mHuBERT-147: A Compact Multilingual HuBERT Model

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, Ioan Calapodescu

07/2024

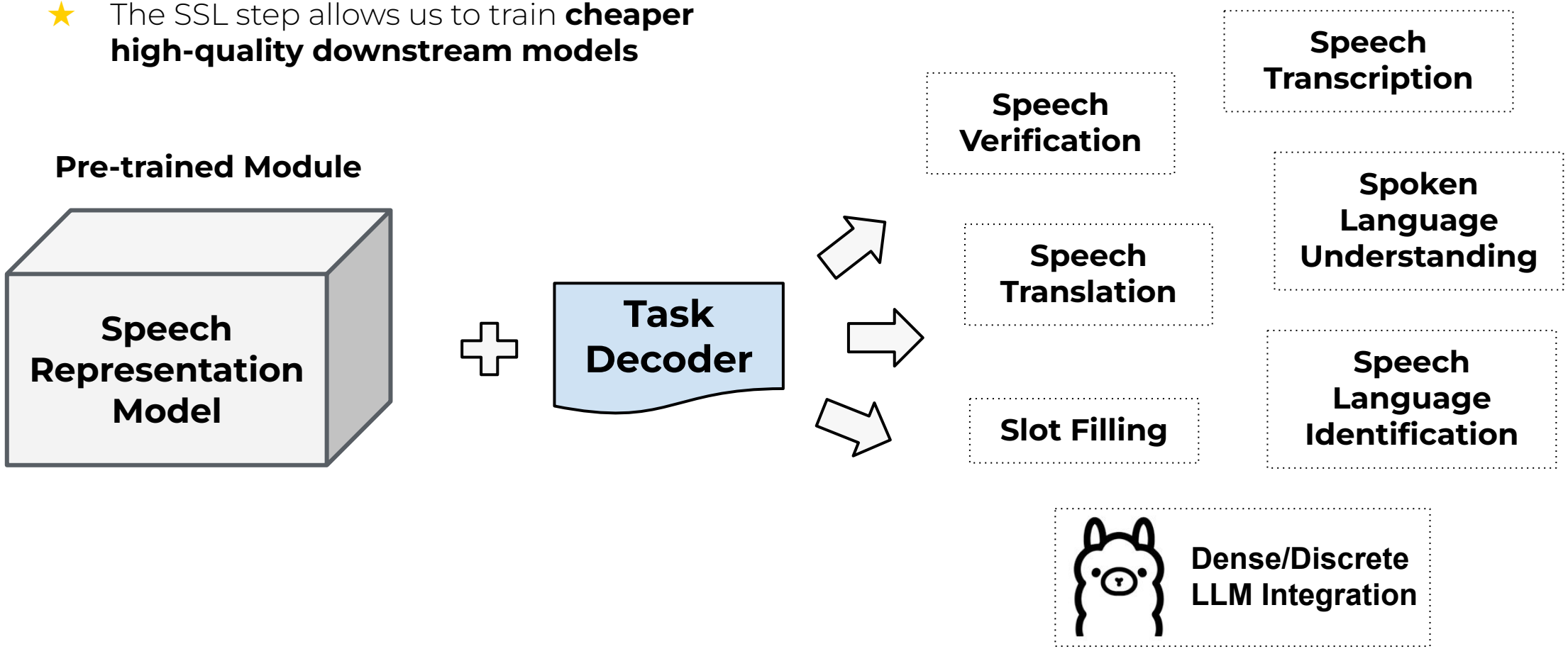
Contact: marcely.zanon-boito@naverlabs.com

NAVER LABS

Speech Representation Learning: learning contextualized high-dimensional *speech embeddings* for downstream tasks

Application: Any speech-*to-something* application

★ The SSL step allows us to train **cheaper high-quality downstream models**



Multilingual Speech Representation Models:

- ★ One single back-bone for (multilingual) speech applications
- ★ Zero-shot unseen languages

Model	Training Approach	# Languages	# Datasets	# Hours
XLSR-53 (Conneau et al. 2020)	wav2vec 2.0	53	3	56K
XLS-R (Babu et al. 2021)	wav2vec 2.0	128	5	436K
MMS (Pratap et al. 2023) **[SOTA]**	wav2vec 2.0	1,362	6	491K
WavLabLM (Chen et al. 2023)	WavLM	136	6	40K

