Interspeech 2024
1-5 September 2024, Kos, Greece

Speech-MASSIVE: A Multilingual Speech Dataset for SLU and Beyond

Beomseok Lee[1,2,3], *Ioan Calapodescu*[2], *Marco Gaido*[3], *Matteo Negri*[3], *Laurent Besacier*[2]

[1]University of Trento, Italy
[2]NAVER LABS Europe, France
[3]Fondazione Bruno Kessler, Italy

beomseok.lee@unitn.it, {ioan.calapodescu,laurent.besacier}@naverlabs.com,
{mgaido,negri}@fbk.eu

Abstract

We present Speech-MASSIVE, a multilingual Spoken Language Understanding (SLU) dataset comprising the speech counterpart for a portion of the MASSIVE textual corpus. Speech-MASSIVE covers 12 languages from different families and inherits from MASSIVE the annotations for the intent prediction and slot-filling tasks. Our extension is prompted by the scarcity of massively multilingual SLU datasets and the growing need for versatile speech datasets to assess foundation models (LLMs, speech encoders) across languages and tasks. We provide a multimodal, multitask, multilingual dataset and report SLU baselines using both cascaded and end-to-end architectures in various training scenarios (zero-shot, few-shot, and full fine-tune). Furthermore, we demonstrate the suitability of Speech-MASSIVE for benchmarking other tasks such as speech transcription, language identification, and speech translation. The dataset, models, and code are publicly available at: https://github.com/hlt-mt/Speech-MASSIVE

Index Terms: spoken language understanding, speech recognition, speech resources, multi-task model

1. Introduction

Multilingual speech corpora have limited coverage of speech-related tasks, primarily focusing on automatic speech recognition (ASR) [1, 2, 3, 4] and speech translation (ST) [5, 6, 7, 8], while neglecting spoken language understanding (SLU – the task of extracting semantic information from spoken utterances, which typically involves subtasks like intent detection and slot filling). Unlike text processing, where extensive efforts in natural language understanding (NLU) have led to resources covering a wide range of languages [9, 10, 11, 12], SLU datasets are mainly English-centric [13], with few exceptions [14, 15, 16].

Our goal is to bridge the gap in multilingual SLU drawing inspiration from [16] and collecting speech recordings in multiple languages. We start with the MASSIVE NLU (i.e. textual) dataset [12], an ideal foundation due to its size, domain diversity, and broad coverage of languages, intent, and slot types. Developed by commissioning professional translators to localize the English SLURP dataset [13] into 51 languages, MASSIVE comprises 1M labeled utterances spanning 18 domains, with 60 intents and 55 slots. Our contribution, Speech-MASSIVE, spans 12 languages from diverse families: Arabic, German, Spanish, French, Hungarian, Korean, Dutch, Polish, European Portuguese, Russian, Turkish, and Vietnamese. It also facilitates evaluation across various speech tasks beyond SLU, including ASR, ST, and language identification (LID). We release Speech-MASSIVE publicly under CC-BY-SA license.[1]

Besides detailing the creation process involving a crowdsourcing-based protocol for data collection and quality control, this paper presents baseline SLU results on Speech-MASSIVE. Our results with both cascade and end-to-end architectures trained in different conditions (zero-shot, few-shot, full fine-tune) will enable future comparisons and tracking SLU advancements compared to the more mature field of NLU. Lastly, we showcase Speech-MASSIVE's versatility through additional experiments on ASR, LID, and ST.

2. Speech-MASSIVE

2.1. Speech data collection and validation process

We created the speech counterpart of textual MASSIVE data by recruiting native speakers through the Prolific crowdsourcing platform.[2] A first group of workers was instructed to record the spoken version of MASSIVE sentences with guidelines emphasizing the importance of accurate and natural reading, as well as proper recording conditions and strict adherence to the corresponding text. To ensure high final data quality, a second group of native speakers validated the recorded utterances. During validation, participants were directed to read the original text, listen to the recording, and label it as *valid* or *invalid*. Those marked as invalid underwent a second iteration of this two-step (recording and validation) process. After the second iteration, the process concluded, irrespective of the outcome of the second validation phase, to avoid potentially endless cycles. This decision was also informed by the observation that, upon inspecting the invalid recordings, we found some were marked as such not due to a lack of adherence of the speech to the text but because of grammatical errors in the original MASSIVE dataset text. Correcting these errors was beyond the scope of our work.

