Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D3.4 Intermediate progress report on speech and natural language processing

Work Package	3
Responsible Partner	LIA
Author(s)	Yannick Estève
Contributors	Antoine Caubrière, Gaëlle Laperrière, Guntis Barzdins, Jarod Duret, Jean-François Bonastre, Marcely Zanon Boito, Natalia Tomashenko, Salima Mdhaffar, Titouan Parcollet, Roberts Dargis, Yannick Estève
Reviewer	Christoph Schmidt
Version	1.0
Contractual Date	31 December 2022
Delivery Date	22 December 2022
Dissemination Level	Public

Version History

Version	Date	Description
0.1	03/11/2022	ToC
0.2	15/11/2022	Initial Content
0.3	15/12/2022	Review Ready
1.0	22/12/2022	Publishable version

Executive Summary

This report presents the progress made during the second year in the SELMA project on speech and language processing. For speech processing, the research work focused mainly on the use of end-to-end neural models, especially based on model pretrained under self-supervision and, this year, on the use of some very recent evolutions of wav2vec 2.0 models like SAMU-XLSR (Semantically-Aligned Multimodal Utterance-level Cross-Lingual Speech Representation) for cross lingual transfer.

> For speech processing, the work focused mainly on the use of models trained under self-supervision, especially to address cross-lingual knowledge transfer

The SELMA project was strongly involved in the LeBenchmark initiative that permitted to pretrained wav2vec 2.0 models on 7K hours of speech in French language, and compare them to wav2vec 2.0 models pretrained on English-only data or multilingual data (containing 53 different languages).

These models have also been fine-tuned on downstream tasks directly related to the SELMA project: speech recognition, speech translation, semantic concept extraction from speech, named entity recognition from speech.

In parallel, during the first year, some baseline automatic speech recognition systems driven by hybrid Hidden Markov Model and Deep Neural Network (HMM/DNN) acoustic models have been developed for some languages (English, French, Latvian). Some of these ASR systems have been integrated to the SELMA platform as NLP components delivered as Docker containers.

A first speech synthesis engine has been built on Brazilian Portuguese broadcast news provided by Deutsche Welle.

During the second year, we prepared the data to pretrain a SELMA wav2vec 2.0, and offered solutions to deal with low resource scenario for spoken language understanding (SLU) and to port an SLU model from a language to another one.

Table of Contents

Executive Summary *3*

1. Introduction *7*

2. The LeBenchmark initiative: end-to-end speech recognition and translation based on speech unit representation learned through self-supervised training *9*

2.1 Background **10**

2.2 Gathering a Large and Heterogeneous Speech Collection in French **11**

2.3 Training and Sharing SSL Models **13**

3. LeBenchmark results on speech recognition, speech translation and other downstream tasks *15*

3.1 Automatic Speech Recognition (ASR) Results **15**

3.2 Automatic Speech Translation (AST) Results **16**

3.3 Spoken Language Understanding (SLU) Results **19**

3.4 Named Entity Recognition (NER) Results **22**

4 Speech Synthesis *25*

4.1 Architecture **25**

4.2 Data **25**

4.3 Evaluation **26**

5 Hybrid ASR system *28*

5.1 French ASR **28**

5.2 Latvian ASR **28**

5.3 English, German, Spanish, Arabic ASR **29**

6 Preparing the data to pretrain the SELMA multilingual wav2vec 2.0 model *30*

7 Low resource spoken language understanding scenario *33*

8 Language portability of spoken language understanding model *36*

9 Conclusion 38

10 References 39

Table of Figures

FIGURE 1 TRAINING SAMU-XLSR 36

Table of Tables

TABLE 1 STATISTICS FOR THE SPEECH CORPORA USED TO TRAIN SSL MODELS ACCORDING TO GENDER INFORMATION (MALE / FEMALE / UNKNOWN). THE SMALL DATASET IS FROM MLS ONLY. EVERY DATASET IS COMPOSED OF THE PREVIOUS ONE + ADDITIONAL DATA; MPF, TCOF AND CFPP2000 APPEAR TWICE WITH DIFFERENT STATS AS DATA EXTRACTION CHANGED; DURATION: HOUR(S):MINUTE(S) 12

TABLE 2 HYPERPARAMETERS OF OUR PRE-TRAINED SSL MODELS 14

TABLE 3 ASR RESULTS (WER%) ON COMMON VOICE AND ETAPE CORPORA, WITH PRE-TRAINED WAV2VEC2.0 MODELS FURTHER FINE-TUNED ON LABELED ASR DATA. GRAY NUMBERS INDICATE 95% CONFIDENCE INTERVALS COMPUTED USING BOOTSTRAP RE-SAMPLING AS PROPOSED IN BISANI AND NEY, 2004 16

TABLE 4 BLEU ON VALID AND TEST SETS OF MULTILINGUAL TEDX (MTEDX). THE HIGHEST VALUE IN EACH GROUP (TASK-AGNOSTIC PRE-TRAINING, TASK-SPECIFIC SELF-SUPERVISED, AND SUPERVISED FINE-TUNING) IS UNDERLINED WHILE THE BEST VALUE IN EACH COLUMN IS HIGHLIGHTED IN BOLD. GRAY NUMBERS DENOTE THE STANDARD DEVIATION COMPUTED USING BOOTSTRAP RE-SAMPLING (KOEHN ET AL. 2004) 19

TABLE 5 END-TO-END SLU DECODING RESULTS (CONCEPT ERROR RATE %) ON THE MEDIA CORPUS 22

TABLE 6 END-TO-END NER DECODING RESULTS (ENTITY ERROR RATE %) ON THE QUAERO DATASET 24

TABLE 7 REPARTITION OF UTTERANCES AND HOURS PER SPEAKER 26

TABLE 8 STATISTICS OF RAW DATA SHARED BY DEUTSCHE WELLE TO BE USED TO PRETRAINED A MULTILINGUAL WAV2VEC 2.0 MODEL 31

TABLE 9 EVALUATION IN NEER (%) OF OUR APPROACH TO TRAIN AN END-TO-END NER MODEL WITHOUT PAIRED TRAINING DATA COMPARED TO OTHER APPROACHES USING SPEECH SYNTHESIS, AND COMPARED TO THE IDEAL SCENARIO WHEN PAIRED DATA IS AVAILABLE 35

1. Introduction

Work Package 3 aims to develop and make advances in state-of-the-art natural language processing technologies, with a special focus on speech processing. In the last decade, such technologies have made considerable progress through the emergence of the deep learning paradigm, but in many tasks, these approaches are still far from solving the most relevant research questions.

One very current hot topic in the speech and language research community is the use of models pretrained by self-supervision. Such deep neural models are trained on huge amounts of unlabeled data. The BERT model, which is dedicated to text processing, has been introduced by Google (Devlin 2019, https://arxiv.org/abs/1810.04805) in 2018: the main state-of-art systems for any NLP tasks are based on the use of deep neural models derived from BERT. The use of BERT-like models consists of first pretraining a model through self-supervised learning on a very huge amount of unlabeled data and then fine-tuning it on (small) in-domain labeled data by supervised learning.

Such an approach has been proposed for speech processing with the introduction of the wav2vec models in 2019 by Facebook (Schneider 2019, https://arxiv.org/abs/1904.05862). Significant improvements were proposed in 2020 with the wav2vec 2.0 models (Baevski 2020, https://arxiv.org/abs/2006.11477): it was shown that it is possible to reach low word error rates (<10%) by exploiting only 10 minutes of manually transcribed speech (audiobook), after pretraining on 960 hours of untranscribed audio.

Pretraining such model needs a lot of computation power and lot of questions are still open about their robustness to acoustic conditions and languages. In the framework of the SELMA project, we brought strong efforts during this first year to master this approach and to pretrain French wav2vec 2.0 models and fine-tuned them into several downstream task. This work was made in association to external partners (University of Grenoble-Alpes, France) and was possible thanks to the use of the French Jean Zay supercomputer. Some convincing results are presented in this report and, taking benefit from this experience, a SELMA model dedicated to Deutsche Welle multilingual broadcast news audio is under construction: during Y2 we have collected and prepared the training data that are described in this report.

In addition to this study on wav2vec 2.0 models, we work on speech synthesis on Deutsche Welle data (from Brazilian Portuguese broadcast news): our first architecture is presented in this report (the same architecture as Y1), that has been kept during Y2 for new experiments.