Multilingual Speech Representation Models:

- ★ One single back-bone for (multilingual) speech applications
- ★ Zero-shot unseen languages

Model	Training Approach	# Languages	# Datasets	# Hours
XLSR-53 (Conneau et al. 2020)	wav2vec 2.0	53	3	56K
XLS-R (Babu et al. 2021)	wav2vec 2.0	128	5	436K
MMS (Pratap et al. 2023) **[SOTA]**	wav2vec 2.0	1,362	6	491K
WavLabLM (Chen et al. 2023)	WavLM	136	6	40K
mHuBERT-147	**HuBERT**	**147**	**17**	**90K**

mHuBERT-147: A Compact Multilingual HuBERT Model

A. **HIGH-QUALITY DATA:** Prioritizing open-license smaller collections and dataset diversity over data quantity

B. **HuBERT TRAINING:** Most of the multilingual models follow the wav2vec 2.0 training approach

C. **COMPACT SIZE:** Existing models are quite large (317M to 2B parameters), resulting in large downstream applications

DATA: 90,430 hours of diverse high-quality data

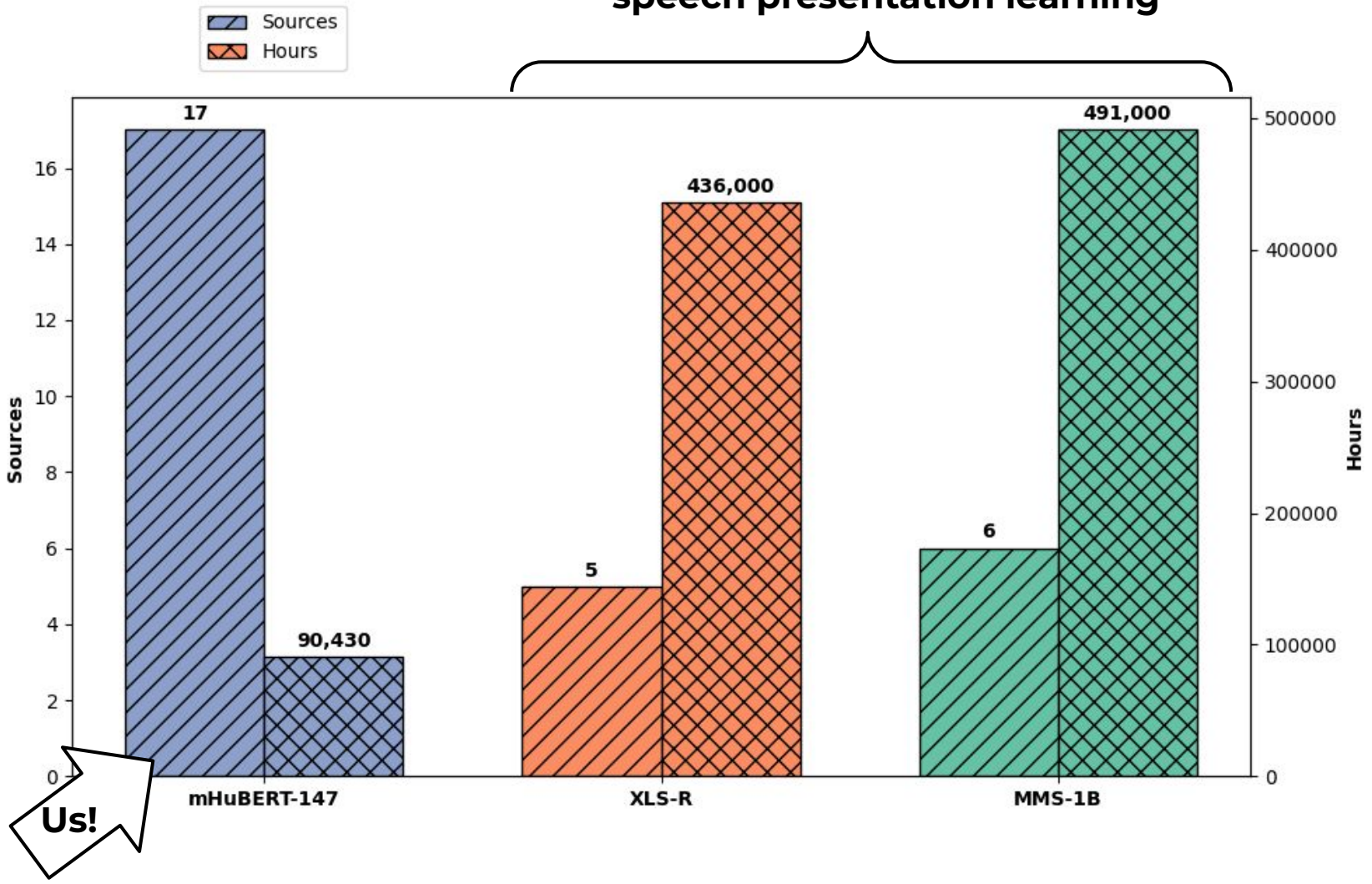
- ➔ We gather **17 datasets** across **147 languages**
- ➔ We filter popular large datasets **removing noise/music**
- ➔ We **downsample high-resource** languages (>2,000 hours per source)

