mHuBERT-147:
A Compact and Powerful Multilingual Speech Foundation Model

Marcely Zanon Boito

10/2024

Contact: marcely.zanon-boito@naverlabs.com

NAVER LABS

INTERSPEECH 2024

UTTER

- (2021) **PhD in Computer Science at University Grenoble Alpes**
  "Models and Resources for Attention-based Unsupervised Word Segmentation: an application to computational language documentation"

- (2021-2022) **Postdoc at Avignon University**
  Low-resource Speech Translation and Self-Supervised Learning for Speech (LeBenchmark project)

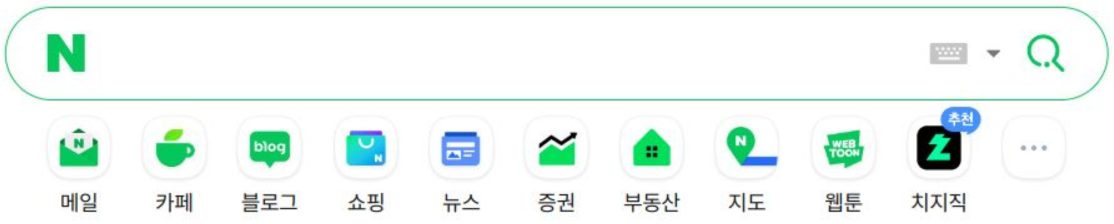
- (Since 2022) **Research Scientist at NAVER LABS Europe**
  Multimodality and Speech Processing

UGA
Université Grenoble Alpes

AVIGNON UNIVERSITÉ

NAVER LABS Europe

NAVER



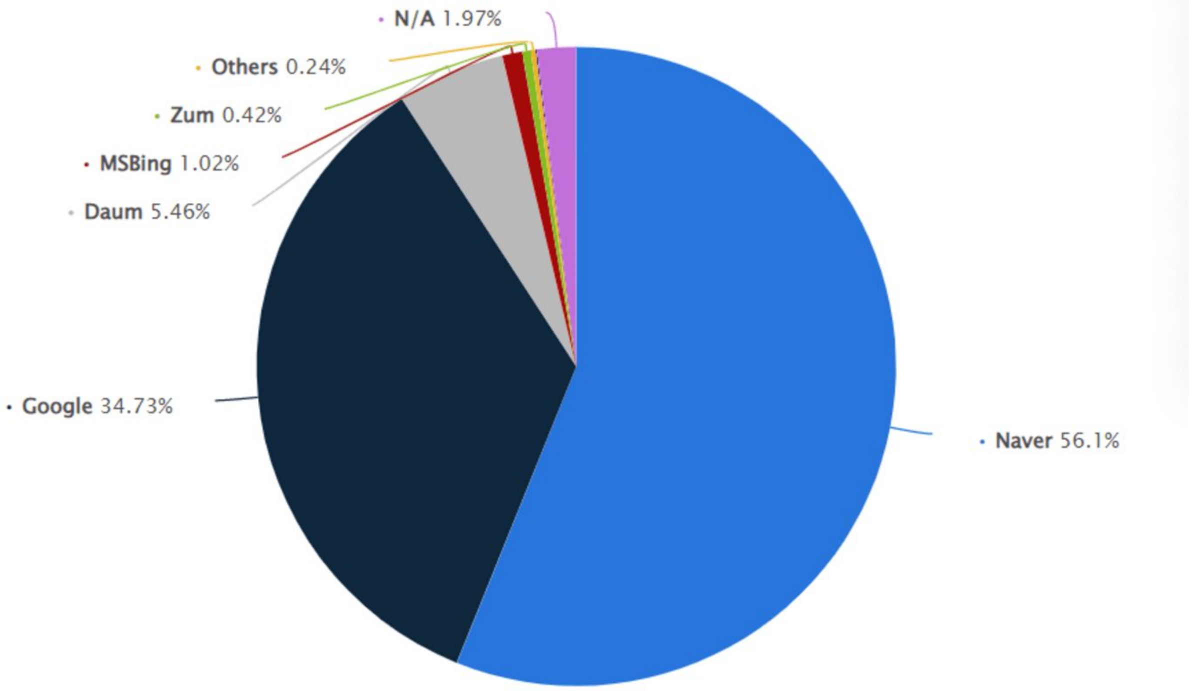
Huge collection of services.
Popular examples:

WEB TOON

NAVER Cloud

LINE

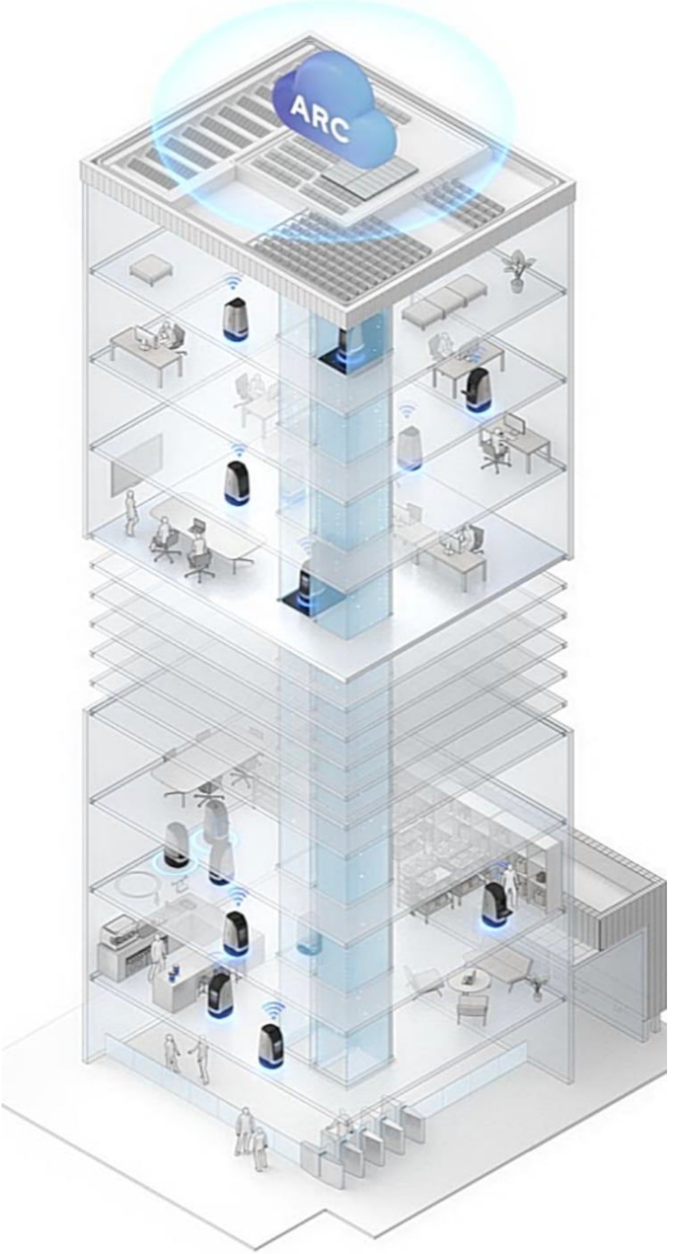
Search engine usage in South Korea:



2021, Source: https://www.link-assistant.com/news/naver-vs-google-in-korea.html

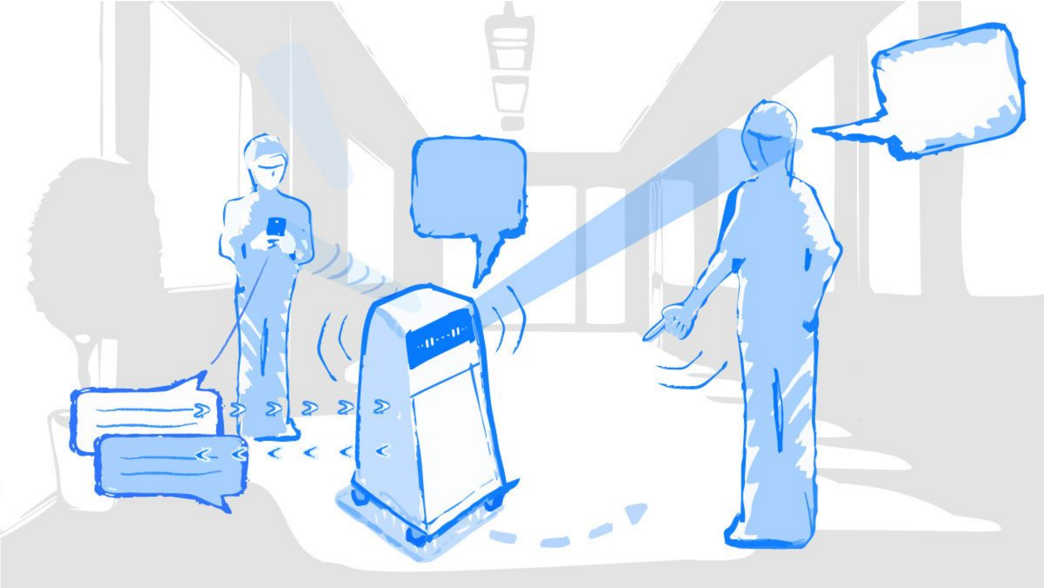
NAVER LABS

Adaptable robots for human environments



NAVER LABS → NAVER LABS Europe

- NAVER LABS Europe is a **fundamental research center**
- **Interactive Systems group** aims to equip robots with interaction (speech, text, other)



