On the Use of Semantically-Aligned Speech Representations for Spoken Language Understanding

Gaelle Laperriere, Valentin Pelloin, Mickael Rouvier, Themos Stafylakis, Yannick Esteve

ON THE USE OF SEMANTICALLY-ALIGNED SPEECH REPRESENTATIONS FOR SPOKEN LANGUAGE UNDERSTANDING

Gaëlle Laperrière[1], *Valentin Pelloin*[2], *Mickaël Rouvier*[1], *Themos Stafylakis*[3], *Yannick Estève*[1]

[1] LIA - Avignon Université, France
[2] LIUM - Le Mans Université, France
[3] Omilia - Conversational Intelligence, Greece

ABSTRACT

In this paper we examine the use of semantically-aligned speech representations for end-to-end spoken language understanding (SLU). We employ the recently-introduced SAMU-XLSR model, which is designed to generate a single embedding that captures the semantics at the utterance level, semantically aligned across different languages. This model combines the acoustic frame-level speech representation learning model (XLS-R) with the Language Agnostic BERT Sentence Embedding (LaBSE) model. We show that the use of the SAMU-XLSR model instead of the initial XLS-R model improves significantly the performance in the framework of end-to-end SLU. Finally, we present the benefits of using this model towards language portability in SLU.

Index Terms— Spoken language understanding, speech representation, language portability, cross modality

1. INTRODUCTION

Spoken language understanding (SLU) refers to natural language processing tasks related to semantic extraction from speech [1]. Different tasks can be addressed as SLU tasks, such as named entity recognition from speech, call routing, slot filling task in a context of human-machine dialogue.

To our knowledge, end-to-end neural approaches have been proposed four years ago in order to directly extract the semantics from speech signal, by using a single neural model [2, 3, 4], instead of applying a classical cascade approach based on the use of an automatic speech recognition (ASR) system, followed by a natural language understanding processing (NLU) module applied to the automatic transcription [1]. Two are the main advantages of end-to-end approaches. The first one is related to the joint optimization of the ASR and NLU part, since the unique neural model is optimized only for the final SLU task. The second one is the mitigation of error propagation: when using a cascade approach, errors generated by the first modules propagate to the following ones.

Since 2018, end-to-end approaches have became very popular in the SLU literature [5, 6, 7, 8, 9]. A main issue of these approaches is the lack of bimodal annotated data (speech audio recordings with semantic manual annotation). Several methods have been proposed in order to address this issue, e.g. transfer learning techniques [10, 11], [12] or artificial augmentation of the training data using speech synthesis [13, 14].

Self-supervised learning (SSL), that benefits from unlabelled data, recently opened new perspectives for automatic speech recognition and natural language processing [15, 16]. SSL has been successfully applied to several SLU tasks, especially through cascade approaches [17]: the ASR system benefits from learning better speech unit representations [18, 19, 20] while the NLU module benefits from BERT-like models [16]. The use of an end-to-end approach exploiting directly both speech and text SSL models is limited by the difficulty to unify the speech and textual representation spaces, in addition to the complexity of managing a huge number of model parameters. Some approaches have been proposed in order to exploit the BERT-like capabilities within an end-to-end SLU model, e.g. by projecting some kinds of sequences of embeddings extracted by an ASR sub-module to a BERT model [21, 22], or by tying at the sentence level the acoustic embeddings to a SLU fine-tuned BERT model for a speech intent detection task [12, 23]. In [24], a similar approach is extended in order to build a multilingual end-to-end SLU model, again for speech intent detection.

Earlier this year, a new promising model was introduced in [25]. The model combines a state-of-the-art multilingual acoustic frame-level speech representation learning model XLS-R [26] with the Language Agnostic BERT Sentence Embedding [27] (LaBSE) model to create an utterance-level multimodal multilingual speech encoder. This model is named SAMU-XLSR, for Semantically-Aligned Multimodal Utterance-level Cross-Lingual Speech Representation learning framework.

In this paper, we analyze the performance and the behavior of the SAMU-XLSR model using the French MEDIA benchmark dataset, which is considered as a very challenging benchmarks for SLU [28]. Moreover, by using the Italian PortMEDIA corpus [29], we also investigate the potential of

porting an existing end-to-end SLU model from one language (French) to another (Italian) through two scenarios concerning the target language: zero-shot or low-resource learning.

