INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech

Solène Evain[1,*], Ha Nguyen[1,2,*], Hang Le[1,*], Marcely Zanon Boito[1,*], Salima Mdhaffar[2,*], Sina Alisamir[1,3], Ziyi Tong[1], Natalia Tomashenko[2], Marco Dinarelli[1,*], Titouan Parcollet[2,*], Alexandre Allauzen[4], Yannick Estève[2], Benjamin Lecouteux[1], François Portet[1], Solange Rossato[1], Fabien Ringeval[1], Didier Schwab[1] and Laurent Besacier[1,5]

[1]Univ. Grenoble Alpes, CNRS, Inria, G-INP, LIG, France
[2]LIA, Avignon Université, France
[3]Atos, Échirolles, France
[4]ESPCI, CNRS LAMSADE, PSL Research University, France
[5]Naver Labs Europe, France

yannick.esteve@univ-avignon.fr, francois.portet@univ-grenoble-alpes.fr

Abstract

Self-Supervised Learning (SSL) using huge unlabeled data has been successfully explored for image and natural language processing. Recent works also investigated SSL from speech. They were notably successful to improve performance on downstream tasks such as automatic speech recognition (ASR). While these works suggest it is possible to reduce dependence on labeled data for building efficient speech systems, their evaluation was mostly made on ASR and using multiple and heterogeneous experimental settings (most of them for English). This questions the objective comparison of SSL approaches and the evaluation of their impact on building speech systems. In this paper, we propose *LeBenchmark*: a reproducible framework for assessing SSL from speech. It not only includes ASR (high and low resource) tasks but also spoken language understanding, speech translation and emotion recognition. We also focus on speech technologies in a language different than English: French. SSL models of different sizes are trained from carefully sourced and documented datasets. Experiments show that SSL is beneficial for most but not all tasks which confirms the need for exhaustive and reliable benchmarks to evaluate its real impact. *LeBenchmark* is shared with the scientific community for reproducible research in SSL from speech.

Index Terms: Self-Supervised Representation Learning, ASR, SLU, Speech Translation, Automatic Emotion Recognition.

1. Introduction

Self-Supervised Learning (SSL) based on huge unlabeled data has been explored successfuly for image processing [1, 2] and Natural Language Processing (NLP) [3]. Recently, pioneering work investigated SSL from speech, and successfully improved performance on downstream tasks such as speech recognition in low-resource scenarios [4, 5]. One observation that can be made about those recent studies on SSL for speech is that, as common benchmarks are not experimented, comparison of different SSL approaches are difficult to make. In addition, contributions have mostly been done on English, with a few recent studies related to multilingual SSL [6, 7]. We propose to remedy these shortcomings by providing a reproducible benchmark[1] that includes:

- a large and heterogeneous collection of French speech utterances (read, prepared, and spontaneous);
- pre-trained SSL models learnt on collections of 1k and 3k hours of French speech;
- assessments on Speech Recognition (ASR), Spoken Language Understanding (SLU), Speech Translation (AST) and Emotion Recognition (AER) in French.

2. Background

Most deep learning methods highly rely on large quantities of labeled training data. Particularly, current acoustic models require thousands of hours of transcribed speech to achieve state-of-the-art performance. However, this requirement cannot be fullfiled by the majority of the nearly 7,000 languages spoken worldwide. To overcome this, SSL has been recently proposed as an interesting alternative for data representation learning, as it requires less or no annotated data. Such learnt representations have been very successful in vision [1, 2] and NLP [3, 22]. Self-supervised learning 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 [23, 24], whereas Contrastive Predictive Coding (CPC) distinguishes a future speech frame from distractor samples [4, 25], which is an easier learning objective compared to APC. Such representations have been shown to improve performance in several speech tasks [26], while being less sensitive to domain and/or language mismatch [5]. It has also been shown that features extracted through a CPC pre-training can be transfered to other languages, with performance being on par or superior to a supervised pre-training [27].