We also built more classical hybrid HMM/DNN ASR systems that have been integrated into the SELMA platform.

During the second year, we also proposed a new approach for low resource scenario in the context of named entity recognition from speech through an end-to-end neural approach – a scientific publication has been submitted, accepted and presented at Interspeech 2022 (Mdhaffar et al., 2022).

Taking benefit to the work made during Y1 in the SELMA project, we also proposed new contributions on language portability of end-to-end models dedicated to semantic extraction from speech – a scientific publication has been submitted and accepted at the IEEE Workshop on Spoken Language Technologies (Laperrière et al., 2023).

2. The ***LeBenchmark*** initiative: end-to-end speech recognition and translation based on speech unit representation learned through self-supervised training

Self-Supervised Learning (SSL) based on huge amounts of unlabeled data has been explored successfully for image and natural language processing (Bachman et al., 2019; Chen et al., 2020; Devlin et al., 2018; Raffel et al., 2019). Recently, researchers investigated SSL from speech as well and successfully improved performance on downstream tasks such as speech recognition (Baevski et al., 2019; Kawakami et al., 2020).

As SSL from speech is a rapidly evolving domain, new models are unfortunately evaluated on different datasets, most of which focus on the English language. In order to carefully assess the progress of speech SSL model-wise and application-wise, common benchmarks are needed. While NLP benchmarking is now widely discussed (Ruder, 2021), multi-task benchmarks are less common in speech despite the fact that the field has a long tradition of evaluation (see for instance long-term NIST and DARPA shared tasks for ASR).

In our papers Evain et al., 2021-A and Evain et al., 2021-B, we propose to contribute to this by providing a reproducible and multifaceted benchmark for evaluating speech SSL models. By *benchmark*, and following the definition of Schlangen, 2021, we mean an ensemble of tasks that allow to discriminate learners (*i.e.,* SSL models) based on their ability to perform well on those tasks.

We propose an initial set of four main tasks (10 sub-tasks overall), measuring specific speech challenges in the French language: Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), Speech Translation (AST), and Emotion Recognition (AER). In this document, we present the main results obtained for the first three tasks, while also including results on Named Entity Recognition (NER). The totality of our results can be found in the original papers (Evain et al., 2021-A and Evain et al., 2021-B), as well as in the website's leader board: http://lebenchmark.com .

In summary, our investigation enables to assess the impact of pre-trained speech models that differ along several dimensions: language used for pre-training (French, English, multilingual), amount of raw speech used for SSL pre-training (1k, 3k, or 7k hours), model size (base, large). For reproducibility, we also provide pre-trained SSL models learned on a large and heterogeneous collection of speech utterances and believe this is a strong contribution to speech technologies in French.

2.1 Background

SSL has been recently proposed as an interesting alternative for data representation learning, as it requires no annotated data. Such learned representations have been very successful in computer vision (Bachman et al., 2019; Chen et al., 2020), and language (Devlin et al., 2018, Peters et al., 2018). SSL from speech consists of resolving *pseudo-tasks*, which do not require human annotation, as a pre-training for the real tasks to solve. These *pseudo-tasks* target predicting the next samples, or solving ordering problems. For instance, Autoregressive Predictive Coding (APC) considers the sequential structure of speech and predicts information about a future frame (Chung et al., 2019; Chung and Glass, 2020-A), whereas Contrastive Predictive Coding (CPC) distinguishes a future speech frame from distractor samples (Baevski et al., 2019, Schneider et al., 2019), which is an easier learning objective compared to APC. Such representations have been shown to improve performance in several speech tasks (Chung and Glass, 2020-B), while being less sensitive to domain and/or language mismatch (Kawakami et al., 2020) and being transferable to other languages (Riviere et al., 2020).

In 2020, a strong speech SSL baseline appeared: the Wav2Vec2.0 model (Baevski et al., 2020) which relies on the CPC idea of Baevski et al., 2019 and Schneider et al., 2019 but with *discrete* speech units that are used as latent representations and fed to a Transformer network to build contextualized representations. Several other bi-directional encoders were also proposed recently: Speech-XLNet (Song et al., 2019), Mockingjay (Liu et al., 2019) and Wang et al., 2020. A few recent studies were also related to multilingual SSL models trained on very large multilingual corpora (Conneau et al., 2020, Wang et al., 2021).

While there are multiple evaluation benchmarks to assess pre-trained models in NLP (for instance *lue* for English, *flue* for French, and *klue* for Korean), we are aware of only one similar initiative for speech SSL model evaluation: the Speech processing Universal PERformance

Benchmark (SUPERB) (Yang et al., 2021) which however targets English only and does not share pre-trained SSL models as we do.

2.2 Gathering a Large and Heterogeneous Speech Collection in French

Recently, large multilingual corpora that include French have been made available, such as MLS (Pratap et al., 2020, 1,096 hours) and Voxpopuli (Wang et al., 2021, +4,500 hours). However, these are restricted to either read or well-prepared speech, failing to provide diversity in the speech samples, such as accented, spontaneous and/or affective speech.

We gathered a large variety of speech corpora in French that cover:

- Different accents: MLS (Pratap et al., 2020), African Accented Speech (SLR57), CaFE (Gournay et al., 2018);
- Acted emotions: GEMEP (Bänziger et al., 2012), CaFE (Gournay et al., 2018), Att-Hack (Le Moine et al., 2020);
- Telephone dialogues : PORTMEDIA (Lefèvre et al., 2012);
- Read sentences: MLS (Pratap et al., 2020), African Accented French (SLR57), MaSS (Boito et al., 2020);
- Spontaneous sentences: CFPP2000 (Branca-Rosoff et al., 2012), ESLO2 (Eshkol-Taravella et al., 2012), MPF (ORTOLANG-MPF), TCOF (ORTOLANG-TCOF), NCCFr (Torreira et al., 2010);
- Broadcast speech: EPAC (Estève et al., 2010);
- Professional speech: Voxpopuli (Wang et al., 2021).

Compared to MLS and Voxpopuli, our dataset is more diverse, carefully sourced and contains detailed metadata (speech type, and speaker gender). Moreover, compared to these, it has a more realistic representation of speech turns in real life. Statistics are reported in Table 1.

Corpus$_{License}$	# Utterances	Duration	# Speakers	Mean Utt. Duration	Speech type
		Small dataset – 1K			
MLS French$_{CCBY4.0}$	**263,055** 124,590 / 138,465 / –	**1,096:43** 520:13 / 576:29 / –	**178** 80 / 98 / –	**15 s** 15 s / 15 s / –	Read
		Medium dataset – 3K			
African Accented French$_{Apache2.0}$	**16,402** 373 / 102 / 15,927	**18:56** – / – / 18:56	**232** 48 / 36 / 148	**4 s** – / – / –	Read
Att-Hack$_{CCBYNCND}$	**36,339** 16,564 / 19,775 / –	**27:02** 12:07 / 14:54 / –	**20** 9 / 11 / –	**2.7 s** 2.6 s / 2.7 s / –	Acted Emotional
CaFE$_{CCNC}$	**936** 468 / 468 / –	**1:09** 0:32 / 0:36 / –	**12** 6 / 6 / –	**4.4 s** 4.2 s / 4.7 s / –	Acted Emotional
CFPP2000$_{CCBYNCSA}$*	**9853** 166 / 1,184 / 8,503	**16:26** 0:14 / 1:56 / 14:16	**49** 2 / 4 / 43	**6 s** 5 s / 5 s / 6 s	Spontaneous
ESLO2$_{NC}$	**62,918** 30,440 / 32,147 / 331	**34:12** 17:06 / 16:57 / 0:09	**190** 68 / 120 / 2	**1.9 s** 2 s / 1.9 s / 1.7 s	Spontaneous
EPAC**$_{NC}$	**623,250** 465,859 / 157,391 / –	**1,626:02** 1,240:10 / 385:52 / –	**Unk** – / – / –	**9 s** – / – / –	Radio Broadcasts
GEMEP$_{NC}$	**1,236** 616 / 620 / –	**0:50** 0:24 / 0:26 / –	**10** 5 / 5 / –	**2.5 s** 2.4 s / 2.5 s / –	Acted Emotional
MPF	**19,527** 5,326 / 4,649 / 9,552	**19:06** 5:26 / 4:36 / 9:03	**114** 36 / 29 / 49	**3.5 s** 3.7 s / 3.6 s / 3.4 s	Spontaneous
PORTMEDIA$_{NC}$ (French)	**19,627** 9,294 / 10,333 / –	**38:59** 19:08 / 19:50 / –	**193** 84 / 109 / –	**7.1 s** 7.4 s / 6.9 s / –	Acted telephone dialogue
TCOF (Adults)	**58,722** 10,377 / 14,763 / 33,582	**53:59** 9:33 / 12:39 / 31:46	**749** 119 / 162 / 468	**3.3 s** 3.3 s / 3.1 s / 3.4 s	Spontaneous
Medium dataset total	**1,111,865** 664,073 / 379,897 / 67,895	**2,933:24** 1,824:53 / 1,034:15 / 74:10	-	-	-
		Large dataset – 7K			
MaSS	**8,219** 8,219 / – / –	**19:40** 19:40 / – / –	**Unk** – / – / –	**8.6 s** 8.6 s / – / –	Read
NCCFr$_{NC}$	**29,421** 14,570 / 13,922 / 929	**26:35** 12:44 / 12:59 / 00:50	**46** 24 / 21 / 1	**3 s** 3 s / 3 s / 3 s	Spontaneous
Voxpopuli$_{CC0}$ *Unlabeled*	**568,338** – / – / –	**4,532:17** – / – / 4,532:17	**Unk** – / – / –	**29 s** – / – / –	Professional speech
Voxpopuli$_{CC0}$ *transcribed*	**76.281** – / – / –	**211:57** – / – / 211:57	**327** – / – / –	**10 s** – / – / –	Professional speech
Large dataset total***	**1,814,242** 682,322 / 388,217 / 99,084	**7,739:22** 1,853:02 / 1,041:07 / 4,845:07	-	-	-