2. SAMU-XLSR

Self-supervised representation learning (SSL) approaches such as Wav2Vec-2.0 [15], HuBERT [20], and WavLM [30] aim to provide powerful deep feature learning (speech embedding) without requiring large annotated datasets. Speech embeddings are extracted at the acoustic frame-level i.e. for short speech segments of 20 ms duration, and they can be used as input features to a model that is specific for the downstream task. These speech encoders have been successfully used in several tasks, such as automatic speech recognition [15], speaker verification [31, 32] and emotion recognition [33, 34]. Self-supervision learning for such speech encoders is designed to discover speech representations that encode pseudo-phonetic or phonotactic information rather than high-level semantic information [35]. On the other hand, high-level semantic information is particularly useful in some tasks such as Machine Translation (MT) or Spoken Language Understanding (SLU). In [25], the authors propose to address this issue using a new framework called SAMU-XLSR, which learns semantically-aligned multimodal utterance-level cross-lingual speech representations.

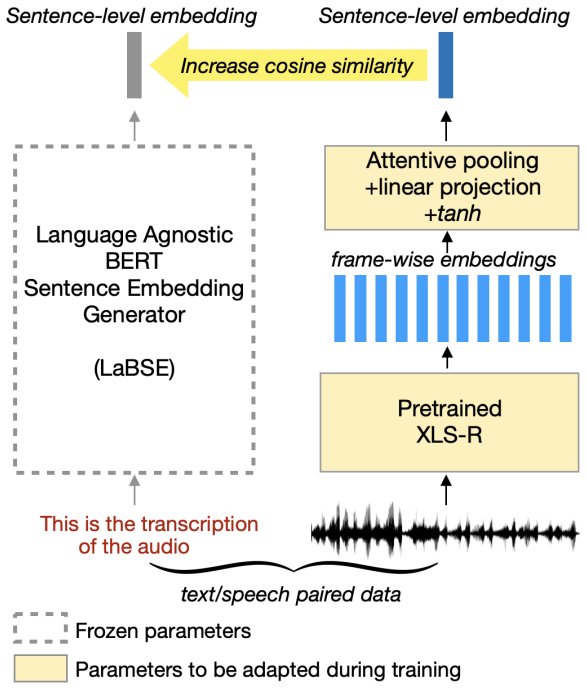

Fig. 1. Training process of SAMU-XLSR.

SAMU-XLSR is based on the pre-trained multilingual XLS-R [1] [26] on top of which all the embeddings generated by processing an audio file are connected to an attentive pooling module.

Thanks to this pooling mechanism (which is followed by linear projection layer and the *tanh* function), the frame-level contextual representations are transformed into a single utterance-level embedding vector. Figure 1 summarizes the training process of the SAMU-XLSR model. Notice than the weights from the pre-trained XLS-R model continue being updated during the process.

The utterance-level embedding vector of SAMU-XLSR is trained via knowledge distillation from the pre-trained language agnostic LaBSE model [27]. The LaBSE model[2] has been trained on 109 languages and its text embedding space is semantically aligned across these 109 languages. LaBSE attains state-of-the-art performance on various bi-text retrieval/mining tasks, while yielding promising zero-shot performance for languages not included in the training set (probably thanks to language similarities). Thus, given a spoken utterance, the parameters of SAMU-XLSR are trained to accurately predict a text embedding provided by the LaBSE text encoder of its corresponding transcript. Because LaBSE embedding space is semantically-aligned across various languages, the text transcript would be clustered together with its text translations.

By pulling the speech embedding towards the anchor embedding, cross-lingual speech-text alignments are automatically learned without ever seeing cross-lingual associations during training. This property is particularly interesting in the SLU context in order to port an existing model built on a well-resourced language to another language with zero or low resources for training.

3. APPLICATION TO SPOKEN LANGUAGE UNDERSTANDING

As defined in [1], *spoken language understanding is the interpretation of signs conveyed by a speech signal.* This interpretation refers to a semantic representation manageable by computers. Usually, this semantic representation is dedicated to an application domain that restricts the semantic field. With the massive deployment of voice assistants like Apple's Siri, Amazon Alexa, Google Assistant, etc. a lot of recent papers aim to process speech intent detection as an SLU task [12, 13, 14, 24, 23]. In such a task, only one speech intent is generally expected by sentence: the speech intent detection task could be considered as a classification task at the sentence-level and, in addition, the SLU model has to fill some expected slots corresponding to the detected intent.