To further enhance the reliability of the collected dataset, we implemented two additional precautions. During the recording phase, we instructed participants to review their own recordings before proceeding to the next sample, allowing them to re-record if the audio was not properly acquired. Additionally, in the validation step, four speech utterances were chosen from Common Voice [1] and inserted among the samples for validation. Out of these four quality control samples, two intentionally featured audio-transcript mismatches to be marked as invalid. The other two cases had perfect audio-transcript alignment to be marked as valid. Care was taken to select quality control samples with clear and intelligible audio. Validation results from a Prolific user were retained only if they accurately assessed all four quality control samples. Any mistakes led to the disregarding of their validations, requiring the entire set of samples from that user to be re-validated by other participants.

[1] https://hf.co/datasets/FBK-MT/Speech-MASSIVE

[2] https://www.prolific.com, Compensated £9 per hour.

10.21437/Interspeech.2024-957

Table 1: *Speech-MASSIVE's overall statistics. '# hrs' displays the recording duration for all samples (including invalid), while '# spk (Male/Female/Unknown)' indicates the number of speakers for all the samples (including invalid). The last 2 columns ('WER', and 'CER') measures Whisper ASR performance.*

lang	split	# sample	# valid	# hrs	total # spk (M/F/U)	WER	CER
ar	dev	2033	2027	2.12	36 (22/14/0)	31.75	14.43
	test	2974	2962	3.23	37 (15/17/5)	34.19	15.85
de	train-full	11514	11201	12.61	117 (50/63/4)	-	-
	dev	2033	2032	2.33	68 (35/32/1)	11.24	3.96
	test	2974	2969	3.41	82 (36/36/10)	11.84	4.16
es	dev	2033	2024	2.53	109 (51/53/5)	7.61	3.00
	test	2974	2948	3.61	85 (37/33/15)	8.95	3.76
fr	train-full	11514	11481	12.42	103 (50/52/1)	-	-
	dev	2033	2031	2.20	55 (26/26/3)	10.20	4.42
	test	2974	2972	2.65	75 (31/35/9)	11.09	4.71
hu	dev	2033	2019	2.27	69 (33/33/3)	25.96	10.93
	test	2974	2932	3.30	55 (25/24/6)	20.98	6.01
ko	dev	2033	2032	2.12	21 (8/13/0)	25.29	7.13
	test	2974	2970	2.66	31 (10/18/3)	26.42	8.04
nl	dev	2033	2032	2.14	37 (17/19/1)	11.03	3.98
	test	2974	2959	3.30	100 (48/49/3)	10.52	3.82
pl	dev	2033	2024	2.24	105 (50/52/3)	9.94	4.88
	test	2974	2933	3.21	151 (73/71/7)	12.58	6.22
pt	dev	2033	2031	2.20	107 (51/53/3)	11.73	5.10
	test	2974	2967	3.25	102 (48/50/4)	12.11	5.13
ru	dev	2033	2032	2.25	40 (7/31/2)	8.55	4.06
	test	2974	2969	3.44	51 (25/23/3)	8.99	4.57
tr	dev	2033	2030	2.17	71 (36/34/1)	16.65	4.56
	test	2974	2950	3.00	42 (17/18/7)	18.06	5.05
vi	dev	2033	1978	2.10	28 (13/14/1)	16.65	10.5
	test	2974	2954	3.23	30 (11/14/5)	14.94	9.77

2.2. Overall statistics

We chose 12 languages based on various criteria. Initially, we considered the number of registered users on Prolific, sorting the 51 languages covered in MASSIVE. Languages with fewer than 200 users were excluded to ensure sufficient worker participation to complete the entire acquisition and validation process in reasonable time. Italian was also excluded due to the availability of the full dataset elsewhere [16]. Finally, with an eye at the balance between budget considerations and linguistic diversity, from the remaining 18 languages we selected Arabic, German, Spanish, French, Hungarian, Korean, Dutch, Polish, European Portuguese, Russian, Turkish, and Vietnamese.