This presentation: mHuBERT-147

- Speech representation model
- Competitive with SOTA
- Compact size and multilingual
- Open-source

Outline:

1. Self-supervised learning for speech
2. Creating mHuBERT-147
3. Results on 3 settings

Self-supervised Learning for Speech

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

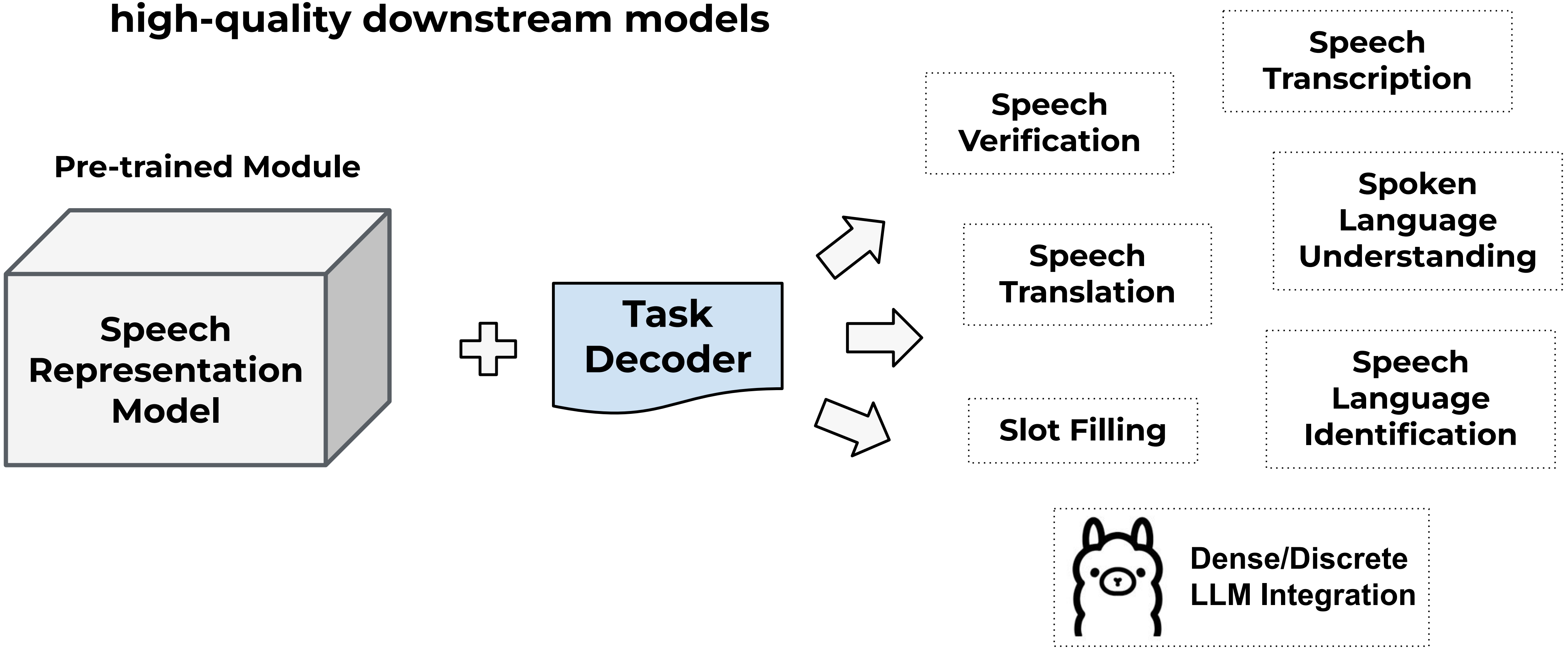
Transcriptions are **NOT** used!

e.g.:

- **wav2vec 2.0** (Baevski et al. 2020)
- **HuBERT** (Hsu et al. 2021)
- **WavLM** (Chen et al. 2021)
- **BEST-RQ** (Chiu et al. 2022)

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	# Parameters	# Languages	# Datasets	# Hours
XLSR-53 (Conneau et al. 2020)	wav2vec 2.0	300M	53	3	56K
XLS-R (Babu et al. 2021)	wav2vec 2.0	300M	128	5	436K
MMS (Pratap et al. 2023) **[SOTA]**	wav2vec 2.0	1B	1,362	6	491K
WavLabLM (Chen et al. 2023)	WavLM	300M	136	6	40K

Multilingual Speech Representation Models:

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

Model	Training Approach	# Parameters	# Languages	# Datasets	# Hours
XLSR-53 (Conneau et al. 2020)	wav2vec 2.0	300M	53	3	56K
XLS-R (Babu et al. 2021)	wav2vec 2.0	300M	128	5	436K
MMS (Pratap et al. 2023) **[SOTA]**	wav2vec 2.0	1B	1,362	6	491K
WavLabLM (Chen et al. 2023)	WavLM	300M	136	6	40K

First limitation: Training approach hardly differs!

Multilingual Speech Representation Models:

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

Model	Training Approach	# Parameters	# Languages	# Datasets	# Hours
XLSR-53 (Conneau et al. 2020)	wav2vec 2.0	300M	53	3	56K
XLS-R (Babu et al. 2021)	wav2vec 2.0	300M	128	5	436K
MMS (Pratap et al. 2023) **[SOTA]**	wav2vec 2.0	1B	1,362	6	491K
WavLabLM (Chen et al. 2023)	WavLM	300M	136	6	40K

Second limitation: The data diversity is small, but the amount of data is huge!
(WavLabLM seems like the exception, but this model performs very poorly)

Multilingual Speech Representation Models:

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

Model	Training Approach	# Parameters	# Languages	# Datasets	# Hours
XLSR-53 (Conneau et al. 2020)	wav2vec 2.0	300M	53	3	56K
XLS-R (Babu et al. 2021)	wav2vec 2.0	300M	128	5	436K
MMS (Pratap et al. 2023) **[SOTA]**	wav2vec 2.0	1B	1,362	6	491K
WavLabLM (Chen et al. 2023)	WavLM	300M	136	6	40K

Third limitation: blocks are getting larger and larger.
Particularly challenging for low-resource and online applications.

Multilingual Speech Representation Models:

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

Model	Training Approach	# Parameters	# Languages	# Datasets	# Hours
XLSR-53 (Conneau et al. 2020)	wav2vec 2.0	300M	53	3	56K
XLS-R (Babu et al. 2021)	wav2vec 2.0	300M/2B	128	5	436K
MMS (Pratap et al. 2023) **[SOTA]**	wav2vec 2.0	1B	1,362	6	491K
WavLabLM (Chen et al. 2023)	WavLM	300M	136	6	40K
mHuBERT-147	**HuBERT**	**95M** ⇩	**147**	**17** ⇧	**90K** ⇩

Creating mHuBERT-147

NAVER LABS

mHuBERT-147: A Compact Multilingual HuBERT Model

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

RQ1: Do we need half a million hours of (mostly english) speech to train?

mHuBERT-147: A Compact Multilingual HuBERT Model

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

RQ1: Do we need half a million hours of (mostly english) speech to train?

B. **TRAINING PROCEDURE:** Scaling up HuBERT to multilingual settings

RQ2: How to effectively balance multilingual sources?

RQ3: Can we train a general-purpose multilingual HuBERT?

mHuBERT-147: A Compact Multilingual HuBERT Model

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

  RQ1: Do we need half a million hours of (mostly english) speech to train?

B. **TRAINING PROCEDURE:** Scaling up HuBERT to multilingual settings

  RQ2: How to effectively balance multilingual sources?

  RQ3: Can we train a general-purpose multilingual HuBERT?

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

  RQ4: Do we really need to scale size for multilingual settings?