SLU benchmarks related to task-driven human-machine spoken dialogue can be more or less complex, depending on the richness of the semantic representation. In this study, we

[1] https://huggingface.co/facebook/wav2vec2-xls-r-300m
[2] https://huggingface.co/sentence-transformers/LaBSE

focus on a hotel booking scenario through a telephone conversation, where the semantic representation is not related to speech intent detection, but based on a more complex ontology that derives from frames [36].

Our experimental work is carried out on the MEDIA SLU benchmark, described in section 4.1. We first expect to evaluate the performance of SAMU-XLSR used as a frame-wise feature extractor in comparison to the use of the initial XLS-R. Then we analyse the quality of the semantic encoding for each layer of the SAMU-XLSR and XLS-R model, to better understand the impact of the SAMU-XLSR training on the XLS-R model. We continue this investigation by fine-tuning the SAMU-XLSR and the XLS-R models on the downstream task. We also investigate the capability of SAMU-XLSR to transfer the semantic knowledge captured on French data to Italian data related to the same SLU task, thanks to the PortMEDIA corpus described in section 4.2. Last, we focus on the sentence-level embedding produced by the SAMU-XLSR model in order to measure the relevance of its semantic content to the target task, including in a language portability scenario and in a cross-modal setting.

4. DATA

4.1. The MEDIA benchmark

The French MEDIA benchmark [37] was created in 2002 as a part of a French governmental project named Technolangue. The *MEDIA Evaluation Package*[3] [4] is distributed by ELRA and freely accessible for academic research. Apart from the data itself, it defines a protocol for evaluating SLU modules, with a task of semantic extraction from speech in a context of human-machine dialogues.

The Wizard-of-Oz method has been used to create the dataset, consisting of hotel reservation phone call recordings. A human plays the role of a machine, by interacting with the user who believes that he is actually speaking to an intelligent machine. 1258 official recorded dialogues were generated from around 250 speakers. Only the user's turns are semantically annotated with both semantic annotation and transcription. Table 1 presents the data distribution, in hours of speech and number of words, into the official training, development and test corpora.

	Train	**Dev**	**Test**
Hours	10h52m	01h13m	03h01m
Words	94.5k	10.8k	26.6k

Table 1. Data distribution of the MEDIA corpus.

The semantic dictionary defined in MEDIA includes 83 basic attributes (renamed as *concepts* in our study) – including 73 database attributes, 4 modifiers, and 6 general attributes – and 19 specifiers [38]: *room-number, hotel-name, location* are examples of database attributes, *comparative, relative-distance* are examples of modifiers, *proposition-connector* and *attribute-connector* are examples of general attributes, and *address, travel* are examples of specifiers, that specializes the attribute role in a dialogue. Some complex linguistic phenomena, like co-references, are also managed thanks to this mechanism. By combining attributes and specifiers, the total number of possible attribute/specifier pairs is 1121. In this study, the "full" MEDIA version has been used for all experiments. Compared to the "relax" one, the full's semantic annotations includes the use of specifiers and modifiers. 150 different semantic concepts (*i.e.* attribute/specifier pairs) are present in this version of MEDIA.

Each semantic concept is associated to values, also called word-support in the MEDIA documentation. The following translated sentence is an example of utterance present in the MEDIA dataset: "I would like to book one double room in Paris". In this paper, we use annotations containing the transcript, the concepts and the location information of their values: "I (would like to book, *reservation*), (one, *room-number*), (double room, *room-type*) in (Paris, *city*)".

4.2. The Italian PortMEDIA corpus

The Italian PortMEDIA corpus [29] has been produced on the same hotel reservation task as MEDIA, and follows the same specifications. ELRA was in charge of the data collection and is currently distributing the corpus, as for MEDIA. [5]

	Train	**Dev**	**Test**
Hours	07h18m	02h32m	04h51m
Words	21.7k	7.7k	14.7k

Table 2. Data distribution of the PortMEDIA corpus.

