Promises and Limitations of Self-supervised Learning for Automatic Speech Processing

Lucas Maison, Marcely Zanon Boito, Yannick Estève

Promises and Limitations of Self-supervised Learning for Automatic Speech Processing

Lucas Maison
Laboratoire Informatique d'Avignon (LIA)
Avignon Université
lucas.maison at univ-avignon.fr

Marcely Zanon Boito
Laboratoire Informatique d'Avignon (LIA)
Avignon Université
marcely.zanon-boito at univ-avignon.fr

Yannick Estève
Laboratoire Informatique d'Avignon (LIA)
Avignon Université
yannick.esteve at univ-avignon.fr

***Abstract*—Self-supervised learning (SSL) has recently been successfully introduced as a training strategy for Transformer-based neural models. Thanks to this approach, these models are now able to construct speech representations by using only audio data, without any manual labels (i.e. no supervision). Once trained, they can be leveraged for training competitive end-to-end models for speech processing with smaller amounts of annotated data. Moreover, when the available annotated data is plenty, automatic speech recognition (ASR) and translation (AST) systems based on these SSL models are now the new state of the art. In this work, we are interested in their application in challenging settings that are relevant for security. We measure the robustness of a French-based SSL model to African accent, and we present some promising but limited results for speech translation without the use of transcriptions.**

***Index Terms*—automatic speech recognition, speech translation, self-supervised learning, speech processing, security**

I. Introduction

Speech recognition has been dominated by data-driven approaches for almost four decades. From the 80s until a few years ago, automatic speech recognition (ASR) systems were based on the use of three kinds of knowledge. The first one was captured by the acoustic models, usually based on a Hidden Markov Model (HMM) combined to a Gaussian Mixture Model (GMM) or more recently to a Deep Neural Network (DNN). The acoustic models were designed to compute the likelihood of the presence of a phoneme (the speech unit that discriminates a word in a language) according to the audio signal. The second knowledge was represented by a pronunciation dictionary, in order to map a sequence of phonemes to one or several words. The last knowledge was captured by a language model, in order to compute the probability to observe a sequence of words in a language.

At that time, acoustic and language models were mainly statistical models, and that is why we say that such approaches are data-driven. To get such models accurate and robust, a large amount of training data is needed, following the "there is no better data than more data" paradigm. During the last decade, DNNs for both acoustic and language models have replaced the GMMs, and neural end-to-end approaches eliminated the HMMs, but the need for data remained, at least, the same.

Among this data, the most *important* modality is having speech audio data with its manual transcription. This paired audio/text data is necessary to train ASR systems. It is also a very costly data that can be rare for many languages.

Self-supervised Learning (SSL) has been recently proposed as an interesting alternative for data representation learning. Proven useful learned representations can be found both in vision [1], [2] and in NLP [3], [4]. The attractiveness of SSL in general, and SSL from speech in particular, is that it can leverage huge amounts of unannotated data, which is cheaper than the audio/text data used by classical systems. This leveraging can be done by resolving pseudo-tasks, which do not require human annotation, as pre-training a feature extractor, which is then used to extract useful speech representations for the real (downstream) tasks. The two most commonly used approaches for SSL from speech are *Autoregressive Predictive Coding* (APC) and *Contrastive Predictive Coding* (CPC). The former's pseudo-task is considering the sequential structure of speech, and predicting information about a future frame [5], [6], whereas the latter's consists of distinguishing a future speech frame from distractor samples [7]–[9] which is an easier learning objective compared to APC. These representations have been proven to improve the performance in several speech tasks [10], while being less sensitive to domain and/or language mismatch [11] and being transferable to other languages [12].

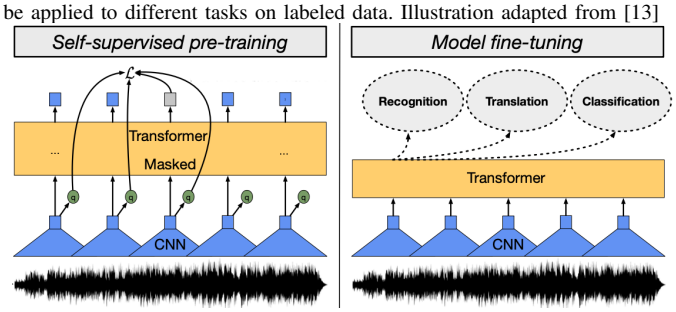
SSL opens new perspectives to build and deploy ASR for low-resource languages, or low-resource domains. Such an approach speeds up the creation of a new ASR system, and reduces its cost, since a significantly smaller amount of annotated data is necessary to get competitive results in comparison to the previous state of the art.

Speech recognition can be involved in many tasks for security purposes. It is also the case for speech translation, on which SSL is also useful, especially for neural end-to-end architectures.