We collected speech recordings for MASSIVE's development and test splits. Acquiring the full training dataset (11,514 utterances for each of the 12 languages) exceeded our budget. In a concession, our emphasis was placed on acquiring comprehensive training data for French and German, while we obtained limited few-shot training data consisting of 115 utterances from the training set for the remaining 10 languages (*train-115* split). Columns 1-6 of Table 1 provide statistics for the collected dataset, including, for each language, the available data splits, the number of recordings, hours of speech, and speakers (total, male, female and unknown). The "# valid" column indicates the count of human-validated utterances for each data split after the two iterations. As a few speech recordings remained invalidated after our two recording-validation cycles, we retained for each utterance the candidate with the lowest Word Error Rate (WER) as transcribed using Whisper [17]. This ensures speech availability for all MASSIVE utterances, even if some may not perfectly align with the reference transcript. Additional information regarding this is included in the corpus metadata.

2.3. ASR assessment

To assess Speech-MASSIVE in multilingual ASR, we used Whisper, since it is one of the recent state-of-the-art multilingual speech recognition models. We selected Whisper-large-v3,[3] utilizing it without additional fine-tuning for our ASR evaluation. Table 1 shows WER and character error rate (CER) across languages and data splits. We compared ASR error rates to those obtained on the FLEURS dataset [2].[4] FLEURS generally yields lower WERs/CERs compared to Speech-MASSIVE. The same observation was made for Italian in [16], which followed a recording methodology similar to ours. This suggests that the higher WERs are likely due to the inherent difficulty of MASSIVE utterances compared to those in FLEURS. Furthermore, there are still discrepancies between our Whisper model's hypotheses and the references in the MASSIVE dataset (e.g., numbers reported in letters in MASSIVE references), which we did not address as optimizing ASR WER was not our main goal. Finally, we calculated the correlation coefficient between WERs (CER for Korean) on Speech-MASSIVE and FLEURS, resulting in a value of 0.96. This shows that Whisper consistently performs across both datasets, despite Speech-MASSIVE being more challenging than FLEURS for ASR.

3. SLU Baselines and Beyond

In this section, we establish several SLU baselines, evaluating them with different training conditions and metrics described in §3.1. Firstly (§3.2), we build a NLU model, serving as an upper bound free from ASR errors. Secondly, we build a cascaded SLU system (§3.3), in which an ASR component transcribes input audio and the NLU model utilizes ASR output for inference. Thirdly, to complete the inventory of SLU baselines, we introduce an end-to-end (E2E) model (§3.4). We conclude by showcasing the versatility of Speech-MASSIVE beyond SLU, computing additional baselines for tasks such as speech translation and language identification (§3.5).

3.1. NLU/SLU training conditions and metrics

To simulate different training resource scenarios, we report performance in three different settings: *(a) Zero-shot:* we train the model only with one language data from the train split (11,514 utterances) and evaluate in all other different languages; *(b) Few-shot:* we employ subsets (115 examples) for each of the 12 non-English languages, aligning with our train-115 split.[5] Additionally, we integrate the full zero-shot training split to enrich the multilingual training dataset, totaling 12.8k samples for training; *(c) Full fine-tune (NLU only):* 11,514 training examples of all 12 languages are pooled (138k samples for training).

We assess intent prediction in a given text or speech with *intent accuracy*.[6] We report the average result (and standard deviation) of three runs with different seeds. All experiments were executed on 1 A100 80GB GPU.

[3] `https://hf.co/openai/whisper-large-v3`

[4] Accessible for our 12 languages except Arabic at `https://github.com/openai/whisper/discussions/1762`

[5] train-115 covers all 18 domains, 60 intents, and 55 slots (including empty slots).

[6] Due to space limitations, we report only intent accuracy scores. However, additional SLU metrics (e.g., micro-averaged slot F1, exact match accuracy, slot-type F1, slot-value CER) exhibit a similar trend and are available in the GitHub repository.