Dataset	Full Names and References	# Languages	# Hours (filtered)	License
Aishells	Aishell [4] and AISHELL-3 [41]	1	212	Apache License 2.0
B-TTS	BibleTTS [20]	6	358	CC BY-SA 4.0
Clovacall	ClovaCall [17]	1	38	MIT
CV	Common Voice version 11.0 [1]	98	14,943	CC BY-SA 3.0
G-TTS	High quality TTS data for Javanese, Khmer, Nepali, Sundanese, and Bengali Languages [42] High quality TTS data for four South African languages [46]	9	27	CC BY-SA 4.0
IISc-MILE	IISc-MILE Tamil and Kannada ASR Corpus [26,27]	2	406	CC BY 2.0
JVS	Japanese versatile speech [44]	1	26	CC BY-SA 4.0
Kokoro	Kokoro Speech Dataset [19]	1	60	CC0
kosp2e	Korean Speech to English Translation Corpus [9]	1	191	CC0
MLS	Multilingual LibriSpeech [32]	8	50,687	CC BY 4.0
MS	MediaSpeech [21]	1	10	CC BY 4.0
Samrómur	Samrómur Unverified 22.07 [43]	1	2,088	CC BY 4.0
TH-data	THCHS-30 [48] and THUYG-20 [37,38]	2	46	Apache License 2.0
VL	VoxLingua107 [45]	107	5,844	CC BY 4.0
VP	VoxPopuli [47]	23	15,494	CC0

DATA: How much is 90K hours of speech?

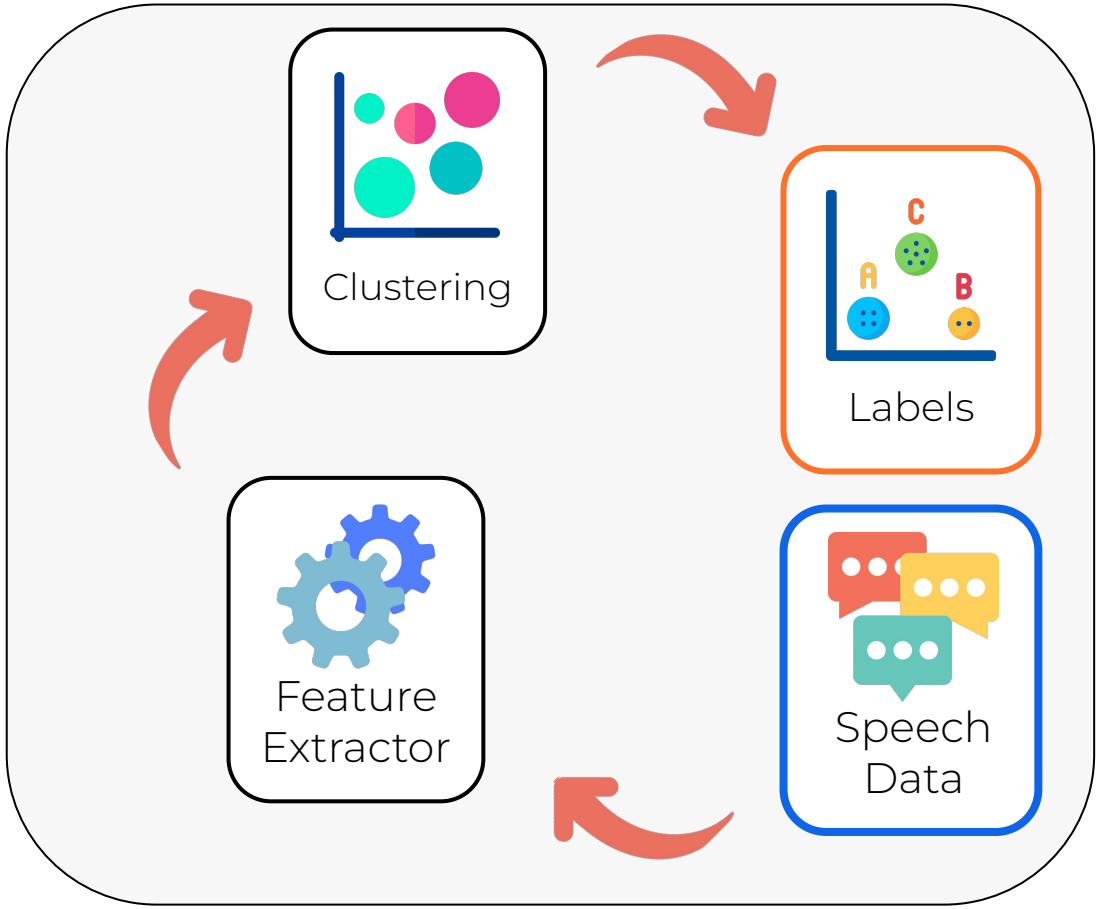
LESS training data compared to SOTA approaches

