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 successfully 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 benchmark1 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 fulfilled 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 transferred 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 vox-populi [7] (44,500 h). However, these are restricted to either read or well-prepared speech, failing to provide diversity in

∗ equal contribution
1 https://github.com/LeBenchmark/
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 gender/style specific models. Moreover, our dataset has a more realistic representation of speech turns in real life, compared to gender/style specific models. Furthermore, our dataset has a more realistic representation of speech turns in real life, compared to gender/style specific models.

We detail below the necessary steps for producing the dataset MLS (see average utterance duration per speaker in Table 1).

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).

<table>
<thead>
<tr>
<th>Corpus</th>
<th># Utterances</th>
<th>Duration</th>
<th># Speakers</th>
<th>Mean Utt. Duration</th>
<th>Speech type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Att-Hack [9]</td>
<td>10,529</td>
<td>18:37</td>
<td>374</td>
<td>4 s</td>
<td>Read</td>
</tr>
<tr>
<td>CaFE [10]</td>
<td>10,715</td>
<td>18:38</td>
<td>374</td>
<td>4 s</td>
<td>Read</td>
</tr>
<tr>
<td>ESLO2 [13], [14]</td>
<td>62,918</td>
<td>34:12</td>
<td>190</td>
<td>1.9 s</td>
<td>Spontaneous</td>
</tr>
<tr>
<td>EPAC** [15]</td>
<td>623,250</td>
<td>1,269:02</td>
<td>1,133</td>
<td>95</td>
<td>Radio</td>
</tr>
<tr>
<td>GEMEP [16]</td>
<td>1,126</td>
<td>2:30</td>
<td>10</td>
<td>2.5 s</td>
<td>Acted</td>
</tr>
<tr>
<td>MLS French [17]</td>
<td>26,105</td>
<td>14:06:45</td>
<td>178</td>
<td>15:0 s</td>
<td>Read</td>
</tr>
<tr>
<td>MPF [18], [19]</td>
<td>15,287</td>
<td>9:56</td>
<td>114</td>
<td>3.3 s</td>
<td>Spontaneous</td>
</tr>
<tr>
<td>PORTMEDIA (French) [20]</td>
<td>87,724</td>
<td>5:59</td>
<td>749</td>
<td>3.3 s</td>
<td>Spontaneous</td>
</tr>
<tr>
<td>TCOF (Adult s) [21]</td>
<td>10,377</td>
<td>3:35:49</td>
<td>119</td>
<td>3:33 s</td>
<td>Spontaneous</td>
</tr>
<tr>
<td>ALL</td>
<td>1,111,865</td>
<td>2:93:18</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

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: 7 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...
Table 2: ASR results (WER,%) on the ETAPE corpus for hybrid DNN-HMM acoustic models with TDNN-F topology.

<table>
<thead>
<tr>
<th>Language Model</th>
<th>ETAPE</th>
<th>ESTER-L2 + EPAC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>bases MFCC</td>
<td>39.28</td>
<td>40.89</td>
</tr>
<tr>
<td>W2V2-En-Large</td>
<td>32.19</td>
<td>33.87</td>
</tr>
<tr>
<td>W2V2-En-Large</td>
<td>39.93</td>
<td>42.30</td>
</tr>
<tr>
<td>XLSR-53-Large</td>
<td>36.36</td>
<td>38.19</td>
</tr>
</tbody>
</table>

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 Speech-Brain toolkit [30] using an encoder/decoder architecture with attention: the encoder is a Convolutional 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-En-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 significantly the best results, while on Common Voice, W2V2-En-M-large is very close. In all (hybrid and end-to-end) ASR experiments, among the wav2vec models, W2V2-En-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.

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 @ 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-En-M-large features. SLU results (SLU decoding in Table 4) follow the same trend. The best performance is obtained again with W2V2-En-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: CoVoST2-2 [44] and multilingual TEDx [45]. Our benchmark covers translation directions from French to three target languages, English (en), Portuguese (pt), and Spanish (es), with following training sizes: 50h (TEDx/en), 38 h (TEDx/es), 25 h (TEDx/pt), and 180h (CoVoST2/en).

Features

We compared models using 80-dimensional MFN 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/slarge), French [7] (W2V2-En-VP-base/large), and the multilingual model XLSR-53--large (XL5R-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 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%.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>CommonVoice</th>
<th>ETAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>MFB</td>
<td>20.19</td>
<td>23.40</td>
</tr>
<tr>
<td>W2V2-En-M-large</td>
<td>20.23</td>
<td>23.06</td>
</tr>
<tr>
<td>W2V2-En-Large</td>
<td>34.07</td>
<td>37.29</td>
</tr>
<tr>
<td>XLSR-53-Large</td>
<td>30.07</td>
<td>32.72</td>
</tr>
</tbody>
</table>
before convolutional layers not only to reduce the number of parameters [48], but also because we preliminarily 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 \( \frac{p}{c} \), 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.

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

<table>
<thead>
<tr>
<th>Input features</th>
<th>Dev/Valid data</th>
<th>Test data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CV2</td>
<td>mTEDx</td>
</tr>
<tr>
<td></td>
<td>en</td>
<td>en</td>
</tr>
<tr>
<td>MFB</td>
<td>23.37</td>
<td>1.14</td>
</tr>
<tr>
<td>W2V2-En-base</td>
<td>19.24</td>
<td>0.90</td>
</tr>
<tr>
<td>W2V2-En-large</td>
<td>17.07</td>
<td>0.75</td>
</tr>
<tr>
<td>W2V2-Fr-S-base</td>
<td>19.86</td>
<td>2.64</td>
</tr>
<tr>
<td>W2V2-Fr-S-large</td>
<td>19.62</td>
<td>5.12</td>
</tr>
<tr>
<td>W2V2-Fr-M-base</td>
<td>19.47</td>
<td>6.98</td>
</tr>
<tr>
<td>W2V2-Fr-M-large</td>
<td>20.17</td>
<td>8.35</td>
</tr>
<tr>
<td>W2V2-Fr-VP-base</td>
<td>18.44</td>
<td>0.81</td>
</tr>
<tr>
<td>W2V2-Fr-VP-large</td>
<td>20.72</td>
<td>7.43</td>
</tr>
<tr>
<td>XLSR-53-large</td>
<td>20.54</td>
<td>0.59</td>
</tr>
</tbody>
</table>

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.

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 \( D = \llbracket 32, 64 \rrbracket \), 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 MFCs (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


