Performance Analysis of Speech Encoders for Low-Resource SLU and ASR in Tunisian Dialect

Salima Mdhaffar[1], Haroun Elleuch[1,2], Fethi Bougares[2], Yannick Estève[1]

[1]LIA, Avignon University, France
[2]Elyadata, Paris, France
salima.mdhaffar@univ-avignon.fr

Abstract

Speech encoders pretrained through self-supervised learning (SSL) have demonstrated remarkable performance in various downstream tasks, including Spoken Language Understanding (SLU) and Automatic Speech Recognition (ASR). For instance, fine-tuning SSL models for such tasks has shown significant potential, leading to improvements in the SOTA performance across challenging datasets. In contrast to existing research, this paper contributes by comparing the effectiveness of SSL approaches in the context of (i) the low-resource spoken Tunisian Arabic dialect and (ii) its combination with a low-resource SLU and ASR scenario, where only a few semantic annotations are available for fine-tuning. We conduct experiments using many SSL speech encoders on the TARIC-SLU dataset. We use speech encoders that were pre-trained on either monolingual or multilingual speech data. Some of them have also been refined without in-domain nor Tunisian data through multimodal supervised teacher-student paradigm. This study yields numerous significant findings that we are discussing in this paper.

1 Introduction

Self-supervised learning methods aim to train a representational model, also called upstream model, that benefits a collection of downstream tasks. SSL models are trained by using information extracted from the input data itself as the label to target. Various techniques have been introduced in the literature in order to learn powerful representations from the speech signal, including those based on autoregressive predictive coding (Chung et al., 2019), contrastive losses (Schneider et al., 2019; Baevski et al., 2020) and masked prediction (Liu et al., 2021; Chen et al., 2022; Hsu et al., 2021). Other works explored the combination of contrastive learning and masked language modeling (Chung et al., 2021). The learned SSL models like wav2vec 2.0 (Baevski et al., 2020), wavLM (Chen et al., 2022), data2vec (Baevski et al., 2022b), data2vec 2.0 (Baevski et al., 2022a), ArTST (Toyin et al., 2023), w2v-BERT (Chung et al., 2021) have been proven effective for a wide range of tasks as they considerably lighten the amount of annotated speech data normally required for downstream tasks, while exploiting very large amounts of unlabeled data. Some models have been extended to a cross-lingual setting through XLS-R-128 (Babu et al., 2021), MMS (Pratap et al., 2023) and the recently released w2v-BERT 2.0 model (Barrault et al., 2023). Several efforts went further to enrich the frame-level speech representations with textual semantic information. In 2022, Khurana et al. (2022) proposed such a model called SAMU-XLSR allowing the obtention of a better semantic encoding in the speech representations of a pre-trained SSL model. SAMU-XLSR approach follows a teacher-student supervised learning framework that uses multilingual text/audio paired data and the Language-agnostic BERT Sentence Embedding (LaBSE) model (Feng et al., 2022). This teacher-student approach is used to refine an SSL pre-trained model dedicated to speech processing to make it able to produce sentence-level embeddings from speech similar to the embbedings given by the LaBSE model from the speech transcriptions. They use the CommonVoice version 8 corpora applied to 53 languages to refine the XLS-R-128 model (Babu et al., 2021).

More Recently, a multilingual sentence embedding space for 200 text and 37 speech languages called SONAR (Duquenne et al., 2023) was introduced by META AI. The authors adopted a two-step training strategy: They first build a sentence embedding space for the multilingual text before extending this to the speech modality via a teacher-student approach. Applied to many downstream tasks, SSL speech encoders have shown their great

Proceedings of The Second Arabic Natural Language Processing Conference, pages 130–139

potential by improving the state-of-the-art performances on challenging benchmark datasets. However, understanding speech encoder capabilities requires a benchmarking effort to compare and draw insights across the techniques.