This paper discusses current limitations of the wav2vec 2.0 models, focusing on two applications relevant for security: automatic speech recognition (ASR) and automatic speech translation (AST). It is organized as follows: Section II presents a high-level summarization of the technology behind these SSL models. Section III discusses their application to ASR, especially to process accented French. Section IV discusses their application to AST in an extreme low resource scenario. Section V presents our final remarks.

II. Self-supervised Models for Speech: the wav2vec 2.0 architecture

Fig. 1. Illustration of the wav2vec 2.0 framework (left), which jointly learns contextualized speech representations and an inventory of discretized speech units during pre-training on unlabeled data. The fine-tuning step (right) can be applied to different tasks on labeled data. Illustration adapted from [13]



The *wav2vec 2.0* proposed by [14] is an extension of [8], [9], [15]. Depicted in Figure 1, it consists of a multi-layer convolutional feature encoder $g_{enc} : \mathcal{X} \rightarrow \mathcal{Z}$, which transforms raw input audio x into latent speech representations $z = \{z_1, z_2, ..., z_T\}$ for T time-steps. These latent features are then fed into a Transformer $g : \mathcal{Z} \rightarrow \mathcal{C}$ for building contextualized representations $c = \{c_1, c_2, ..., c_T\}$ that capture the information of the whole sequence.

The wav2vec 2.0 also performs discretization on the output of the feature encoder z_t to q_t by using a quantization module $\mathcal{Z} \rightarrow \mathcal{Q}$. The model's Transformer network learns contextualized representations directly from continuous speech representations (z) via time-step masking and a contrastive task (CPC) which identifies the true quantized latent audio representation in a set of distractors for each masked time step. This consequently allows [14] to train wav2vec 2.0 in an end-to-end fashion, in which all its components are trained jointly toward minimizing an objective (Equations 1, 2, 3).

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d \tag{1}$$

$$\mathcal{L}_m = -\log \frac{\exp(sim(c_t, q_t)/\kappa)}{\sum_{\tilde{q} \sim Q_t} \exp(sim(c_t, \tilde{q})/\kappa)} \tag{2}$$

$$\mathcal{L}_d = \frac{1}{GV} \sum_{g=1}^{G} -H(\bar{p}_g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v} \tag{3}$$

In Equation 1, the training objective is defined as the sum of two components:

- **Contrastive Loss** $\mathcal{L}_m$ which is defined in Equation 2. Particularly, given c_t centered over the masked time step t, the model is trained to contrast the true quantized latent speech representation q_t from K quantized latent distractors $\tilde{q} \in Q_t$ uniformly sampled from other masked time steps of the same utterance. $sim(a, b) = a^\top b / ||a|| ||b||$ is the *cosine similarity* between context representations c_t and quantized latent speech representations q_t.
- **Diversity Loss** $\mathcal{L}_d$ which is defined in Equation 3. It helps to increase the use of the quantized codebook representations, encouraging the model to equally use all the V entries in each of G codebooks by maximizing the entropy of the averaged softmax distribution over the codebook entries for each codebook $\bar{p}_g$ across a batch of utterances. In Equation 1, $\mathcal{L}_d$ is scaled by α, which is a tunable hyperparameter.

Masking: time-step masking mentioned earlier is done by randomly sampling without replacement a certain proportion p of all time steps to be starting indices and then mask the subsequent M consecutive time steps for every sampled index. Note that spans may overlap, and inputs to the quantization module are not masked.

Fine-tuning: the wav2vec 2.0 framework also allows fine-tuning the pre-trained model directly on ASR (or AST, or speech classification) labeled data by stacking a linear projection layer initialized randomly on top of the context network. Fine-tuned models are optimized by minimizing a Connectionist Temporal Classification (CTC) loss. It has been shown [14] that fine-tuning even on only 10 minutes of labeled training data (48 recordings of 12.5 seconds on average) helps achieve a respective Word Error Rate (WER) of 4.8% and 8.2% on the test-clean and test-other sets of the Librispeech corpus (read speech).

III. ASR application: from standard to accented speech

In recent years, huge progress has been achieved in the domain of automatic speech recognition (ASR). Neural models now reach human performance on the English ASR task: best systems reach a Word Error Rate (WER) of 5.8% and 11% on Switchboard and CallHome datasets respectively, whereas performance of human annotators is estimated to be around 5.9% and 11.9% [16].

However, it has been shown that the performance of these models can decrease drastically when they are used in new or non-ideal conditions. For example, a noisy environment or the accent of a speaker can both impact the quality of the transcription [17], [18]. It is important to study the robustness of systems to such conditions, because these are likely to be encountered in real settings. Furthermore, support of accented speech is mandatory if one aspire to build an inclusive, general-purpose ASR system. It could also be of interest for security or intelligence agencies wanting to support a large spectrum of accents. In this section, we focus on the case of accented speech in French, presenting results for ASR models trained on standard and accented French speech.