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

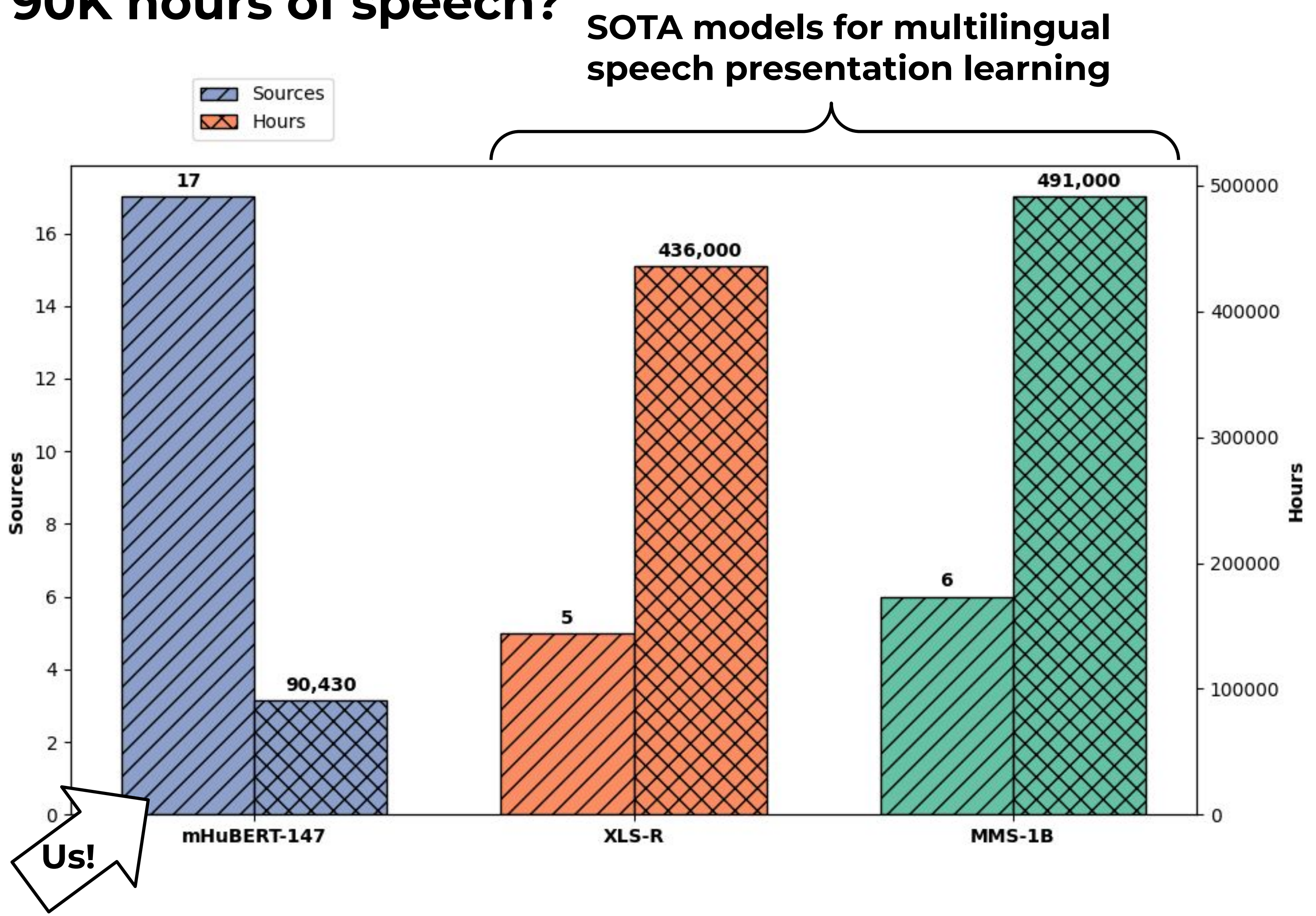
- ➔ We gather **17 datasets** across **147 languages**
- ➔ We filter popular large datasets **removing noise/music**
- ➔ We **downsample high-resource** languages (>2,000 hours per source)

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)

1. **Why HuBERT?**
   Consistent better performance compared to wav2vec 2.0 (Yang et al. 2021)

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

1. **Why HuBERT?**
   Consistent better performance compared to wav2vec 2.0 (Yang et al. 2021)

2. **Why not WavLM?**
   WavLM trains on top of an already existing HuBERT model, so we first need the mHuBERT model!

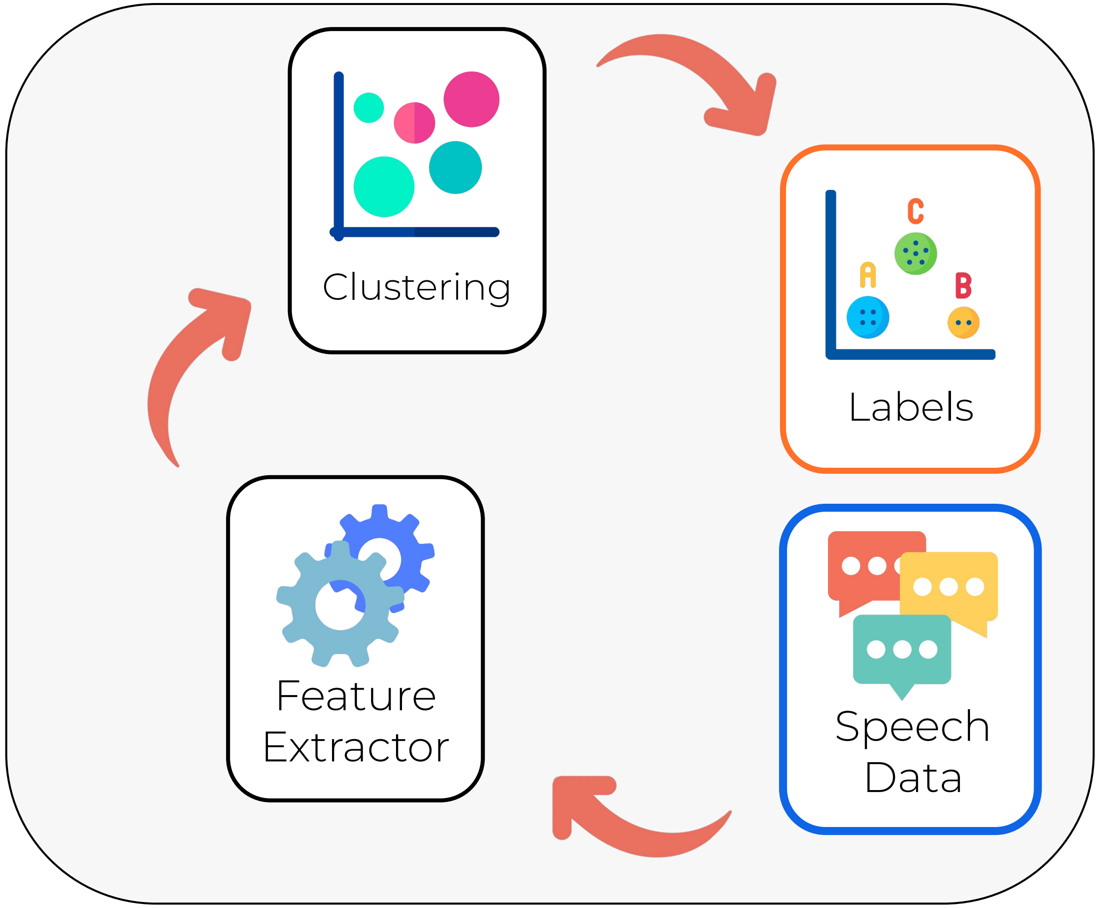
TRAINING APPROACH: Hidden Units BERT (Hsu et al. 2021)

1. **Why HuBERT?**
   Consistent better performance compared to wav2vec 2.0 (Yang et al. 2021)

2. **Why not WavLM?**
   WavLM trains on top of an already existing HuBERT model, so we first need the mHuBERT model!