3. Gathering a Large and Heterogeneous Speech Collection in French

Recently, large multilingual corpora that include French have been made available, such as MLS [17] (1,096 h), or voxpopuli [7] (+4,500 h). However, these are restricted to either read or well-prepared speech, failing to provide diversity in

* equal contribution
[1] https://github.com/LeBenchmark/

Table 1: *Statistics for the speech corpora used to train SSL models according to gender information (male / female / unknown). The small dataset (1k hours) is from MLS only, and the medium dataset (2.9k hours) is from all of them; duration: hour(s):minute(s).*

Corpus	# Utterances	Duration	# Speakers	Mean Utt. Duration	Speech type
African Accented French [8]	**16,402** 373 / 102 / 15,927	**18:56** – / – / 18:56	**232** 48 / 36 / 148	**4 s** – / – / –	Read
Att-Hack [9]	**36,339** 16,564 / 19,775 / 0	**27:02** 12:07 / 14:54 / 0:00	**20** 9 / 11 / 0	**2.7 s** 2.6 s / 2.7 s / –	Acted Emotional
CaFE [10]	**936** 468 / 468 / 0	**1:09** 0:32 / 0:36 / 0:00	**12** 6 / 6 / 0	**4.4 s** 4.2 s / 4.7 s / –	Acted Emotional
CFPP2000* [11] [12]	**9853** 166 / 1,184 / 8,503	**16:26** 0:14 / 1:56 / 14:16	**49** 2 / 4 / 43	**6.0 s** 5.0 s / 5.0 s / 6.0 s	Spontaneous
ESLO2 [13], [14]	**62,918** 30,440 / 32,147 / 331	**34:12** 17:06 / 16:57 / 0:09	**190** 68 / 120 / 2	**1.9 s** 2.0 s / 1.9 s / 1.7 s	Spontaneous
EPAC** [15]	**623,250** 465,859 / 157,391 / 0	**1,626:02** 1,240:10 / 385:52 / 0:00	**1,935** – / – / –	**9 s** – / – / –	Radio Broadcasts
GEMEP [16]	**1,236** 616 / 620 / 0	**0:50** 0:24 / 0:26 / 0:00	**10** 5 / 5 / 0	**2.5 s** 2.4 s / 2.5 s / –	Acted Emotional
MLS French [17]	**263055** 124,590 / 138,465 / 0	**1,096:43** 520:13 / 576:29 / 0:00	**178** 80 / 98 / 0	**15.0 s** 15.0 s / 15.0 s / –	Read
MPF [18], [19]	**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 (French) [20]	**19,627** 9,294 / 10,333 / 0	**38:59** 19:08 / 19:50 / 0:00	**193** 84 / 109 / 0	**7.1 s** 7.4 s / 6.9 s / –	Acted telephone dialogue
TCOF (Adult s) [21]	**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
ALL	**1,111,865** 664,073 / 379,897 / 67,895	**2,933:18** 1,824:53 / 1034:15 / 74:10	-	-	-

*version without the CEFC corpus v2.1, 02/2021; **speakers are not uniquely identified.

the speech samples, such as accented, spontaneous and/or affective speech. As a consequence, SSL models trained only on these corpora may present poor generalisation abilities on spontaneous or affective speech. In this work, we gathered a large variety of speech corpora in French that cover different accents (MLS, African Accented Speech, CaFE), acted emotions (GEMEP, CaFE, Att-Hack), telephone dialogues (PORTMEDIA), read (MLS, African Accented French) and spontaneous sentences (CFPP2000, ESLO2, MPF, TCOF), as well as broadcast speech (EPAC). Compared to MLS and Voxpopuli, our dataset is more diverse, carefully sourced and contains detailed metadata (speech type, and speaker gender), which would facilitate future fine-grained analysis of SSL such as training gender/style specific models. Moreover, our dataset has a more realistic representation of speech turns in real life, compared to MLS (see average utterance duration per speaker in Table 1). We detail below the necessary steps for producing the dataset whose statistics are reported in Table 1.