A. Models and datasets

Focusing on the French language, the *LeBenchmark* initiative [19], [20] is a proeminent work in the area of SSL benchmarking. It provides evaluation recipes for four downstream tasks (ASR, AST, automatic emotion recognition, spoken language understanding) alongside with wav2vec 2.0 models

Fig. 2. ASR results (WER, the lower the better) over the two test sets for models fine-tuned on CommonVoice.

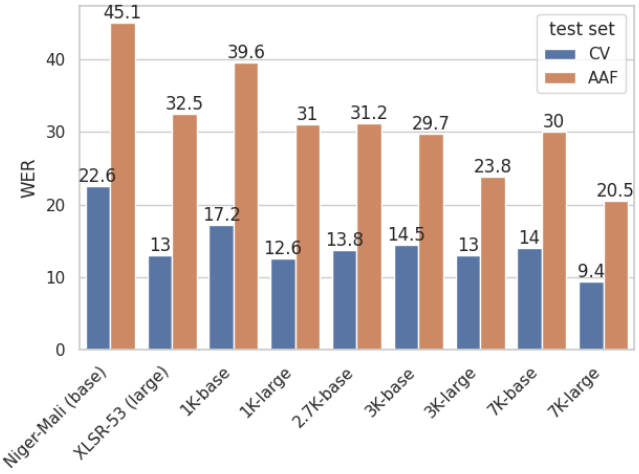
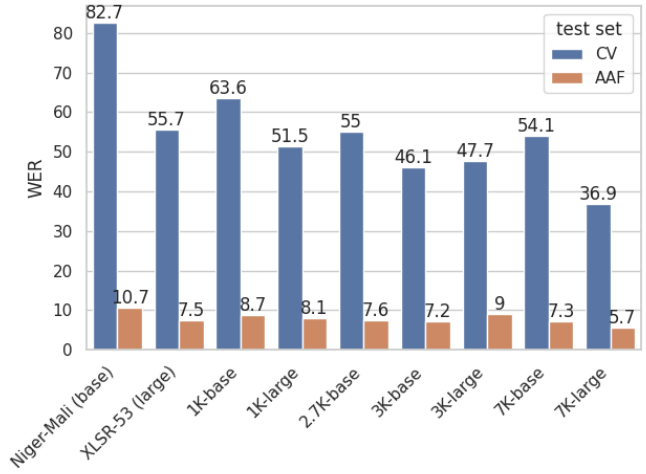


Fig. 3. ASR results (WER, the lower the better) over the two test sets for models fine-tuned on African Accented French.



of various sizes, and pre-trained using different amounts of speech audio.

For our ASR models, we use the following models from the *LeBenchmark*: LB-1K-base/large, LB-2.7K-base, LB-3K-base/large, and LB-7K-base/large, which were pre-trained on respectively 1,096, 2,773, 2,933 and 7,739 hours of French audio [20]. The "base" refers to the standard model architecture from [14] that has 95 millions parameters, while the "large" refers to their larger architecture that presents greater capacity (317 millions parameters).

In addition to these French models, two multilingual models were tested. The first one, Niger-Mali [21], is a base model pre-trained on 641 h of speech in 5 languages, including 111 h of accented French. The second one, XLSR-53 [22], is a large model pre-trained on 56k hours of speech in 53 languages, including 1,429 h of French.

Each pre-trained wav2vec 2.0 model acts as a speech encoder, which is optimized for the ASR task together with an additional feed-forward network. This head network consists of three linear layers with 768 or 1,024 neurons for a *base* or *large* model, respectively. Each linear layer is followed by batch normalization and a Leaky ReLU [23] activation function. We use dropout with $p = 0.15$ between each linear layer. At last, a final linear layer projects the output into token space, and log-softmax is applied to obtain probabilities of each token. We use individual characters as tokens. Note that we do not apply any language model besides our end-to-end model.

We employ the `SpeechBrain` [24] toolkit for all our experiments. All models are fine-tuned during 50 epochs using the CTC loss, and Adam [25] and Adadelta [26] optimizers are used to update the weights, one for the wav2vec 2.0 model and one for the additional top layers.

We use two different datasets in this study. The first one is the French subset of CommonVoice (CV) 3.0 [27] that comprises 56 h of recordings. It represents our reference dataset of unaccented speech. The second one is the African Accented French (AAF) dataset [28]. It is composed of 13 h of speech [1]. Speakers are from Cameroon, Chad, Congo, Gabon and Niger, and they speak French with a strong African accent. It should be noted that this latter dataset was used as part of the pre-training data for the following *LeBenchmark* models: LB-3K-base/large, LB-7K-base/large.

