Spoken MASSIVE: A Multilingual Spoken Language Understanding Dataset

Chutong Meng **Milton Lin**

Abstract

Multilingual natural language understanding (NLU) is an important task for assisting users speaking different languages. However, in practice, many virtual assistants take speech rather than text as input. Yet there is currently no publicly available dataset with multilingual spoken commands, making it hard to develop and evaluate multilingual spoken language understanding (SLU) models. We therefore present Spoken MASSIVE, the first multilingual SLU dataset, by synthesizing speech using the text in MASSIVE, the largest multilingual NLU dataset so far. We also train SLU models on this dataset and share our findings.

1 Introduction

Spoken language understanding (SLU) is the foundation of voice-based virtual assistants like Alexa, Siri, and Google Assistant, which currently support only a small fraction of the world's 7,000+ languages (FitzGerald et al., 2023a). A typical SLU task is intent classification, which takes an utterance as input, such as `"does dominoes do takeaway"`, and outputs its intent, `"takeaway_query"`. Another task is slot filling, generating an annotated utterance `"does [food_type : dominoes] do [order_type : takeaway]"`, where two slots `food_type` and `order_type` are filled with `dominoes` and `takeaway` respectively.

One of the challenges the SLU community faces is that automatic speech recognition (ASR) and natural language understanding (NLU) communities are often disconnected (Faruqui and Hakkani-Tür, 2021). The ASR model transcribes speech into text, and then the NLU model predicts the intent or slots. And the ASR and NLU models are developed separately.

They have the following disadvantages:

- Cascaded models have a longer inference time, which is not desirable in an interactive context.
- ASR transcriptions often contain errors, especially for low-resource languages, which will propagate to the NLU module.
- The rich information of speech (e.g., tempo, pitch, and intonation) is lost after ASR.

End-to-End (E2E) SLU models have the potential to alleviate these issues (Lugosch et al., 2019a; Serdyuk et al., 2018; Lugosch et al., 2019b; Denisov and Vu, 2023). They aim to take speech as input and complete intent classification and slot filling without relying on intermediate text. However, most of current works are based on English-only SLU dataset. To the best of our knowledge, there are no works on multilingual SLU and there are no publicly available multilingual SLU dataset.

In this work, we propose two methods to attempt to create the first multilingual SLU dataset.

1. Utilize publicly available multilingual ASR datasets and automatically label some audio samples based on transcriptions.
2. Synthesize multilingual speech for text in multilingual NLU datasets.

Our main contributions are:

- This is the first multilingual SLU dataset. See Table 1 for English speech statistics.
- We provide some baseline methods and results based on this dataset.

2 Related works

NLU dataset. MASSIVE (FitzGerald et al., 2023b) is the largest multilingual NLU dataset. It contains 1M realistic, parallel, labeled virtual assistant utterances spanning 51 languages, 18 domains,

Subset	Speakers per sample	Duration (hours)	#Samples
Train	random 4 out of (10M+10F)	31.25	46,056
Dev	2M+2F	5.41	8,132
Test	2M+2F	8.09	1,1896

Table 1: Statistics for English. M refers to male speaker and F refers to female speaker.

60 intents, and 55 slots. It can be used to accomplish 3 tasks: slot filling, intent classification, and virtual assistant evaluation. Its first iteration was NLU evaluation Benchmark dataset, (Liu et al., 2019), which was subsequently updated as SLURP(Bastianelli et al., 2020).

SLU dataset. There are a few monolingual SLU datasets available. S2IDataset (Rajaa et al., 2022) consists of Indian accented English speech corpus with 14 coarse-grained intents from Banking domain. Timers and Such (Lugosch et al., 2021) is an open source dataset of spoken English commands, with an emphasis on numbers. SLURP (Bastianelli et al., 2020) contains 58h of English audios and spans 18 domains. CATSLU (Zhu et al., 2019) provides the first Chinese SLU dataset. Spoken SQuAD (Li et al., 2018) is created by converting the passage part of SQuAD (Rajpurkar et al., 2016) dataset into speech by using Google Text-to-Speech. MEDIA (Bonneau-Maynard et al., 2006) is a French SLU dataset with 41.5h of training data. But there is currently *no* multilingual SLU dataset.

Multilingual ASR dataset. There are multilingual ASR datasets available. VoxPopuli (Wang et al., 2021) is a large collection of parallel European Parliament plenary session recordings in 23 European Union languages. Common Voice (Ardila et al., 2020) is a crowdsourcing dataset covering 38 languages. MMS (Pratap et al., 2023) provides New Testament recordings in 1107 languages.