Indeed, there has been a considerable amount of work and effort in order to benchmark such SSL models. SUPERB (Yang et al., 2021) is a good example of this effort. It provides a comprehensive speech SSL benchmark including tasks such as phoneme detection, ASR, slot filling, intent detection, keyword spotting, etc. The SUPERB benchmark has been extended to evaluate multilingual speech systems in the ML-SUPERB benchmark (Shi et al., 2023). XTREME-S (Conneau et al., 2022) is another example that focuses only on SSL models trained with multiple languages. Despite the large number of high-quality benchmarks that evaluate SSL models on various downstream tasks, a limited number of studies have probed their effectiveness for downstream tasks in spoken Arabic dialects.

In this paper, we propose to broaden the SSL benchmarking effort to the Automatic Speech Recognition (ASR) and Spoken Language Understanding (SLU) of spoken Arabic Dialects. The contributions of this work are as follows:

1. We benchmark the effectiveness of various state-of-the-art SSL speech encoders in the very challenging context of a low-resource spoken Arabic dialect (Tunisian dialect) with a limited training data;

2. We compare the performances of mono vs. multilingual/bilingual SSL models, and the impact of a semantic encoding refinement through a multimodal supervised teacher-student approach;

3. We explore the use of recently released SSL models (w2v-BERT 2.0 and SONAR) for both ASR and SLU. To our knowledge, this is the first work in the literature to evaluate w2v-BERT 2.0 and SONAR for an SLU task;

4. We release our code and models[1] for reproducibility and to encourage future research on ASR and SLU of Arabic dialects.

[1] https://github.com/speechbrain/speechbrain/tree/develop/recipes/TARIC

2 Benchmarking protocol

In this section, we formulate the downstream tasks and describe the Tunisian dialect dataset together with the speech encoders used in our study.

2.1 Downstream tasks

In this work, we address two tasks: Automatic Speech Recognition and Spoken Language Understanding of the Tunisian Arabic spoken dialect.

(a) ASR: Automatic Speech Recognition has the ultimate goal of providing the correct transcription given spoken utterance. It is used to assess the ability of SSL models to extract content information from audio inputs. The ASR task will be evaluated in terms of Word Error Rate (WER).

(b) SLU: Spoken language understanding refers to natural language processing tasks that aim to extract semantic information from speech (Tur and De Mori, 2011). Different tasks can be addressed as SLU tasks, such as named entity recognition from speech, dialog state tracking, intent recognition, slot filling, etc... In the context of a conventional dialogue system, information is typically represented through a semantic frame structure. For each utterance, constructing the semantic representation primarily involves (i) classifying the user's utterance in terms of 'speech acts' (SA) (Searle, 1969) or 'intents' and (ii) slot filling (Wang et al., 2005). We consider these two SLU tasks following the available annotations: (1) **Speech Act classification** classifies speech utterances into predefined classes to determine the intent of the speaker; (2) **Slot filling** is a natural language processing and information extraction technique that entails the identification and extraction of specific pieces of information or attributes, referred to as 'slots', from unstructured text or spoken language. These slots are typically associated with predefined categories or entities. The SLU speech act recognition will be evaluated in terms of Speech Act Error Rate (SAER), which is a standard classification error rate. The SLU slot filling task will be evaluated in terms of Concept Error Rate (COER) and Concept/Value Error Rate (CVER). COER is

computed similarly to WER by taking into account only the semantic labels in the reference and hypothesis annotations. The CVER computation is identical, but the occurrences of concept/value pairs are taken into account instead of the concept alone. CVER implies that if any character within the word's support prediction or the concept tag prediction differs from the reference, the entire prediction for that concept is considered as an error. For the calculation of these metrics, we have drawn on the detailed description in Laperrière et al. (2022).

2.2 Dataset

The ASR dataset TARIC (Masmoudi et al., 2014) has been used for this work along with its recent SLU enrichment, TARIC-SLU (Mdhaffar et al., 2024). The acquisition of the TARIC dataset was carried out in train stations in Tunisia. The dataset is made of human-human recordings with their manual transcriptions and semantic annotations. It is composed of more than 2,000 dialogues from 109 different speakers. The dataset[2] is split into three parts (train, dev and test) as described in Table 1.

Table 1: TARIC-SLU data set split into Train, Dev and Test

	Train	Dev	Test
#Utterance	15752	771	1249
#Dialog	1713	103	173
Duration	7.5 hrs	29 min	53 min

TARIC-SLU was annotated using 62 semantic concept tags such as *city name arrival*, *departure time*, *ticket price*, etc. and 3 speech acts (*directives-answer*, *directives-query* and *politeness*). The example below shows the original sentence (a) together with its English translation (b) and the corresponding semantic annotation (c).