B. Fine-tuning ASR models: from non-accented to accented speech

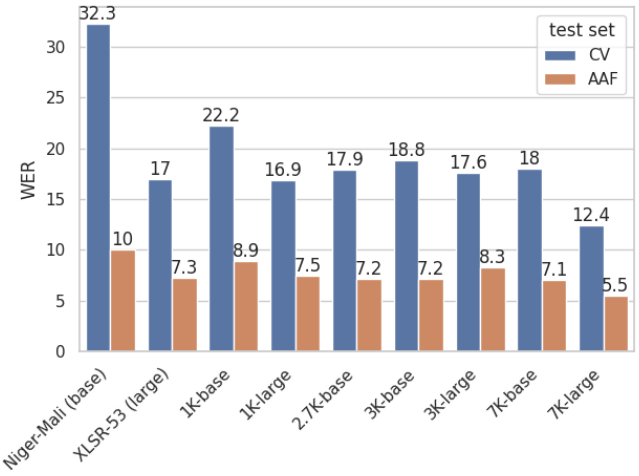
To assess the robustness of the pre-trained wav2vec 2.0 models with respect to accent variability, we fine-tune each model on the train split of CV. Then, we evaluate the resulting models on both the test split of CV and AAF. Results are shown on Figure 2. The trend we observe is that, the more speech data we use for pre-training, the best the model performs, meaning that it better specialized its speech representations.

We also notice that, thanks to their increased capacity, "large" models perform much better than "base" ones. Moreover, multilingual models perform rather badly when compared with French models trained with similar amounts of speech. The best model (LB-7K-large) obtains a WER of 9.37% on CV, which is comparable to the best scores reported in [20] on the same dataset (CV). However, this same model scores 20.47% on AAF. Moreover, all the tested models follow a similar trend, with a WER that doubles on AAF compared to CV. This means that these models are likely to make twice as many transcription errors when used by non-native speakers rather than native speakers.

In order to improve the robustness of the models on accented speech, we restart the experiment and fine-tune each pre-

[1]The original dataset is larger, but we excluded portions containing annotation errors.

Fig. 4. ASR results (WER, the lower the better) over the two test sets for models fine-tuned on a mixed dataset (CV+AAF).



trained model on the train split of AAF, and evaluate the resulting models on the same test sets as before. We can see on Figure 3 that doing so greatly reduce the WER on AAF (-75% on average), with our best model now scoring 5.72%. However, this large improvement comes at the cost of a similar performance degradation on CV (+282% on average). This demonstrates that fine-tuning directly on accented speech is beneficial if we desire to transcribe a particular accent, but should not be done if the goal is to build an all-purpose system.

Finally, we created a mixed dataset of accented and native speech by taking the full AAF training set and an equal amount of speech from the CV training set. We use this new dataset to fine-tune the models. Results of evaluation are shown on Figure 4. We can see that fine-tuning on this mixed dataset allows the models to reach good performance on both accented and non-accented speech. Our best model scores 12.38% and 5.47% on CV and AAF respectively. Compared to the models fine-tuned on CV only, these models reach much lower WER on AAF (-75% on average), while only suffering mild performance degradation on CV (+32% on average).

C. Discussion of results

It may seem surprising that the "Niger-Mali" model does not obtain a good performance despite being pre-trained on a large amount of accented French. We believe that the most important factor contributing to low WER is the amount of French audio in the pre-training dataset. The Niger-Mali model is the one with the lowest quantity of French seen during pre-training (111 h), thus explaining its poor score.

The presence of accented French in the pre-training dataset may still play an important role: we can see on Figures 2, 3, and 4 that LB-3K-base is achieving slightly better results on AAF compared to LB-2.7K-base, the main difference between these two models being the presence of accented speech in the pre-training dataset of the former.

The second multilingual model has been pre-trained on much more data (56k hours), but only 1,429 h of French. Its scores no better than the other *large* models pre-trained on much smaller (but French only) datasets. This seems to indicate that multilingual models, despite obtaining good performances on a variety of languages, are not suited for recognizing accented speech.

In summary, in this section we illustrated existing limitations of ASR models for transcribing accented speech in French. We experimented with two multilingual wav2vec 2.0 models, and seven French models, comparing the performance of obtained ASR systems in standard and accented French speech. We find that models fine-tuned on native speech only are not robusts to accent variation, but that incorporating accented data in the fine-tuning dataset greatly improve robustness.

IV. AST Application: SSL models for low-resource end-to-end AST

Traditionally, the speech translation task is defined in a cascaded fashion: the speech is first transcribed by an ASR model, and then a text-to-text machine translation (MT) module produces the final translation in the target language. The limitations of this approach for AST includes the error propagation between the ASR and MT modules, the omission of speech cues that could disambiguate the information given to the MT module, and the need for both a considerable amount of transcribed *and* translated data.

Going beyond the practical time and money constrains for producing this data in non-mainstream languages in order to train and deploy cascaded systems, and the cost of training the systems themselves, it is also important to be aware that not all languages present a standard written form. Indeed, most of the world's languages are not actively written, even the ones with an official writing system [29]: these are called *oral languages*.