But **MORE** dataset diversity, **BETTER** language ratio



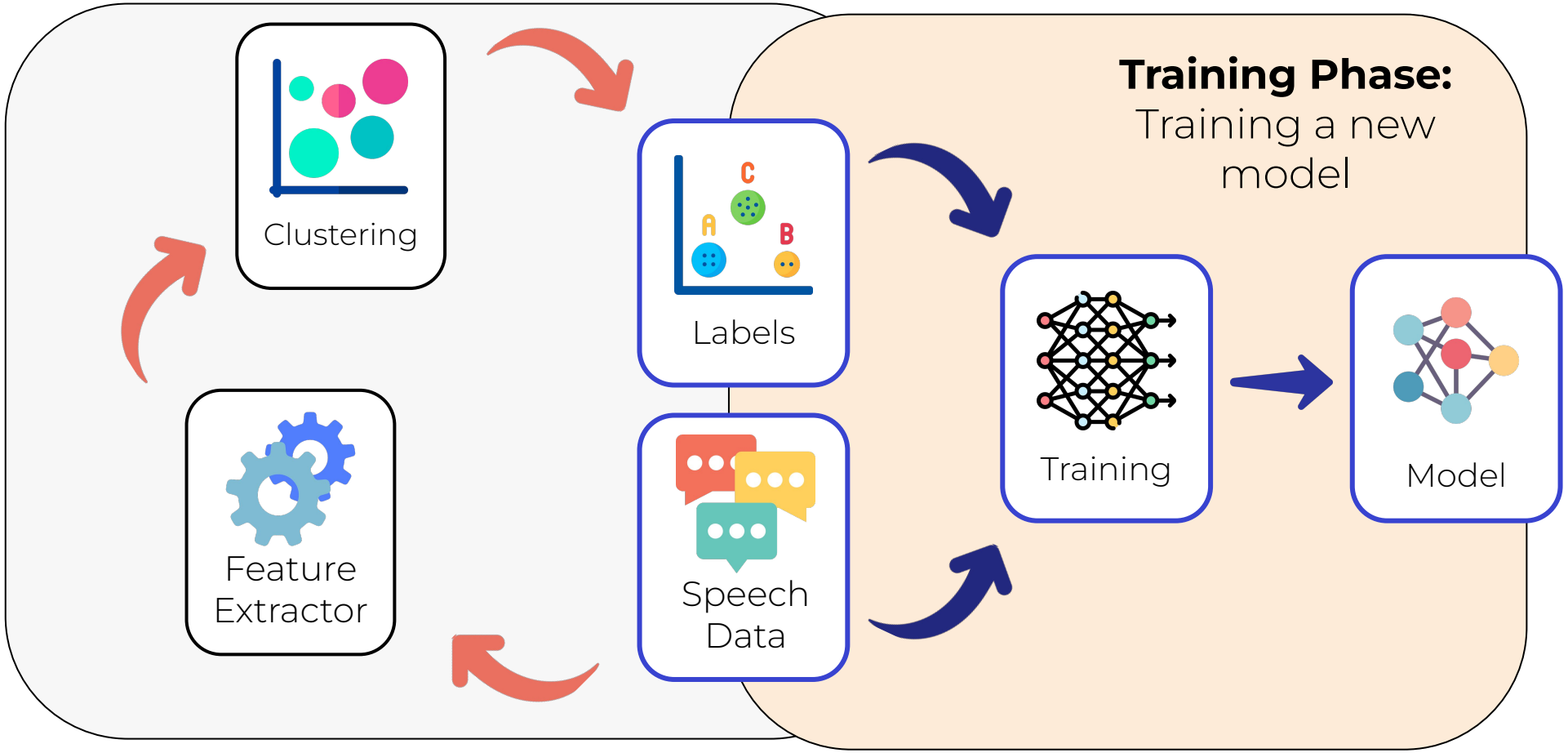
TRAINING APPROACH: **Hidden Units BERT** (Hsu et al. 2021)