As for the annotation tags, the first word of each sentence represents the speech act tag (**Directives-query** in this example). '**<city-name-departure** ' is an opening tag starting the support word sequence 'Sfax' and expressing that this word sequence is associated with the *city-name-departure* semantic concept. The character '>' represents the closing tag, and it is used to close all concept tags.

[2] https://demo-lia.univ-avignon.fr/taric-dataset/

The same schema is used for the other semantic concepts. For speech act tags, there is no closing tag, since each sentence has a single speech act.

(a) باللاهي أعطيني زوز تكايات من صفاقس لتونس مع الثمنية و نصف

(b) Please give me two tickets from Sfax to Tunis at eight thirty

(c) **Directives-query** please **<command-task** give me > **<number-of-tickets** two > **<object** tickets > from **<city-name-departure** Sfax > to **<city-name-arrival** Tunis > at **<departure-time** eight thirty >

2.3 SSL speech encoders

For our study, we used different types of speech encoders. We considered monolingual SSL speech encoders (French: wav2vec 2.0 LeBenchmark-7K; English: wav2vec 2.0 LV60, HuBERT, wavLM, data2vec 2.0 and SONAR-ENG) as well as cross-lingual SSL speech encoders (wav2vec 2.0 VP-100K, XLS-R-128, MMS, MMS-1B, w2v-BERT 2.0 and SAMU-XLSR) and two bilingual speech encoders (SONAR-ARB and SONAR-FRA). As explained in the introduction, SAMU-XLSR is a modified version of XLS-R-128 and SONAR models are a modified version of w2v-BERT [3]. We put the URLs for all these models in the Appendix A.

SAMU-XLSR is based on the pre-trained multilingual XLS-R-128. SAMU-XLSR will process audio and text paired data. The XLS-R-128 model used in this approach was designed to generate speech representations for short 20 milliseconds speech frames. To make use of this model, SAMU-XLSR performs pooling and projection to create a single sentence-level representation. In parallel, LaBSE (Feng et al., 2022) sentence-level textual representations are simply extracted. Both representations being on the same semantic space, SAMU- XLSR's is then being pulled towards LaBSE's with the help of a cosine similarity loss function. This means the parameters of all SAMU-XLSR's components are optimized to predict the textual representations generated by the frozen

[3] Notice that w2v-BERT and w2v-BERT 2.0 are two different models. Both of them are trained using the same architecture, but w2v-BERT is an English model and w2v-BERT 2.0 is a multilingual model. w2v-BERT has not been released to the research community, so we cannot evaluate its performance.

Table 2: Architecture details of benchmarked SSL encoders. * SAMU-XLSR is a modified version of XLS-R-128, ♣ all SONAR speech encoders models are a modified version of w2v-BERT, ⋆ represents the size of the paired dataset used in the refinement step.

Speech encoder	#Param	#Lang	Hours	Network architecture
LeBenchmark-7K (Evain et al., 2021)	317M	1	7K	7CNN-24Trans-Enc
English lv60 (Baevski et al., 2020)	317M	1	960	7CNN-24Trans-Enc
HuBERT (Hsu et al., 2021)	316M	1	60K	7CNN-24Trans-Enc
wavLM (Chen et al., 2022)	316M	1	94K	7CNN-24Trans-Enc
data2vec 2.0 (Baevski et al., 2022a)	314M	1	960	7CNN-24Trans-Enc
VP-100K (Wang et al., 2021)	317M	23	100K	7CNN-24Trans-Enc
XLS-R-128 (Babu et al., 2021)	317M	128	436K	7CNN-24Trans-Enc
MMS(Pratap et al., 2023)	317M	1024	23K	7CNN-24Trans-Enc
MMS-1B (Pratap et al., 2023)	1B	1024	23K	7CNN-48Trans-Enc
w2v-BERT 2.0 (Barrault et al., 2023)	600M	143	4.5M	2CNN-24Conformer
SAMU-XLSR (Khurana et al., 2022) *	317M	128	436K+12.7K⋆	7CNN-24Trans-Enc
SONAR-ARB (Duquenne et al., 2023) ♣	600M	2	60K+822⋆	2CNN-24Conformer
SONAR-ENG (Duquenne et al., 2023) ♣	600M	1	60K+N/A⋆	2CNN-24Conformer
SONAR-FRA (Duquenne et al., 2023) ♣	600M	2	60K+2K⋆	2CNN-24Conformer