This is one reason behind the recent motivation of the speech community to investigate *end-to-end* approaches for AST [30], [31]. We define end-to-end AST as a single optimized model that receives as input speech and produces as output textual translations. Optionally, these models can be jointly optimized for producing transcriptions as well, when these are available during training. This joint training was shown to increase translation performance [32], [33].

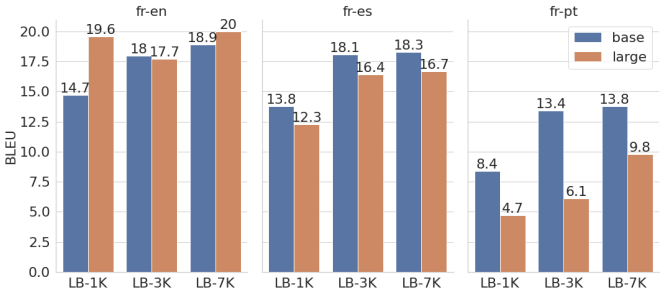
In this section we shed light on some limitations of SSL-based end-to-end AST models for processing oral-languages, for which the amount of available data is limited. This is relevant in the context of security because a government or an organization might aim to deploy AST models for minority or dialect languages in areas of particular interest. In these cases, the amount of available data is usually limited. Ideal AST systems for security should thus be able to work in *low-resource settings.*

This section is organized as follows. We first validate our AST architecture by producing results for three language pairs in the `mTEDx` dataset [34] (Section IV-A).

TABLE I
STASTISTICS FOR THE MTEDX FR-{EN,ES,PT} DATASET.

train						valid		test	
en		es		pt					
# spk	duration	# spk	duration	# spk	duration	# spk	duration	# spk	duration
250	45:04	196	32:30	112	20:01	12	1:38	10	1:33

Fig. 5. AST results (BLEU scores, the higher the better) over the test set for the three language pairs, and using the base and large wav2vec 2.0 models.



These languages have decreasing amounts of available parallel data: French-English (48 h), French-Spanish (35 h), French-Portuguese (23 h).

Having defined these *mid-to-low-resource baselines*, in Section IV-B, we explore the case of the Tamasheq dataset presented in this year's IWSLT campaign [35]. The challenge is producing translation, without available transcription, having only 17 hours of speech in Tamasheq aligned to French translations. In this case, we illustrate how general purpose SSL models fail to produce exploitable representations. Lastly, in Section IV-C we summarize our findings on the use of SSL models for low-resource end-to-end AST.

A. AST in mid-to-low-resource settings

The presented end-to-end AST models are similar to the end-to-end ASR model architecture presented in Section III. They are implemented on `SpeechBrain` [24], being made of a wav2vec 2.0 as a *foundation block*, followed by a linear projection, and a Transformer Decoder [36]. The weights for the wav2vec 2.0 speech encoder block are initialized from the pre-trained SSL models available in the *LeBenchmark* model collection [19], [20].[2] The model is trained on the negative log likelihood loss, and two different instances of the Adam optimizer manage the weight updates: one dedicated to the wav2vec 2.0 block, the other one to the following layers.

As aforementioned, we train mid-to-low-resource baselines using the `mTEDx` dataset. We use as source the French language (speech), and as target languages (text): English (en), Spanish (es) and Portuguese (pt). The resulting language pairs share validation and test sets, but they vary on the amount of available training data. This information (duration), together with the number of speakers (# spk), is presented in Table I.

[2]Available at https://huggingface.co/LeBenchmark

Figure 5 presents the AST results[3] using the three language pairs, and three different wav2vec 2.0 models: LB-1K, LB-3K and LB-7K. These models differ on the amount of training data used during SSL pre-training, with 1K corresponding to approximately 1,000 hours of speech. For each model, we experiment with both *base* and *large* architecture sizes.[4]

Looking at the results using **base** wav2vec 2.0 models (Figure 5, darker bars), we notice that AST models trained in all language pairs benefit from having SSL wav2vec 2.0 models trained using more data: using LB-7K-base and LB-3K-base as foundation blocks seem to be clearly superior compared to AST models that used LB-1K-base. We however, do not observe a very clear distinction between LB-7K-base and LB-3K-base, which might be due to the wav2vec 2.0 model reaching the limits of its own capacity [20].

Focusing on the **large** models (Figure 5, brighter bars), we notice that the trend is not the same for all languages. For English, it seems to still exist some benefit on employing these larger models, compared to their base counterparts: BLEU scores are higher or equivalent. For the other two languages, we notice that performance using large models is inferior to the one reached by their base counterparts.