*Composed of audio files not included in the CEFC corpus v2.1, 02/2021; **speakers are not uniquely identified.; ***Stats of CFPP2000, MPF and TCOF have changed a bit due to a change in data extraction; License: CC=Creative Commons; NC=non-commercial; BY= Attribution; SA= Share Alike; ND = No Derivative works; CC0 = No Rights Reserved

Table 1 *Statistics for the speech corpora used to train SSL models according to gender information (male / female / unknown). The small dataset is from MLS only. Every dataset is composed of the previous one + additional data; MPF, TCOF and CFPP2000 appear twice with different stats as data extraction changed; duration: hour(s):minute(s)*

- **Pre-processing for SSL training:** Recordings were segmented using time stamps from transcriptions. We retrieved, when available, speaker labels and gender information. Following Baevski et al., 2020, we removed utterances shorter than 1s, and longer than 30s. When possible, overlapping speech sentences were also removed. When necessary, audio segments were converted to mono PCM 16bits, 16kHz.
- **Small dataset (approximately 1k hours):** It is only composed of the MLS corpus for comparison with Wav2Vec2.0 Baevski et al., 2020 which uses only read English speech. It is also gender balanced.

- **Medium dataset (approximately 3k hours):** It includes 2,933 hours of speech, from which 1,115 hours is read speech, 1,626 hours broadcast speech, 123 hours spontaneous speech, 38 hours acted telephone dialogues, and 29 hours acted emotional speech. Regarding gender, we collected 1,824 hours of speech from male speakers, 1,034 hours from female speakers, and 74 hours from unknown gender.
- **Large dataset (approximately 7.7k hours):** It has 4 additional corpora: MaSS, NCCFr and Voxpopuli (unlabeled + transcribed). It includes 7,739 hours of speech, from which 1,135 hours is read speech, 1,626 hours broadcast speech, 165 hours spontaneous speech, 38 hours acted telephone dialogues, 29 hours acted emotional speech, and 4744 hours professional speech. Except for NCCFr, no info about gender is given in the added datasets.

2.3 Training and Sharing SSL Models

The *LeBenchmark* provides seven Wav2Vec2.0 models pretrained on the gathered French data described above. Following Baevski et al., 2020, two different Wav2Vec2.0 architectures (*large* and *base*) are coupled with our small (1K), medium (3K) and large (7K) corpora to form our set of Wav2Vec2.0 models: W2V2-Fr-1K-base, W2V2-Fr-1K-large, W2V2-Fr-3K-base, W2V2-Fr-3K-large, W2V2-Fr-7K-base, W2V2-Fr-7K-large.

Hyperparameters and architectures for base and large are identical to the ones first introduced in Baevski et al., 2020. *W2V2-Fr-1K*, *W2V2-Fr-3K* and *W2V2-Fr-7K* are trained respectively for 200K, 500K, 500K and 500K updates on 4, 32, 32 and 64 Nvidia Tesla V100 (32GB), with one update corresponding to a call to the *.backward()* function in PyTorch. Detailed summary of the hyperparameters used to train our SSL models can be found in Table 2. In practice, training is stopped at a round number of updates once the loss observed on the development set of the MLS corpus reaches a stable point. Pre-trained Wav2Vec2.0 models are shared with the community via HuggingFace for further integration with well-known toolkits such as SpeechBrain, Fairseq or Kaldi.

Pre-existing Wav2Vec2.0 models obtained from Fairseq are also considered in downstream experiments. First, *XLSR-53-large* is used as a comparison to multilingual models. Then, *W2V2-En-base* and *W2V2-En-large* (LS960) are used to assess English representations from

LibriSpeech. For the sake of conciseness, we remove the prefix W2V2- in all our results tables in the next section.

Model	Training Data	Transformer Blocks	Model Dimension	Inner Dimension	Heads	Updates
Fr-1K-base	1,096 h	12	768	3,072	8	200K
Fr-1K-large	1,096 h	24	1024	4,096	16	200K
Fr-3K-base	2,933 h	12	768	3,072	8	500K
Fr-3K-large	2,933 h	24	1024	4,096	16	500K
Fr-7K-base	7,739 h	12	768	3,072	8	500K
Fr-7K-large	7,739 h	24	1024	4,096	16	500K

Table 2 *Hyperparameters of our pre-trained SSL models*

3. *LeBenchmark* results on speech recognition, speech translation and other downstream tasks

We benchmark SSL models on four different tasks: Automatic Speech Recognition (ASR), Speech Language Understanding (SLU), Automatic Speech Translation (AST), and Named Entity Recognition (NER). Since our goal is to evaluate the impact of SSL for the best baselines for each task addressed, we have a different architecture for each task, and it corresponds to the best baseline performance we could obtain using MFCC/MFB features. As a different architecture/approach is used for each task, we evaluate the different SSL models as feature extractors for these tasks. These 'SSL extractors' are either 'task agnostic' or 'task specific' (SSL models fine-tuned on the task data), as further explained below.

3.1 Automatic Speech Recognition (ASR) Results

Automatic Speech Recognition (ASR) consists in transcribing the content of a speech utterance. In this section, we present ASR results using an end-to-end model and two datasets. Results focus on larger Wav2vec2.0 models (3K and 7K), as these are the ones for which we notice the most expressive improvements.

- **Datasets:** The ASR tasks target two different types of corpora: Common Voice (Ardila et al. 2020) and ETAPE (Gravier et al. 2012). Common Voice is a very large crowd-sourced corpus (477 hours) of read speech in French with transcripts (train: 428h, dev: 24h, and test: 25h), while ETAPE is a smaller (36 hours) but more challenging corpus composed of diverse French TV broadcast programs (train: 22h, dev: 7h, and test: 7h).

- **Architecture:** Our models are implemented with the SpeechBrain toolkit (Ravanelli et al., 2021). The baseline system is fed by 80-dimension log Mel filterbank (MFB) features and is based on an encoder/decoder architecture with attention. When used with an SSL pre-trained Wav2Vec2.0 model, the system simply adds an additional hidden layer and an output layer on top of a Wav2Vec2.0 architecture.

- **Results:** Table 3 presents the results achieved with ASR systems on French Common Voice 6.1 and on ETAPE. Before the use of Wav2vec2.0 models for ASR, the baseline MFB-based system (first line) was the state-of-the-art e2e model on CommonVoice/French. Other lines of the table present different Wav2vec2.0 models fine-tuned on labeled ASR data from CommonVoice or ETAPE. Wav2vec2.0 *base* and *large* models provided by ***LeBenchmark*** outperform clearly *En-large* and *XLSR-53-large* models. The best model is *Fr-3K-large*, pretrained on a smaller training dataset than *Fr-7K-large*, and it provides the best results on all the experiments.