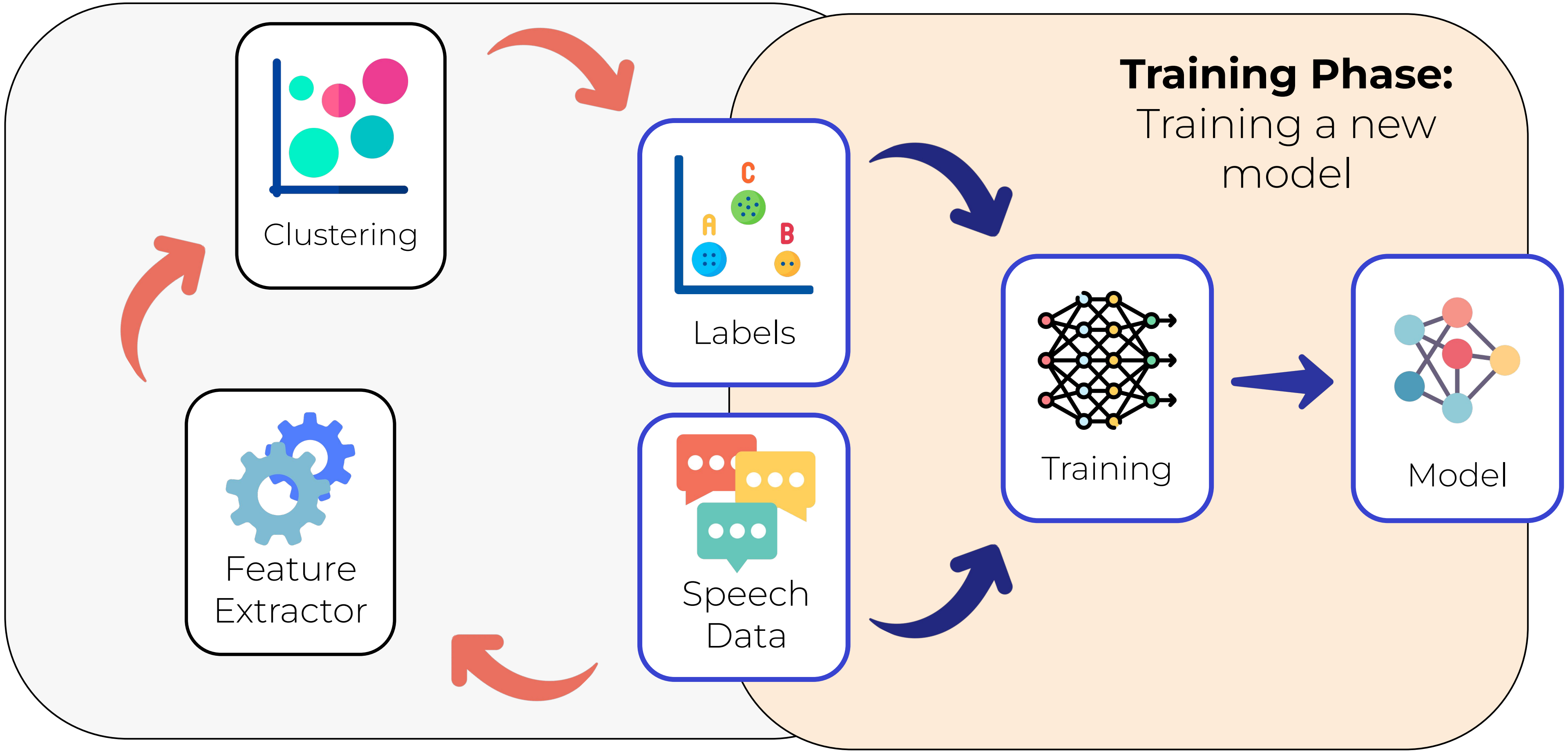
3. **Why from scratch?**
   - WavLabLM is an example of "recycling" monolingual features
   - Our own preliminary experiments using XLS-R models were not encouraging

