Topic 5: ML-SUPERB: Multilingual Speech Universal PERformance Benchmark

May 19, 2025

Dani BOUCH
ENS Paris-Saclay
Final Project Report
dani.bouch@ens-paris-saclay.fr

Abstract

Self-supervised speech representation learning (SSL) models have shown remarkable potential in low-resource settings, offering rich contextual embeddings from unlabeled audio. In this work, we evaluate the performance of the multilingual SSL model XLS-R 128, based on the wav2vec 2.0 architecture, for Automatic Speech Recognition (ASR) in three typologically distinct languages: French, Occitan, and Kabyle. We fine-tune the model on limited training data (1 hour, 30 minutes, and 10 minutes) and assess performance using Character Error Rate (CER) and Word Error Rate (WER). Our results show that XLS-R achieves strong CER even with as little as 10 minutes of audio. We further test the model under extremely low-resource conditions (5 and 1 minute) and observe meaningful results, especially for Latin-script languages. Finally, we demonstrate that bilingual training with linguistically related languages (French and Occitan) improves ASR performance, highlighting the benefit of cross-lingual transfer in low-resource ASR.

1. Introduction

Self-supervised methods have emerged as new paradigms for training deep learning models, demonstrating significant results in text processing tasks [1] and extending their effectiveness to the image modality as well [2]. The speech processing community has similarly benefited from these advances, primarily due to the availability of vast amounts of unlabeled speech data, enabling models to learn rich contextual representations. SSL models such as wav2vec 2.0 [3] and HuBERT [4] have demonstrated substantial improvements across various downstream tasks when fine-tuned on labeled datasets. Among these, Automatic Speech Recognition (ASR) is particularly noteworthy, focusing on accurately converting spoken language into text. This has a wide array of applications of that include but not limited to real-time captioning for online meetings and interaction with voice-controlled assistants.

The goal of this project is to assess the performance of multilingual SSL models, specifically a multilingual model based on the wav2vec architecture, when fine-tuned for ASR using limited labeled data following the ML-SUPERB [5] protocol. The aim is to evaluate performance across different languages with varying degrees of proximity to English and each other, quantify the impact of training data volume on performance, and determining whether meaningful results can be achieved even in extremely low-data scenarios. Additionally, we explore whether jointly performing multilingual ASR on typologically and phonetically similar languages can enhance individual language performance, thus addressing key challenges faced by low-resource language communities.

2. Materials and methods

2.1. Languages and datasets:

We utilized the Common Voice dataset [6] , a crowdsourced speech corpus containing recordings from speakers across various languages. Each language-specific subset of the dataset comprises audio clips paired with corresponding transcriptions, and frequently includes additional metadata such as the speaker's age and gender, which we did not utilize in this project. Specifically, we selected three distinct languages to evaluate the performance of our multilingual ASR model: French, a widely-spoken which we included as a baseline of a forgein language different than English; Occitan, a rare regional language primarily spoken in the south of France; and Kabyle, a dialect of the Tamazight language spoken predominantly in North Africa. For each of these languages, we created training datasets of varying lengths: 1 hour, 30 minutes, 10 minutes, 5 minutes, and 1 minute. Additionally, an evaluation set and a testing set of 10 minutes each were prepared for assessing model performance.

2.2. Data preparation and vocabulary creation:

We began by preparing our data through discarding unnecessary metadata. Given that Common Voice is a crowd-

sourced speech corpus, the transcriptions often include extensive textual content with special characters irrelevant to the speech recognition task. To ensure consistency and facilitate model training, we removed all occurrences of special characters and normalized the transcriptions by converting text to lowercase and appending a word separator token at the end of each transcription. Additionally, for each language, we extracted the unique characters from the cleaned datasets to construct the language-specific vocabularies to further build the model's tokenizer. Careful consideration was taken in this part to include only meaningful characters, ensuring no essential linguistic symbols were omitted depending on the language. Finally, we appended [PAD] and [UNK] tokens to our vocabulary, which are necessary for utilizing the Connectionist Temporal Classification (CTC) [7] loss during model training, as explained later in this work.

2.3. Pretrained Model:

Initially, our goal was to compare various SSL approaches as outlined in the ML-SUPERB framework. However, due to computational constraints, we chose to focus on a single method, specifically the wav2vec approach. Within this category, we selected the XLS-R 128 [8] model, specifically the 300 million-parameter variant, known for achieving excellent performance in monolingual and multilingual ASR tasks.