-

Corpus	**CommonVoice**		**ETAPE**	
Features	**Dev**	**Test**	**Dev**	**Test**
MFB	17.67 (0.37)	20.59 (0.41)	54.03 (1.33)	54.36 (1.32)
En-large	12.05 (0.23)	14.17 (0.52)	42.14 (0.72)	44.82 (0.74)
XLSR-53-large	16.41 (0.27)	19.40 (0.29)	58.55 (0.65)	61.03 (0.70)
Fr-3K-base	11.25 (0.23)	13.22 (0.24)	26.14 (0.70)	28.86 (0.79)
Fr-3K-large	**8.34** (0.18)	**9.75** (0.20)	**23.51** (0.68)	**26.14** (0.77)
Fr-7K-base	10.84 (0.21)	12.88 (0.24)	25.13 (0.68)	28.16 (0.79)
Fr-7K-large	8.55 (0.18)	9.94 (0.21)	24.14 (0.70)	27.25 (0.78)

Table 3 *ASR results (WER%) on Common Voice and ETAPE corpora, with pre-trained Wav2vec2.0 models further fine-tuned on labeled ASR data. Gray numbers indicate 95% confidence intervals computed using bootstrap re-sampling as proposed in Bisani and Ney, 2004*

3.2 Automatic Speech Translation (AST) Results

Automatic speech-to-text translation (AST) consists in translating a speech utterance in a source language to a text in a target language. In this work, we are interested in translating directly from French speech to text in another language.

- **Dataset:** We selected subsets having French as the source in the multilingual TEDx dataset (Salesky et al., 2021). Our benchmark covers translation directions from French to three target languages: English (*en*), Spanish (*es*), and Portuguese (*pt*), with the following training sizes: 50h (*en*), 38h (*es*), and 25h (*pt*).

- **Experiments:** Our baselines are models using 80-dimensional MFB features. For learned representations derived from SSL models, we focused on the feature extraction approach where features are extracted from either task-agnostic or task-specific pre-training. Task-agnostic pre-training refers to the direct use of SSL models as feature extractors whereas the task-specific method consists of one additional phase where the SSL models are further trained on the in-domain task data, with (supervised fine-tuned) or without (self-supervised fine-tuned) labels.

  We performed supervised fine-tuning with speech transcriptions as labels and leave supervised fine-tuning with AST data for future work. In the task-specific scenario, we only considered three SSL models: two best French SSL models (*Fr-3K-large* and *Fr-7K-large*) and one best non-French SSL model (*XLSR-53-large*). Since the French speech is overlapped between the language pairs, we selected the pair having the most speech data (fr-en) to perform task-specific pre-training and used the obtained models to extract features for the remaining pairs (fr-es and fr-pt). For a fair comparison, we did not use additional data augmentation technique nor ASR encoder pre-training in the experiments.

- **Architecture:** We used a small Transformer (Vaswani et al., 2017) architecture having 6 layers of encoders, 3 layers of decoders, and hidden dimension 256 in all experiments. Following previous work (Nguyen et al. 2020; Evain et al. 2021-A), we inserted a block of Linear-ReLU before convolutional layers in the speech encoder for parameter efficiency and model performance reasons.

- **Results:** Table 4 displays the results of the AST experiments. One can observe that SSL features, whether task-agnostic or task-specific and whether being pre-trained on English, French, or multilingual data, outperform the baselines using MFB features by a large margin (except for the task-agnostic multilingual model XLSR-53 on the two pairs fr-es and fr-pt, which are in very low-resource settings).

Comparing blocks: Among the three groups using SSL features (task-agnostic pre-training, task-specific self-supervised, and task-specific fine-tuned for ASR), the ASR fine-tuning approach (c) yields the best results. We observe considerable improvements from task-specific self-supervised (b) to task-specific fine-tuned (c) (+6.19, +8.50, +8.53 on average for en, es, and pt, respectively) while the benefits of using self-supervised fine-tuning compared to task-agnostic pre-training are only marginal or even slightly negative.

The substantial gains when using the supervised fine-tuning approach (even with the somehow indirect signal of transcripts for the AST downstream task) shows that giving more signals of the task-specific data to the SSL models is helpful. In particular, in the case of task-specific self-supervised fine-tuning (b), we further trained the SSL models for 20k more steps on the raw task-specific data, whereas in ASR fine-tuned scenario (c), we used raw data plus the transcripts to guide the SSL models.

Task-agnostic SSL: Focusing on task-agnostic block (a), we see that French SSL models clearly outperform those pre-trained on English and multilingual data. Multilingual XLSR-53 model surpasses the English models on fr-en, yet all of them fail to generate meaningful translations on fr-es and fr-pt where little training data is available.

Comparing across different French SSL model sizes (base vs large), the large architecture yields considerable improvements (nearly 3 to 6 BLEU points) over its base counterpart. When looking into the French SSL models with different amounts of pre-training data (1K, 3K, and 7K), we observe large gains for the base architecture from using 1K to using 3K or more pre-training data. There is, however, no significant difference between base models using 3K and 7K data. Using 7K data even hurts the performance on the pair fr-pt. On the other hand, for the large network, using more data consistently improves the performance on all language pairs.

Task-specific SSL: Finally, moving on to task-specific models, Fr-7K-large is the best-performing model (or being on par with the best one) in each group. Noticeably, there is a huge improvement when using the ASR fine-tuning approach (c) for the multilingual XLSR-53 model. The method considerably boosts the performance of the

multilingual model (compared to using it directly or further pre-training it on the task data) and makes it even on par with the best French SSL models.

Features	Valid			Test		
	en	es	pt	en	es	pt
MFB	1.15 (0.27)	0.67 (0.15)	0.61 (0.13)	1.10 (0.14)	0.87 (0.12)	0.32 (0.03)
(a) Task agnostic pre-training						
En-base	5.54 (0.27)	1.30 (0.17)	0.54 (0.11)	5.20 (0.28)	1.47 (0.15)	0.38 (0.05)
En-large	4.11 (0.25)	1.67 (0.20)	0.32 (0.03)	3.56 (0.22)	2.29 (0.18)	0.43 (0.05)
Fr-3K-base	15.05 (0.49)	13.19 (0.25)	4.44 (0.29)	14.80 (0.47)	14.27 (0.44)	4.72 (0.25)
Fr-3K-large	17.94 (0.51)	16.40 (0.49)	8.64 (0.34)	18.00 (0.51)	18.12 (0.48)	9.55 (0.36)
Fr-7K-base	15.13 (0.45)	12.78 (0.40)	2.65 (0.20)	14.50 (0.45)	13.61 (0.44)	2.66 (0.23)
Fr-7K-large	19.23 (0.54)	17.59 (0.49)	9.68 (0.37)	19.04 (0.53)	18.24 (0.49)	10.98 (0.41)
XLSR-53-large	7.81 (0.33)	0.49 (0.13)	0.43 (0.07)	6.75 (0.29)	0.52 (0.08)	0.36 (0.05)
(b) Task specific pre-training (self-supervised on mTEDx)						
Fr-3K-large	18.54 (0.53)	16.40 (0.48)	8.81 (0.36)	18.38 (0.52)	17.84 (0.48)	10.57 (0.41)
Fr-7K-large	19.65 (0.55)	17.53 (0.47)	9.35 (0.36)	19.36 (0.54)	18.95 (0.53)	10.94 (0.38)
XLSR-53-large	6.83 (0.33)	0.54 (0.14)	0.34 (0.03)	6.75 (0.32)	0.34 (0.03)	0.29 (0.03)
(c) Task specific pre-training (fine-tuned for ASR on mTEDx)						
Fr-3K-large	21.09 (0.53)	19.28 (0.53)	14.40 (0.47)	21.34 (0.58)	21.18 (0.52)	16.66 (0.49)
Fr-7K-large	**21.41** (0.51)	20.32 (0.49)	**15.14** (0.48)	**21.69** (0.58)	**21.57** (0.52)	**17.43** (0.52)
XLSR-53-large	21.09 (0.54)	**20.38** (0.56)	14.56 (0.45)	20.68 (0.53)	21.14 (0.55)	17.21 (0.54)

Table 4 *BLEU on valid and test sets of multilingual TEDx (mTEDx). The highest value in each group (task-agnostic pre-training, task-specific self-supervised, and supervised fine-tuning) is underlined while the best value in each column is highlighted in bold. Gray numbers denote the standard deviation computed using bootstrap re-sampling (Koehn et al. 2004)*

3.3 Spoken Language Understanding (SLU) Results

Spoken Language Understanding (SLU) aims at extracting a semantic representation from a speech signal in human-computer interaction applications (De Mori, 1997). Given the difficulty

of creating an open-domain SLU application, many works focus on specific domains. We focus on the hotel information and reservation domain provided within the French corpus MEDIA (Bonneau Maylard et al., 2006; Quarteroni et al., 2009).