The Italian PortMEDIA corpus is made of 604 dialogues from more than 150 Italian speakers. Table 2 gives information about the number of recorded hours and number of words distributed into training, development and test datasets. The PortMEDIA training corpus is more than four times smaller than the MEDIA one in terms of words. If speech duration seems not so low in comparison, this is due to a less precise speech segmentation that includes large portions of silence.

The PortMEDIA corpus is only available with "full" semantic annotations. 139 different concepts are present in the whole Italian PortMEDIA dataset.

The PortMEDIA corpus is used in our study in order to conduct experiments on language portability from French to Italian for SLU.

[3] http://catalog.elra.info/en-us/repository/browse/ELRA-E0024/

[4] International Standard Language Resource Number: 699-856-029-354-6

[5] http://www.elra.info/en/projects/archived-projects/port-media/

4.3. MEDIA and PortMEDIA Metrics

Historically, on the MEDIA corpus, two metrics are jointly used: the Concept Error Rate (CER) and the Concept-Value Error Rate (CVER). The CER is computed similarly to Word Error Rate (WER), by only taking into account the concepts occurrences in both the reference and the hypothesis files. The CVER metrics is an extension of the CER. It considers the correctness of the complete concept/value pair.

Since our models generate transcript with semantic concepts, we also evaluate our systems in terms of Character Error Rate (ChER) and WER. The ChER is computed by taking into account all the characters in the prediction that are not related to concepts (tags or concepts themselves). The same holds for WER but for words instead of characters.

For the semantic analysis of the sentence levels embeddings in section 5.3, we evaluate our bag-of-concepts outputs with the micro F_1-score, i.e. the harmonic mean between the precision and recall.

5. EXPERIMENTS

5.1. Layer-wise analysis of frame-level embeddings

Figure 2 presents the general architecture of the end-to-end model used for this study.

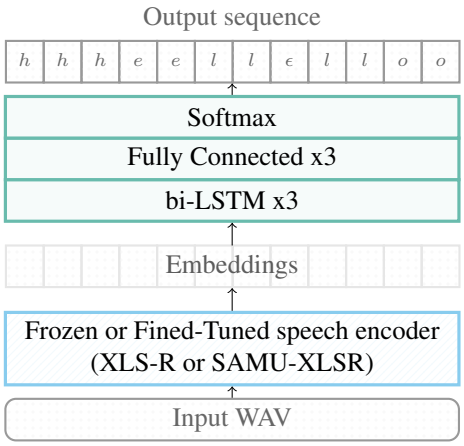

Fig. 2. Neural Architecture for an SLU layer-wise analysis of speech encoders with the MEDIA dataset.

This model aims to produce a transcription with semantic labels, such as: *I* <reservation> *would like to book* </reservation> <room-number> *one* </room-number> <room-type> *double room* </room-type> *in* <city> *Paris* </city>.

For comparison purposes, we used as an encoder the original XLS-R or the SAMU-XLSR, taking WAV signal as input. Both have been frozen or fine-tuned during our experiments. The encoder is followed by three bi-LSTM layers of 1024 neurons, to keep the context of the entire segment when decoding the speech embedding. Then, the bi-LSTM outputs are fed to three linear layers of the same number of neurons, with a separate Adadelta optimizer starting with a 1.0 learning rate. When fine-tuning the encoder, the Adam optimizer is used, with a initial learning-rate of 0.0001. The bi-LSTM layers have a similar optimizer. All the decoding layers are activated with a LeakyReLU function. The outputs are finally passing through a Softmax function and a CTC loss function is applied. For early stopping and for the optimizers, the Concept Error Rate is the value we aim to minimize.

To make a layer-wise analysis, we removed the upper layers of each encoder, one by one, and extracted our speech embeddings.The encoder kept layers are frozen with their initial weights for the "Frozen" architecture, or fine-tuned by supervision to solve the MEDIA task, leading to the "Fine-Tuned" results.

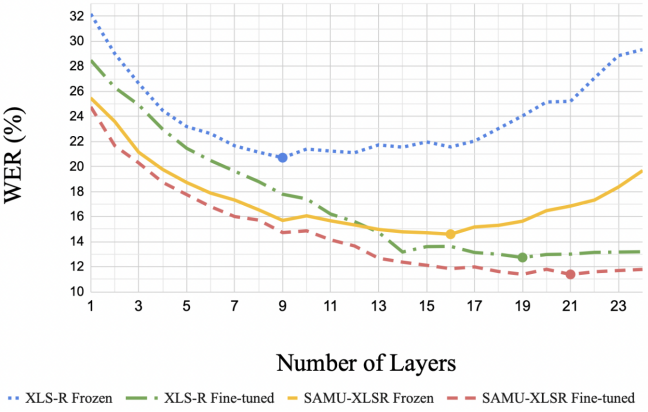

Fig. 3. Layer-wise analysis of WER on the test data.