Pre-processing for SSL training Audio recordings were segmented using time stamps from speech transcriptions. We also retrieved, when available, speaker labels and gender information. Following [28], we removed utterances shorter than 1 s, and longer than 30 s. Finally, when necessary, audio segments were converted to mono PCM files using 16 bits and a sampling frequency of 16 kHz.

Final dataset To the best of our knowledge, this is the very first study that explores such a diverse and large ensemble of datasets for SSL training. It includes 2,933 h of speech, from which 1,115 h is read speech, 1,626 h broadcast speech, 124 h spontaneous speech, 38 h acted telephone dialogues, and 29 h acted emotional speech. Regarding gender, we collected 1,824 h of speech from male speakers, 1,034 h from female speakers, and 74 h from unknown gender. This full corpus is refered as *medium*, and the subset made of MLS only is refered as *small*.

4. Training SSL Models

LeBenchmark provides four Wav2Vec2.0 models [28] pre-trained on the gathered French data described in Section 3. Following [28], two different Wav2Vec2.0 architectures (*large* and *base*) are coupled with our *small* (S) and *medium* (M) corpus to form our set of Wav2Vec2.0 models: W2V2-Fr-S-*base*, W2V2-Fr-S-*large*, W2V2-Fr-M-*base*, W2V2-Fr-M-*large*. Hyperparameters and architectures for *base* and *large* are identical to the ones first introduced in [28]. All models are trained on four Nvidia Tesla V100 (32GB) until the loss observed on the validation set of the MLS corpus (Section 3) reaches a stable point. Pretrained Wav2Vec2.0 models are shared with the community via HuggingFace[2] for further integration with well-known toolkits such as Fairseq [29], SpeechBrain [30] or Kaldi [31]. In some downstream experiments mentioned below, we also use two Wav2Vec2.0 (*base* / *large*, no finetuning) English models pre-trained on the full LibriSpeech (960 h) corpus by Fairseq,[3] and refer to them as *W2V2-En base* and *W2V2-En large*. The XLSR-53-*large* [6] multilingual model is also used.

5. Benchmarking our SSL Models

5.1. Automatic Speech Recognition

We evaluate the contribution of SSL for ASR using a hybrid DNN-HMM and an end-to-end approach.

Datasets The ASR tasks target two different types of corpora: Common Voice [32] and ETAPE [33]. Common Voice is a very large crowdsourced corpus (477 h) of read speech in French with transcripts – training: 428 h, development: 24 h, and test: 25 h, while ETAPE is a smaller (36 h) but more challenging corpus composed of diverse French TV broadcast programs – training: 22 h, development: 7 h, and test: 7 h.

Hybrid DNN-HMM The baseline acoustic models (AM) have been trained on 40-dimensional high-resolution (*hires*) MFCC features using the Kaldi [31] toolkit with a state-of-the-art factorized time delay neural network (TDNN-F) [34, 35] on the ETAPE training corpus [33] only. The model has 12 TDNN-F layers (1,024-dimensional, with projection dimension of 128) and a 3,432-dimensional output layer. It was trained using lattice-free maximum mutual information (LF-MMI) [36] and

[2] `https://huggingface.co/LeBenchmark`

[3] `https://github.com/pytorch/fairseq/tree/master/examples/wav2vec`

Table 2: *ASR results (WER,%) on the ETAPE corpus for hybrid DNN-HMM acoustic models with TDNN-F topology.*

Language Model	**ETAPE**		**ESTER-1.2 + EPAC**	
Features	**Dev**	**Test**	**Dev**	**Test**
hires MFCC	39.28	40.89	35.60	37.73
W2V2-Fr-M-*large*	**32.19**	**33.87**	**28.53**	**30.77**
W2V2-En-*large*	39.93	42.30	36.18	38.75
XLSR-53-*large*	36.36	38.19	32.81	35.17

cross-entropy criteria. Speed and volume perturbations have been applied for data augmentation. We used a similar topology to train three other systems with different types of input features extracted by W2V2-Fr-M-*large*, W2V2-En-*large* [28], and XLSR-53-*large* models. 100-dimensional speaker i-vectors were appended to the input features for all the models. Two trigram LMs were used in evaluation: (1) a larger one with a 82k vocabulary and (2) a smaller one trained on ETAPE training data only with a 17.5k vocabulary.