XLS-R was pretrained on approximately 436,000 hours of unlabeled audio data spanning 128 different languages, significantly surpassing previous models in scale and multilingual coverage (see Figure 1). This extensive training set combines publicly available datasets, including VoxPopuli (parliamentary speech), Multilingual LibriSpeech (audiobooks), CommonVoice (read speech), VoxLingua107 (YouTube speech), and BABEL (telephone conversations), providing diverse acoustic conditions and speaking styles to enhance the robustness and cross-lingual capabilities of the learned representations.

For practical implementation, we accessed and utilized the pretrained XLS-R 128 model provided by Hugging Face, where the pretrained weights were directly accessible.

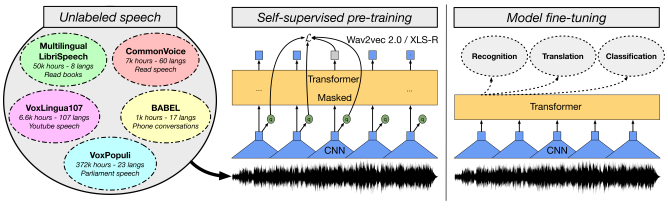


Figure 1. **Self-supervised cross-lingual representation learning with XLS-R [8]**

2.4. Architecture for ASR:

Wav2vec-based models, such as XLS-R, generally consist of a convolutional neural network (CNN) encoder that extracts audio features from raw waveform data, followed by a transformer-based encoder that projects these features into context-rich representations. Finally, a linear layer serves as the language modeling head, mapping representations into vocabulary tokens. More detail on how the Wav2vec architecture works can be found in the Appendix 5.1.

To adapt the pretrained XLS-R model for Automatic Speech Recognition (ASR), we fine-tuned it using Connectionist Temporal Classification (CTC) loss. CTC is a powerful algorithm specifically designed for sequence-to-sequence problems like ASR and handwriting recognition, enabling the model to align input sequences (audio) with corresponding output sequences (transcriptions) without requiring explicit alignment at each timestep.

In our case, we used the **Wav2Vec2FeatureExtractor** module from Hugging Face to process the raw audio and extract meaningful features. These features are then passed through the XLS-R encoder, and we trained only the final linear layer on top using the CTC loss. This layer takes the high-dimensional contextual embeddings and projects them onto a smaller dimensional space that corresponds to the vocabulary size of the target language (i.e., one logit per character in the vocabulary, plus [PAD] and [UNK]). Each time step of the output sequence thus represents a probability distribution over possible characters, which are then decoded into the final transcription using CTC decoding. For this training we leveraged the **Trainer** package from Hugging Face and defined a special padding data collator that dynamically pads the training examples to the longest example in the batch and handles the Inputs (Audio waveforms) and the Labels (transcriptions) padding in a different way, while making sure the different padding tokens are ignored in the loss calculation.

2.5. Evaluation metrics:

For evaluation, we primarily used the Character Error Rate (CER), following the ML-SUPERB benchmark setup. Additionally, we computed the Word Error Rate (WER) to better capture recognition accuracy across the different languages. Both metrics were computed using the `evaluate` library provided by Hugging Face.

3. Experiments and Results

3.1. Training setup

As discussed before, training was conducted on three languages—French, Occitan, and Kabyle—using the same setup and procedures across all cases. Due to limited computational resources, we leveraged Google Colab GPUs, primarily using Tesla T4 instances and occasionally A100

GPUs when Colab credits were available. In total we used over 30 dollars of Google collab credits

We set the learning rate to 7e-5, which we selected after testing several values. The optimizer used was AdamW, with a warmup ratio of 0.1 and a batch size of 16. While in ML-SUPERB monolingual ASR experiments run for 15,000 steps, this was not feasible given our time and resource constraints. Instead, we trained all models for 5,000 steps, except for the 5-minute and 1-minute data settings, which were trained for 2,000 and 500 steps respectively. Evaluation using both CER and WER was performed every 500 training steps. We also used Weights and Biases to track the different experiments. You can find some of the training loss curves in Appendix 5.2.

3.2. Monolingual ASR results

As done in the ML-SUPERB paper, we trained monolingual ASR models using 1 hour and 10 minutes of audio. We also included an additional training condition using 30 minutes of data for each of the target languages—French, Occitan, and Kabyle—to have a sort of middle-ground experiment. The results, in terms of Word Error Rate (WER) and Character Error Rate (CER), are presented in Table 1.