Figure 3 illustrates how the linguistic information is encoded through each layer of both encoders. First, we observe that in terms of WER, SAMU-XLSR gets better results than XLS-R. We also can see that the minimum WER is achieved with higher layers for SAMU-XLSR than it is for XLS-R, both frozen and fine-tuned. We assume this is due to the fine-tuning made on SAMU-XLSR by forcing its highest representations to be aggregated to LaBSE's text embeddings.

Figure 4 presents results measured with the Concept Error Rate metric, relevant to the specific semantics of the downstream task.

We observe that the original frozen XLS-R model lost almost 7 points of CER between its best generated embeddings for semantic extraction task, layer 15, and its final generated embeddings, layer 24. On the other hand, since learning SAMU-XLSR consists on projecting its sentence-level embedding into the semantic multi-lingual LaBSE's encoding space, the highest layers of its encoder tend to capture and encode the semantics until the top layer. Both speech encoders give best CER results in middle layers. As expected, fine-tuning the speech encoders allows the models to extract as much semantic information as possible from the audio sig-

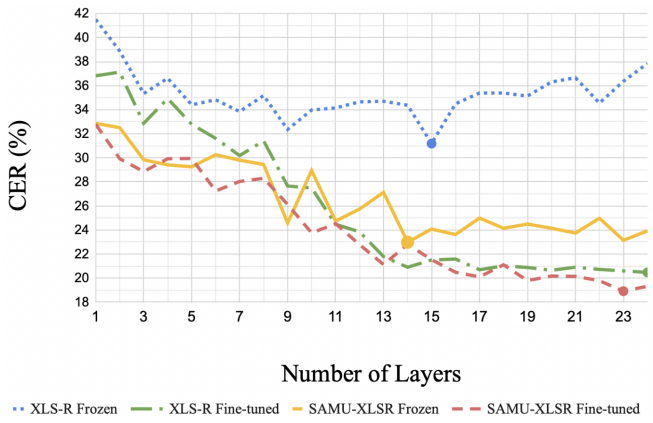

Fig. 4. Layer-wise analysis of Concept Error Rate on the test data.

nal. Even if the semantics extracted from the frozen SAMU-XLSR middle layers were already mostly kept through upper layers, performances were enhanced by fine-tuning the encoder.

5.2. Language portability

5.2.1. Zero shot

To evaluate the multilingual portability of the SAMU-XLSR encoder compared to the original XLS-R, we apply zero-shot learning on our French (MEDIA) and Italian (PortMEDIA) data. We first train each end-to-end SLU model on the French data, by freezing or fine-tuning the speech encoder, and then make a simple inference on the Italian data. We also aim to measure how fine-tuning the speech encoder on the French data impacts the language portability capabilities.

		ChER	WER	CER	CVER
XLS-R	Frozen	68.77	129.08	88.24	100.44
	Fine-T.	63.22	123.94	85.36	101.54
SAMU-XLSR	Frozen	49.35	100.13	54.62	99.83
	Fine-T.	59.10	124.49	83.45	101.63

Table 3. Zero-shot results (%) from French MEDIA training to Italian PortMEDIA inference.

Results in Table 3 show that the use of a frozen SAMU-XLSR speech encoder gives strongly better performance than other setups for concept recognition in these conditions: a CER of 54.62% is attained while with the other configurations performs at more than 83% error rate. We notice that, as expected, the performance related to the transcription itself is very bad: SAMU-XLSR is able to extract general semantics, but is not designed to provide language-dependent information useful to transcribe speech. It also appears that fine-tuning SAMU-XLSR on French degrades the capability of the module to generate good semantic embeddings on Italian.

5.2.2. Low resource

In these experiments, we exploit the training data in Italian to train the models. IT in Table 4 means the SLU model has been trained from scratch on the Italian data. FR→IT means the SLU model weights have been initialized with the French model before being trained on the Italian data.