- **Dataset:** The MEDIA corpus is made of 1~250 human-machine dialogues acquired with a *Wizard-of-Oz* approach, where 250 users followed 5 different reservation scenarios. Spoken data were manually transcribed and annotated with domain concepts, following a rich ontology. The official corpus split is made up of 12,908 utterances (41.5 hours) for training, 1,259 utterances (3.5 hours) for development and 3,005 utterances (11.3 hours) for test. We note that, while all turns have been manually transcribed and can be used to train ASR models, only user turns have been annotated with concepts and can be used to train SLU models. This results in only 41.5 hours of speech training data for ASR models, and only 16.8 hours for SLU models.

- **Architecture:** All our models are based on LSTM (Hochreiter and Schmidhuber, 1997) seq2seq with attention (Bahdanau et al., 2014), being similar to the one proposed in previous works (Dinarelli et al., 2017; Dinarelli et al., 2020, Evain et al., 2021-A). In particular we use a similar speech encoder employing a pyramidal hierarchy of RNN layers like Chan et al., 2016 and Evains et al., 2021.

  The decoder has been also improved, integrating two attention mechanisms: one as usual for attending the encoder's hidden states; the other for attending all previous decoder prediction's embeddings, instead of the previous prediction only like in the original LSTM-based encoder-decoder models (Bahdanau et al., 2014). Our model is implemented using the *Fairseq* library (Ott et al., 2019).

- **Experiments:** We use a total of 3 bidirectional LSTM layers of size 256 stacked in a pyramidal fashion in our encoder and the LSTM decoder has 2 layers of size 256. In addition to using spectrogram features and features from task agnostic SSL models, we also use features from task specific models (SLU on MEDIA). Two types of task-specific pre-training are performed: *self-supervised* which consists in resuming the SSL model training using the MEDIA training data and minimizing the Wav2Vec 2.0 loss ((b) self-supervised on MEDIA in the results table, also called task-adaptive pre-training in Gururangan et al., 2020); and *ASR supervised* ((c) fine-tuned for ASR on MEDIA in

the results table) which consists in fine-tuning the full SSL model for a supervised downstream task with a CTC loss minimization objective (Graves et al., 2006).

Finally, in this work we chose to fine-tune models with respect to the ASR task on MEDIA (not the SLU one) to see how it compares to self-supervised fine-tuning. We leave fine-tuning with respect to SLU for future work.

- **Results:** The results for SLU obtained with different speech representations are shown in Table 5. They are given in terms of Concept Error Rate (CER), computed the same way as Word Error Rate (WER) but on concept sequences. CER are accompanied by standard deviations (in gray), computed with the bootstrap method of Bisani and Ney, 2004 .

  We first note that our *spectrogram* baseline obtains a substantial improvement over the one in Evain et al., 2021-A . Such gain is due to the slightly different settings and model architecture. Using SSL model features as input resulted in an impressive drop in CER, even when using English SSL models (CER from 31.10 to 20.84 on the test set with the *base* model).

  Task-agnostic SSL: At best, among task-agnostic pre-trained models, we achieve a CER of 15.95 on the test data with *Fr-3K-large* features. Surprisingly, using features from the model trained with 7k hours of speech (*Fr-7K-large*), results are worse on both dev and test. In contrast, we also evaluated these models in terms of ASR performance, finding that the 7k-model led to the best results.

  Task-specific SSL: We performed task-specific pre-training only with the most effective SSL models: French 3k and 7k models and multi-lingual *XLSR-53-large*. The best overall pre-trained model is the 7k-model fine-tuned for ASR on MEDIA, though results are close to those obtained with features from the 3k-model (13.97 vs. 13.78). Indeed, our significance tests confirm that these two models are equivalent and they are significantly better than all the others. This shows that pre-trained SSL speech models can be specialized using task specific pre-training with either self-supervised learning on raw speech (block (b) in the table), or fine-tuning on raw speech and associated transcripts (block (c) in the table), the latter being slightly better than the former.

Features	Dev	Test
Spectrogram from Evain et al., 2021-A	33.63 (1.28)	34.76 (0.83)
spectrogram	**29.07** (1.31)	**31.10** (0.83)
(a) Task agnostic pre-training		
En-base	22.38 (1.24)	20.84 (0.68)
En-large	23.31 (1.31)	25.26 (0.77)
Fr-1K-base	22.89 (1.26)	23.27 (0.76)
Fr-1K-large	20.10 (1.10)	20.66 (0.72)
Fr-3K-base	19.44 (1.11)	18.56 (0.67)
Fr-3K-large	**15.96** (1.02)	**15.95** (0.62)
Fr-7K-base	20.70 (1.07)	18.86 (0.68)
Fr-7K-large	17.25 (1.02)	16.35 (0.66)
XLSR-53-large	18.45 (1.15)	18.78 (0.66)
(b) Task specific pre-training (self-supervised on MEDIA)		
Fr-3K-large	15.93 (1.01)	**14.94** (0.60)
Fr-7K-large	**15.42** (1.03)	15.17 (0.60)
XLSR-53-large	16.77 (1.09)	15.56 (0.61)
(c) Task specific pre-training (fine-tuned for ASR on MEDIA)		
Fr-3K-large	**14.49** (1.06)	13.97 (0.59)
Fr-7K-large	14.58 (1.01)	**13.78** (0.58)
XLSR-53-large	16.05 (1.05)	15.46 (0.60)

Table 5 *End-to-end SLU decoding results (Concept Error Rate %) on the MEDIA corpus*

3.4 Named Entity Recognition (NER) Results

Named Entity Recognition (NER) aims to locate and classify named entity mentions in speech transcripts into pre-defined categories (such as person names, organizations, locations, …).

- **Dataset:** The QUAERO data has been developed during the research project QUAERO (2008-2013). It consists in the manual annotation of named entities of the manual transcription of the ESTER1 corpus. ESTER1 Graves et al., 2004 is an evaluation campaign focusing on the evaluation of orthographic transcription, event detection and tracking, and information extraction. An official QUAERO test dataset has also been added. This entire corpus is composed of data recorded from French radio and TV stations between 1998 and 2004. The official corpus split is made up of 93.5 hours for training and 6.5 hours for testing. Named Entities often include seven major groups: person, location, organization, amount, time, production and function. Within the framework of the QUAERO project, an extended named entity annotation with compositional and hierarchical structure has been proposed (Galibert et al., 2011). The QUAERO dataset does not contain a development dataset. So, we use the ETAPE development part. ETAPE is a French dataset composed of data recorded from French radio and TV stations between 2010 and 2011. It is annotated with the same pre-defined categories of entities used in the QUAERO annotation.
- **Architecture:** Our model is based on end-to-end approaches. The end-to-end system is composed of a large pre-trained French wav2vec model (LeBenchmark Fr-7K-large), a linear hidden layer of 1024 units, and a softmax output layer. The loss function used for the supervised fine-tuning step is the Connectionist Temporal Classification (CTC) loss function (Graves et al., 2006).
- **Results:** The obtained results for NER are shown in Table 6. They are given in terms of Entity Error Rate (EnER), computed in the same way as Word Error Rate (WER) but only on entity sequences, exactly like the Concept Error Rate used for SLU. We compute the total EnER '*All Entities*' and an EnER for each entity category. We report also results in term of WER for the transcription without entities. The results are obtained by using the flat version of the named entity representation retained in the QUAERO dataset (i.e., not a structured representation).

Features	**Test**
(a) Word Error Rate (WER)	
WER	10.9%
(b) Entity Error Rate (EnER)	
All Entities	32.24%
Entity 'Person'	27.60%
Entity 'function'	52.84%
Entity 'organisation'	46.24%
Entity 'location'	27.09%
Entity 'production'	70.87%
Entity 'amount'	24.66%
Entity 'time'	28.8%

Table 6 *End-to-end NER decoding results (Entity Error Rate %) on the QUAERO dataset*

4 Speech Synthesis

Text to speech (TTS), or speech synthesis, which aims to synthesize intelligible and natural speech given text, is a hot research topic in speech, language, and machine learning communities. Thanks to the advances in deep learning and artificial intelligence, neural network-based TTS has significantly improved the quality of synthesized speech in recent years.