As shown in Table 1, the results are overall promising. Notably, even with as little as 10 minutes of training data, the model was able to achieve a CER of **0.1834** for Occitan. Despite Kabyle using non-Latin characters, it still reached a CER of **0.3884** under the same 10-minute setup. Interestingly, Occitan consistently outperformed French across all durations, reaching a CER as low as **0.0780** with 1 hour of data and even **0.1834** with as little as 10 minutes of audio. This was somewhat surprising, and we hypothesize that the simpler grammar and vocabulary of Occitan compared to French may have contributed to this performance.

Kabyle, on the other hand, clearly showed weaker performance compared to the other two languages, especially in terms of WER. This was expected due to the script differences and possible mismatch between character boundaries and word boundaries. From a qualitative analysis (Appendix 5.3), we observed that the model often predicted the correct characters but split one word into two, or merged two words into one. Despite these challenges, the model still achieved a CER of **0.2133** for Kabyle using the 1-hour dataset, which is a strong result given the divergence of the language compared to the other two. Overall, we clearly see that the amount of data we have influences the performance of the model and that the more data we use, the better the results are which is expected. With only 1 hour of data we can get very decent results for the CER and achieve accurate character transcriptions, but WER still struggles quite often and could definitely benefit from training with more data to achieve better and usable results for real-life ASR systems.

Qualitative prediction examples for each language on the test dataset can be found in Appendix 5.3.

Language	Duration	WER	CER
Kabyle	1h	0.6931	0.2133
	30m	0.7623	0.2470
	10m	0.8553	0.3884
French	1h	0.4262	0.1360
	30m	0.4646	0.1517
	10m	0.6437	0.2345
Occitan	1h	0.2622	0.0780
	30m	0.4045	0.1144
	10m	0.6216	0.1834

Table 1. Monolingual ASR results in terms of WER and CER for each language and training duration.

3.3. Monolingual ASR with extremely limited data

We also wanted to evaluate how well the model performs in extremely low-resource settings, using only 5 minutes and even as little as 1 minute of training data for each language. We followed the same training procedure outlined in Section 3.1. While adjusting the training steps to 2000 for the 5 minutes experiments and to only 500 for the 1 min experiments to avoid extreme overfitting on the limit amounts of data. The results, presented in Table 2, show that despite the very limited data, the model is still able to achieve a CER <0.35 for both French and Occitan, even with only 1 minute of training data. This is likely due to the shared Latin script and more abundant pretraining data in related languages. However, the data is clearly insufficient to produce meaningful WER scores, as accurate word segmentation requires more training. Interestingly enough, French outperforms Occitan in terms of WER in this extremely low data regime. For Kabyle, the model struggles significantly more, with CER values of **0.8326** and **0.7086** for 1-minute and 5-minute datasets, respectively. This performance gap is probably due to several factors, including the use of non-Latin characters, less pretraining exposure to similar languages, and potentially more complex or less standardized orthography in the dataset. However this still shows the strong capabilites of such SSL models especially on abundent languages like French and similar ones such as Occitan.

3.4. Bilingual Finetuning

One of the reasons we chose Occitan as a language to evaluate is its close proximity to French. We hypothesized that bilingual ASR—training on both languages simultaneously—could improve performance compared to monolingual finetuning on each language individually. To test this, we designed an experiment that mimics a realistic low-

Language	Duration	WER	CER
Kabyle	5m	0.9836	0.7086
	1m	0.9824	0.8326
French	5m	0.7362	0.2911
	1m	0.7939	0.3400
Occitan	5m	0.8529	0.2826
	1m	0.8612	0.2835

Table 2. Monolingual ASR performance using extremely limited training data (5 minutes and 1 minute).

resource scenario, where we only have 10 minutes of Occitan data and 1 hour of French data. These two datasets were concatenated, and a joint vocabulary was created across both languages. We followed the same training procedure as outlined in Section 3.1 and evaluated the model separately on the French and Occitan test sets. The goal was to see if adding an hour of French data would improve the results when having 10 minutes of Occitan and also if the 10 minutes of Occitan could improve the performence on French.

The results, summarized in Table 3, show that bilingual finetuning outperforms the monolingual baseline for both languages in terms of both WER and CER. As we hypothesized, it could potentially be benefical to add a similar language to the existing rare language to improve results. However, it's worth noting that this isn't always true and such training could lead to diminishing results for both languages.

Language	Setup	WER	CER
French	Monolingual (1h)	0.4262	0.1360
	Bilingual (1h + 10m)	**0.3825**	**0.1137**
Occitan	Monolingual (10m)	0.6216	0.1834
	Bilingual (1h + 10m)	**0.5469**	**0.1549**

Table 3. Comparison of monolingual and bilingual finetuning results on French and Occitan test sets.