TRAINING APPROACH: Hidden Units BERT (Hsu et al. 2021)

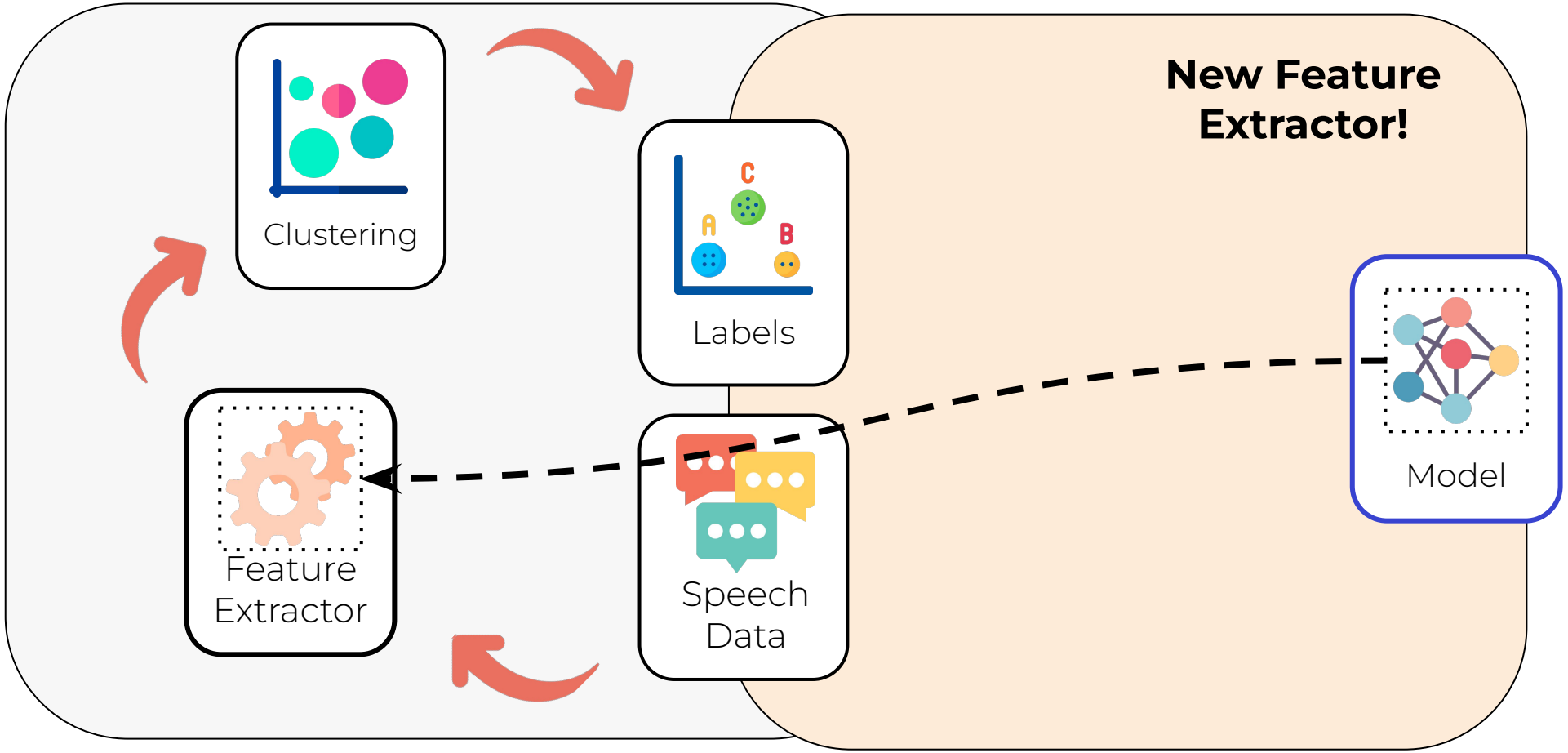


Labeling Phase: Creating fake discrete labels