End-to-End Our end-to-end system is defined with the SpeechBrain toolkit [30] using an encoder/decoder architecture with attention: the encoder is a Convolutionnal Recurrent Deep Neural Network CRDNN (VGG + RNN + DNN), and the decoder is a joint CTC/Attention LSTM neural network. When used with the Wav2Vec2.0 features (same from hybrid DNN-HMM ASR experiments), the CNN blocks are removed from the CDRNN encoder. For end-to-end ASR experiments, the neural network output corresponds to 500 byte pair encoding (BPE) units [37] computed on the manual transcriptions of the respective training datasets. No additional language model is used in these experiments, neither data augmentation. For comparison purposes, we also use 80-dimension log Mel filterbank (MFB) features.

Results The WER results on the ETAPE development and test data sets for the hybrid DNN-HMM models are given in Table 2. Among the models trained on SSL features, two models provide improvement over the baseline AM trained on MFCC features: the model trained on XLSR-53 features (7–8% of relative WER reduction) and the model trained on W2V2-Fr-M-*large* features (17–20% of relative WER reduction). To our knowledge, this is the first time SSL features are used for hybrid DNN-HMM ASR. Actually, the hybrid DNN-HMM ASR system is much better than its end-to-end counterpart on ETAPE (see next paragraph). This is partly due to the use of speaker adaptation (i-vectors) and hand-crafted pronunciation dictionary which might be particularly beneficial to the hybrid DNN-HMM system, compared to end-to-end ASR, for the low resource ETAPE task.

Table 3 presents the results achieved with end-to-end ASR on Common Voice 6.1 and on ETAPE datasets. On the ETAPE, filterbank parameters (MFB) got significanly the best results, while on Common Voice, W2V2-Fr-M-*large* is very close. In all (hybrid and end-to-end) ASR experiments, among the wav2vec models, W2V2-Fr-M-*large* got the best results.

5.2. Spoken Language Understanding

Dataset The MEDIA corpus [38, 39] is used for the French SLU benchmark. The corpus is made up of 12,908 utterances (41.5 h) for training, 1,259 utterances (3.5 h) for development and 3,005 utterances (11.3 h) for test.

Model Our end-to-end model has a pyramidal LSTM encoder similar to [40]. The decoder integrates, in addition to the usual attention mechanism for attending the encoder hidden states, an attention mechanism for attending all previous decoder prediction's embeddings, instead of just the previous one [41]. We use an incremental training strategy similar to [39], by first training an ASR model from scratch which is used to initialize parameters of a SLU model using a simple linear layer as decoder; and then using this simple SLU to initialize parameters of our final SLU model, which uses a LSTM decoder. The model, which is implemented using *Fairseq* [29], has the same settings as [39] to allow direct and fair comparison.

Table 3: *End-to-end ASR results (WER,%) on Common Voice and ETAPE corpora.* $(*)$ *means the training algorithm did not converge to a WER smaller than 100%.*

Corpus	**CommonVoice**		**ETAPE**	
Features	**Dev**	**Test**	**Dev**	**Test**
MFB	**20.19**	**23.40**	**54.55**	**56.17**
W2V2-Fr-M-*large*	20.23	24.06	55.56	57.04
W2V2-En-*large*	34.07	37.29	98.79	99.10
XLSR-53-*large*	30.07	32.72	$(*)$	$(*)$