Multilingual Text-to-Speech (TTS). MMS (Pratap et al., 2023) also built TTS models that support 1107 languages, but it has only a single speaker. XTTS[1] is a multilingual TTS model. It supports multiple speakers in 17 languages and has a voice cloning ability. VCTK[2] dataset is usually used in multi-speaker TTS and voice cloning. It consists of 110 English speakers with different accents.

Synthetic speech generation. Previous works have used TTS models to improve speech recognition training. Some works focus on creating more training data when the amount of real data is limited, or out of distribution word recognition.

[1] https://coqui.ai/blog/tts/open_xtts

[2] https://datashare.ed.ac.uk/handle/10283/3443

3 Methodologies

Below we outline two methods we have tried so far. In section 3.1, we talk about how we tried to annotate real speech data based on their transcriptions, or "mining" from real speech. In section 3.2, we talk about how we synthesize speech using the text in MASSIVE (FitzGerald et al., 2023a).

3.1 Mining real speech from multilingual ASR datasets

3.1.1 Slot filling

We extract all the slots and corresponding phrases from MASSIVE. Some phrases can have multiple possible slots, we discard these phrases as we cannot be sure which slot to use in different contexts. Then, we search for the phrases from transcriptions of multilingual ASR datasets. If any phrases are matched, we can annotate them with the same slots. We have applied this approach to the English subset of VoxPopuli and Common Voice, because it is easier for us to inspect the results.

However, by manual inspections, the result can be really misleading. A lot of the phrases in MASSIVE require more context to determine their slot types. Simply using the slots themselves as matching criteria will lead to many wrong annotations. A lot of work is needed to design hand-crafted rules for filtering the annotated data. So, we give up on this approach.

3.1.2 Intent detection

We choose CommonVoice (Ardila et al., 2020) as raw data for mining transcriptions that are semantically close to MASSIVE sentences. Considering computing resources, we download around 1000 hours of CommonVoice speech for English, Chinese, and German. For Vietnamese, we download the largest version which has 19 hours. We then apply LASER2 sentence embeddings (Heffernan et al., 2022) to embed both MASSIVE sentences and ASR transcriptions. Note that, in contrast to the normal usage that source and target sentences are in different languages, here they are in the same language, as we do not need them to be translations. To measure the semantic similarity, we compute the margin-based cosine similarities (Artetxe and Schwenk, 2019) between sentence embeddings, as in Equation 1.

$$\text{margin-score}(x, y) = \frac{\cos(x, y)}{\sum_{z \in NN_k(x)} \frac{\cos(x,z)}{2k} + \sum_{z \in NN_k(y)} \frac{\cos(y,z)}{2k}} \quad (1)$$

where $NN_k(x)$ denotes the k nearest neighbors of x in the other side. In our experiments, we use $k = 16$. A higher margin-score means the source sentence x and the target sentence y are semantically closer.

Text mining is implemented based on Stopes[3] (Andrews et al., 2022). We show some mined examples in Table 2. We decide to use these mined samples for intent detection.

margin score	MASSIVE sentence	CommonVoice transcript
1.5656	hey i missed you	I missed you.
1.4087	good evening	Good evening.
1.3398	i can't hear you	I couldn't hear you.
1.2819	please mute the sound	Please mute the sound.
1.2110	what is the definition for this object	What is the dimension of this object?

Table 2: Some mined examples between MASSIVE sentences and CommonVoice transcriptions.

3.2 Generating synthesized speech

The advancement of TTS technologies, particularly from projects like MMS (Pratap et al., 2023), enables the creation of highly diverse and multilingual synthetic speech datasets. We first tried to apply TTS models from MMS to synthesize the text in MASSIVE dataset. The benefit of using MMS models is that they support more than 1000 languages, enabling them to be easily applied to other multilingual NLU datasets. However, the drawback is that they are single-speaker models, which are not suitable for training applicable SLU models.

We then chose to use XTTS model. It is a multilingual TTS model supporting 17 languages. More importantly, it is a multi-speaker model and also supports voice cloning. We chose to use its voice cloning functionality instead of multi-speaker functionality, because voice cloning is easier to support much more different speakers as long as we can find audios from different speakers. We use a multi-speaker English speech dataset VCTK to provide sources of speakers. It has 110 speakers in total. We carefully split the speakers for train, dev, and test sets so that there are no overlaps between them in order to measure the model performance fairly. For training set, we ramdomly choose 20 speakers (10 females + 10 males). For each MASSIVE training sample, we randomly select 4 out of 20 speakers to synthesize the sentence. For dev and test set, we select 4 different speakers (2 females + 2 males) and use them all to synthesize the MASSIVE sentences. The statistics for English subset can be found in Table 1. Other languages have similar statistics.