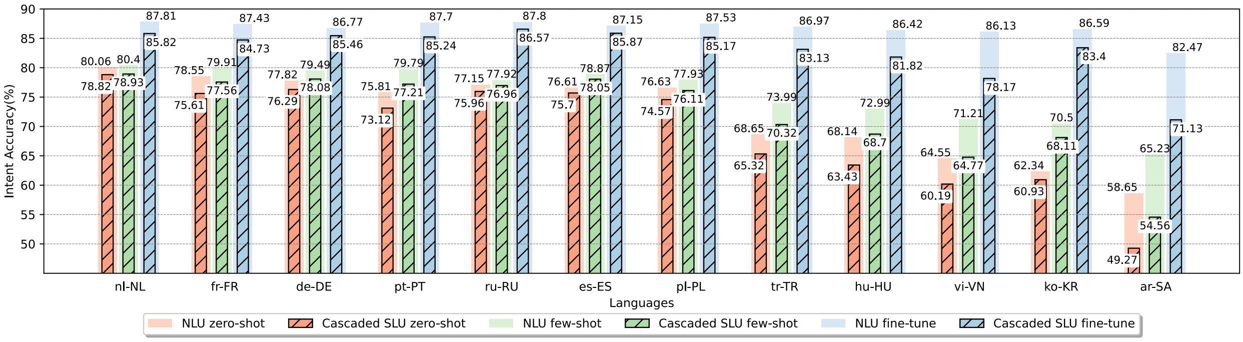

Figure 1: *NLU vs Cascaded SLU (Intent Accuracy) on our Speech-MASSIVE Dataset.*

```
[Original text in MASSIVE]
En) where can i go tonight
En-Annot) where can i go [timeofday :  tonight]
En-Slots) Other Other Other Other timeofday
Fr) où puis-je aller ce soir
Fr-Annot) où puis-je aller [timeofday:ce soir]
Fr-Slots) Other Other Other timeofday timeofday
Intent) recommendation_events
[NLU]
Fr-Src) Annotate:  où puis-je aller ce soir
Fr-Tgt) Other Other Other timeofday timeofday
recommendation_events
[Cascaded SLU]
Fr-ASR) où puis je aller ce soir
Fr-Src) Annotate:  où puis je aller ce soir
Fr-Tgt) Other Other Other timeofday timeofday
recommendation_events
[E2E SLU]
Fr-Tgt) où puis-je aller ce soir | Other Other
Other timeofday timeofday | recommendation_events
```

Figure 2: *Input/Output formatting across NLU/SLU tasks. En: original English text. Fr: French translation in MASSIVE. Annot, Slots and Intent: slot and intent annotation of MASSIVE.*

3.2. NLU model

Our NLU system uses the mT5 encoder-decoder architecture [18], selected for its superior performance as demonstrated in [12], where the mT5 *text-to-text* model outperformed both the mT5 encoder-only model and the XLM-R model [19]. We use a pre-trained mT5-base model,[7] and fine-tune both the encoder and decoder in a sequence-to-sequence manner. We supply source and target texts as described in [12] and shown in Figure 2. For instance, the French sentence (*Fr*) *"où puis-je aller ce soir"* is annotated in slots (*Fr-Slots*) as *'Other Other Other timeofday timeofday"* and intent (*Intent*) as *"recommendation_events"* in MASSIVE. We adapt those annotations to create source and target texts to be used in training: for the source text (*Fr-Src in [NLU]*), we prepend *"Annotate:"* to the French sentence (*Fr*); for the target text (*Fr-Tgt in [NLU]*), we concatenate slots (*Fr-Slots*) and intent (*Intent*).

Figure 1 displays the intent accuracy results of our NLU system across all languages and modes (zero-shot, few-shot, full fine-tune), along with those of the cascaded SLU models discussed in §3.3. Unsurprisingly, NLU performance increases when moving from zero-shot to full fine-tune regimes. Also, as expected, higher scores are observed for languages (*Nl*, *Fr*, *De*, *Pt*, *Ru*, *Es* and *Pl*) that are better represented in the mC4 multilingual dataset used to train mT5 model [18]. Finally, the highest results align with those reported in the MASSIVE paper [12], serving as a suitable reference upper bound for comparisons with the SLU models discussed in the following section.

3.3. Cascaded SLU model

We develop a cascaded SLU system in which an ASR model based on Whisper-large-v3 transcribes the speech, and the same NLU models of §3.2 (zero-shot, few-shot, full fine-tune) predict slots and intent from the transcribed texts.