	Train Data	ChER	WER	CER	CVER
Frozen	IT	14.91	36.90	42.66	54.31
	FR→IT	12.78	32.41	35.39	49.60
Fine-T.	IT	13.36	37.02	42.72	57.47
	FR→IT	7.55	20.01	26.92	40.11

Table 4. XLS-R PortMEDIA results (%) of PortMEDIA (IT) training and MEDIA training followed by PortMEDIA fine-tuning (FR→IT).

Tables 4 and 5 illustrate the potential of portability of both encoders from French to Italian, with or without fine-tuning. We can observe that using SAMU-XLSR as a speech encoder still outperforms XLS-R, with a CER of 33.01% (resp. 42.66% for XLS-R) without fine-tuning and without the use of French data. With both fine-tuning and use of French data, SAMU-XLSR is able to reach 26.18% of CER, but the gap with XLS-R – that reaches 26.92% – is less significant.

	Train Data	ChER	WER	CER	CVER
Frozen	IT	12.62	27.92	33.01	46.99
	FR→IT	11.01	25.09	26.90	42.70
Fine-T.	IT	6.47	16.59	30.66	42.09
	FR→IT	7.04	17.81	26.18	39.28

Table 5. SAMU-XLSR PortMEDIA results (%) with PortMEDIA (IT) training and MEDIA training followed by PortMEDIA fine-tuning (FR→IT).

5.3. Semantic analysis of sentence-level embeddings

In the previous experiments, we analyzed the contents of the sequence embeddings produced by XLS-R and SAMU-XLSR. In this section, we examine if the pooled sentence embeddings produced by SAMU-XLSR contain enough semantic information according to the MEDIA and PortMEDIA tasks, and we analyze their cross-modal and cross-lingual abilities. We simplify the MEDIA and PortMEDIA benchmark tasks to a bag-of-concepts classification task. For each segment, the system has to predict all the concepts that are

present in the speech segment. We use a multi-hot representation for the output. This simplification allows us to use the sentence embeddings without needing a complex decoder, that would bias our semantic analysis of the embeddings.

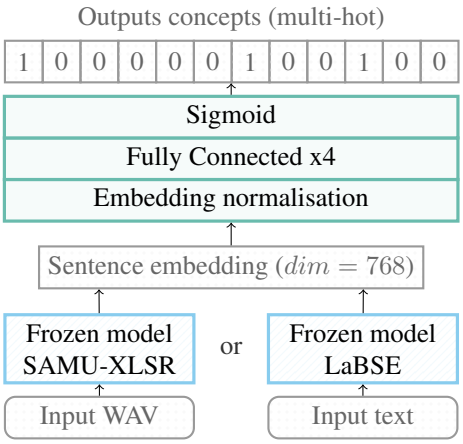

Fig. 5. Neural Architecture for an SLU language portability analysis of speech encoders with the MEDIA and Port-MEDIA datasets.

Figure 5 illustrates the simple architecture we implemented for this sentence-level analysis. We extract the sentence embeddings of either SAMU-XLSR or LaBSE depending on the model we analyse. We apply L_2-normalisation on the embeddings as follows: $\mathbf{x} \leftarrow \frac{\sqrt{d}}{||\mathbf{x}||}\mathbf{x}$, where $d = 768$, i.e. the dimension of the embeddings. Fixing the norm (both in training and evaluation) is critical; unnormalized embeddings with large norm generate many false positives, while unnormalized embeddings with small norm generate many false negatives. Note also that the norm of the SAMU-XLSR embeddings may not be that informative, since the network is trained using cosine similarity between SAMU-XLSR and LaBSE embeddings, which is a norm-invariant objective function. The network consists of four fully connected layers with ReLU activation functions and Dropout, and it is trained with weighted Binary Cross-Entropy loss.

In order to test both the cross-modal and cross-lingual properties of the embeddings, we train our classification models only on the French dataset. The Italian dataset is only used for testing, to obtain cross-lingual results. For the cross-modal properties, we trained a model on SAMU-XLSR (speech) embeddings, and tested it on both SAMU-XLSR and LaBSE (text) embeddings. We also trained a second model on LaBSE embeddings in order to observe the difference when testing it with SAMU-XLSR speech embeddings. The results, in terms of micro F_1-score are given in Table 6. We did not test the XLS-R embeddings, as this model only provides frame-level embeddings. We also report the frame-wise baseline results obtained in the previous experiments, by converting the sequence outputs of the models to a bag-of-concepts output. For these results, the model is trained to extract the sequence of concepts on the sequence of embeddings provided by SAMU-XLSR.