Results for ASR and SLU obtained with different speech representations are shown in Table 4, and they are given in terms of Word Error Rate (WER) and Concept Error Rate (CER) respectively, which is computed the same way as WER but on concept sequences. The ASR results are included because we use token-level models (ASR) to pre-initialize SLU models. The $\oplus$ symbol is used for separating Encoder and Decoder names: *Kheops* is the pyramidal encoder inspired from [40], *Basic* is the linear decoder, and *LSTM* is the more advanced LSTM decoder. For ASR, using SSL features as input resulted in an impressive drop in WER, even when using English SSL models. At best, we achieve a WER of 11.77% on the development data with the W2V2-Fr-M-*large* features. SLU results (*SLU decoding* in Table 4) follow the same trend. The best performance is obtained again with W2V2-Fr-M-*large* features, with a CER of 18.54 on the development data. This result improves previous work by almost 5 points (23.39 vs. 18.54), and stands as the new state-of-the-art result using only MEDIA training data for learning SLU models. Better results have been obtained in [42, 43] by using more transcribed and annotated data, in addition to the MEDIA corpus, via transfer learning.

5.3. Speech-to-text Translation

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 from French to another language. **Datasets** We selected subsets having French as the source language in two large multilingual speech corpora: CoVoST-2 [44] and multilingual TEDx [45]. Our benchmark covers translation directions from French to three target languages, English (`en`), Portugese (`pt`), and Spanish (`es`), with following training sizes: 50 h (TEDx/`en`), 38 h (TEDx/`es`), 25 h (TEDx/`pt`), and 180 h (CoVoST2/`en`).

Features We compared models using 80-dimensional MFB features and SSL representations. In addition to the four French Wav2Vec2.0 models trained in Section 4, we also considered the following off-the-shelf models: English [28] (W2V2-En-*base/large*), French [7] (W2V2-Fr-VP-*base/large*), and the multilingual model XLSR-53 [6] (XLSR-53-*large*). For a fair comparison, we did not use additional data augmentation technique nor ASR encoder pre-training in the experiments.

Models We trained Transformer [46] models using the FAIRSEQ S2T toolkit [47], and using a small architecture with 12-layers encoder, 6-layers decoder, and hidden dimension $D = 256$. For models using SSL features, we inserted a block of Linear-ReLU

Table 4: *End-to-end SLU results on the MEDIA corpus.*

Model	Features	Dev	Test
Token decoding (Word Error Rate %)			
[39] Seq	spectrogram	**29.42**	**28.71**
Kheops⊕Basic	spectrogram	36.25	37.12
Kheops⊕LSTM	spectrogram	**35.37**	**35.98**
Kheops⊕Basic	W2V2-En-*base*	19.80	21.78
Kheops⊕Basic	W2V2-En-*large*	24.44	26.96
Kheops⊕Basic	W2V2-Fr-S-*base*	23.11	25.22
Kheops⊕Basic	W2V2-Fr-S *large*	18.48	19.92
Kheops⊕Basic	W2V2-Fr-M-*base*	14.97	16.37
Kheops⊕Basic	W2V2-Fr-M *large*	**11.77**	**12.85**
Kheops⊕Basic	XLSR-53-*large*	14.98	15.74
SLU decoding (Concept Error Rate %)			
[39] Seq	spectrogram	28.11	27.52
[39] XT	spectrogram	**23.39**	**24.02**
Kheops⊕Basic	spectrogram	39.66	40.76
Kheops⊕Basic +token	spectrogram	34.38	34.74
Kheops⊕LSTM +SLU	spectrogram	**33.63**	**34.76**
Kheops⊕LSTM	W2V2-En-*base*	26.31	26.11
Kheops⊕LSTM	W2V2-En-*large*	28.38	28.57
Kheops⊕LSTM	W2V2-Fr-S-*base*	26.16	26.69
Kheops⊕LSTM	W2V2-Fr-S *large*	22.53	23.03
Kheops⊕LSTM	W2V2-Fr-M-*base*	22.56	22.24
Kheops⊕LSTM	W2V2-Fr-M-*large*	**18.54**	**18.62**
Kheops⊕LSTM	XLSR-53-*large*	20.34	19.73

Table 5: *AST results (BLEU) on dev/valid and test sets of CoVoST-2 (CV2) and multilingual TEDx (mTEDx).*