4. Conclusion

Throughout this project, we encountered several challenges, including limited computational resources, coordination issues within the team, and unstable training behavior for some of the models. Despite these obstacles, we successfully evaluated the performance of the state-of-the-art self-supervised learning model XLS-R 128 on Automatic Speech Recognition (ASR) for three linguistically distinct languages: French, Occitan, and Kabyle.

Our experiments demonstrated that while the amount of training data plays a crucial role in ASR performance, models like XLS-R 128 are still capable of producing surprisingly strong results even under extremely low-resource settings, such as with just 1 or 5 minutes of data. We observed that performance varied significantly across languages. French, being both a high-resource language and shares similarities with English (which dominates pretraining data), achieved strong results even with limited supervision. Occitan, although a rare language, shares substantial linguistic similarity with French and other high resource romance languages present in the model's pretraining and benefited from this proximity—achieving very promising results despite the small dataset size. Kabyle, on the other hand, presented more of a challenge due to its different script (non-Latin), richer morphology, and distinct phonetic system, yet still produced reasonable outputs under the same settings.

Lastly, we explored a bilingual training setup and found that multilingual ASR can offer notable advantages when training on closely related languages. Our bilingual finetuning experiment showed improvements for both French and Occitan, confirming that linguistic proximity can be leveraged to enhance performance, especially for low-resource languages.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, et al. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 1

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, et al. Emerging properties in self-supervised vision transformers, 2021. 1

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. 1

[4] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021. 1

[5] Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, et al. Ml-superb: Multilingual speech universal performance benchmark, 2025. 1

[6] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, et al. Common voice: A massively-multilingual speech corpus, 2020. 1

[7] Alex Graves, Santiago Fernández, Faustino Gomez, Jürgen Schmidhuber, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 369–376, New York, NY, USA, 2006. ACM. 2

[8] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale, 2021. 2

5. Appendix

5.1. Wav2Vec 2.0 Architecture

The Wav2Vec 2.0 architecture, which forms the backbone of XLS-R, consists of three main components as shown in Figure 2:

- **Feature Encoder** ($\mathcal{Z}$): A stack of temporal convolutions that takes in raw waveform $\mathcal{X}$ and outputs latent speech representations $\mathcal{Z}$.
- **Quantization Module** ($\mathcal{Q}$): The latent representations are quantized to discrete tokens using a quantization module. These tokens serve as targets in contrastive learning.
- **Context Network** ($\mathcal{C}$): A Transformer that takes masked latent features and outputs contextualized representations $\mathcal{C}$. A contrastive loss $\mathcal{L}$ is applied between the masked outputs and their corresponding true quantized targets.

This architecture enables the model to learn robust contextual speech representations in a self-supervised manner from large amounts of unlabeled audio.

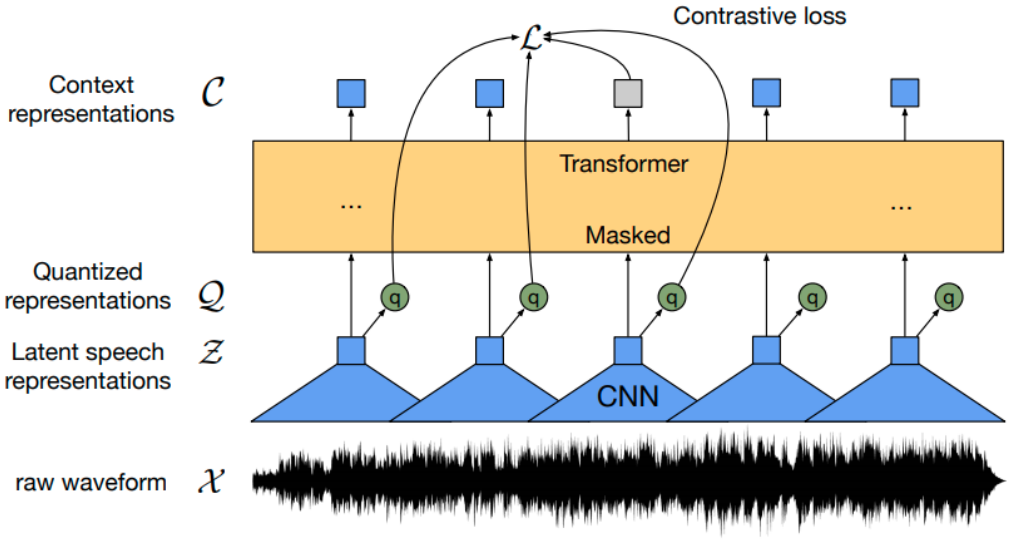
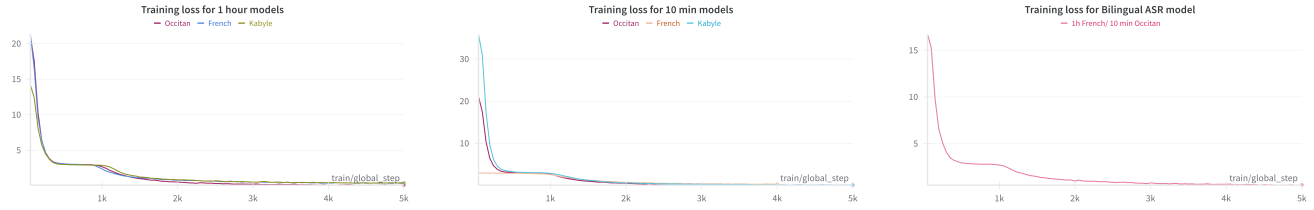