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

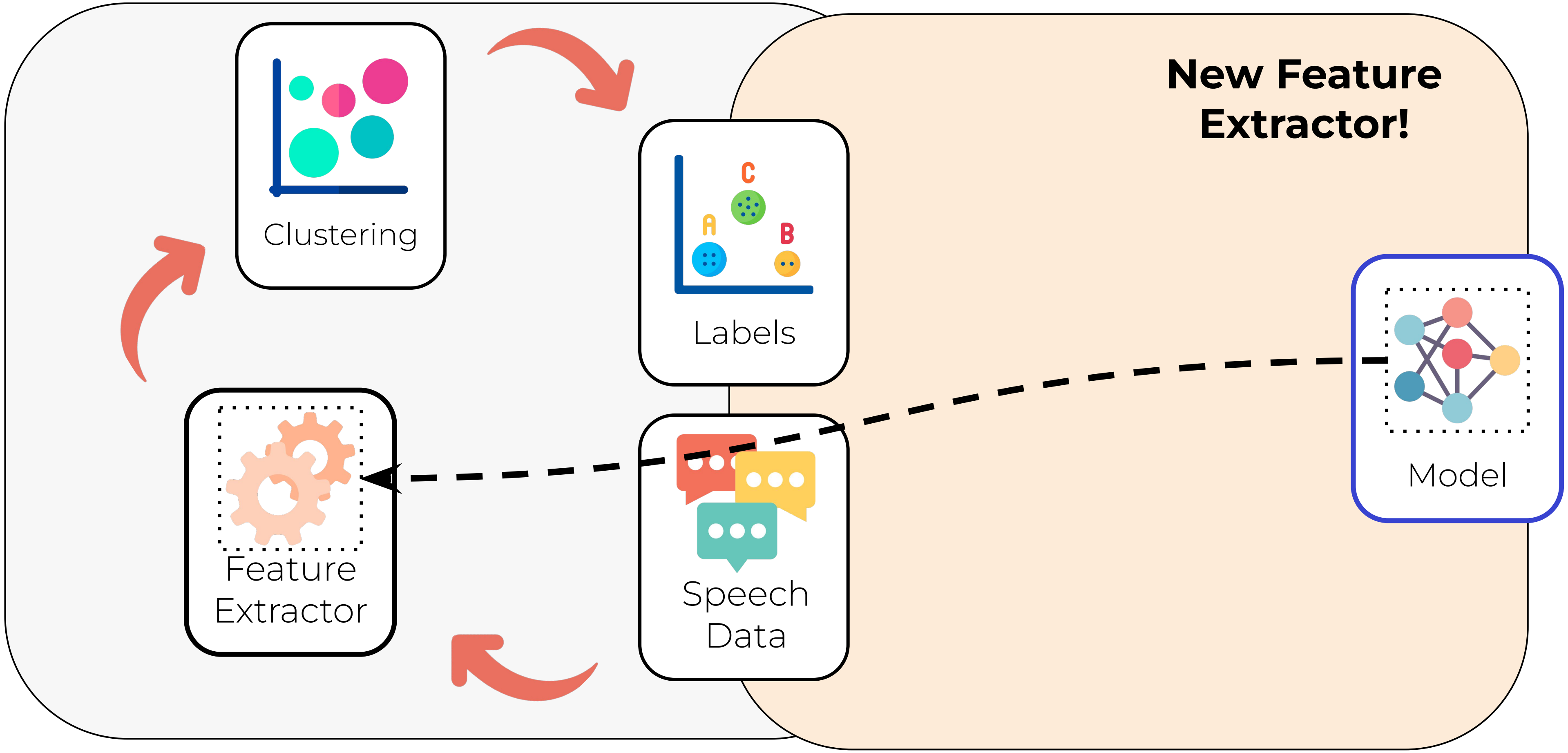


Labeling Phase:
Creating fake discrete labels

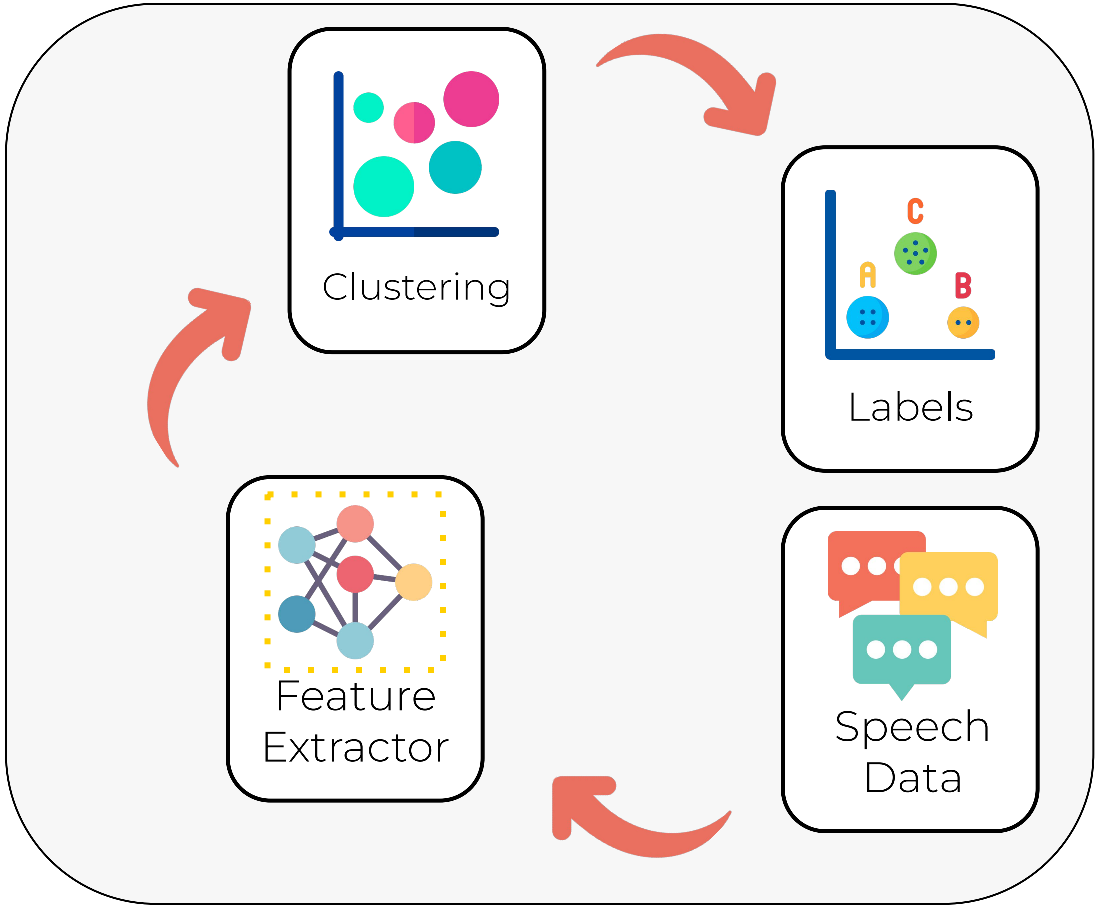
TRAINING APPROACH: Hidden Units BERT (Hsu et al. 2021)



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

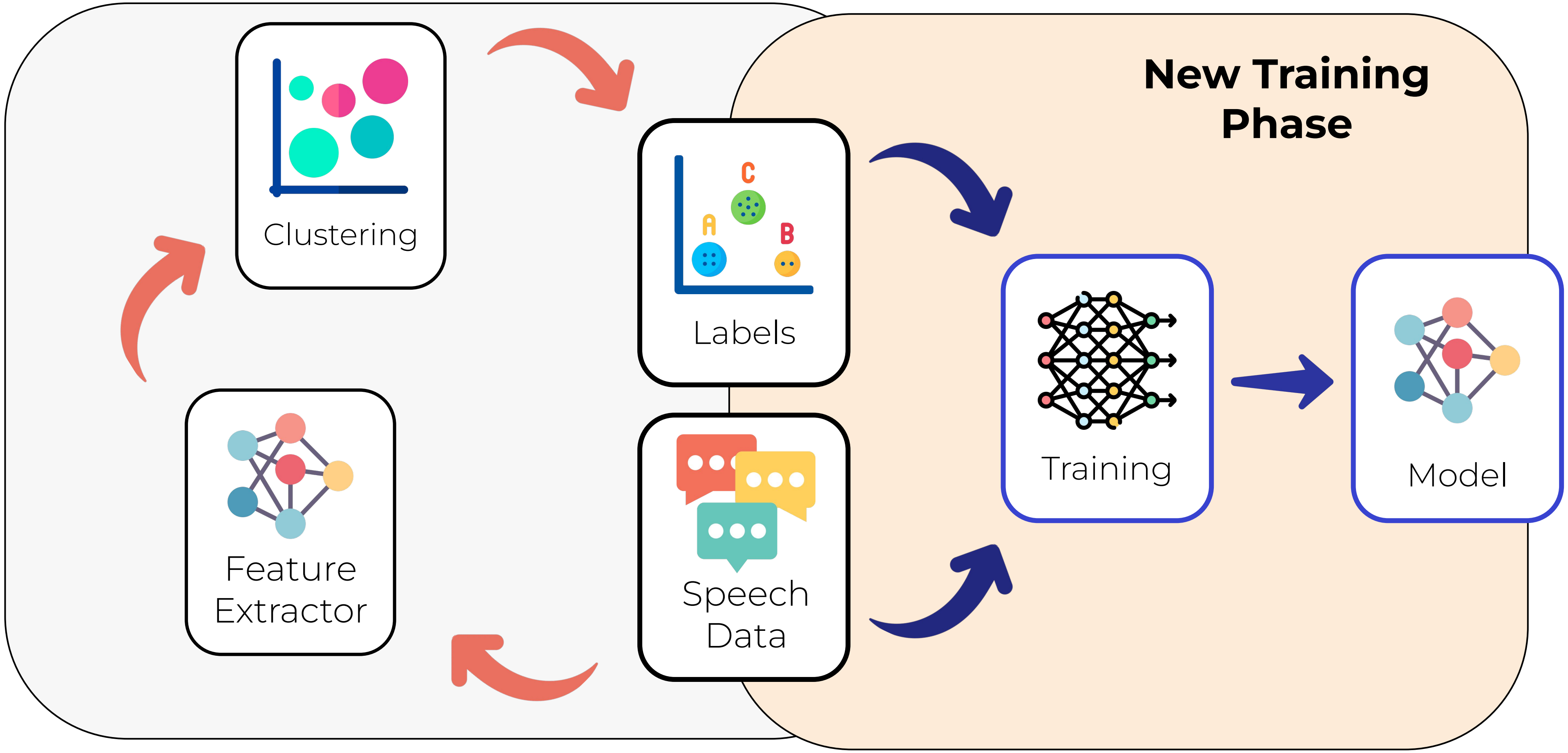


TRAINING APPROACH: Hidden Units BERT (Hsu et al. 2021)



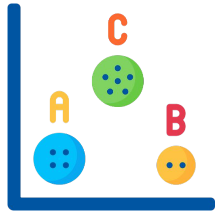
New Labeling Begins!

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



Scaling the Approach for Multilingual Settings

➔ **HuBERT is an expensive SSL approach!**

	Task	RAM/job	CPU/job	GPU/job	Total Disk	Total Processing Time
0	Speech	-	-	-	9.6 TB	-
1	Speech Features iteration 1	-	-	-	4.7 TB	-
1	Speech Features iteration 2-3	50-500 GB	-	1x V100-32 GB	48 TB	3-5 days
2	`faiss` Index Training	2 TB	32x Intel(R) Xeon(R) Platinum 8452Y	-	3.6 GB	2 days
3	`faiss` Index Application	50-500 GB	4-8 cores (any)	-	37 GB	2-3 days
4	Manifest/Labels	-	-	-	65 GB	-
5	Model training (iteration)	2 TB/node	64 x AMD EPYC 7313 16-Core Processor	32x A100-80GB	35 GB	20 days

Table 7: Overview of the hardware requirements necessary per job (RAM; CPU; GPU), total disk required to store the output of the tasks, and total estimated processing time. The numbers in the left illustrate the hierarchy of processes, with larger numbers depending on the completion of the previous processes.

From our extended report at: https://arxiv.org/pdf/2406.06371

Scaling the Approach for Multilingual Settings

➔ **HuBERT is an expensive SSL approach!**

	Task	RAM/job	CPU/job	GPU/job	Total Disk	Total Processing Time
0	Speech	-	-	-	9.6 TB	-
1	Speech Features iteration 1	-	-	-	4.7 TB	-
1	Speech Features iteration 2-3	50-500 GB	-	1x V100-32 GB	48 TB	3-5 days
2	faiss Index Training	2 TB	32x Intel(R) Xeon(R) Platinum 8452Y	-	3.6 GB	2 days
3	faiss Index Application	50-500 GB	4-8 cores (any)	-	37 GB	2-3 days
4	Manifest/Labels	-	-	-	65 GB	-
5	Model training (iteration)	2 TB/node	64 x AMD EPYC 7313 16-Core Processor	32x A100-80GB	35 GB	20 days

Table 7: Overview of the hardware requirements necessary per job (RAM; CPU; GPU), total disk required to store the output of the tasks, and total estimated processing time. The numbers in the left illustrate the hierarchy of processes, with larger numbers depending on the completion of the previous processes.