In this section, the neural network-based architecture developed in SELMA for our first text-to-speech engine is presented, in addition to the data used for the training process. Last, we discuss about how our work on speech synthesis applied to Brazilian Portuguese broadcast news could be evaluated.

4.1 Architecture

Our TTS system consists of two components, an acoustic model and a vocoder. The acoustic model generates acoustic features from linguistic features (text in our case), and the vocoder synthesizes waveform from the acoustic features.

For the acoustic model, we conducted experiments with several architectures. This allowed us to draw the following conclusion: purely in terms of quality and naturalness Tacotron 2 [Shen et al, 2018] + DDC gave us the best performance. Other architectures like GlowTTS [Kim et al., 2020], SpeedySpeech [Vainer and Dusek, 2020] or FastSpeech [Ren et al., 2019] are faster and synthesize intelligible speech but not as good as Tacotron 2.

Considering the vocoder, we also had multiple choices, we mainly worked on two architectures: Hifi-Gan [Kong et al., 2020] and WaveRNN [Kalchbrenner et al., 2018]. The first one did not give us the expected results, so we have decided to go for the second one. From the paper, there is not a significant difference between the two in terms of speech quality, the main difference is about inference time. Since we have no inference real-time constraints, this is not a problem.

4.2 Data

We use the audio news bulletins that are produced by DW's Brasil department to train the speech synthesis engine. The audio files have been downloaded from Youtube and the scripts were

retrieved from github in a repository with all the text scripts that DW uses to produce their weekday news podcasts.

The dataset contains approximately 32 hours of speech from 8 speakers. The repartition of utterances and hours per speaker after cleaning is described below in Table 7.

#	Name	Training utterances	Hours
1	Roberto	3510	8.5
2	Alexandre	3348	7.7
3	Philip	2759	6.0
4	Leila	2077	5.1
5	Bruno	679	1.7
6	Marcio	554	1.3
7	Clarissa	357	0.9
8	Renate	295	0.7

Table 7 *Repartition of utterances and hours per speaker*

4.3 Evaluation

Currently, we are still working on the evaluation part of the speech synthesis engine. The evaluation protocol can be divided into two parts. First, we will evaluate the accuracy of the speech synthesis using a speech recognition model.

Using the original transcription and the output of an ASR model, we can compute the Word Error Rate (WER) which is a common metric for measuring speech-to-text accuracy of automatic speech recognition systems.

As this first evaluation protocol doesn't measure the prosodic aspect of the TTS system, we have to introduce a second one involving human rating. The next step will be to organize a perceptual evaluation campaign where samples are rated by humans on a scale from 1 to 5 with 0.5-point increments, from which a subjective mean opinion score (MOS) is calculated.

A Mean Opinion Score (MOS) is a numerical measure of the human-judged overall quality of an event or experience. In telecommunications, a Mean Opinion Score is a ranking of the quality of voice and video sessions.

A demo webpage with our first system here: click here to access our TTS demonstration webpage

5 Hybrid ASR system

Classical hybrid automatic speech recognition systems are based on HMM/DNN acoustic models of phonemes, a dictionary of words with their explicit pronunciations (sequence of phonemes), and language models.

Kaldi is a popular open-source toolkit designed to build such ASR systems. In SELMA, we implemented ASR systems for different languages using Kaldi, mainly to be integrated into the SELMA platform as the first ASR components.

5.1 French ASR

A Kaldi-based ASR system has been built for the French language. The acoustic models (AM) are trained on 40-dimensional high-resolution (hires) MFCC features with a state-of-the-art factorized time delay neural network (TDNN-F) architecture (Povey et al., 2018 ; Peddinti et al., 2015) on 300 hours of French Broadcast data with manual transcriptions. The acoustic model was trained using lattice-free maximum mutual information (LF-MMI) (Povey et al., 2016) and cross-entropy criteria. Speed and volume perturbation have been applied for data augmentation (Ko et al., 2015). The word error rate got on Broadcast News data not included in the training data is around 17.5%.

5.2 Latvian ASR

The baseline ASR system for Latvian is trained using the Kaldi framework. The acoustic model has been trained on a general-domain Latvian speech corpus containing 100 hours of broadcast recordings (Pinnis at al., 2014) augmented with various noisy recordings and musical recordings from the MUSAN corpus (Snyder, 2015). The TDNN+LSTM neural network is trained on 40-dimension FBANK vectors. Language models (LM) are trained using the SRILM toolkit (Stolcke, 2002). Trigram language models pruned to 1e-8 are used in all experiments. The LM is trained on the Latvian portion of the CommonCrawl. A rule-based system is used to generate the pronunciation lexicon based on 52 phonemes. The word error rate (WER) is measured on 22 minutes of various radio and TV broadcasts and is around 10.5%.

5.3 English, German, Spanish, Arabic ASR

Kaldi-based ASR systems for English, German, Spanish and Arabic have been developed by various partners (University of Edinburgh, IDIAP, QCRI) within the H2020 SUMMA project (Grant agreement: 688139) and released publicly afterwards.

These legacy systems have been adapted for use in the SELMA project as baseline ASR systems, although technical incompatibility with the latest Kaldi versions and high WER around 20% on broadcast news limit the scope of their use.

6 Preparing the data to pretrain the SELMA multilingual wav2vec 2.0 model

Thanks to Deutsche Welle, the SELMA project has access to a high amount of multilingual speech data, that are audio or video documents related to news. Thanks to our experience coming from the LeBenchmark initiative and from the literature (Hsu et al., 2021) we expect that a wav2vec 2.0 pretrain on in-domain data (here, journalistic data) will get a better performance on this kind of data. So, we plan to pretrain a such model.

In order to pretrain by self-supervision a wav2vec, we collected a huge amount of multilingual data provided by the Deutsche Welle partner. Statistics of the raw data are presented in Table 8:

Language	Type	Items	Duration	Date range
Arabic	Audio	151	108 h	2021-08-06 > 2022-07-31
	Video	6,15	965 h	2021-08-06 > 2022-07-31
	All	**6,301**	**1,073 h**	
Brazilian	Audio	963	90 h	2018-07-24 > 2022-07-30
	Video	1,957	209 h	2018-07-24 > 2022-07-30
	All	**2,92**	**299 h**	
Chinese	Audio	640	279 h	2012-11-18 > 2022-07-31
	Video	2,922	129 h	2012-11-18 > 2022-07-31
	All	**3,562**	**408 h**	
Dari	Audio	455	48 h	2017-08-15 > 2022-12-06
	Video	721	85 h	2017-08-15 > 2022-12-06
	All	**1,176**	**133 h**	
English	Article	903	2 h	2021-01-02 > 2022-07-31
	Audio	1,063	433 h	2021-01-02 > 2022-07-31
	All	**1,966**	**435 h**	
French	Audio	1,852	661 h	2020-06-05 > 2022-07-31
	Video	974	92 h	2020-06-05 > 2022-07-31
	All	**2,826**	**752 h**	
German	Audio	99	34 h	2022-04-01 > 2022-07-31
	Video	1,463	244 h	2022-04-01 > 2022-07-31
	All	**1,562**	**279 h**	
Greek	Audio	1,234	52 h	2013-06-05 > 2022-12-11

