xhosa	756	71	177
yor_Latn	Yoruba	686	71	172
zho_Hans	Chinese	751	71	176
zho_Hant	Chinese	624	70	172
zsm_Latn	Standard Malay	713	67	171
zul_Latn	Zulu	739	69	175

Table 5: Number of samples by split and language in SIB-Fleurs.

A.3 Belebele-Fleurs

Language Code	Language Name	Samples
afr_Latn	Afrikaans	309
amh_Ethi	Amharic	782
arb_Arab	Modern Standard	387
asm_Beng	Assamese	824
azj_Latn	North Azerbaijani	759
ben_Beng	Bengali	855
bul_Cyrl	Bulgarian	873
cat_Latn	Catalan	652
ceb_Latn	Cebuano	783
ces_Latn	Czech	802
ckb_Arab	Central Kurdish	842
dan_Latn	Danish	696
deu_Latn	German	804
ell_Grek	Greek	837
eng_Latn	English	844
est_Latn	Estonian	736
fin_Latn	Finnish	826
fra_Latn	French	839
fuv_Latn	Nigerian Fulfulde	848
gaz_Latn	West Central Oromo	25
guj_Gujr	Gujarati	880
hau_Latn	Hausa	838
heb_Hebr	Hebrew	878
hin_Deva	Hindi	515
hrv_Latn	Croatian	896
hun_Latn	Hungarian	879
hye_Armn	Armenian	861
ibo_Latn	Igbo	838
ind_Latn	Indonesian	783
isl_Latn	Icelandic	81
ita_Latn	Italian	851
jav_Latn	Javanese	835
jpn_Jpan	Japanese	590
kan_Knda	Kannada	606
kat_Geor	Georgian	372
kaz_Cyrl	Kazakh	870
kea_Latn	Kabuverdianu	770
khk_Cyrl	Halh	869
khm_Khmr	Khmer	439
kir_Cyrl	Kyrgyz	811
kor_Hang	Korean	535
lao_Laoo	Lao	346
lin_Latn	Lingala	778
lit_Latn	Lithuanian	834
lug_Latn	Ganda	703
luo_Latn	Luo	512
lvs_Latn	Standard Latvian	555
mal_Mlym	Malayalam	809

Language	Name	Samples
mar_Deva	Marathi	869
mkd_Cyrl	Macedonian	667
mlt_Latn	Maltese	816
mri_Latn	Maori	877
mya_Mymr	Burmese	864
nld_Latn	Dutch	674
nob_Latn	Norwegian Bokmål	635
npi_Deva	Nepali	876
nso_Latn	Northern Sotho	569
nya_Latn	Nyanja	752
ory_Orya	Odia	220
pan_Guru	Eastern Panjabi	396
pbt_Arab	Southern Pashto	628
pes_Arab	Western Persian	673
pol_Latn	Polish	765
por_Latn	Portuguese	791
ron_Latn	Romanian	815
rus_Cyrl	Russian	819
slk_Latn	Slovak	513
slv_Latn	Slovenian	724
sna_Latn	Shona	735
snd_Arab	Sindhi	878
som_Latn	Somali	874
spa_Latn	Spanish	659
srp_Cyrl	Serbian	766
swe_Latn	Swedish	681
swh_Latn	Swahili	780
tam_Taml	Tamil	714
tel_Telu	Telugu	567
tgk_Cyrl	Tajik	632
tgl_Latn	Tagalog	505
tha_Thai	Thai	745
tur_Latn	Turkish	706
ukr_Cyrl	Ukrainian	773
urd_Arab	Urdu	482
uzn_Latn	Northern Uzbek	812
vie_Latn	Vietnamese	847
wol_Latn	Wolof	495
xho_Latn	Xhosa	900
yor_Latn	Yoruba	652
zho_Hans	Chinese	888
zho_Hant	Chinese	527
zsm_Latn	Standard Malay	749
zul_Latn	Zulu	838

Table 6: Number of samples per language in Belebele-Fleurs.