For these languages, we have less training examples compared to English: for Spanish we have only 35 h, and for Portuguese only 23 h. We thus believe that this discrepancy in performance between base and large models for these languages might be related to the amount of available data, since for large wav2vec 2.0 models we have an additional 221.2 million trainable parameters.[5] The lack of available data might result on these extra parameters not being properly fine-tuned, thus resulting in the observed deterioration in performance. We believe that the fine-tuning of base models might be more *realistic* in settings of data scarcity, as the overhead caused by the extra parameters in large architectures seem to be excessive for models working with less than 50 hours of speech.

Finally, it is also important to highlight that the results obtained for our baselines in this work are considerably higher compared to results obtained with AST models trained *without* SSL models as a foundation block. In [34], and for the same dataset, they reach BLEU scores of 8.9, 10.6 and 7.9 for English, Spanish and Portuguese respectively.

Summarizing, in this section we presented end-to-end AST models in mid-to-low-resource settings. For the language pair with the most available speech data (fr-en) we observe benefits on having large pre-trained SSL models, and we reach acceptable BLEU scores compared to the literature [38], [39]. This finding however does not hold as we reduce the amount of trainable data: even 35 h in Spanish seems not to be enough to fully fine-tune a large wav2vec 2.0 architecture, and results for large models drag behind the results for base models.

[3]BLEU4 scores computed using `sacreBLEU` [37].

[4]There are approximately 221.2 million extra parameters in the large architecture.

[5]In this work we do not explore partially freezing the wav2vec 2.0 blocks.

The challenge of low-resource AST is illustrated with the clear performance drop between our three setups: *fr-en* reaches higher performance than *fr-es* (data reduction of 13 h), and the latter outperforms *fr-pt* models (data reduction of 12 h compared to Spanish, 25 h compared to English). This highlights the existence of a minimal amount of data needed in order to make the training of end-to-end AST architectures based on SSL models exploitable.

B. Use case: AST for Tamasheq

We now present our experiments for the Tamasheq-French dataset in the context of the IWSLT 2022 low-resource speech translation track. The dataset contains 17 h of speech in the Tamasheq language, which corresponds to 5,829 utterances translated to French [40]. Additional audio data was also made available through the *Niger-Mali audio collection*: 224 h in Tamasheq and 417 h in geographically close languages (French from Niger, Fulfulde, Hausa, and Zarma).[6] For all this data, the speech style is radio broadcasting, and the dataset presents no transcription.

We start by training AST models that use wav2vec 2.0 models simply as feature extractors: the output of these SSL architectures replaces commonly used mel filterbank (MFB). This was shown to result in a considerable performance boost using the mTEDx dataset in [19], and we choose this approach since our results from the previous session hint that end-to-end fine-tuning for the AST task requires more data than the available 17 h.

In these settings, we compare two general purpose wav2vec 2.0 models – the multilingual XLSR-53 [22] and the French LB-7K-large [20] – against two smaller base models trained in the target language: Tamasheq-only, trained on 243 h of Tamasheq, and Niger-Mali, trained on the totality of the Niger-Mali audio collection (641 h).

Our AST models that use wav2vec 2.0 models as feature extractors are very close to the recipe for low-resource ST from wav2vec 2.0 features described in [20]. We use the *fairseq s2t* toolkit [41] for training an end-to-end AST Transformer model [36] with 4 heads, dimensionality of 256, inner projection of 1,024, 6 encoder and 3 encoder layers. The Transformer is preceded by a 1D convolutional layer (k=5, stride=2) for down-projecting the wav2vec 2.0 large (1,024) or base (768) features into the Transformer input dimensionality. These models are trained for 500 epochs using the Adam optimizer with 10k warm-up steps. For decoding, we use beam search with a beam size of 5. We generate a 1k unigram vocabulary for the French text using *Sentencepiece* [42], with no pre-tokenization. Lastly, we include baseline results that replace wav2vec 2.0 features by 80-dimensional MFB features. In this setting, the CNN preceding the transformer encoder is identical from the one in [20].

AST results using the four wav2vec 2.0 models as feature extractors are presented in Table II. For each model other than Tamasheq-only, we investigate fine-tuning on a *task-agnostic* fashion for approximately 20,000 updates on all available Tamasheq speech (243 h). This fine-tuning is supposed to reduce performance issues related to domain shift, however, in our setting we do not notice a significant performance gap between fine-tuned models and their pre-trained counterparts.

TABLE II
BLEU4 RESULTS FOR TAMASHEQ-FRENCH AST.

wav2vec 2.0 model	Fine-tuning	valid	test
None (MFB)	-	2.22	1.80
LB-FR-7K	-	2.36	1.80
LB-FR-7K	Task-agnostic	2.48	1.92
XLSR-53	-	2.05	1.42
XLSR-53	Task-agnostic	1.99	1.91
Niger-Mali	-	2.81	**2.68**
Niger-Mali	Task-agnostic	2.94	**2.57**
Tamasheq-only	-	2.99	**2.42**

Regarding the overall very low AST performance,[7] it is notable that the results obtained by using general purpose wav2vec 2.0 models are not very different from the baseline results (MFB). This was also observed in the literature: it seems wav2vec 2.0 models tend to perform poorly as feature extractors in low-resource settings, compared to MFB [43]. Finally, although performance is poor regardless of the feature extractor, the wav2vec 2.0 models trained on target data seem to output superior features for Tamasheq-French AST. This is despite the fact they are trained with considerably smaller quantities of data compared to the large and general purpose models.