Table 3: Architecture details of Whisper models

Model	Network architecture	#Params
Whisper-small (Radford et al., 2023)	2-conv 12 Enc-Dec	244 M
Whisper-medium (Radford et al., 2023)	2-conv 24-Enc-Dec	769 M

LaBSE model. Figure 1 illustrates the training process of SAMU-XLSR.

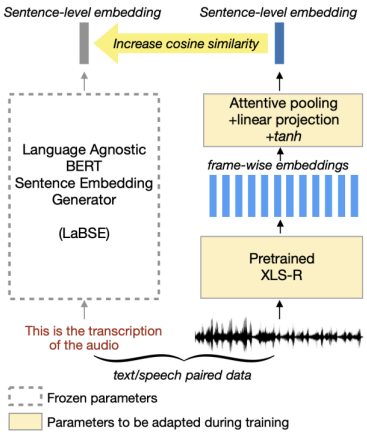

Figure 1: SAMU-XLSR Training and specialization

SONAR is a new multilingual and -modal text embedding space trained in encoder-decoder architecture for 200 languages, which substantially outperforms existing approaches like LASER3 (Heffernan et al., 2022) or LaBSE in multilingual similarity search. Authors apply a teacher-student approach to extend this embedding space to the speech modality and currently cover 36 languages. Mining is performed in data from publicly available repositories of web data (tens of billions of sentences) and speech (4 million hours). In total, to train the model, they align more than 443,000 hours of speech with texts and create about 29,000 hours of speech-to-speech alignments. Figure 2 illustrates the training process of SONAR.

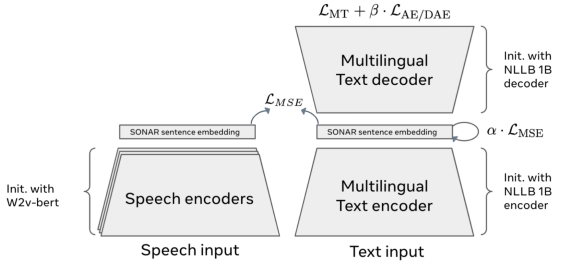

Figure 2: Sonar architecture

Overall, we experimented with 14 different speech encoders. Table 2 shows all the details of SSL speech encoder models.

In addition to the aforementioned SSL models, we also evaluated using the recently released Whisper models from OpenAI (Radford et al., 2023). Unlike self-supervised speech models, Whisper is

a multilingual ASR model trained using a large amount of labelled audio transcription data (680k hours). Using Whisper was motivated by the recent achievements reported by Wang et al. (2023). The detailed properties of Whisper models used in this study are presented in Table 3.

3 Experiments and Results

3.1 Training details

For comparison purposes, we set the same parameters for all the models using the different speech encoders.

3.1.1 ASR

In addition to the speech encoder model, we incorporate an extra layer with 1024 neurons and LeakyReLU as the activation function, followed by a fully-connected layer and a final 40-dimensional softmax layer, each dimension corresponding to a character. The weights of these two additional layers were randomly initialized, while the weights of the speech encoder part for SSL models of the neural architecture were initialized using the pre-trained weights. The fine-tuning is done with the TARIC training set using a character-level CTC loss function. We optimize the loss with an Adam optimizer of learning rate = 0.0001 for both speech encoder, and Adadelta with learning rate = 1.0 for the linear layer.