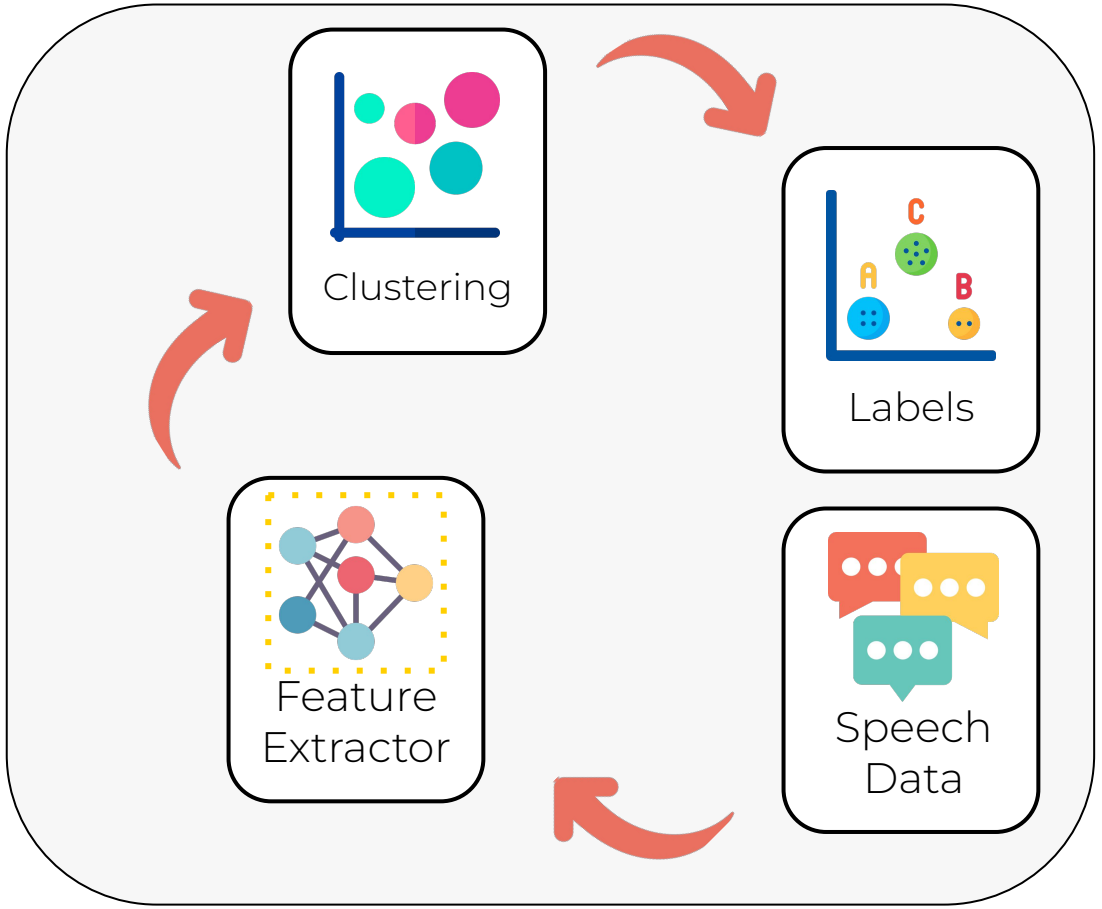
TRAINING APPROACH: Hidden Units BERT (Hsu et al. 2021)



TRAINING APPROACH: Hidden Units BERT (Hsu et al. 2021)