Input features	Dev/Valid data				Test data			
	CV2	mTEDx			CV2	mTEDx		
	en	en	es	pt	en	en	es	pt
MFB	**23.37**	1.14	0.84	0.49	**22.66**	1.33	0.98	0.68
W2V2-En-*base*	19.24	0.90	0.65	0.43	18.19	0.88	0.34	0.27
W2V2-En-*large*	17.07	0.75	0.61	0.45	16.45	0.85	0.67	0.32
W2V2-Fr-S-*base*	19.86	2.64	0.49	0.50	19.04	1.66	0.67	0.61
W2V2-Fr-S-*large*	19.62	5.12	4.62	**2.06**	18.61	2.97	3.19	**2.25**
W2V2-Fr-M-*base*	19.47	6.98	1.87	0.63	18.32	6.37	1.99	0.54
W2V2-Fr-M-*large*	20.17	**9.35**	**7.72**	1.58	19.35	**6.76**	**6.63**	1.63
W2V2-Fr-VP-*base*	18.44	0.81	0.45	0.56	17.40	0.89	0.58	0.75
W2V2-Fr-VP-*large*	20.72	7.43	4.66	0.43	19.88	5.39	3.62	0.49
XLSR-53-*large*	20.54	0.59	0.41	0.49	19.93	0.44	0.62	0.29

before convolutional layers not only to reduce the number of parameters [48], but also because we preliminary observed improved performance with this block.

Results shown in Table 5 highlight the benefit of SSL features only in medium and low-resource settings, namely mTEDx: our W2V2-Fr-M-*large* produces the best results across all language pairs, except for `pt` which is too low-resourced to obtain decent BLEU whatever the features used. In the higher-resource scenario (CV2), however, the best-performing SSL features are still 2.65 BLEU point below the MFB ones.

5.4. Automatic Emotion Recognition

Automatic emotion recognition aims at detecting human's apparent emotions from sensors such as microphones and cameras. Affective computing has many useful applications in the domain of health, education, art and entertainment.

Datasets We used the RECOLA dataset [49], which contains 3.8 h of noise-free recordings of spontaneous interactions between French-speaking subjects solving a collaborative task in remote condition – training, development and test partitions include each one third of the data, and AlloSat [50], a more recent corpus containing 37 h of real-life call center conversations in French – training: 25.6 h, development: 5.8 h, and test: 6.0 h. Both datasets are annotated by several annotators using time-continuous dimensions which are averaged to define an emotion *gold-standard*: arousal (from passive to active) and valence (from negative to positive) for RECOLA, and a dimensional axis ranging from frustration to satisfaction for AlloSat.

Table 6: *AER results (Concordance Correlation Coefficient of emotion predictions) on the RECOLA and AlloSat test sets.*

Corpus		RECOLA		AlloSat
Model	**Feature**	**Arousal**	**Valence**	**Satisfaction**
Linear-Tanh	MFB	0.192	0.075	0.065
Linear-Tanh	W2V2-Fr-M-*base*	**0.385**	**0.090**	**0.193**
Linear-Tanh	W2V2-En-*large*	0.328	-0.005	0.075
Linear-Tanh	XLSR-53-*large*	0.155	0.024	0.093
GRU-32	MFB	0.654	0.252	0.437
GRU-32	W2V2-Fr-M-*base*	**0.767**	**0.376**	**0.507**
GRU-32	W2V2-En-*large*	0.629	-0.032	0.334
GRU-32	XLSR-53-*large*	0.605	0.320	0.446
GRU-64	MFB	0.712	0.307	0.400
GRU-64	W2V2-Fr-M-*base*	**0.760**	**0.352**	**0.507**
GRU-64	W2V2-En-*large*	0.635	0.015	0.367
GRU-64	XLSR-53-*large*	0.585	0.280	0.434

Features We extracted 40-dimensional MFB features that were standardized to zero mean and unit standard deviation according to the training set, and SSL features that were pre-processed by a normalisation layer. Annotations were resampled to match the sampling frequency of the features, which was 100 Hz for MFB and 50 Hz for the Wav2Vec models.