A.4 Analysis of Speech-To-English-Text-Translation Performance

Language	Whisper-v3-Large	SeamlessM4Tv2-Large
AVG	**13.4**	**23.6**
afr_Latn	32.9	42.5
amh_Ethi	0.8	18.5
arb_Arab	19.3	33.4
asm_Beng	2.3	21.9
ast_Latn	25.7	27.7
azj_Latn	10.7	18.3
bel_Cyrl	10.6	17.9
ben_Beng	8.1	27.2
bos_Latn	27.6	35.6
bul_Cyrl	26.4	33.2
cat_Latn	32.5	39.9
ceb_Latn	6.2	9.1
ces_Latn	24.5	33.1
ckb_Arab	1.6	23.2
cym_Latn	9.2	33.2
dan_Latn	32.6	38.9
deu_Latn	32.6	36.8
ell_Grek	20.5	27.7
est_Latn	15.3	30.4
fin_Latn	19.2	27.4
fra_Latn	34.0	36.1
fuv_Latn	0.2	0.9
gaz_Latn	0.3	0.7
gle_Latn	1.5	18.6
glg_Latn	27.9	35.6
guj_Gujr	10.7	30.2
hau_Latn	0.5	1.3
heb_Hebr	16.1	31.3
hin_Deva	19.3	28.5
hrv_Latn	24.9	31.7
hun_Latn	17.5	27.9
hye_Armn	9.0	31.2
ibo_Latn	0.6	2.6
ind_Latn	26.3	30.7
isl_Latn	7.5	25.7
ita_Latn	23.8	27.6
jav_Latn	4.1	23.6
jpn_Jpan	16.0	17.8
kam_Latn	0.8	2.5
kan_Knda	6.7	24.9
kat_Geor	2.3	21.7
kaz_Cyrl	3.8	25.0
kea_Latn	26.5	28.7
khk_Cyrl	0.8	18.5
khm_Khmr	4.5	22.3
kir_Cyrl	2.4	19.0
kor_Hang	19.0	22.7
lao_Laoo	7.1	25.9
lin_Latn	0.5	1.5
lit_Latn	12.7	24.3

Language	Whisper-v3-Large	SeamlessM4Tv2-Large
ltz_Latn	16.0	17.6
lug_Latn	0.7	18.3
luo_Latn	0.9	1.3
lvs_Latn	13.3	29.0
mal_Mlym	9.9	25.0
mar_Deva	9.9	27.2
mkd_Cyrl	26.0	35.6
mlt_Latn	11.1	40.1
mri_Latn	6.5	1.3
mya_Mymr	0.4	19.4
nld_Latn	22.1	25.8
nob_Latn	29.7	34.7
npi_Deva	10.6	9.4
nso_Latn	0.7	2.6
nya_Latn	0.9	19.3
oci_Latn	17.6	23.5
ory_Orya	4.8	26.8
pan_Guru	12.9	30.2
pbt_Arab	1.7	18.2
pes_Arab	16.0	30.5
pol_Latn	20.7	24.2
por_Latn	37.5	38.9
ron_Latn	30.6	36.5
rus_Cyrl	26.2	30.3
slk_Latn	25.1	33.6
slv_Latn	18.0	27.4
sna_Latn	1.0	3.8
snd_Arab	3.6	8.7
som_Latn	0.4	18.1
spa_Latn	22.0	24.6
srp_Cyrl	29.8	37.3
swe_Latn	34.3	38.5
swh_Latn	5.6	32.4
tam_Taml	6.0	22.5
tel_Telu	10.3	26.5
tgk_Cyrl	8.7	26.9
tgl_Latn	21.2	25.6
tha_Thai	13.4	23.5
tur_Latn	22.2	30.1
ukr_Cyrl	27.6	32.8
umb_Latn	0.2	1.1
urd_Arab	15.9	25.5
uzn_Latn	5.0	25.5
vie_Latn	19.2	26.4
wol_Latn	1.2	1.6
xho_Latn	0.8	7.6
yor_Latn	0.6	14.6
zho_Hans	14.6	22.4
zho_Hant	8.7	18.5
zsm_Latn	24.7	31.0
zul_Latn	0.6	12.4

Table 7: Per-language average sacreBLEU for Speech-to-English-Text-Translation of Fleurs utterances to their original English sentences, computed over the pooled train, dev, and test splits. Results are shown for Whisper-v3-Large and SeamlessM4Tv2-Large (cf. §4).