Test Data	**Test Encoder**	**Train Encoder**	
		SAMU-XLSR	LaBSE
FR	SAMU-XLSR	77.52%	71.77%
	LaBSE	78.04%	82.15%
	*frame-wise**	*84.69%*	-
IT	SAMU-XLSR	68.55%	65.14%
	LaBSE	62.05%	69.58%
	*frame-wise**	*59.76%*	-

Table 6. Micro F_1-scores for sentence-level semantic analysis with classification model trained on French and tested on French and Italian data. *Baseline results are obtained with the language portability models on frame-wise embeddings, and converted in bag-of-concepts outputs for evaluation.

We can observe that both LaBSE and SAMU-XLSR models obtain comparable results with their corresponding test embeddings when tested on the MEDIA dataset (77.52% for SAMU-XLSR, 82.15% for LaBSE). The capacity of SAMU-XLSR to reproduce sentence-level embeddings close to the LaBSE ones is noticeable, and validates the strategy used to train SAMU-XLSR.

Finally, by comparing the results from the previous experiments with the frame-wise embeddings and this sentence-level embedding analysis, we observe that the sentence-level embeddings are better in extracting cross-lingual representations than the frame-level ones.

6. CONCLUSION

In this work, we investigated the capacity of the recently introduced SAMU-XLSR in addressing a challenging SLU task. SAMU-XLSR is a speech encoder is a fine-tuned version of the XLS-R model, using LaBSE embeddings as targets. In addition to its promising performance, we demonstrated how this speech encoder differs from the XLS-R model in the way it encodes the semantic information in its intermediate hidden representations. We also showed the real potential of the SAMU-XLSR for language portability. Finally, we showed its capacity to build a sentence-level embedding able to highlight the semantic information of the task and its promising cross-lingual and cross-modal properties.

7. ACKNOWLEDGEMENTS

This work was performed using HPC resources from GENCI–IDRIS (grant 2022-AD011012565) and received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie ESPERANTO project (grant agreement N°101007666), through the SELMA project (grant N°957017) and from the French National Research Agency through the AISSPER project (ANR-19-CE23-0004).

8. REFERENCES

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

[2] Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin, "End-to-end named entity and semantic concept extraction from speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 692–699.

[3] Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters, "From audio to semantics: Approaches to end-to-end spoken language understanding," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 720–726.

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

[5] Thierry Desot, François Portet, and Michel Vacher, "SLU for voice command in smart home: comparison of pipeline and end-to-end approaches," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 822–829.

[6] Marco Dinarelli, Nikita Kapoor, Bassam Jabaian, and Laurent Besacier, "A data efficient end-to-end spoken language understanding architecture," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8519–8523.

[7] Martin Radfar, Athanasios Mouchtaris, and Siegfried Kunzmann, "End-to-end neural transformer based spoken language understanding," *arXiv preprint arXiv:2008.10984*, 2020.

[8] Elisavet Palogiannidi, Ioannis Gkinis, George Mastrapas, Petr Mizera, and Themos Stafylakis, "End-to-end architectures for ASR-free spoken language understanding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7974–7978.

[9] Jakob Poncelet, Vincent Renkens, et al., "Low resource end-to-end spoken language understanding with capsule networks," *Computer Speech & Language*, vol. 66, pp. 101142, 2021.

[10] Swapnil Bhosale, Imran Sheikh, Sri Harsha Dumpala, and Sunil Kumar Kopparapu, "End-to-End Spoken Language Understanding: Bootstrapping in Low Resource Scenarios.," in *Interspeech*, 2019, pp. 1188–1192.

[11] Antoine Caubrière, Natalia Tomashenko, Antoine Laurent, Emmanuel Morin, Nathalie Camelin, and Yannick Estève, "Curriculum-Based Transfer Learning for an Effective End-to-End Spoken Language Understanding and Domain Portability," in *Proc. Interspeech 2019*, 2019, pp. 1198–1202.