3.1.2 SLU

We formulate the end-to-end SLU task as a character level prediction where slots are delimited by tag-specific special characters, as in Yadav et al. (2020); Ghannay et al. (2018); Mdhaffar et al. (2022). We also added the speech act token to the reference annotation as the first token of each sequence of words. This way, the end-to-end model learns to both classify the utterances in terms of speech act, and recognize slot/value pairs present in the speech segment. As input, the neural network receives a WAV audio file (PCM, 16 bits, 16kHz, signed integer), and the output is a transcription enriched with semantic labels and speech acts tags. After processing through the softmax layer (which has the size of 106[4]), the outputs are generated by a simple greedy decoder. We optimize the loss with an Adam optimizer of learning rate = 0.0001 for both speech encoder, and Adadelta with learning rate = 1.0 for the linear layer.

3.1.3 Training details for ASR and SLU

We employed a batch size of 4 samples, distributed across 4 NVIDIA V100 32GB GPU cards. For MMS-1B, we used 4 NVIDIA A100 80GB GPU cards. We utilized two optimizers: Adadelta for updating the additional layers weights and Adam for fine-tuning the self-supervised learning (SSL) model. The initial learning rate for Adadelta was 1.0, while for Adam it was set to 0.0001. Our models were implemented using the SpeechBrain toolkit (Ravanelli et al., 2021). The speech encoders are obtained through the fairseq (Ott et al., 2019) framework for SAMU-XLSR [5], SONAR and data2vec 2.0 and through HuggingFace for the remaining of models. More details about our models and the configuration files will be publicly available as a part of the Speechbrain toolkit.

3.2 Results

In this section, we report, analyze and discuss the performance of our models across various dimensions. All the results, including both ASR and SLU evaluations, are presented in Table 4. ASR models are evaluated with WER and SLU models are evaluated using WER (after removing the semantic tags and the speech act tag from the system output), COER, SAER and CVER. The first part of the table shows results by using SSL models, and the second part shows results by using Whisper models. The first part is divided into three sub-parts according to the type of SSL models: (i) the first sub-part, colored blue for monolingual models, (ii) the second sub-part, colored pink for cross-lingual models, and (iii) the third sub-part, colored yellow for the cross-lingual model after a teacher-student multi-modal semantic training. The second part of this table is dedicated to the results obtained when using Whisper small and medium models.

Overall, except for the SLU SAER score on the dev set, the best ASR and SLU results are obtained when our models are trained using w2v-BERT 2.0 model. Below, an analysis of the obtained results according to the type of the used SSL models.

Monolingual models: When we compare the performances of models trained using monolingual SSL (blue section of the results table), we observe that WavLM is by far the best-performing model

[4] 40 characters that cover the alphabet of the TARIC dataset, 62 characters for slots, one character for closing slots and three characters for speech acts

[5] SAMU-XLSR is not yet publicly available; it has been kindly shared by the authors of Khurana et al. (2022)

Table 4: Comparison of speech encoders performance across ASR and SLU tasks. Results are reported in WER for ASR and SLU system transcripts. SLU results are reported using COER and CVER for slot filling detection and in terms of SAER for speech act classification. Bold numbers are the best results in all the table. Underlined numbers represent the best results for each colored block (mono-lingual models, cross-lingual models without teacher-student multi-modal training, cross-lingual models with teacher-student multi-modal training, whisper models). FR-7K refers to LeBenchmark-7K, Whisper-S refers to Whisper-Small, Whisper-M refers to Whisper-Medium

