mHuBERT-147:
A Compact Multilingual HuBERT Model

Marcely Zanon Boito♦, Vivek Iyer♣, Nikolaos Lagos♦, Laurent Besacier♦, Ioan Calapodescu♦
♦NAVER LABS Europe - FR ♣ University of Edinburgh - UK

NAVER LABS Europe

europe.naverlabs.com

In this paper we introduce mHuBERT-147

A. We train a multi-iteration multilingual HuBERT model from scratch using **high-quality data from 147 languages.**

B. For **faster data pre-processing**, we leverage faiss indices for multilingual clustering. **For balanced multilingual training,** our **two-level up-sampling approach** considers both language and data source.

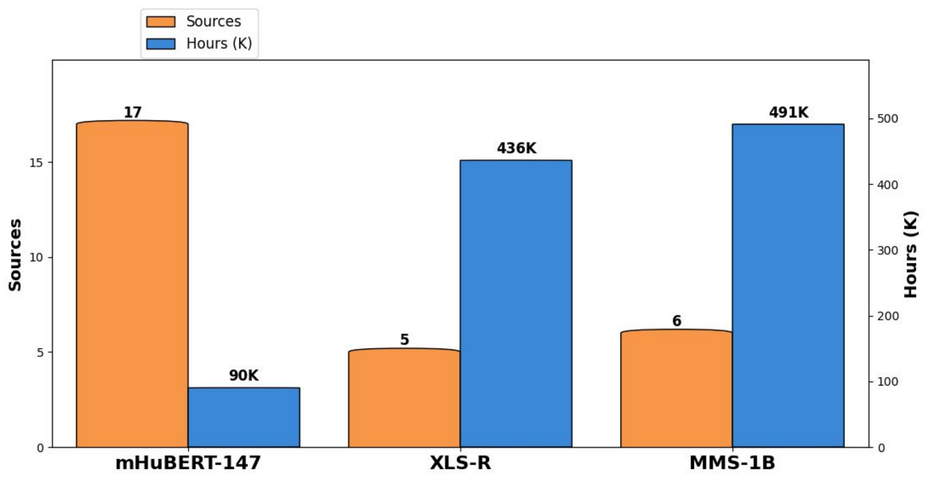
C. The result is a compact **95M parameter model** that is able to beat models 3-10x larger on ML-SUPERB, offering an **unprecedented balance between high performance and parameter efficiency.**

A. *Better* data for SSL pre-training

Focus on source diversity and quality instead of sheer amount. We use open license data only.

Key characteristics:

- 90K hours of speech after filtering,
- 17 datasets, 19 linguistic families, 147 languages.

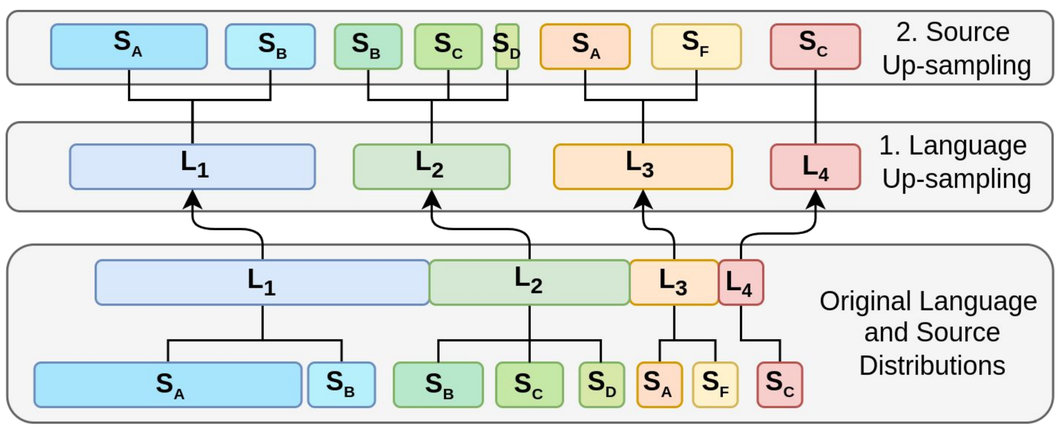


B. mHuBERT-147 Training

➔ **Faster pre-processing:** Validation of faiss indices for k-means clustering (5.2x times faster application).

➔ **Longer training:** 2M updates per iteration, 3 iterations.

➔ **Two-level language and source up-sampling:**



C. mHuBERT-147: Compact yet Powerful

We evaluate **mHuBERT-147** on the ML-SUPERB benchmark.

- It matches the performance of MMS-1B, **despite being 10x smaller.** We achieved 2nd and 1st place rankings.
- It reaches 3 new SOTA scores in LID tasks.

Table 1 - ML-SUPERB leaderboard (10min/1h)

SSL Block	Parameters	SUPERB 10min (↑)	SUPERB 1h (↑)
MMS-1B	965 M	**983.5** (1st)	948.1 (2nd)
mHuBERT-147	**95 M**	949.8 (2nd)	**950.2** (1st)
MMS-300M	317 M	824.9	844.3
XLS-R-300M	317 M	730.8	850.5
WavLabLM-large-MS	317 M	707.5	740.9

mHuBERT-147: A *reproducible* pipeline

- We share all the code necessary to reproduce or continuous pre-train the **mHuBERT-147** model.
- We include manifest files listing all files used for training.

Data pre-processing

➔ Dataset specific scripts
➔ VAD filtering
➔ Faiss clustering scripts

License: Apache

Code for Training

➔ Data loading optimizations
➔ Two-step language and source up-sampling

License: MIT

Trained Models

➔ Weights for all iterations
➔ Trained faiss indices
➔ Manifest files

License: CC-BY-NC-4.0

Acknowledgements: This is an output of the **European Project UTTER** (Unified Transcription and Translation for Extended Reality) funded by European Union's Horizon Europe Research and Innovation programme under grant agreement number 101070631.

UTTER

Contact: marcely.zanon-boito@naverlabs.com