[12] Yinghui Huang, Hong-Kwang Kuo, Samuel Thomas, Zvi Kons, Kartik Audhkhasi, Brian Kingsbury, Ron Hoory, and Michael Picheny, "Leveraging unpaired text data for training end-to-end speech-to-intent systems," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7984–7988.

[13] Thierry Desot, François Portet, and Michel Vacher, "Corpus generation for voice command in smart home and the effect of speech synthesis on End-to-End SLU," in *12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 6395–6404.

[14] Loren Lugosch, Brett H Meyer, Derek Nowrouzezahrai, and Mirco Ravanelli, "Using speech synthesis to train end-to-end spoken language understanding models," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8499–8503.

[15] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, Minneapolis, Minnesota, June 2019, Association for Computational Linguistics.

[17] Gaëlle Laperriere, Bassam Jabaian, and Yannick Esteve, "Where Are We in Semantic Concept Extraction for Spoken Language Understanding?," in *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings*. Springer Nature, 2021, vol. 12997, p. 202.

[18] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6419–6423.

[19] Andy T Liu, Shang-Wen Li, and Hung-yi Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[20] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman 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.

[21] Pengwei Wang, Liangchen Wei, Yong Cao, Jinghui Xie, and Zaiqing Nie, "Large-scale unsupervised pre-training for end-to-end spoken language understanding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7999–8003.

[22] Yu-An Chung, Chenguang Zhu, and Michael Zeng, "SPLAT: Speech-Language Joint Pre-Training for Spoken Language Understanding," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1897–1907.

[23] Bhuvan Agrawal, Markus Müller, Samridhi Choudhary, Martin Radfar, Athanasios Mouchtaris, Ross McGowan, Nathan Susanj, and Siegfried Kunzmann, "Tie your embeddings down:

Cross-modal latent spaces for end-to-end spoken language understanding," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7157–7161.

[24] Markus Müller, Samridhi Choudhary, Clement Chung, Athanasios Mouchtaris, and Siegfried Kunzmann, "In Pursuit of Babel-Multilingual End-to-End Spoken Language Understanding," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 1042–1049.

[25] Sameer Khurana, Antoine Laurent, and James Glass, "SAMU-XLSR: Semantically-Aligned Multimodal Utterance-level Cross-Lingual Speech Representation," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[26] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al., "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[27] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang, "Language-agnostic BERT Sentence Embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 878–891.

[28] Frédéric Béchet and Christian Raymond, "Benchmarking Benchmarks: Introducing New Automatic Indicators for Benchmarking Spoken Language Understanding Corpora," *Proc. Interspeech 2019*, pp. 4145–4149, 2019.

[29] Fabrice Lefèvre, Djamel Mostefa, Laurent Besacier, Yannick Estève, Matthieu Quignard, Nathalie Camelin, Benoit Favre, Bassam Jabaian, and Lina 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'12)*, 2012, pp. 1436–1442.

[30] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[31] Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6147–6151.

[32] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Zhuo Chen, Peidong Wang, Gang Liu, Jinyu Li, Jian Wu, Xiangzhan Yu, et al., "Why does Self-Supervised Learning for Speech Recognition Benefit Speaker Recognition?," *arXiv preprint arXiv:2204.12765*, 2022.

[33] Manon Macary, Marie Tahon, Yannick Estève, and Anthony Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 373–380.

[34] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[35] Ramon Sanabria, Wei-Ning Hsu, Alexei Baevski, and Michael Auli, "Measuring the Impact of Individual Domain Factors in Self-Supervised Pre-Training," *arXiv preprint arXiv:2203.00648*, 2022.

[36] Collin F Baker, Charles J Fillmore, and John B Lowe, "The berkeley framenet project," in *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998.

[37] Hélene Bonneau-Maynard, Matthieu Quignard, and Alexandre Denis, "MEDIA: a semantically annotated corpus of task oriented dialogs in French," *Language Resources and Evaluation*, vol. 43, no. 4, pp. 329–354, 2009.

[38] H. Bonneau-Maynard, C. Ayache, F. Bechet, A. Denis, A. Kuhn, F. Lefevre, D. Mostefa, M. Quignard, S. Rosset, C. Servan, and J. Villaneau, "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, May 2006, European Language Resources Association (ELRA).