Speech encoder	COER (SLU)		WER (SLU)		CVER (SLU)		SAER (SLU)		WER (ASR)	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
FR-7K	37.78	33.16	36.04	28.61	57.57	49.57	**23.87**	21.14	34.80	29.69
English lv60	36.77	32.69	37.39	30.49	57.7	50.04	26.07	21.94	35.04	30.2
HuBERT	39.84	33.76	39.41	31.74	61.16	52.80	25.94	21.86	34.25	28.75
WavLM	35.85	32.25	34.22	27.23	55.95	50.92	24.9	21.14	32.28	26.7
Data2vec 2.0	34.50	31.8	39.41	31.71	58.94	51.15	25.16	22.9	38.58	32.1
VP-100K	35.77	31.56	35.46	27.56	56.37	48.90	24.64	22.9	32.71	26.68
XLS-R-128	35.62	31.24	34.65	26.7	56.24	48.73	24.64	**20.9**	33.82	28.06
MMS	36.73	31.77	43.97	37.98	62.07	56.47	24.9	24.66	35.91	29.46
MMS-1B	43.82	36.13	44.36	38.46	66.91	58.03	28.4	23.83	41.97	34.76
w2v-BERT 2.0	**32.29**	**29.13**	**26.08**	**20.84**	**49.55**	**46.22**	25.6	**20.9**	**25.11**	**21.47**
SAMU-XLSR	32.73	30.11	31.10	23.95	51.93	48.06	24.12	22.5	28.56	24.66
SONAR-ENG	36.58	33.59	38.34	31.58	57.15	52.33	36.71	26.10	39.77	33.67
SONAR-FRA	36.88	33.8	38.29	30.38	58.83	53.43	34.46	27.79	39.16	32.71
SONAR-ARB	35.93	31.62	35.6	28.17	55.98	49.77	32.68	23.38	35.24	29.05
Whisper-S	39.52	34.81	39.79	32.85	64.83	56.57	32.99	29.56	38.5	32.1
Whisper-M	39.1	33.96	33.37	29.56	59.13	54.02	25.94	21.86	32.5	29.05

for the ASR task with **32.28%** and **26.7%** WER for dev and test sets respectively. This is confirmed as well when evaluating WER of the SLU model output (column WER (SLU) in table 4). Regarding the SLU task, while we did not observe any emerged trend, we observed a better performance of data2vec 2.0 in terms of COER evaluation.

Cross-lingual models without teacher-student multi-modal semantic training: Results indicate that w2v-BERT 2.0 yields the best performance in both ASR and SLU tasks (expect for the SAER).

All cross-lingual models: When it comes to cross-lingual models, with (colored in pink) or without teacher-student multi-modal semantic training (colored in yellow), setting apart the w2v-BERT 2.0 model, SAMU-XLSR clearly outperforms VP-100k and MMS based models.

Mono-lingual vs. cross-lingual models: If we compare performances between monolingual and cross-lingual SSL models, cross-lingual models achieve better results for the SLU task. Indeed, data2vec 2.0 shows competitive COER scores while its WER is higher compared to other models. This leads to the conclusion that while data2vec 2.0 demonstrates the ability to identify semantic concepts, it falls short in providing accurate transcriptions.

Cross-lingual models with teacher-student refinement: As stated earlier, the SLU performances of VP-100K, XLSR-128, and MMS exhibit a consistent trend, with XLSR-128 showing the highest performance among them. Combining this multilingual SSL (XLS-R-128) and teacher-student multi-modality training in one model (SAMU-XLSR) provides better results for both ASR and SLU. Shifting the focus to SONAR models, as expected, SONAR-ARB demonstrates superior performance compared to both SONAR-ENG and SONAR-FRA for both ASR and SLU tasks. SONAR models are based on the English version of w2v-BERT. Unfortunately, the w2v-BERT model used to initialize the speech encoders of SONAR is not available to assess its contribution.

Whisper models: Results show that Whisper-medium outperform results obtained using Whisper-small.

Whisper models vs. all SSL models: If we compare performances between Whisper models (second part of the table) and all SSL models (first part of the table), almost SSL models achieve better

results for the slot filling SLU task. To transcribe audio files, WER obtained by the ASR or the SLU show competitive results. For example, ASR results for the dev set (32.5%) obtained by Whisper medium outperform almost all SSL models except for w2v-BERT 2.0, SAMU-XLSR and wavLM. This leads to the conclusion that while Whisper models demonstrates the ability to transcribe audio files, it falls short in providing accurate semantic extraction compared to the use of SSL speech encoders.

4 Discussion

We carried out some error analyses to quantify some SLU systems performance across different conditions. We focus on 6 SSL models: LeBenchmark, XLS-R-128, data2vec 2.0, SAMU-XLSR, w2v-BERT 2.0 and SONAR-ARB.