From our extended report at: https://arxiv.org/pdf/2406.06371

Scaling the Approach for Multilingual Settings

➔ **HuBERT is an expensive SSL approach!**

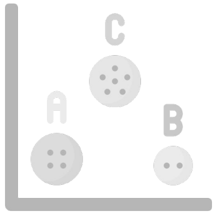
	Task	RAM/job	CPU/job	GPU/job	Total Disk	Total Processing Time
0	Speech	-	-	-	9.6 TB	-
1	Speech Features iteration 1	-	-	-	4.7 TB	-
1	Speech Features iteration 2-3	50-500 GB	-	1x V100-32 GB	48 TB	3-5 days
2	faiss Index Training	2 TB	32x Intel(R) Xeon(R) Platinum 8452Y 4-8 cores (any)	-	3.6 GB	2 days
3	faiss Index Application	50-500 GB		-	37 GB	2-3 days
4	Manifest/Labels	-	-	-	65 GB	-
5	Model training (iteration)	2 TB/node	64 x AMD EPYC 7313 16-Core Processor	32x A100-80GB	35 GB	20 days

Table 7: Overview of the hardware requirements necessary per job (RAM; CPU; GPU), total disk required to store the output of the tasks, and total estimated processing time. The numbers in the left illustrate the hierarchy of processes, with larger numbers depending on the completion of the previous processes.

From our extended report at: https://arxiv.org/pdf/2406.06371

Scaling the Approach for Multilingual Settings

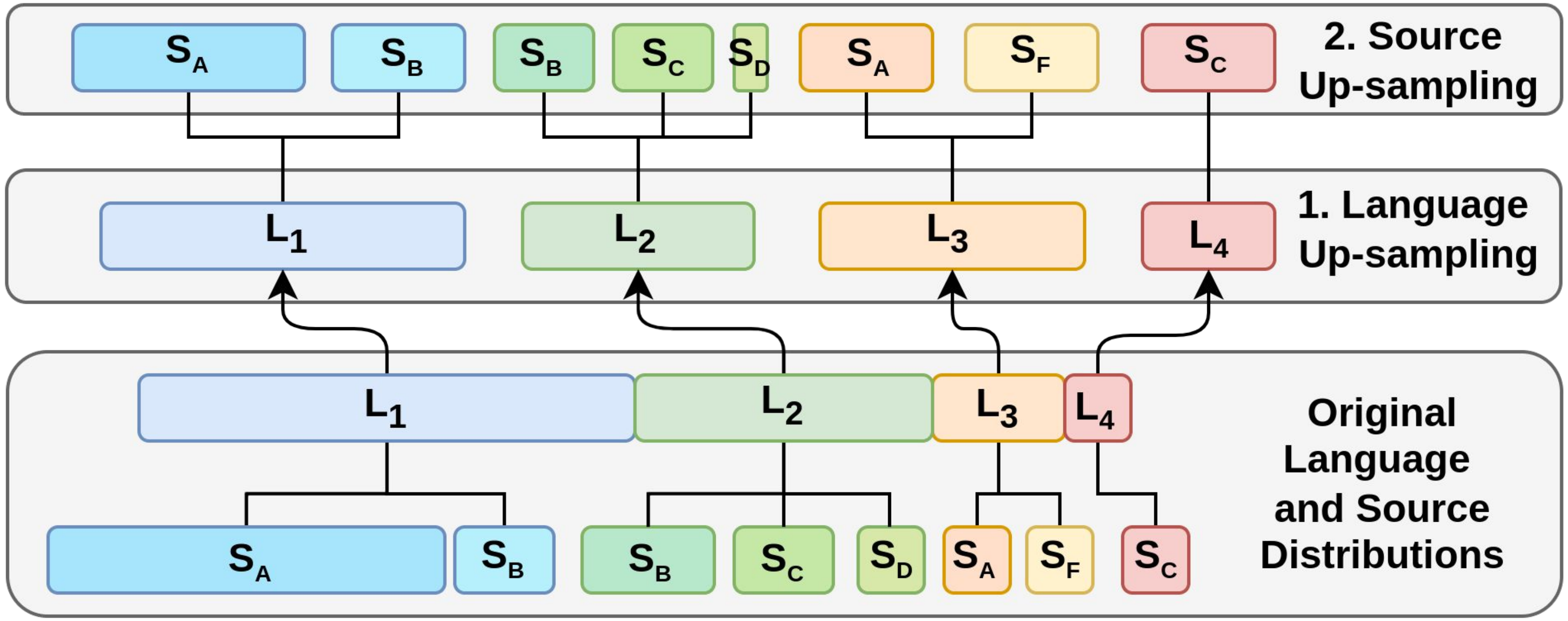
A. **Clustering/Labeling step are extremely expensive and slow**
   - We make training set smaller, but higher-quality
   - We make clustering faster by using faiss + graph search for labeling **(5.2x times faster)**



B. **Multilingual data distribution**
   - We propose a multilingual batching approach: two-step language, data source up-sampling

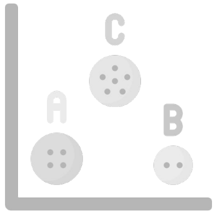
Scaling the Approach for Multilingual Settings

➔ **Two-step language, data source up-sampling**
Two different hyper-parameters control language and data up-sampling

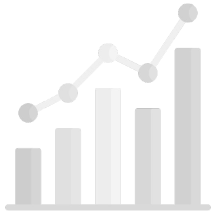


Scaling the Approach for Multilingual Settings

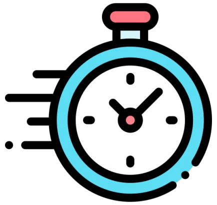
A. **Clustering/Labeling step are extremely expensive and slow**
   - We make training set smaller, but higher-quality
   - We make clustering faster by using faiss + graph search for labeling **(5.2x times faster)**



B. **Multilingual data distribution**
   - We propose a multilingual batching approach: two-step language, data source up-sampling



C. **Training iterations for longer!**
   - We find that existing speech representation models tend to be quite undertrained (accounting for some grokking)



Scaling the Approach for Multilingual Settings

➔ **Why don't we stop at ~400k updates?**

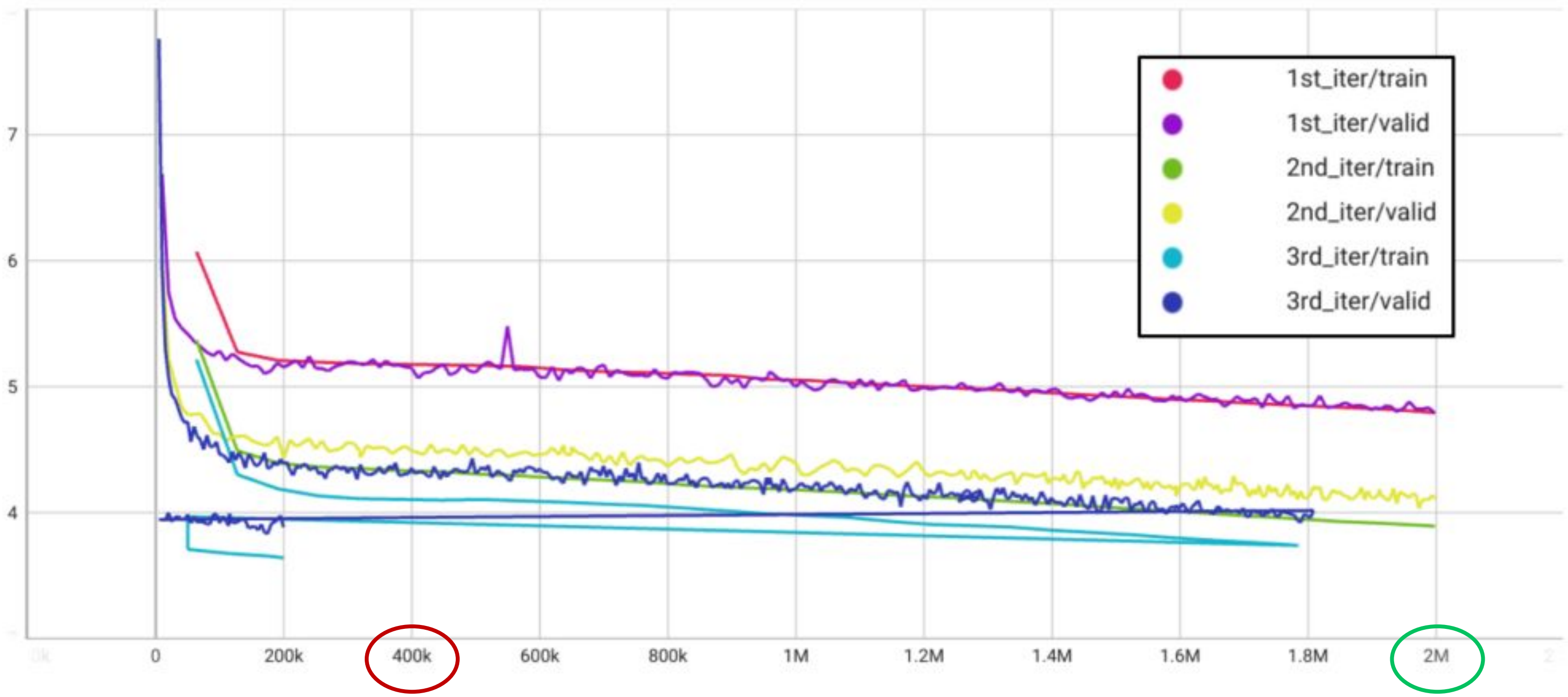


Figure 2: Loss curves for the 3 iterations of mHuBERT-147 training. For the 3rd iteration, the step jump from 1.8M to 0 is due to the optimizer re-initialization: this model crashed in fp16 and had to be reinitialized using fp32 for the last 200 K updates. Best seen in color.