	Video	1,333	69 h	2013-06-05 > 2022-12-11
	All	**2,567**	**121 h**	
Hausa	Audio	10,763	6,252 h	2013-12-10 > 2022-07-31
	Video	675	32 h	2013-12-10 > 2022-07-31
	All	**11,438**	**6,284 h**	
Hindi	Audio	50	6 h	2013-01-06 > 2022-07-31
	Video	3,985	297 h	2013-01-06 > 2022-07-31
	All	**4,035**	**303 h**	
Indonesian	Video	3,019	232 h	2013-06-10 > 2022-07-31
	All	**3,019**	**232 h**	
Pashto	Audio	1,06	89 h	2013-06-28 > 2022-12-13
	Video	847	89 h	2013-06-28 > 2022-12-13
	All	**1,907**	**177 h**	
Persian	Audio	1,824	375 h	2012-02-27 > 2022-07-31
	Video	2,41	112 h	2012-02-27 > 2022-07-31
	All	**4,234**	**487 h**	
Polish	Audio	46	8 h	2012-12-20 > 2022-12-13
	Video	2,414	139 h	2012-12-20 > 2022-12-13
	All	**2,46**	**148 h**	
Russian	Audio	98	33 h	2011-07-26 > 2022-07-31
	Video	12,98	1,281 h	2011-07-26 > 2022-07-31
	All	**13,078**	**1,314 h**	
Spanish	Audio	70	110 h	2021-01-01 > 2022-07-31
	Video	7,109	951 h	2021-01-01 > 2022-07-31
	All	**7,179**	**1,061 h**	
Turkish	Audio	5,257	713 h	2011-08-05 > 2022-07-31
	Video	7,485	672 h	2011-08-05 > 2022-07-31
	All	**12,742**	**1,385 h**	
Ukrainian	Video	9,233	551 h	2012-12-05 > 2022-07-29
	All	**9,233**	**551 h**	
Urdu	Audio	2,769	302 h	2012-05-29 > 2022-10-14
	All	**2,769**	**302 h**	
All	**All**	**94,974**	**15,743 h**	

Table 8 *Statistics of raw data shared by Deutsche Welle to be used to pretrained a multilingual wav2vec 2.0 model*

These files have been processed in order to extract only speech segments and to specify the gender of the speaker involved for each speech segment. By the way, we aim to build a

gender-balanced and language-balanced pretraining data: we kept a maximum of 250 hours of speech for each language.

The first SELMA multilingual wav2vec 2.0 is training. The model will be available in 2/3 weeks and first experiments will be carried out to evaluate this model. The SpeechBrain toolkit, interfaced to the HuggingFace *transformers* library is used for this SSL training.

Notice that this model will be released under a (free and) very permissive licence in order to contribute to the advances of the research community.

7 Low resource spoken language understanding scenario

In our low resource SLU scenario, an end-to-end model for ASR and a corpus of textual documents with named entity annotations but without the corresponding audios are available.

Our approach (Mdhaffar et al., 2022) is based on the use of an external model trained to generate a sequence of vectorial representations from text. These representations mimic the hidden representations that could be generated inside an end-to-end automatic speech recognition model by processing a speech signal. A SLU neural module is then trained to use these representations as input and the annotated text as output. Last, the SLU module replaces the top layers of the ASR model to achieve the construction of the end-to-end model.

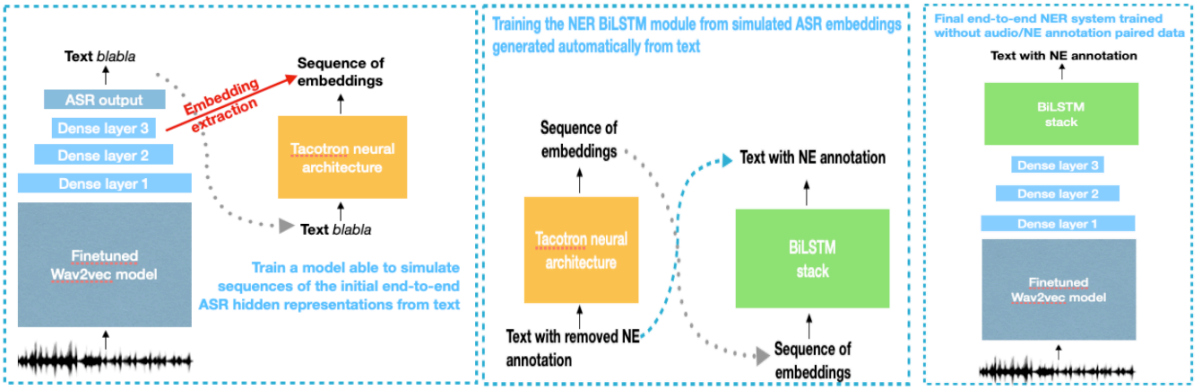
To generate the simulated ASR hidden representations (or ASR embeddings), we train a sequence-to-sequence neural model, called *Text-to-ASR-Embeddings* model. Such an approach can be compared to propositions in literature that use synthetic voices to feed an ASR end-to-end model.

We motivated our proposition for different reasons. First, the use of synthetic speech introduces some artifacts in the input of the ASR model. If the ASR model is fine-tuned on such synthetic voices, these artifacts will degrade the capability of the model to process natural voices. A solution to avoid this consists of freezing the weights of the bottom layers and only update the weights of the higher layers, in which the semantic is better encoded. Since the bottom layers were optimized to process natural speech, the quality of the embeddings computed from synthetic speech is not guaranteed, and can introduce a gap between embedding computed from natural and computed from synthetic speech.

With our approach, we aim to reduce this gap. In addition, our approach needs less computation at training time than the ones based on synthetic speech, since we avoid the use of a consequent number of lower layers.

To train this *Text-to-ASR-Embeddings* neural model, we must produce a training dataset composed of pairs of transcriptions, used as input, and sequences of ASR embeddings, used as output. To produce this training dataset, the end-to-end ASR model is used to transcribe its

training dataset. For each transcribed utterance, we extract a sequence of ASR embeddings from a hidden layer, and associate this ASR embedding sequence to the automatic transcription. When the entire ASR training data has been processed, the ASR embedding sequences and their associated automatic transcriptions are used to train the *Text-to-ASR-Embeddings* model, as illustrated in (A) in the following figure.



(A) (B) (C)

At this stage, we obtain a module able to simulate ASR embeddings from text. Our objective is then to train a neural SLU sub-module able to convert such a sequence of ASR embedding into an automatic transcription with SLU annotation, like annotation of named entities.

For this purpose, we exploit the textual dataset with semantic annotation. For each sentence in this dataset, we first remove the semantic annotation to keep only the sequence of words. Thanks to the *Text-to-ASR-Embeddings* model, we transform this sequence of words to a sequence of ASR embeddings (B). We iterate this process for all the annotated sentences in the semantic textual dataset. We get a set of pairs composed of a sequence of ASR embeddings and the corresponding text sequence of words semantically annotated. Once the entire textual dataset has been processed, we use this data to train an SLU sub-module able to generate a sequence of words semantically annotated from a sequence of ASR embeddings (B).

Finally, we plug the end-to-end ASR and the SLU sub-module (C). In order to merge the ASR model with the SLU sub-module, we keep all the ASR hidden layers needed to generate the ASR embeddings that can be mimicked by the *Text-to-ASR-Embeddings* model. The mimicked hidden layer is then connected to the SLU sub-module.

The final model is an end-to-end model able to transcribe and extract semantic information from speech, while no real paired training data exists.

Our approach, based on artificial ASR embeddings generated from text, exhibits highly promising results outperforming alternative approaches based on the use of synthetic speech. These results, computed in terms of Name Entity Error Rate (NEER) are presented in the table 9.

Training data	**Dev**	**Test**
ASR embeddings simulation (ours)	47.6	39.1
Synthetic speech (all weights are updated)	65.2	62.7
Synthetic speech (frozen speech encoder)	86.4	92.5
Oracle (real audio)	45.9	34.1

Table 9 *Evaluation in NEER (%) of our approach to train an end-to-end NER model without paired training data compared to other approaches using speech synthesis, and compared to the ideal scenario when paired data is available*

We consider that this approach can be extended to similar SLU tasks such as slot filling, and opens new perspectives in different use cases where enriching or adapting the linguistic knowledge captured by an end-to-end ASR model is needed.

8 Language portability of spoken language understanding model

SAMU-XLSR is based on the pre-trained multilingual XLSR 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.