Figure 2. Wav2Vec 2.0 architecture: SSL learning using a CNN encoder, quantization, and masked transformer with contrastive loss.

5.2. Training loss curves:

Below, in Figure 3, we present the training loss curves obtained using Weights Biases (W&B) for three different settings: models trained on 1 hour of data, models trained on 10 minutes of data, and the bilingual model trained on 1 hour of French and 10 minutes of Occitan.



1 hour models | 10 minute models | Bilingual (1h French + 10m Occitan)

Figure 3. Training loss curves for the different experimental setups.

5.3. Qualitative results:

To better understand the model behavior beyond aggregated metrics, we include a few qualitative examples from the predictions made by the 1-hour trained models on the test dataset for each of the three languages.

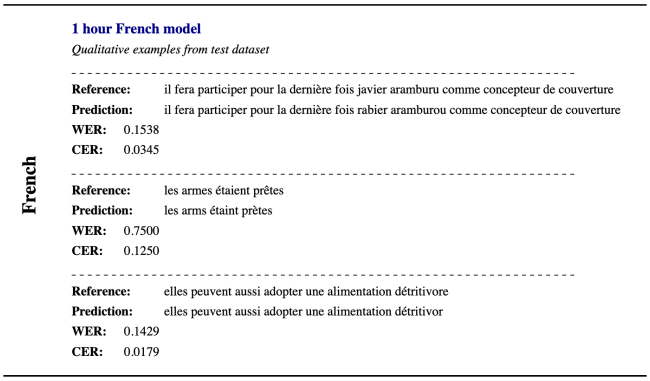


Figure 4. Qualitative examples from the French ASR model (1-hour training).

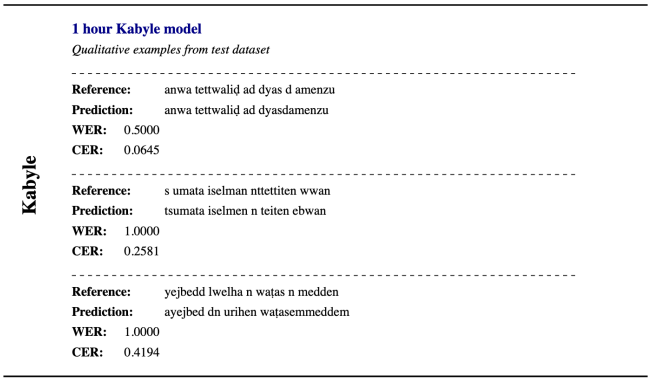


Figure 5. Qualitative examples from the Kabyle ASR model (1-hour training).

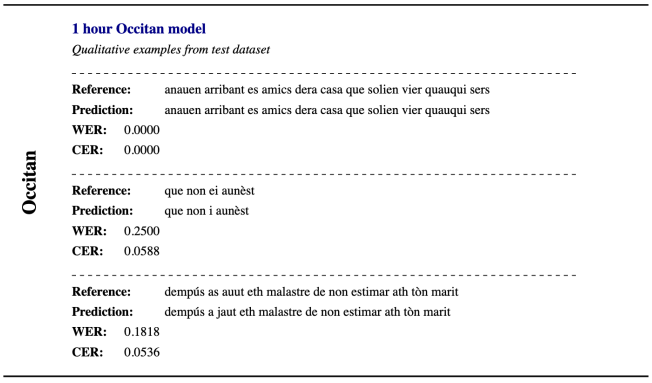


Figure 6. Qualitative examples from the Occitan ASR model (1-hour training).