Based on this finding, the best AST results we reported on [21] used these *smaller* wav2vec 2.0 models on an end-to-end fashion. However, as expected by our mid-to-low-resource baseline results from last section, there are not enough training hours to successfully fine-tune an entire wav2vec 2.0-based AST architecture for Tamasheq-French.[8]

We circumvent this by exploring the representation from intermediate layers: previous work [44] has shown that the middle layers inside the Transformer Encoder inside the wav2vec 2.0 architecture contain a higher abstraction level with respect to the speech signal, being more useful for end-to-end ASR fine-tuning compared to the last layers. Inspired by that finding, we experimented pruning the last layers of the wav2vec 2.0 model, which reduced the amount of trainable parameters. Our final model, and the best result for the IWSLT 2022 low-resource task, was an end-to-end wav2vec 2.0-based AST model using the Tamasheq-only model. It comprised only 7 Transformer layers (out of 12) on its wav2vec 2.0 foundation block, and **it achieved a BLEU4 of 6.0**.

C. SSL models for low-resource end-to-end AST

Throughout this section, we illustrated how the performance of wav2vec 2.0-based AST models drops in settings of data scarcity. One important aspect of our results is the complete

[6]https://demo-lia.univ-avignon.fr/studios-tamani-kalangou/

[7]High-resource end-to-end speech translation BLEU4 scores range from 19 to 30 in the last editions of IWSLT [38], [39].

[8]The best BLEU score for an end-to-end wav2vec 2.0 AST model was of 2.34, by using the Tamasheq-only wav2vec 2.0 model.

lack of transcription we impose to our experimental setup: by doing this we reduce the final translation scores, but we are able to assess performance for situations where this information is not available (e.g. the processing of oral dialects).

The main finding of our AST experiments is the overall under-performance of off-the-shelf wav2vec 2.0 large models in low-resource settings, even after fine-tuning them on target data. We also notice that the wav2vec 2.0 base models we trained with considerable less data were more effective in these same settings. We believe this happens because these task-specific SSL models better inform downstream tasks such as AST, since there is no domain shift in the data representation. This hints that massive multi-purpose wav2vec 2.0 models might not be the adequate solution for low-resource speech-to-text approaches, and that instead, smaller and better informed SSL blocks, trained on target data and/or domain, should be favored.

V. Final Remarks

This paper presented some premises and limitations of wav2vec 2.0 models. These are promising because they allow us to build ASR and AST models with less data than the previous approaches: this is an important issue since in-domain manually labeled data is very rare. This offers a new mean of action in order to address new languages.

Regarding our ASR experiments, we have seen that in order to improve recognition of accented speech, it is necessary to include accented speech in the fine-tuning data. Fine-tuning on both accented and non-accented speech seems to be a promising method for building general-purpose systems. In this work we used an equal amount of accented and native speech in our mixed dataset. The variation of the amount of accented speech in the fine-tuning dataset, the inclusion of additional accents (Swiss French, Quebec French), and the use of data augmentation to increase artificially the amount of accented speech are left as future work.

Regarding our AST experiments, we illustrated how building speech translation models without the existence of transcriptions is a challenging topic, and that models based on off-the-shelf wav2vec 2.0 models fail performance-wise in low-resource settings. Future research will focus on increasing performance in challenging settings: techniques such as speech augmentation, the production of *dummy transcriptions* [21], and multilingual pre-training and adaptation are promising topics.

The LIA will continue to study these approaches with the goal of making them highly accurate for real-world data.

Acknowledgment

This work used HPC resources from GENCI-IDRIS (grants 2020-A0111012991, 2021-AD011013317 and 2021-AD011013331). It was also funded by the European Commission through the SELMA project under grant number 957017.

References

[1] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," *arXiv preprint arXiv:1906.00910*, 2019.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://aclanthology.org/N18-1202

[5] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *arXiv preprint arXiv:1904.03240*, 2019.

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

[7] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[8] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," *arXiv preprint arXiv:1911.03912*, 2019.

[9] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[10] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3497–3501.

[11] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. v. d. Oord, "Learning robust and multilingual speech representations," *arXiv preprint arXiv:2001.11128*, 2020.

[12] M. Riviere, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7414–7418.

[13] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv*, vol. abs/2111.09296, 2021.

[14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[15] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations (ICLR)*, 2020.

[16] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, 2017.