4.1 Acoustic complexity

First, we used the WER of the transcripts produced by the model giving the best ASR system (w2v-BERT 2.0) to quantify the general complexity of the utterance. We defined three groups of segments with low (WER <= 20), medium (20 < WER <= 50) and high (WER > 50) complexity. Figure 3 shows that, as expected, the SLU performances are overall better for spoken utterances belonging to the low complexity group. The x-axis represents different levels of WER, and the y-axis represents COER. However, when we compare the COER of the SLU systems by group, we observe that w2v-BERT performs better for segments with low WER. For segments with medium WER, we observe that all the systems have a comparable behavior except for LeBenchmark: the system shows a higher COER. For utterances hard to transcribe, SAMU-XLSR is the best one followed by w2v-BERT 2.0.

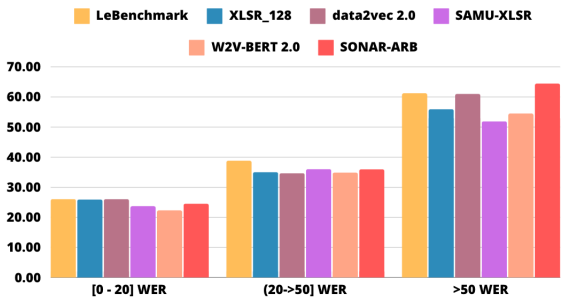

Figure 3: SLU performance (COER) across different general complexity levels in test utterances.

4.2 Semantic complexity

In the semantic complexity analysis, we used the number of semantic tags per utterance in the reference as a proxy of the semantic complexity. Figure 4 shows the results obtained by the evaluated models. The x-axis represents the number of semantic tags in test utterances, and the y-axis represents COER. Across the board, utterances with two to six concepts seem to be the easiest. While data2vec 2.0 and SONAR-ARB perform worse when there are fewer semantic concepts to extract. They perform best when there are more than six.

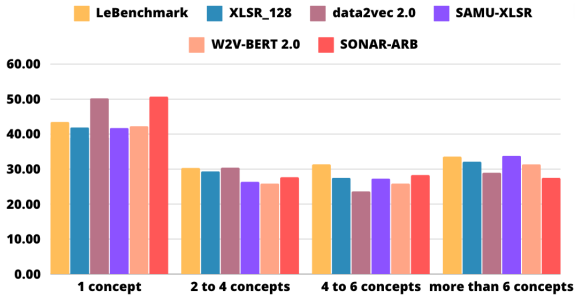

Figure 4: SLU performance (COER) across different numbers of semantic tags in test utterances.

5 Conclusion

Our study investigates the usage of various SSL speech encoders for a Spoken Language Understanding task in challenging circumstances characterized by a scarcity of both SLU and ASR training data and the low-resource characteristics of the targeted Tunisian Arabic dialect. Our findings emphasize the efficacy of SSL pre-trained speech encoders in such conditions, with notable success observed when employing the w2v-BERT 2.0 model, with 600 millions parameters trained on 4.5M hours of speech from 143 languages. Additionally, we highlight the noteworthy performance of data2vec 2.0, pre-trained on English monolingual data, particularly excelling in handling semantically complex utterances. These outcomes collectively provide valuable information for advancing SLU methodologies, especially in resource-limited linguistic contexts. Last, SAMU-XLSR provides very competitive results thanks to semantic enrichment made by the teacher student approach and in future work we plan to train a SAMU-w2vBERT 2.0 model to take benefit of the joint SAMU and w2v-BERT 2.0 capabilities.

6 Acknowledgements

This work was performed using HPC resources from GENCI-IDRIS (grant AD011012108R1) and received funding from the EU H2020 SELMA (grant No 957017), ANR TRADEF project (ANR-22-ASGC-0003) and ESPERANTO research and innovation programme under the Marie Skłodowska-Curie (grant No 101007666). We would like to especially thank our colleagues Sameer Khurana and Antoine Laurent for sharing their SAMU-XSLR model with us.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. pages arXiv–2111.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. 2022a. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022b. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An Unsupervised Autoregressive Model for Speech Representation Learning. In *Proc. Interspeech 2019*.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *ASRU 2021*.

Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, et al. 2022. XTREME-S: Evaluating Cross-lingual Speech Representations. In *Proc. Interspeech 2022*, pages 3248–3252.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Solène Evain, Manh Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, et al. 2021. Task agnostic and task specific self-supervised learning from speech with lebenchmark. In *NeurIPS*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

S. Ghannay, A. Caubrière, Y. Estève, A. Laurent, and E. Morin. 2018. End-to-end named entity extraction from speech. *CoRR*, abs/1805.12045.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. *arXiv preprint arXiv:2205.12654*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM TASLP*.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *Journal of Selected Topics in Signal Processing*.

Gaëlle Laperrière, Valentin Pelloin, Antoine Caubrière, Salima Mdhaffar, Nathalie Camelin, Sahar Ghannay, Bassam Jabaian, and Yannick Estève. 2022. The spoken language understanding media benchmark dataset in the era of deep learning: data updates, training and evaluation tools. In *LREC*, pages 1595–1602.

Andy T Liu, Shang-Wen Li, and Hung-yi Lee. 2021. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM TASLP*.

Abir Masmoudi, Yannick Estève, Mariem Ellouze Khmekhem, Fethi Bougares, and Lamia Hadrich Belguith. 2014. Phonetic tool for the tunisian arabic. In *Spoken Language Technologies for Under-Resourced Languages*. Citeseer.

Salima Mdhaffar, Fethi Bougars, Renato De mori, Salah Zaiem, Mirco Ravanelli, and Yannick Estève. 2024. Taric-slu: A tunisian dataset for spoken language understanding. *LREC*.

Salima Mdhaffar, Valentin Pelloin, Antoine Caubrière, Gaëlle Laperrière, Sahar Ghannay, Bassam Jabaian, Nathalie Camelin, and Yannick Estève. 2022. Impact analysis of the use of speech and language models pretrained by self-supersivion for spoken language understanding. In *LREC 2022*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, et al. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *Interspeech 2019*.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, et al. 2023. Ml-superb: Multilingual speech universal performance benchmark. *arXiv*.

Hawau Olamide Toyin, Amirbek Djanibekov, Ajinkya Kulkarni, and Hanan Aldarmaki. 2023. Artst: Arabic text and speech transformer. *arXiv preprint arXiv:2310.16621*.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Changhan Wang, Morgane Rivière, 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 *ACL*.

Minghan Wang, Yinglu Li, Jiaxin Guo, Xiaosong Qiao, Zongyao Li, Hengchao Shang, Daimeng Wei, Shimin Tao, Min Zhang, and Hao Yang. 2023. Whislu: End-to-end spoken language understanding with whisper. In *Proc. Interspeech*, volume 2023, pages 770–774.

Ye-Yi Wang, Li Deng, and Alex Acero. 2005. Spoken language understanding. *Signal Processing Magazine*.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. *Interspeech*.

Shu Wen Yang, Po Han Chi, Yung Sung Chuang, Cheng I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pages 3161–3165. International Speech Communication Association.

A Appendix: URLs of all models used in this study

For reproducibility of all results obtained in this paper, we put in the table 5 the url for each model.

Table 5: URLs of all models used in this study

Speech encoder	URL
LeBenchmark-7K	https://huggingface.co/LeBenchmark/wav2vec2-FR-7K-large
English lv60	https://huggingface.co/facebook/wav2vec2-large-lv60
HuBERT	https://huggingface.co/facebook/hubert-large-ll60k
WavLM	https://huggingface.co/microsoft/wavlm-large
Data2vec 2.0	https://github.com/facebookresearch/fairseq/tree/main/examples/data2vec
VP-100K	https://huggingface.co/facebook/wav2vec2-large-100k-voxpopuli
XLS-R-128	https://huggingface.co/facebook/wav2vec2-xls-r-300m
MMS	https://huggingface.co/facebook/mms-300m
MMS-1B	https://huggingface.co/facebook/mms-1b
w2v-BERT 2.0	https://huggingface.co/facebook/w2v-bert-2.0
SAMU-XLSR	SAMU-XLSR is not yet publicly available; it has been shared by Khurana et al. (2022)
SONAR-ENG	https://github.com/facebookresearch/SONAR
SONAR-FRA	https://github.com/facebookresearch/SONAR
SONAR-ARB	https://github.com/facebookresearch/SONAR
Whisper-small	https://huggingface.co/openai/whisper-small
Whisper-medium	https://huggingface.co/openai/whisper-medium