The SLU intent accuracy scores in Figure 1 reveal that processing automatically transcribed utterances introduces performance drops of varying magnitude across the different languages and training modes. This is especially notable for languages with lower ASR quality (i.e., higher WER), such as *Ar*, *Hu*, *Ko*, *Tr*, and *Vn*. This supports our expectations about the difficulty for the downstream textual NLU component of the SLU cascade to handle unrecoverable transcription errors. As a matter of fact, in zero-shot mode, the distance with the text-only upper-bound NLU system is considerably smaller for languages featuring higher ASR quality. Similar to what we observed for NLU (§3.2), cascaded SLU performance in few-shot mode improves thanks to the additional multilingual data. The gains are particularly significant for languages with lesser representation in mT5 model, such as *Tr*, *Vn*, *Ko*, and *Ar*. Lastly in full fine-tune mode, leveraging a larger multilingual training dataset leads to substantial performance enhancements. While the gains are variable, we observe that: *i)* for some languages (i.e. *De*, *Ru*, and *Es*), the gap with the highest results of the textual NLU upper bound shrinks to less than two points, while *ii)* for all languages, the scores are significantly higher than those achieved by the textual NLU models dealing with clean input not only in zero-shot, but also in few-shot mode.

3.4. E2E SLU model

To complete the inventory of SLU baselines for comparison, we introduce an end-to-end (E2E) SLU model: a direct solution that bypasses intermediate text representations (ASR transcripts). We utilize Whisper, following the approach proposed in [20], which showed superior performance compared to cascaded systems and other speech encoders like wav2vec2.0 [21] and HuBERT [22]. Model training follows a sequence-to-sequence approach, with predictions extended to include transcript, slots, and intent. This allows us to leverage both speech and text information in the model's predictions. We intro-

[7] `https://huggingface.co/google/mt5-base`

Table 2: *LID accuracy and ST BLEU results with Whisper-large-v3 on Speech-MASSIVE.*

lang	ar		de		es		fr		hu		ko		nl		pl		pt		ru		tr		vi	
split	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
LID accuracy	90.9	89.5	98.9	98.4	99.0	98.6	98.7	98.9	94.6	95.8	99.1	98.7	94.8	94.9	95.3	94.6	95.9	96.0	99.1	98.8	96.1	96.0	90.7	93.2
ST BLEU	17.2	16.6	36.7	38.2	38.5	38.2	38.7	40.1	19.4	20.6	19.7	19.5	40.0	38.9	29.9	28.8	32.4	32.3	28.4	28.2	26.7	26.0	18.9	20.2

duce an additional separator ``|'' between the tasks, allowing Whisper's tokenizer to tokenize the target text as is, without the need to add slots or intents to the original vocabulary. Two specific tokens, ``|'' and ``_'', are removed from Whisper's suppressed token list, as they are required for predicting SLU outputs as task separators and in certain intent values. In zero-shot mode, we fine-tune Whisper-large-v3 with either a) the English train set of [13], or b) the French train set of Speech-MASSIVE. These two conditions (*En vs Fr*) allow us to investigate the impact of the training language on zero-shot E2E SLU across all other languages. Additionally, in few-shot mode, we fine-tune Whisper-large-v3 with the English or French train sets, along with train-115 splits from other languages. We do not provide a full fine-tune E2E SLU mode since only two languages in Speech-MASSIVE are supported by full train splits.

Table 3 compares cascaded and E2E SLU performance in both zero-shot and few-shot modes. It is worth noting that the comparison between the two approaches is fair only when using the English train set (*En*), since they utilize the same training utterances albeit in different modalities (written form for cascade and spoken form for E2E). In this condition (*En*), for zero-shot mode, cascaded SLU outperforms E2E SLU for all languages. In few-shot mode, we note a different trend, with cascaded and E2E models exhibiting similar average performance. Employing the French training set from Speech-MASSIVE (*Fr*), E2E SLU surpasses models trained on the English dataset from [13] (*En*) in both zero-shot and few-shot modes. In zero-shot mode, we observe improvements of more than 5 points for 9 out of 11 languages. In few-shot mode, although the influence of the training language (*En vs Fr*) diminishes due to multilingual training, using French as the majority language still yields better performance than using English. These results highlight the significant influence of the 'training language' on the performance of E2E SLU models in zero/few-shot settings. Speech-MASSIVE provides a unique opportunity to explore this intriguing observation further. Finally, examining French (*Fr*) results representing the full fine-tune mode for this language, E2E SLU achieves intent accuracy of 85.87%, compared to 84.73% for cascaded SLU and 87.43% for NLU given in Fig.1.