[17] Q.-S. Zhu, J. Zhang, Z.-Q. Zhang, M.-H. Wu, X. Fang, and L.-R. Dai, "A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3174–3178.

[18] T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, "Data Augmentation Improves Recognition of Foreign Accented Speech," in *Proc. Interspeech 2018*, 2018, pp. 2409–2413.

[19] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, Y. Estève, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and L. Besacier, " *LeBenchmark*: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech," in *Interspeech*, 2021, pp. 1439–1443.

[20] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet *et al.*, "Task agnostic and task specific self-supervised learning from speech with *LeBenchmark*," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[21] M. Z. Boito, J. Ortega, H. Riguidel, A. Laurent, L. Barrault, F. Bougares, F. Chaabani, H. Nguyen, F. Barbier, S. Gahbiche, and Y. Estève, "ON-TRAC consortium systems for the IWSLT 2022 dialect and low-resource speech translation tasks," in *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Dublin, Ireland (in-person and online): Association for Computational Linguistics, May 2022, pp. 308–318. [Online]. Available: https://aclanthology.org/2022.iwslt-1.28

[22] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[23] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[24] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[26] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *ArXiv*, vol. abs/1212.5701, 2012.

[27] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[28] "African accented french," https://www.openslr.org/57/, accessed : 22-06-2022.

[29] S. Bird, "Bootstrapping the language archive: New prospects for natural language processing in preserving linguistic heritage," *Linguistic Issues in Language Technology*, vol. 6, no. 4, 2011.

[30] A. Berard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *CoRR*, vol. abs/1612.01744, 2016.

[31] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-Sequence Models Can Directly Translate Foreign Speech," in *Proc. Interspeech 2017*, 2017, pp. 2625–2629.

[32] H. Le, J. Pino, C. Wang, J. Gu, D. Schwab, and L. Besacier, "Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation," *arXiv preprint arXiv:2011.00747*, 2020.

[33] M. Sperber, H. Setiawan, C. Gollan, U. Nallasamy, and M. Paulik, "Consistent transcription and translation of speech," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 695–709, 2020.

[34] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and M. Post, "Multilingual tedx corpus for speech recognition and translation," in *Proceedings of Interspeech*, 2021.

[35] A. Anastasopoulos, L. Barrault, L. Bentivogli, M. Z. Boito, O. Bojar, R. Cattoni, A. Currey, G. Dinu, K. Duh, M. Elbayad, Y. Estéve, M. Federico, C. Federmann, S. Gahbiche, H. Gong, R. Grundkiewicz, B. Haddow, B. Hsu, D. Javorský, V. Kloudová, S. M. Lakew, X. Ma, P. Mathur, P. McNamee, K. Murray, M. Nădejde, S. Nakamura, M. Negri, J. Niehues, X. Niu, J. Ortega, J. Pino, E. Salesky, J. Shi, S. Stüker, K. Sudoh, M. Turchi, Y. Virkar, A. Waibel, C. Wang, and S. Watanabe, "FINDINGS OF THE IWSLT 2022 EVALUATION CAMPAIGN," in *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Dublin, Ireland: Association for Computational Linguistics, 2022.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[37] M. Post, "A call for clarity in reporting BLEU scores," in *Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: https://www.aclweb.org/anthology/W18-6319

[38] E. Ansari, A. Axelrod, N. Bach, O. Bojar, R. Cattoni, F. Dalvi, N. Durrani, M. Federico, C. Federmann, J. Gu *et al.*, "Findings of the iwslt 2020 evaluation campaign," in *Proceedings of the 17th International Conference on Spoken Language Translation*, 2020, pp. 1–34.

[39] A. Anastasopoulos, O. Bojar, J. Bremerman, R. Cattoni, M. Elbayad, M. Federico, X. Ma, S. Nakamura, M. Negri, J. Niehues, J. Pino, E. Salesky, S. Stüker, K. Sudoh, M. Turchi, A. Waibel, C. Wang, and M. Wiesner, "FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN," in *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*. Bangkok, Thailand (online): Association for Computational Linguistics, Aug. 2021, pp. 1–29. [Online]. Available: https://aclanthology.org/2021.iwslt-1.1

[40] M. Z. Boito, F. Bougares, F. Barbier, S. Gahbiche, L. Barrault, M. Rouvier, and Y. Estéve, "Speech resources in the tamasheq language," *Language Resources and Evaluation Conference (LREC)*, 2022.

[41] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, "fairseq s2t: Fast speech-to-text modeling with fairseq," *arXiv preprint arXiv:2010.05171*, 2020.

[42] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[43] D. Berrebbi, J. Shi, B. Yan, O. Lopez-Francisco, J. D. Amith, and S. Watanabe, "Combining spectral and self-supervised features for low resource speech recognition and translation," *arXiv preprint arXiv:2204.02470*, 2022.

[44] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," *arXiv preprint arXiv:2107.04734*, 2021.