Scaling the Approach for Multilingual Settings

Because our models keep getting better!

It is important to avoid focusing only on the loss!

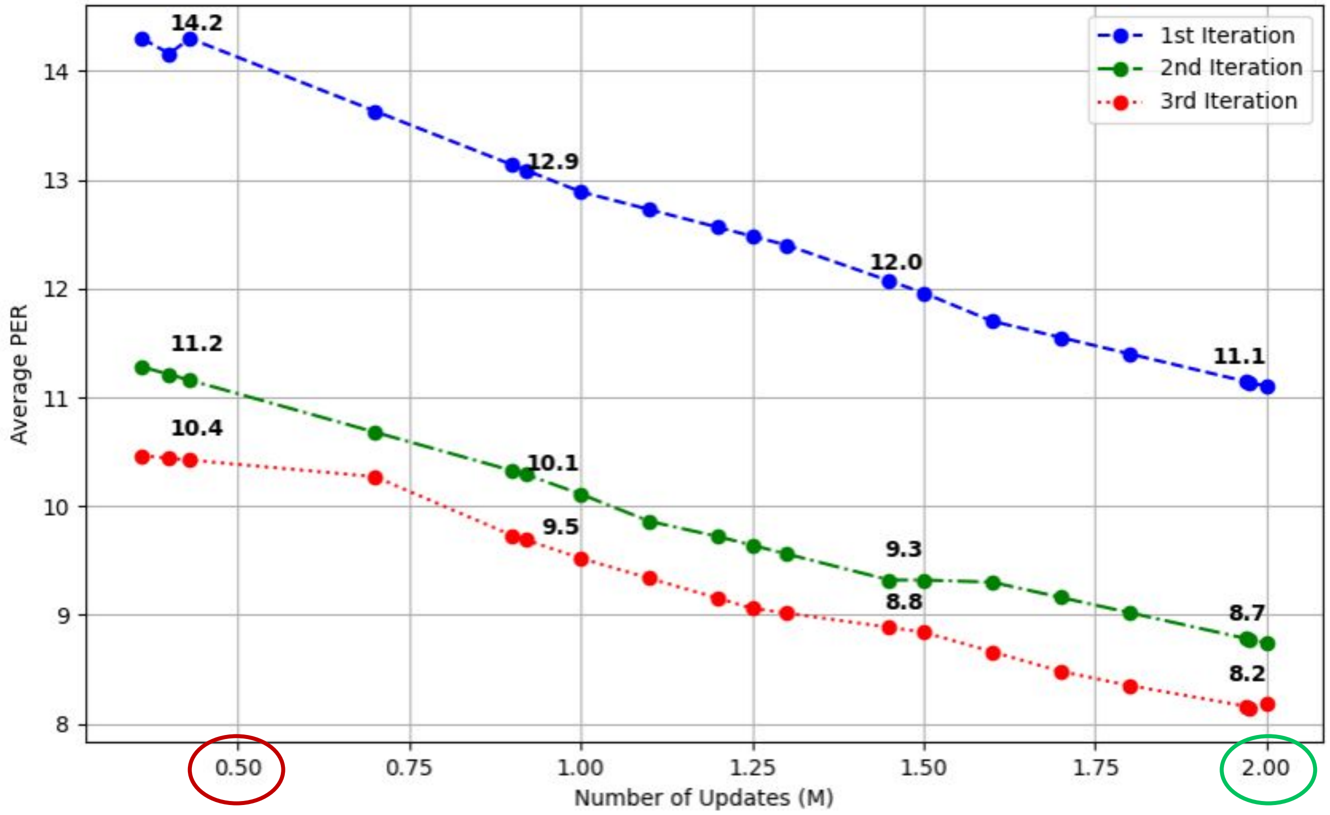


Figure 3: Linear interpolation of average PER performance of mHuBERT-147 across different iterations, and with different number of updates. The bold scores correspond to scores obtained at the following updates, from left to right: 400 K, 1 M, 1.5 M and 2 M.

Proxy task: PER from CV (10 languages); same than XLSR-53

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/

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]

Who's "we"?


Myself! :)


Vivek Iyer


Nikolaos Lagos


Laurent Besacier


Ioan Calapodescu

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

Evaluation

Evaluating mHuBERT-147

1. How good are the **multilingual representations** obtained by this model?

2. How good is this model in **few-shot settings**?

3. How good is this model **compared to the English HuBERT**?

EVALUATION 1: 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

EVALUATION 2: Few-shot ASR with FLEURS-102

A. **Monolingual ASR setup:** Around **12h** per language

B. **Hyperparameter search** on a subset of languages (29) using XLS-R

C. **3-4 runs per language:** 102 languages * ~3 runs * 4 models = **~1300 runs**

Languages grouped by geographic families for easier display.

WE: Western Europe	**EE:** Eastern Europe
CMN: Central-Asia/Middle-East/North-Africa	**SSA:** Sub-Saharan Africa
SA: South-Asia	**SEA:** South-East Asia
CJK: Isolates	

FLEURS-102: Few-shot ASR evaluation

SSL	General Avg (102)	WE (25)	EE (16)	CMN (12)	SSA (20)	SA (14)	SEA (11)	CJK (4)
MMS-1B	22.3	**17.4**	**11.0**	37.8	23.3	27.7	25.9	**17.8**
MMS-300M	24.9	19.5	12.5	39.8	24.8	29.1	29.5	35.8
XLS-R-300M	24.5	18.7	11.8	39.2	24.8	29.7	29.5	33.4
mHuBERT-147	**21.1**	18.7	15.3	**23.4**	**22.7**	**25.5**	**19.8**	31.5

Table 5: FLEURS-102 CER (↓) geographic group averages, with number of languages between parentheses.

WE: Western Europe
EE: Eastern Europe
CMN: Central-Asia/Middle-East/North-Africa
SSA: Sub-Saharan Africa
SA: South-Asia
SEA: South-East Asia
CJK: Isolates

FLEURS-102: A problematic evaluation

SSL	General Avg (102)	WE (25)	EE (16)	CMN (12)	SSA (20)	SA (14)	SEA (11)	CJK (4)
MMS-1B	22.3	**17.4**	**11.0**	37.8	23.3	27.7	25.9	**17.8**
MMS-300M	24.9	19.5	12.5	39.8	24.8	29.1	29.5	35.8
XLS-R-300M	24.5	18.7	11.8	39.2	24.8	29.7	29.5	33.4
mHuBERT-147	**21.1**	18.7	15.3	**23.4**	**22.7**	**25.5**	**19.8**	31.5

Table 5: FLEURS-102 CER (↓) geographic group averages, with number of languages between parentheses.

FLEURS-102: A problematic evaluation

- **Language optimization**

  The few-shot setting makes hyperparameter search very important, but FLEURS-102 monolingual setups are **too expensive to optimize**

FLEURS-102: A problematic evaluation

- **Language optimization**
  The few-shot setting makes hyperparameter search very important, but FLEURS-102 monolingual setups are **too expensive to optimize**

- **Model optimization**
  It is **not fair** to adopt the same hyper-parameters across models (learning rate; warm-up scale; dropout; etc)

FLEURS-102: A problematic evaluation

- **Language optimization**
  The few-shot setting makes hyperparameter search very important, but FLEURS-102 monolingual setups are **too expensive to optimize**

- **Model optimization**
  It is **not fair** to adopt the same hyper-parameters across models (learning rate; warm-up scale; dropout; etc)

MMS-1B is always ~2 CER points below mHuBERT-147 on this setting… **when it converges**

In summary: results are highly variable, highly dependent to the hyperparameters

EVALUATION 3: Monolingual Comparison

A. **Datasets:** all open datasets from ESB Benchmark **downsampled to 50h max using the method from** *Lagos and Calapodescu, 2024*

B. **Task:** English ASR

C. **Models:** English HuBERT vs mHuBERT-147

ESB Benchmark: ASR results (training = 50h)

	HuBERT-base	mHuBERT-147
AMI	**33.1**	33.6 (+0.5)
CV	37.2	**35.4** (-1.8)
Earnings-22	35.3	**34.7** (-0.6)
GigaSpeech	**25.9**	26.3 (+0.4)
LibriSpeech-clean	**7.7**	9.7 (+2.0)
LibriSpeech-other	**13.6**	17.3 (+3.7)
TED-LIUM	**11.7**	13.1 (+1.4)
VP	**18.7**	19.0 (+0.3)
Average WER (↓)	**23.1**	23.6 (+0.5)

Table 15: WER (↓) scores for English ASR systems trained on the different datasets from the ESB Benchmark. Following Lagos and Calapodescu [27], only 50 h of training data are used. The score difference between mHuBERT-147 and HuBERT-base scores is presented between parentheses.