A.5 Analysis of ASR Performance

Language	Whisper-v3-Large	SeamlessM4Tv2-Large
afr_Latn	$13.1_{8.8}$	$12.6_{13.6}$
amh_Ethi	$207.8_{97.7}$	$23.9_{13.0}$
arb_Arab	$7.6_{7.8}$	$8.0_{11.3}$
asm_Beng	$99.0_{34.9}$	$15.1_{9.2}$
ast_Latn	$15.5_{6.8}$	$19.8_{15.6}$
azj_Latn	$7.2_{9.4}$	$8.3_{9.6}$
bel_Cyrl	$11.9_{5.6}$	$6.9_{10.8}$
ben_Beng	$33.9_{34.9}$	$9.0_{6.4}$
bos_Latn	$4.6_{5.2}$	$7.8_{11.4}$
bul_Cyrl	$5.2_{5.9}$	$9.6_{13.7}$
cat_Latn	$3.3_{4.6}$	$5.3_{9.4}$
ceb_Latn	$17.4_{47.7}$	$19.0_{11.5}$
ces_Latn	$4.2_{10.0}$	$8.0_{11.8}$
ckb_Arab	$98.3_{140.7}$	$13.9_{15.7}$
cym_Latn	$15.8_{18.6}$	$14.4_{19.6}$
dan_Latn	$4.8_{5.2}$	$10.0_{13.2}$
deu_Latn	$2.2_{3.3}$	$6.5_{9.7}$
ell_Grek	$6.4_{7.4}$	$10.8_{12.5}$
eng_Latn	$3.1_{4.9}$	$6.2_{9.1}$
est_Latn	$7.3_{8.7}$	$11.0_{15.6}$
fin_Latn	$3.2_{4.3}$	$10.2_{13.3}$
fra_Latn	$2.8_{4.1}$	$6.6_{9.9}$
fuv_Latn	$194.3_{368.0}$	$57.0_{88.7}$
gaz_Latn	$35.8_{14.7}$	$47.8_{40.9}$
gle_Latn	$85.1_{111.7}$	$21.4_{15.3}$
glg_Latn	$4.3_{4.0}$	$6.5_{10.3}$
guj_Gujr	$21.3_{17.1}$	$10.3_{8.4}$
hau_Latn	$33.8_{39.0}$	$51.8_{69.7}$
heb_Hebr	$12.2_{14.9}$	$14.5_{18.3}$
hin_Deva	$11.7_{15.3}$	$10.2_{10.3}$
hrv_Latn	$4.5_{17.2}$	$8.5_{12.5}$
hun_Latn	$5.2_{12.8}$	$8.5_{11.3}$
hye_Armn	$18.0_{32.4}$	$9.1_{10.5}$
ibo_Latn	$42.9_{36.4}$	$59.3_{52.3}$
ind_Latn	$4.0_{9.8}$	$9.2_{13.8}$
isl_Latn	$16.0_{26.3}$	$10.3_{12.9}$
ita_Latn	$2.1_{2.9}$	$5.2_{9.3}$
jav_Latn	$26.3_{52.3}$	$11.3_{11.0}$
jpn_Jpan	$6.3_{12.1}$	$16.4_{10.5}$
kam_Latn	$45.1_{84.7}$	$53.9_{43.7}$
kan_Knda	$20.1_{26.5}$	$10.9_{9.2}$
kat_Geor	$19.8_{13.9}$	$7.1_{9.7}$
kaz_Cyrl	$8.9_{8.5}$	$9.7_{12.1}$
kea_Latn	$35.8_{37.2}$	$38.1_{13.9}$
khk_Cyrl	$37.6_{36.8}$	$14.6_{22.9}$
khm_Khmr	$144.0_{61.7}$	$29.8_{11.2}$
kir_Cyrl	$28.3_{21.5}$	$8.8_{11.5}$
kor_Hang	$7.4_{9.0}$	$10.3_{11.3}$
lao_Laoo	$109.7_{45.3}$	$30.3_{11.7}$
lin_Latn	$22.0_{22.2}$	$58.7_{55.7}$
lit_Latn	$8.3_{10.6}$	$11.9_{17.8}$
ltz_Latn	$29.5_{16.0}$	$40.1_{15.0}$
lug_Latn	$44.9_{88.9}$	$14.1_{13.9}$
luo_Latn	$42.7_{110.1}$	$58.8_{59.5}$
lvs_Latn	$7.3_{20.8}$	$8.2_{13.5}$
mal_Mlym	$104.3_{71.4}$	$14.2_{14.1}$
mar_Deva	$24.3_{14.3}$	$10.7_{9.5}$
mkd_Cyrl	$5.8_{8.2}$	$8.6_{12.3}$
mlt_Latn	$26.0_{29.9}$	$12.1_{11.3}$
mri_Latn	$13.4_{14.7}$	$54.7_{54.4}$
mya_Mymr	$132.9_{78.6}$	$22.2_{9.5}$
nld_Latn	$3.1_{3.8}$	$6.8_{9.7}$
nob_Latn	$4.6_{5.7}$	$9.0_{10.2}$
npi_Deva	$25.6_{14.0}$	$74.1_{52.2}$
nso_Latn	$95.7_{162.5}$	$66.9_{63.9}$
nya_Latn	$35.6_{56.8}$	$12.4_{11.9}$
oci_Latn	$25.9_{13.9}$	$33.8_{16.5}$
ory_Orya	$93.8_{12.5}$	$11.4_{7.7}$
pan_Guru	$44.2_{49.1}$	$10.3_{8.4}$
pbt_Arab	$36.8_{13.9}$	$21.6_{25.3}$
pes_Arab	$8.6_{7.2}$	$7.7_{9.8}$
pol_Latn	$2.3_{3.8}$	$7.9_{13.8}$
por_Latn	$2.8_{4.5}$	$7.6_{10.4}$
ron_Latn	$3.2_{4.1}$	$6.6_{11.5}$
rus_Cyrl	$2.6_{4.4}$	$6.1_{10.7}$
slk_Latn	$3.9_{5.4}$	$7.0_{11.8}$
slv_Latn	$6.0_{7.4}$	$9.2_{12.2}$
sna_Latn	$26.7_{24.2}$	$32.7_{20.7}$
snd_Arab	$94.5_{36.4}$	$33.5_{17.8}$
som_Latn	$34.9_{28.7}$	$17.9_{18.8}$
spa_Latn	$2.2_{3.6}$	$6.3_{10.2}$
srp_Cyrl	$76.6_{26.5}$	$7.6_{12.0}$
swe_Latn	$3.7_{5.7}$	$9.2_{11.3}$
swh_Latn	$11.0_{22.2}$	$8.9_{12.0}$
tam_Taml	$13.3_{15.2}$	$11.6_{13.2}$
tel_Telu	$90.1_{91.9}$	$13.3_{14.3}$
tgk_Cyrl	$28.8_{37.2}$	$8.8_{9.9}$
tgl_Latn	$5.0_{10.3}$	$12.0_{10.8}$
tha_Thai	$9.8_{11.6}$	$11.4_{12.5}$
tur_Latn	$6.2_{25.6}$	$7.5_{9.0}$
ukr_Cyrl	$3.0_{4.1}$	$8.9_{12.0}$
umb_Latn	$156.7_{303.2}$	$55.5_{71.9}$
urd_Arab	$9.4_{6.9}$	$9.6_{8.0}$
uzn_Latn	$25.0_{32.4}$	$8.4_{11.4}$
vie_Latn	$4.7_{5.6}$	$6.7_{8.3}$
wol_Latn	$140.6_{260.8}$	$47.9_{52.8}$
xho_Latn	$48.1_{90.1}$	$32.3_{25.2}$
yor_Latn	$49.3_{18.2}$	$33.2_{13.2}$
zho_Hans	$16.2_{14.8}$	$19.0_{13.4}$
zho_Hant	$30.6_{41.5}$	$35.1_{16.8}$
zsm_Latn	$3.4_{4.0}$	$8.5_{10.7}$
zul_Latn	$42.6_{76.4}$	$18.4_{15.3}$