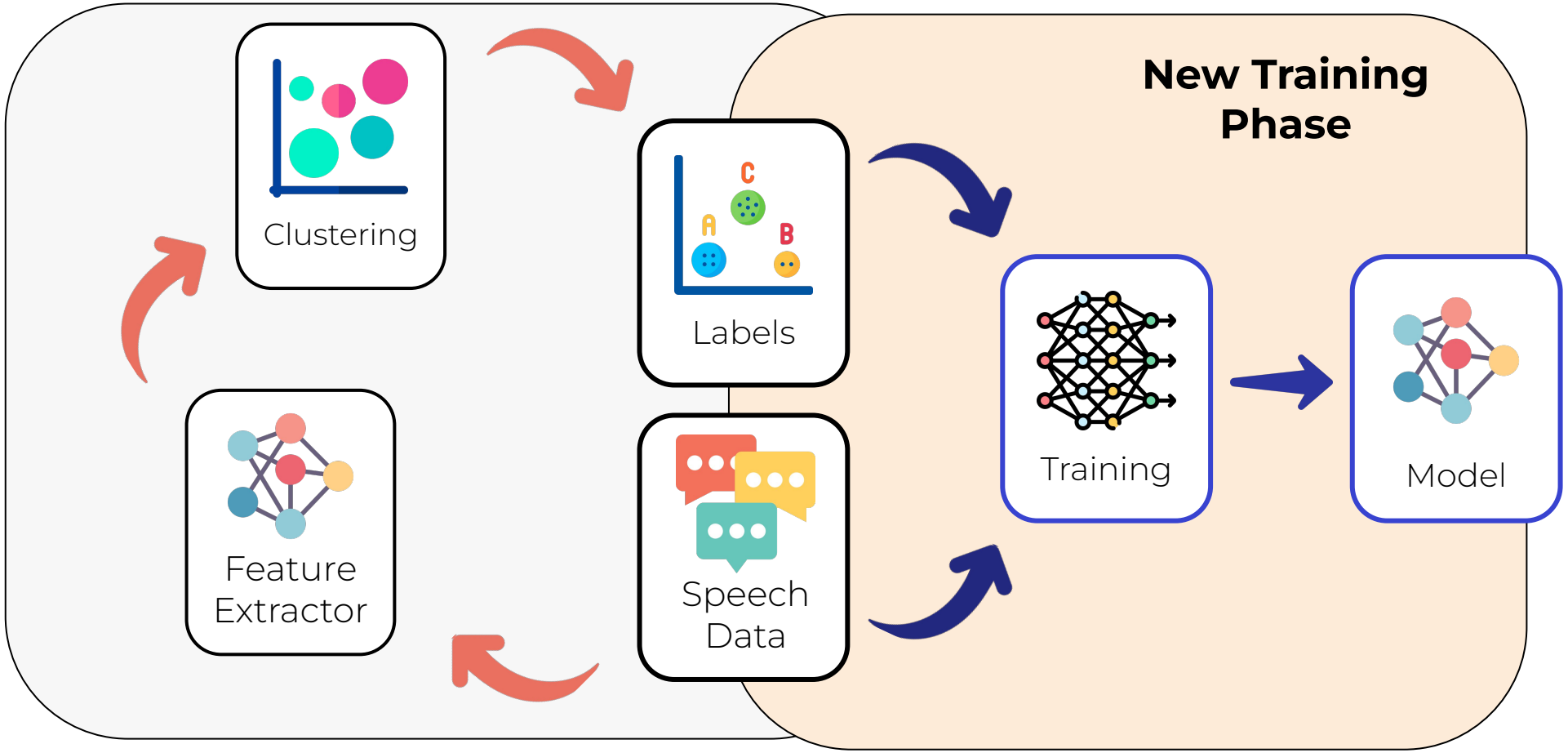


TRAINING APPROACH: Hidden Units BERT (Hsu et al. 2021)



New Labeling Begins!