3.5. Other baselines

We conclude our experiments using Whisper-large-v3 without any finetuning to compute other baselines and demonstrate the versatility of Speech-MASSIVE. We perform Language Identification (LID) and Speech Translation (ST) across $x \rightarrow$en language directions. Different types of tokens are fed to Whisper's decoder depending on the tasks as shown in Figure 3. Table 2 reports Whisper-large-v3 model's LID accuracy and ST BLEU [23] on Speech-MASSIVE. LID is calculated over all the samples in dev and test splits. For ST, instead, BLEU is computed on subsets of dev and test splits identified using meta information from MASSIVE to exclude samples with *localized* translation. This filtering is necessary to ensure an accurate assessment of translation quality, as localized references may introduce discrepancies in word choice (see §1). Besides indicating the versatility of Speech-MASSIVE for evaluation purposes, our additional baselines on speech-related tasks offer valuable reference scores for cross-task comparisons and for exploring collaborative solutions to leverage potential mutual benefits.

```
ASR
[<|startofstranscript|>, <|language_id|>,
<|transcribe|>, <|notimestamps|>]
E2E SLU
[<|startofstranscript|>, <|language_id|>,
<|transcribe|>, <|startoflm|>, <|notimestamps|>]
LID
[<|startofstranscript|>]
ST
[<|startofstranscript|>, <|language_id|>,
<|translate|>, <|notimestamps|>]
```

Figure 3: *Various task control tokens fed to Whisper's decoder.*

Table 3: *Intent accuracy of cascaded and E2E SLU. Both E2E SLU zero-shot and few-shot models are trained either with initial English train set of [13] (En) or with French train set of Speech-MASSIVE (Fr). We exclude French (*) from the average as fr-FR scores are no longer zero/few-shot when French is used as the training language.*

lang	Casc. (En) **zero-shot**	E2E (En) **zero-shot**	E2E (Fr) **zero-shot**	Casc. (En) **few-shot**	E2E (En) **few-shot**	E2E (Fr) **few-shot**
ar	49.27 ± 0.90	33.04 ± 4.74	40.00 ± 2.44	54.56 ± 0.73	57.71 ± 1.46	61.22 ± 1.74
de	76.29 ± 0.14	70.68 ± 1.37	73.91 ± 0.73	78.08 ± 0.50	78.64 ± 0.65	78.45 ± 0.64
es	75.70 ± 0.19	73.12 ± 0.75	78.62 ± 0.41	78.05 ± 0.33	79.79 ± 0.66	80.59 ± 0.31
fr	75.61 ± 0.48	68.43 ± 2.30	*85.87 ± 0.26**	77.56 ± 0.13	77.11 ± 0.77	*85.93 ± 0.35**
hu	63.43 ± 0.92	36.62 ± 1.49	42.28 ± 2.20	68.70 ± 0.80	60.75 ± 2.40	63.93 ± 0.19
ko	60.93 ± 0.84	57.96 ± 2.26	66.09 ± 1.86	68.11 ± 0.04	72.82 ± 0.23	74.09 ± 0.73
nl	78.82 ± 0.45	65.17 ± 0.57	67.24 ± 1.44	78.93 ± 0.34	77.49 ± 0.77	77.37 ± 0.47
pl	74.57 ± 0.37	64.82 ± 1.51	64.38 ± 1.29	76.11 ± 0.39	74.85 ± 0.58	76.88 ± 1.37
pt	73.12 ± 0.49	62.91 ± 1.97	72.60 ± 1.01	77.21 ± 0.65	78.15 ± 1.16	80.02 ± 0.29
ru	75.96 ± 0.19	69.06 ± 1.71	74.75 ± 0.28	76.96 ± 0.08	79.22 ± 0.67	79.51 ± 0.26
tr	65.32 ± 0.61	47.60 ± 3.08	55.08 ± 1.09	70.32 ± 0.48	69.44 ± 1.62	71.14 ± 1.15
vi	60.19 ± 0.39	35.44 ± 1.48	49.67 ± 2.30	64.77 ± 0.98	63.36 ± 1.69	68.71 ± 0.33
avg.	**69.10 ± 0.19**	**57.07 ± 1.82**	**62.24 ± 0.92**	**72.45 ± 0.32**	**72.45 ± 0.53**	**73.81 ± 0.58**