[3] https://github.com/facebookresearch/stopes

4 Experiments

We will first evaluate the quality of synthesized speech. Broadly, there are four approaches (Lu et al., 2023):

1. Human evaluation: this is the most common strategy in speech synthesis. Each human evaluator rates the synthesized speech against real speech in a blind hanner. However, this is error-prone and not scalable (Donahue et al., 2018).

2. Statistical difference evaluation. This is given by evaluating statistical metrics, common in field of health records (Yan et al., 2022).

3. Training on synthetic and testing on real dataset. We were only able to do this in the setting of English dataset.

4. Evaluation using pre-trained models. In our case, we can evaluate the quality of generated speech using currently available ASR models. A low WER would indicate good quality.

We focus on the third approach. Then, we will train SLU models on synthesized and real speech. We will focus on English, Chinese, Spanish, and German.

4.1 Experiment Setup

We follow the experiment setup in Rajaa et al. (2022). We use XLSR (Babu et al., 2021) as the encoder model, and add a linear classifier on top of the representations. We freeze the convolutional layers in XLSR. All models are trained for 20 epochs, and the checkpoint is chosen based on validation accuracy. We use a learning rate of 1e-5 and a batch size of 16 audios.

4.2 Real and synthetic data ratio

In order to quantify the impact of synthetic speech data on model performance, we conduct an ablation study, as proposed by Hu et al. (2021). The ablation study methodology involves:

1. Training baseline models on the original dataset without any synthetic data for baseline.

2. Incrementally increasing the ratio of synthetic to real data in the training set and observing the variations in model performance.

3. Utilizing different sampling algorithms for synthetic data generation and integration.

We use the dev and test sets from SLURP. For training set, we always use all SLURP training set, and we randomly sample a fraction of synthesized MASSIVE speech.

The results are shown in Table 3. The model performs the best when we use all synthetic data, demonstrating its usefulness. Note that SLURP and MASSIVE English subset are not exactly the same. So, it could be the diversity of `(text, intent)` helps, but not the synthetic audios themselves.

Train Set	Test Acc	Test F1
real + 100% synthetic	**0.7793**	**0.7827**
real + 80% synthetic	0.7701	0.7735
real + 60% synthetic	0.7665	0.7692
real + 40% synthetic	0.7592	0.7620
real + 20% synthetic	0.7517	0.7584
real + 0% synthetic	0.7489	0.7538

Table 3: Adding synthetic data on real SLURP training set in English. **Bold** results mean the best.

4.3 Performance across various languages

If we take English performance as a baseline, the performance in other languages is comparably good. These models are trained and tested on **synthetic** data, whilst in subsection 4.2, they were tested on English SLURP test data. Though, the accuracy has improved, which is *not* so significant but still suggests that these synthetic datasets are useful benchmarks. As in Table 4, Engish performs the best while Chinese per forms the worsest.

Language	Accuracy	F-1 Score
English	0.8198	0.8223
Spanish	0.8004	0.8005
German	0.7967	0.7968
Chinese	0.7540	0.7561

Table 4: Performance of models on the synthetic MASSIVE test set across different languages.

There are a few reasons where Chinese performed poorly: this could both due to lack of Chinese dataset representation and the XLSR model. There are only 200 hours Chinese pretraining data (including HK, TW). For English, German and Spanish have 70, and 20k hours, respectively.

4.4 Effect of using mined data

We also tried to add the mined speech into the training data. We choose the training samples from the ones with highest margin-score defined in Eq. 1. It is interesting to observe that our initial experiments show that mined data actually decrease performance. The results are shown in Table 5.

It seems like adding mined data does not improve the baseline. By adding training samples with margin-scores ≥ 1.20 lead to comparable result to baseline. Adding more training samples reduce the model performance. We think it is because (1) the mined samples are highly imbalanced; and (2) the mined quality is not so good.

Training Set	Accuracy	F-1 Score
Synthetic MASSIVE	0.6757	**0.6880**
+ margin-score ≥ 1.20	**0.6778**	0.6836
+ margin-score ≥ 1.15	0.6621	0.6702
+ margin-score ≥ 1.09	0.6695	0.6716
+ margin-score ≥ 1.00	0.6453	0.6494

Table 5: Train on synthetic + mined data. Test on SLURP. **Bold** means the best.