Table 8: Per-language average CER for transcribing Fleurs utterances to their original sentences, computed over the pooled train, dev, and test splits. Results are shown for Whisper-v3-Large and SeamlessM4Tv2-Large (cf. §4). Utterances in unsupported languages are transcribed using the closest manually selected supported language.

A.6 Silent Fleurs Examples

Language	Split	Count
nb_no	train	497
es_419	train	490
cy_gb	train	394
sd_in	train	307
ny_mw	train	15
ckb_iq	train	8
ny_mw	test	8
wo_sn	train	7
nso_za	test	6
ny_mw	dev	6
ur_pk	test	6
ps_af	train	4
fa_ir	train	4
so_so	train	4
ceb_ph	train	3
lg_ug	train	3
kea_cv	train	2
bg_bg	train	2
bn_in	train	2
cy_gb	test	2
ff_sn	train	2
hr_hr	train	2
hy_am	train	2
nso_za	dev	2
ur_pk	dev	2
ar_eg	train	1
da_dk	test	1
da_dk	train	1
en_us	train	1
ff_sn	dev	1
ha_ng	train	1
he_il	dev	1
he_il	test	1
he_il	train	1
ig_ng	train	1
kam_ke	train	1
kn_in	dev	1
kn_in	test	1
kn_in	train	1
mi_nz	dev	1
mn_mn	train	1
ms_my	train	1
or_in	train	1
sk_sk	dev	1
sk_sk	test	1
so_so	test	1
ta_in	train	1
te_in	test	1
te_in	train	1
umb_ao	train	1

Table 9: Number of silent examples in Fleurs by language and split.