Models We used a simple model based on a linear layer mapping features to one emotional dimension, followed by a *Linear-Tanh* function. The other model is a 1-layer GRU with the hidden layer $D = [32, 64]$, followed by the Linear-Tanh function. Adam optimiser was used and patience was set to 15 epochs, and the Concordance Correlation Coefficient [51] was used as loss function to train the models as in [52, 53].

Results Best results are obtained by our W2V2-Fr-M-*base* representation on valence, satisfaction and arousal, c. f. Table 6. With a simpler model, best scores are also achieved on both data sets with the Wav2Vec features, meaning that SSL representations are rich enough to be used with a simple regressor, even for low-quality speech signals (telephone conversations).

6. Discussion

After training our own SSL models for French, we evaluated them for four speech tasks (ASR, SLU, AST, and AER) using different architectures (shallow and deep architectures, end-to-end or not). The learnt SSL models are particularly beneficial for lower resource tasks (SLU, AST/TEDx, AER) or with simpler NN architectures (AER) but they sometimes fail providing a benefit compared to MFB or MFCCs (End-to-end ASR). Fine-tuning of SSL models could probably help bridging the gap remaining for some tasks, but we used SSL features 'as they are' for this paper. Furthermore, efficient data augmentations techniques for Mel Filterbanks such as *SpecAugment* were disabled here to provide a comparison with SSL features, so we should highlight that we did not make the best of our Mel Filterbanks. All of these remarks and findings advocate for more exhaustive and reliable evaluations to assess the real impact of SSL for speech systems. We hope that decentralized projects such as *LeBenchmark* will contribute to this goal.

7. References

[1] P. Bachman *et al.*, "Learning representations by maximizing mutual information across views," in *NeurIPS*, 2019.

[2] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *PMLR*, 2020.

[3] J. Devlin *et al.*, "BERT: pre-training of deep bidirectional transformers for language understanding," *NAACL-HLT*, 2018.

[4] A. Baevski *et al.*, "Effectiveness of self-supervised pre-training for speech recognition," in *ICASSP*, 2020.

[5] K. Kawakami *et al.*, "Learning robust and multilingual speech representations," in *EMNLP*, 2020.

[6] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv:2006.13979*, 2020.

[7] C. Wang *et al.*, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv:2101.00390*, 2021.

[8] "African accented french, slr57," 2003, type: dataset, https://www.openslr.org/57/.

[9] C. Le Moine and N. Obin, "Att-HACK: An Expressive Speech Database with Social Attitudes," in *Speech Prosody*, 2020.

[10] P. Gournay *et al.*, "A Canadian French emotional speech dataset," in *MMSys*, 2018.

[11] S. Branca-Rosoff *et al.*, "Discours sur la ville. Présentation du Corpus de Français parlé Parisien des années 2000 (CFPP2000)," 2012, http://cfpp2000.univ-paris3.fr/CFPP2000.pdf.

[12] "Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000)," http://cfpp2000.univ-paris3.fr/.

[13] I. Eshkol-Taravella *et al.*, "Un grand corpus oral "disponible" : le corpus d'Orléans 1968-2012," *Ressources Linguistiques Libres - TAL*, 2011.

[14] "Eslo, enquêtes sociolinguistiques à orléans," http://eslo.huma-num.fr/index.php.

[15] Y. Estève *et al.*, "The EPAC Corpus: Manual and Automatic Annotations of Conversational Speech in French Broadcast News," in *LREC*, 2010.

[16] T. Bänziger *et al.*, "Introducing the Geneva multimodal expression corpus for experimental research on emotion perception," *Emotion*, 2012.

[17] P. Vineel *et al.*, "MLS: A large-scale multilingual dataset for speech research," *ArXiv:2012.03411*, 2020.

[18] G. Françoise, "Les parlers jeunes dans l'île-de-France multiculturelle," *Paris and Gap, Ophrys*, 2017.