5 Conclusion and future directions

Initial experiments in subsection 4.2 has shown the that our synthetic dataset can improve performance in Englisg further. It can potentially surve as a benchmark for SLU tasks in other languages subsection 4.3. In our future work, we collect data in other languages to test if synthetic data improves performance, justifying that our synthetic dataset can be used in low resource settings.

5.1 Variations in SLU model architecture

We will also run SLU experiments on the dataset. We could compare the performance of cascaded models (ASR + NLU) and E2E models, such study has been conducted in (Qian et al., 2021).

5.2 Robustness and generalization

In order to evaluate the robustness and generalization capabilities of our models, one can subject

them to a variety of test conditions that mimic real-world scenarios, such as noisy environments, varying accents, and speech spoken at different speeds.

5.3 Data augmentation

The synthesized speech is usually too fluent and clean by some manual inspections. However, real speech is usually mixed with background noises and disfluencies like hesitations and repetition of words. To train models for real applications, we will try to apply some data augmentation techniques to the synthesized speech.

- Increase audio diversities. Increasing the variation in audio characteristics (e.g., speed, pitch, tone) to better mimic real-world speech patterns. We can use reverberation methods by convolving the audio with an acoustic impulse response (AIR) randomly selected (Fazel et al., 2021).
- Add more speakers. Introducing models that vary the speaker's characteristics, such as sex or age.
- Generate expressive speech. Leveraging TTS models capable of adding emotional tones to the speech, making the dataset more applicable for emotion recognition tasks.
- Inject background noises (Snyder et al., 2015). We can use simulated real-life audio environments by mixing generated speech with background noises, improving the robustness of speech recognition models trained on the dataset.

References

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

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

H. Bonneau-Maynard, C. Ayache, F. Bechet, A. Denis, A. Kuhn, F. Lefevre, D. Mostefa, M. Quignard, S. Rosset, C. Servan, and J. Villaneau. 2006. Results of the French evalda-media evaluation campaign for literal understanding. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Pavel Denisov and Ngoc Thang Vu. 2023. Leveraging multilingual self-supervised pretrained models for sequence-to-sequence end-to-end spoken language understanding. *arXiv preprint arXiv:2310.06103*.

Chris Donahue, Julian McAuley, and Miller Puckette. 2018. Adversarial audio synthesis. In *International Conference on Learning Representations*.

Manaal Faruqui and Dilek Z. Hakkani-Tür. 2021. Revisiting the boundary between asr and nlu in the age of conversational dialog systems. *Computational Linguistics*, 48:221–232.

Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. 2021. Synthasr: Unlocking synthetic data for speech recognition.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023a. 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)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023b. 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)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ting-Yao Hu, Mohammadreza Armandpour, Ashish Shrivastava, Jen-Hao Rick Chang, Hema Koppula, and Oncel Tuzel. 2021. Synt++: Utilizing imperfect synthetic data to improve speech recognition.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents.

Ying-Cheng Lu, Huazheng Wang, and Wenqi Wei. 2023. Machine learning for synthetic data generation: a review. *ArXiv*, abs/2302.04062.

Loren Lugosch, Brett Meyer, Derek Nowrouzezahrai, and Mirco Ravanelli. 2019a. Using speech synthesis to train end-to-end spoken language understanding models.

Loren Lugosch, Piyush Papreja, Mirco Ravanelli, Abdelwahab Heba, and Titouan Parcollet. 2021. Timers and such: A practical benchmark for spoken language understanding with numbers.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019b. Speech model pre-training for end-to-end spoken language understanding.

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

Yao Qian, Ximo Bian, Yu Shi, Naoyuki Kanda, Leo Shen, Zhen Xiao, and Michael Zeng. 2021. Speech-language Pre-training for End-to-end Spoken Language Understanding.

Shangeth Rajaa, Swaraj Dalmia, and Kumarmanas Nethil. 2022. Skit-s2i: An indian accented speech to intent dataset.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A Music, Speech, and Noise Corpus. ArXiv:1510.08484v1.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Chao Yan, Yao Yan, Zhiyu Wan, Ziqi Zhang, Larsson Omberg, Justin Guinney, Sean D. Mooney, and Bradley A. Malin. 2022. A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications*, 13.

Su Zhu, Zijian Zhao, Tiejun Zhao, Chengqing Zong, and Kai Yu. 2019. Catslu: The 1st chinese audio-textual spoken language understanding challenge. In *2019 International Conference on Multimodal Interaction*, ICMI '19, page 521–525, New York, NY, USA. Association for Computing Machinery.