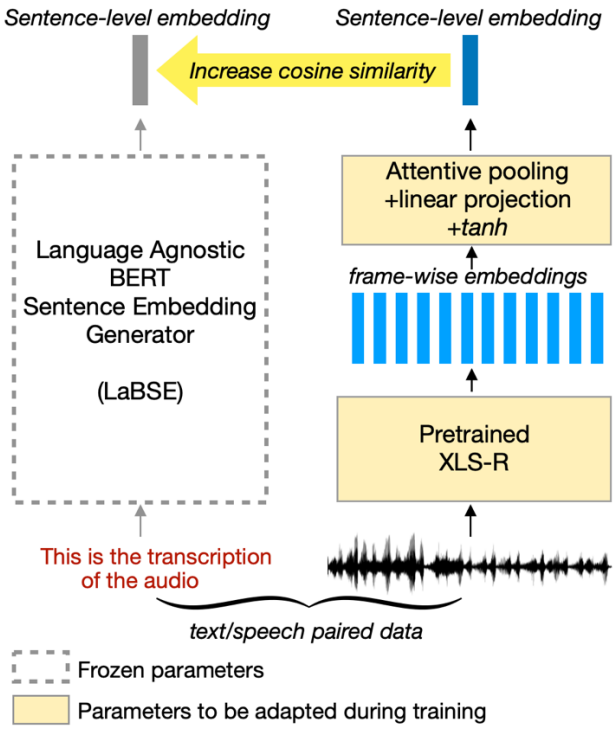


Figure 1 *Training SAMU-XLSR*

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 (Feng et al., 2022). The LaBSE model 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.

During Y2, we investigate the use of SAMU-XLSR in order to train a Spoken Language Understanding neural model in French language and transfer it to the Italian language for which the amount of annotated data related to the SLU task (extraction of semantic concepts/values for hotel reservation task-oriented human/machine dialogue) is very low (less than 8 hours) or zero. In the zero shot scenario (the model is trained on French and evaluated on Italian), our SAMU-XLSR-based model can get a Concept Error Rate (CER) of 54.6% while a classical XSLR model get 85.3%. When a few data in Italian is available, the gain in CER is less significant (26.2% instead of 26.9%), but the gain in Word Error Rate is promising (17.8% instead of 20%).

More details are available in (Laperrière et al., 2023).

9 Conclusion

During the first year of the SELMA project, the use of self-supervised pre-training for end-to-end speech processing tasks has been promising investigated with state-of-the-art results for different tasks like automatic speech recognition, speech translation, spoken language understanding, including named entity recognition from speech.

Such approaches will be extended to other languages, and we are now (end of Y2) pretraining a multilingual SELMA models by self-supervision, in order to investigate cross lingual transfer and domain dependence of such models. News components (for speaker recognition, speech synthesis, end-to-end speech recognition, speech translation) for different languages will be deployed in the SELMA platform.

During the second year, we propose solutions accepted by the international research community on low resource scenario for end-to-end named entity recognition from speech, and language portability.

10 References

A

Ardila et al. 2020 https://arxiv.org/abs/1912.06670

B

Bachman et al., 2019 https://arxiv.org/abs/1906.00910

Baevski et al., 2019 https://arxiv.org/abs/1911.03912

Baevski et al., 2020, https://arxiv.org/abs/2006.11477

Bahdanau et al., 2014 https://arxiv.org/abs/1409.0473

Bänziger et al., 2012 https://pubmed.ncbi.nlm.nih.gov/22081890/

Bisani and Ney, 2004 https://ieeexplore.ieee.org/document/1326009

Boito et al., 2020 https://arxiv.org/abs/1907.12895

Bonneau Maylard et al., 2006 https://aclanthology.org/L06-1385/

Branca-Rosoff et al., 2012 http://cfpp2000.univ-paris3.fr/CFPP2000.pdf

C

Chan et al., 2016 https://ieeexplore.ieee.org/document/7472621

Chen et al., 2020 https://arxiv.org/abs/2002.05709

Chung et al., 2019 https://arxiv.org/abs/1904.03240

Chung and Glass, 2020-A https://arxiv.org/abs/2004.05274

Chung and Glass, 2020-B https://arxiv.org/abs/1910.12607

Conneau et al., 2020 https://arxiv.org/abs/2006.13979

D

Devlin et al., 2018 https://arxiv.org/abs/1810.04805

De Mori, 1997 https://www.elsevier.com/books/spoken-dialogue-with-computers/de-mori/978-0-12-209055-4

Dinarelli et al., 2017 https://hal.archives-ouvertes.fr/hal-01553830v1;

Dinarelli et al., 2020 https://arxiv.org/abs/2002.05955

E

Eshkol-Taravella et al., 2012 https://halshs.archives-ouvertes.fr/halshs-01163053/document

Estève et al., 2010 https://aclanthology.org/L10-1442/

Evain et al., 2021-A https://arxiv.org/abs/2104.11462

Evain et al., 2021-B https://openreview.net/forum?id=TSvj5dmuSd

F

Feng et al., 2022 https://arxiv.org/pdf/2007.01852.pdf

G

Gournay et al., 2018

https://www.researchgate.net/publication/326022359_A_canadian_french_emotional_speech_dataset

Graves et al., 2006 https://dl.acm.org/doi/10.1145/1143844.1143891

Gravier et al. 2012 https://aclanthology.org/L12-1270/

Gururangan et al., 2020 https://arxiv.org/abs/2004.10964

H

Hochreiter and Schmidhuber, 1997 https://dl.acm.org/doi/10.1162/neco.1997.9.8.1735

Hsu et al., 2021 Hsu et al., 2021

K

Kalchbrenner et al., 2018 https://arxiv.org/abs/1802.08435

Kawakami et al., 2020 https://arxiv.org/abs/2001.11128

Koehn et al., 2004 https://aclanthology.org/W04-3250/

Khurana et al. 2022 https://arxiv.org/abs/2205.08180

Kim et al, 2020 https://arxiv.org/abs/2005.11129

Kong et al., 2020 https://arxiv.org/abs/2010.05646

L

Laperrière et al, 2022 https://arxiv.org/pdf/2210.05291.pdf

Lefèvre et al., 2012 https://hal.archives-ouvertes.fr/hal-01434925

Le Moine et al., 2020 https://arxiv.org/abs/2004.04410

Liu et al., 2019 https://arxiv.org/abs/1910.12638

M

Mdhaffar et al., 2022 https://www.isca-speech.org/archive/pdfs/interspeech_2022/mdhaffar22_interspeech.pdf

N

Nguyen et al., 2020 https://hal.archives-ouvertes.fr/hal-02962186

O

ORTOLANG-MPF https://hdl.handle.net/11403/mpf/v3

ORTOLANG-TCOF https://hdl.handle.net/11403/tcof/v2.1

Ott et al., 2019 https://arxiv.org/abs/1904.01038

P

Peters et al., 2018 https://arxiv.org/abs/1802.05365

Peddinti et al., 2015 https://www.isca-speech.org/archive_v0/interspeech_2015/papers/i15_3214.pdf

Pinnis et al., 2014 https://aclanthology.org/L14-1257/

Povey et al., 2011 https://www.danielpovey.com/files/2011_asru_kaldi.pdf

Povey et al., 2016 https://www.isca-speech.org/archive_v0/Interspeech_2016/pdfs/0595.PDF

Povey et al., 2018 https://www.isca-speech.org/archive_v0/Interspeech_2018/pdfs/1417.pdf

Pratap et al., 2020 https://arxiv.org/abs/2012.03411

Q

Quarteroni et al., 2009 http://www.marcodinarelli.it/NewSite/styles/publications/Interspeech09-Ontology.pdf

R

Raffel et al., 2019 https://arxiv.org/abs/1910.10683

Ravanelli et al., 2021 https://arxiv.org/abs/2106.04624

Ren et al., 2019 https://arxiv.org/abs/1905.09263

Riviere et al., 2020 https://arxiv.org/abs/2002.02848

Ruder, 2021 https://ruder.io/nlp-benchmarking/

S

Salesky et al., 2021 https://arxiv.org/abs/2102.01757

Schlangen, 2021 https://arxiv.org/abs/2007.04792

Schneider et al., 2019 https://arxiv.org/abs/1904.05862

Shen et al., 2018 https://arxiv.org/abs/1712.05884

SLR57, https://www.openslr.org/57/

Snyder, 2015 https://arxiv.org/abs/1510.08484

Song et al., 2019 https://arxiv.org/abs/1910.10387

Stolcke, 2002 http://www.speech.sri.com/projects/srilm/papers/icslp2002-srilm.pdf

T

Torreira et al., 2010 https://hal.archives-ouvertes.fr/hal-00608402

V

Vainer and Dusek, 2020 https://arxiv.org/abs/2008.03802

Vaswani et al., 2017 https://arxiv.org/abs/1706.03762

W

Wang et al., 2020 https://arxiv.org/abs/2001.10603

Wang et al., 2021 https://arxiv.org/abs/2101.00390

Y

Yang et al., 2021 https://arxiv.org/abs/2105.01051