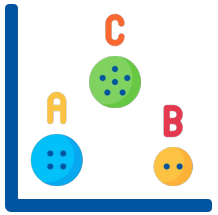
TRAINING APPROACH: Hidden Units BERT (Hsu et al. 2021)



Scaling the Approach for Multilingual Settings

A. Clustering/Labeling step are extremely expensive and slow

- We make training set smaller, but higher-quality
- We make clustering faster by using faiss + graph search for labeling **(5.2x times faster)**

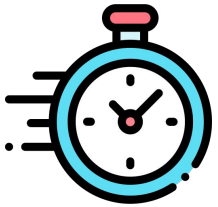


B. Multilingual data distribution

- We propose a multilingual batching approach: two-step language, data source up-sampling

C. Training iterations from longer!

- We find that existing speech representation models tend to be quite undertrained (accounting for some grokking)



The mHuBERT-147 models

- ★ **HIGH QUALITY DATA:** +90K hours in 147 languages
- ★ **COMPACT SIZE:** 95M parameters, 1000 discrete units
- ★ **TRAINED FOR LONGER:** 3 iterations, each for 2M updates (~20 days at 32xA100-80GB) at NAVER Cloud platform[1]

[1] https://www.ncloud.com/

Evaluation

NAVER LABS

EVALUATION: Multilingual Speech Representation Benchmark (ML-SUPERB)

A. **Two setups: 10min** and **1h** per language

B. **Two language settings: normal** (123 languages) and **few-shot** (20 languages)

C. **Four tasks:**
   - monolingual ASR
   - multilingual ASR
   - Language Identification (LID)
   - Multilingual ASR + LID

Final ranking SUPERB scores are computed considering SOTA and baseline scores

ML-SUPERB: Leaderboard Summary

1 1st place
2 2nd place
3 3rd place

SSL backbone	# Parameters	SUPERB Score 10min (↑)	SUPERB Score 1h (↑)
MMS-1B	**965M**	**983.5** 1	**948.1** 2
mHuBERT-147	**95M**	**949.8** 2	**950.2** 1
MMS-300M	**317M**	824.9 3	844.3
NWHC1 (MMS-300M variant)	317M	774.4	876.9 3
NWHC2 (MMS-300M variant)	317M	759.9	873.3
XLS-R-300M	317M	730.8	850.5
WavLabLM-large-MS	317M	707.5	740.9

Table: SUPERB scores (10min/1h).

ML-SUPERB: Leaderboard Summary

★ **mHuBERT-147** is competitive while being much more compact!

SSL backbone	# Parameters	SUPERB Score 10min (↑)	SUPERB Score 1h (↑)
MMS-1B	965M	**983.5** 1	**948.1** 2
mHuBERT-147	**95M**	**949.8** 2	**950.2** 1
MMS-300M	317M	824.9	844.3
NWHC1 (MMS-300M variant)	317M	774.4	876.9
NWHC2 (MMS-300M variant)	317M	759.9	873.3
XLS-R-300M	317M	730.8	850.5
WavLabLM-large-MS	317M	707.5	740.9

Table: SUPERB scores (10min/1h).

1 1st place

2 2nd place

ML-SUPERB: Detailed Scores

SSL backbone	# Params	Monolingual ASR CER (↓)		Multilingual ASR (normal/few-shot) CER (↓)		LID ACC (↑)		Multilingual ASR+LID (normal/few-shot) ACC - CER/CER (↓/↓)	
		10min	1h	10min	1h	10min	1h	10min	1h
MMS-1B	965M	**33.3**	**25.7**	**21.3/30.2**	**18.1/30.8**	84.8	86.1	73.3 - **26.0/25.4**	74.8 - 25.5/**24.8**
mHuBERT-147	**95M**	34.2	26.3	23.6/33.2	22.0/32.9	**85.3**	**91.0**	**81.4** - 26.2/34.9	**90.0** - **22.1**/33.5

Table: Detailed ML-SUPERB scores for the two best models.

★ **mHuBERT-147 reaches three new SOTA scores on ML-SUPERB**

★ Omitted from the table:
- **Competitive to MMS-300M**
  We beat it in all tasks but 3: few-shot CER LID+ASR (10min/1h) and monolingual ASR (10min)
- **mHuBERT-147 > XLS-R and WavLabLM** in all tasks

Summarizing:

- ★ We proposed the first multilingual HuBERT model, covering 147 languages
- ★ We approached data collection differently
- ★ We proposed a two-level language-source up-sampling for training
- ★ We produced a **compact** yet **powerful** foundation model that should be more exploitable for compact applications

mHuBERT-147 is an output of the EU project UTTER[1]

[1]Unified Transcription and Translation for Extended Reality. https://he-utter.eu/

mHuBERT-147: A Compact Multilingual HuBERT Model

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, Ioan Calapodescu

07/2024

Contact: marcely.zanon-boito@naverlabs.com

NAVER LABS