[19] "Mpf," 2019, https://hdl.handle.net/11403/mpf/v3, ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

[20] F. Lefèvre *et al.*, "Robustesse et portabilités multilingue et multi-domaines des systèmes de compréhension de la parole : le projet PortMedia," in *JEP-TALN-RECITAL*, 2012.

[21] ATILF, "TCOF : Traitement de corpus oraux en français," 2020, https://hdl.handle.net/11403/tcof/v2.1, ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

[22] M. E. Peters *et al.*, "Deep contextualized word representations," *NAACL-HLT*, 2018.

[23] Y. Chung *et al.*, "An unsupervised autoregressive model for speech representation learning," in *Interspeech*, 2019.

[24] Y. Chung and J. Glass, "Improved speech representations with multi-target autoregressive predictive coding," *arXiv:2004.05274*, 2020.

[25] S. Schneider *et al.*, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Interspeech*, 2019.

[26] Y. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP*, 2020.

[27] M. Rivière *et al.*, "Unsupervised pretraining transfers well across languages," in *ICASSP*, 2020.

[28] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv:2006.11477*, 2020.

[29] M. Ott *et al.*, "fairseq: A fast, extensible toolkit for sequence modeling," in *NAACL-HLT: Demonstrations*, 2019.

[30] M. Ravanelli *et al.*, "Speechbrain," https://github.com/speechbrain/speechbrain, 2021.

[31] D. Povey *et al.*, "The Kaldi speech recognition toolkit," 2011.

[32] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," in *LREC*, 2020.

[33] G. Gravier *et al.*, "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language," in *LREC*, 2012.

[34] D. Povey *et al.*, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Interspeech*, 2018.

[35] V. Peddinti *et al.*, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Interspeech*, 2015.

[36] D. Povey, Peddinti, and other, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016.

[37] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *ACL*, 2016.

[38] H. Bonneau-Maynard *et al.*, "Results of the french evalda-media evaluation campaign for literal understanding," in *LREC*, 2006.

[39] M. Dinarelli *et al.*, "A data efficient end-to-end spoken language understanding architecture," in *ICASSP*, Spain, 2020.

[40] W. Chan *et al.*, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016.

[41] D. Bahdanau *et al.*, "Neural machine translation by jointly learning to align and translate," *CoRR*, 2014.

[42] A. Caubrière *et al.*, "Curriculum-Based Transfer Learning for an Effective End-to-End Spoken Language Understanding and Domain Portability," in *Proc. Interspeech 2019*, 2019, pp. 1198–1202. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1832

[43] N. Tomashenko *et al.*, "Recent advances in end-to-end spoken language understanding," in *International Conference on Statistical Language and Speech Processing*. Springer, 2019, pp. 44–55.

[44] C. Wang, A. Wu, and J. Pino, "Covost 2 and massively multilingual speech-to-text translation," *arXiv e-prints*, pp. arXiv–2007, 2020.

[45] E. Salesky *et al.*, "The multilingual TEDx corpus for speech recognition and translation," 2021.

[46] A. Vaswani *et al.*, "Attention is all you need," in *NIPS*, 2017.

[47] C. Wang *et al.*, "fairseq S2T: Fast speech-to-text modeling with fairseq," *arXiv preprint arXiv:2010.05171*, 2020.

[48] H. Nguyen, F. Bougares, N. Tomashenko, Y. Estève, and L. Besacier, "Investigating self-supervised pre-training for end-to-end speech translation," in *ICML*, 2020.

[49] F. Ringeval *et al.*, "Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions," in *EmoSPACE, FG*, 2013.

[50] M. Macary *et al.*, "AlloSat: A new call center French corpus for satisfaction and frustration analysis," in *LREC*, 2020.

[51] L. Li, "A Concordance Correlation Coefficient to evaluate reproducibility," *Biometrics*, 1989.

[52] F. Weninger *et al.*, "Discriminatively trained Recurrent Neural Networks for continuous dimensional emotion recognition from audio," in *IJCAI*, 2016.

[53] G. Trigeorgis *et al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*, 2016.