4. Conclusion

We introduced Speech-MASSIVE, a multilingual SLU dataset spanning 12 languages for intent prediction and slot-filling tasks. Alongside dataset creation, we established baselines for SLU across various resource and architecture configurations. Additionally, we showcased Speech-MASSIVE's versatility beyond SLU, extending to tasks such as ASR, LID, and ST. With its diverse array of native speakers and recording environments, Speech-MASSIVE holds promise as a benchmark for multilingual, multimodal, and multi-task speech research. Future research opportunities include exploring further the influence of training languages on zero/few-shot SLU performance, thoroughly comparing cascade and E2E SLU solutions, assess the effect of including multi-task and multilingual corpora in the training of speech foundation models, and pushing the boundaries of E2E multi-task speech systems beyond our baselines.

Acknowledgements. The speech collection was funded by EU Horizon Europe (HE) Research and Innovation programme grant No 101070631. We also acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

5. References

[1] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4218–4222.

[2] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 798–805.

[3] M. Zanon Boito, W. Havard, M. Garnerin, É. Le Ferrand, and L. Besacier, "MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 6486–6493.

[4] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," in *Interspeech 2020*. ISCA, Oct. 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2826

[5] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2012–2017.

[6] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and M. Post, "The Multilingual TEDx Corpus for Speech Recognition and Translation," in *Proc. Interspeech 2021*, 2021, pp. 3655–3659.

[7] J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8229–8233.

[8] C. Wang, A. Wu, and J. M. Pino, "Covost 2: A massively multilingual speech-to-text translation corpus," *CoRR*, vol. abs/2007.10310, 2020. [Online]. Available: https://arxiv.org/abs/2007.10310

[9] P. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk, "MLQA: Evaluating cross-lingual extractive question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7315–7330.

[10] N. Moghe, E. Razumovskaia, L. Guillou, I. Vulić, A. Korhonen, and A. Birch, "Multi3NLU++: A multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue," in *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 3732–3755.

[11] W. Xu, B. Haider, and S. Mansour, "End-to-end slot alignment and recognition for cross-lingual NLU," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 5052–5063.

[12] J. FitzGerald, C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh, S. Ranganath, L. Crist, M. Britan, W. Leeuwis, G. Tur, and P. Natarajan, "MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4277–4302.

[13] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "SLURP: A spoken language understanding resource package," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7252–7262.

[14] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet, and J. Dureau, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *CoRR*, vol. abs/1805.10190, 2018. [Online]. Available: http://arxiv.org/abs/1805.10190

[15] F. Lefèvre, D. Mostefa, L. Besacier, Y. Estève, M. Quignard, N. Camelin, B. Favre, B. Jabaian, and L. M. Rojas-Barahona, "Leveraging study of robustness and portability of spoken language understanding systems across languages and domains: the PORTMEDIA corpora," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*. European Language Resources Association (ELRA), 2012, pp. 1436–1442.

[16] A. Koudounas, M. La Quatra, L. Vaiani, L. Colomba, G. Attanasio, E. Pastor, L. Cagliero, and E. Baralis, "ITALIC: An Italian Intent Classification Dataset," in *Proc. INTERSPEECH 2023*, 2023, pp. 2153–2157.

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[18] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 483–498. [Online]. Available: https://aclanthology.org/2021.naacl-main.41

[19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747

[20] M. Wang, Y. Li, J. Guo, X. Qiao, Z. Li, H. Shang, D. Wei, S. Tao, M. Zhang, and H. Yang, "Whislu: End-to-end spoken language understanding with whisper," in *Proc. Interspeech*, vol. 2023, 2023, pp. 770–774.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.