ESB Benchmark: ASR results (training = 50h)

	HuBERT-base	mHuBERT-147
AMI	**33.1**	33.6 (+0.5)
CV	37.2	**35.4** (-1.8)
Earnings-22	35.3	**34.7** (-0.6)
GigaSpeech	**25.9**	26.3 (+0.4)
LibriSpeech-clean	**7.7**	9.7 (+2.0)
LibriSpeech-other	**13.6**	17.3 (+3.7)
TED-LIUM	**11.7**	13.1 (+1.4)
VP	**18.7**	19.0 (+0.3)
Average WER (↓)	**23.1**	23.6 (+0.5)

Largest gap observed on Librispeech

Table 15: WER (↓) scores for English ASR systems trained on the different datasets from the ESB Benchmark. Following Lagos and Calapodescu [27], only 50 h of training data are used. The score difference between mHuBERT-147 and HuBERT-base scores is presented between parentheses.

ESB Benchmark: ASR results (training = 50h)

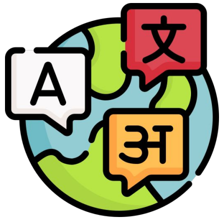
mHuBERT-147 is very competitive to the English model, while covering additional 146 languages

	HuBERT-base	mHuBERT-147
AMI	**33.1**	33.6 (+0.5)
CV	37.2	**35.4** (-1.8)
Earnings-22	35.3	**34.7** (-0.6)
GigaSpeech	**25.9**	26.3 (+0.4)
LibriSpeech-clean	**7.7**	9.7 (+2.0)
LibriSpeech-other	**13.6**	17.3 (+3.7)
TED-LIUM	**11.7**	13.1 (+1.4)
VP	**18.7**	19.0 (+0.3)
Average WER (↓)	**23.1**	23.6 (+0.5)

Table 15: WER (↓) scores for English ASR systems trained on the different datasets from the ESB Benchmark. Following Lagos and Calapodescu [27], only 50 h of training data are used. The score difference between mHuBERT-147 and HuBERT-base scores is presented between parentheses.

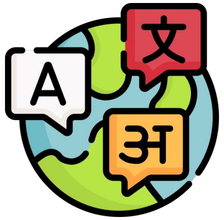
Summarizing:

★ **First multilingual HuBERT model**, covering 147 languages



Summarizing:

★ **First multilingual HuBERT model**, covering 147 languages

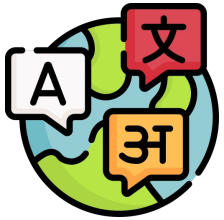


★ **Trained differently:**

- Curated data collection
- Two-level language-source up-sampling for training

Summarizing:

★ **First multilingual HuBERT model**, covering 147 languages



★ **Trained differently:**
- Curated data collection
- Two-level language-source up-sampling for training

★ We produced a **compact** yet **powerful** foundation model
- **a.** SOTA multilingual speech representation
- **b.** Good few-shot ASR model for budgeted settings
- **c.** Competitive to its monolingual equivalent

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

Data Processing and Clustering

License: Apache

Code for Training

License: MIT

Trained Models + Training Data

License: CC-BY-NC-SA-4.0

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

UTTER

Check out our French SLU demo and tutorial!

DEMO: French Spoken Language Understanding with the new speech resources from NAVER LABS Europe

Community Article Published August 28, 2024

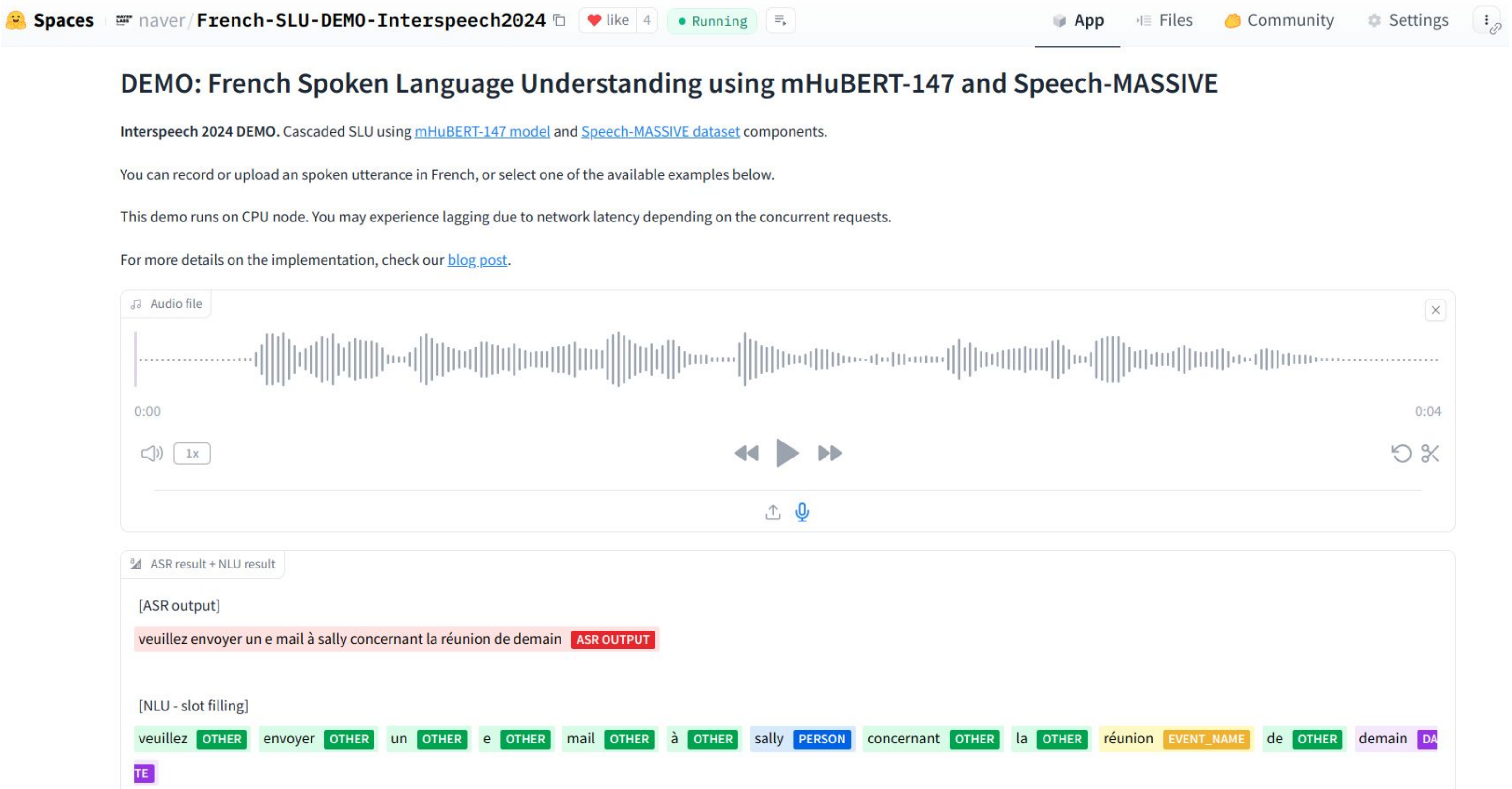
Edit article

mzboito
Marcely Zanon Boito

In this blog post we showcase the recent speech resources released by NAVER LABS Europe that will be presented at Interspeech 2024. The **Speech-MASSIVE** dataset is a multilingual spoken language understanding (SLU) dataset with rich metadata information, and the

Upvoted 8

Demo available here
Tutorial available here



mHuBERT-147: A Compact and Powerful Multilingual Speech Foundation Model

Marcely Zanon Boito

10/2024

Contact: marcely.zanon-boito@naverlabs.com

NAVER LABS

INTERSPEECH 2024

UTTER

UTTER: Unified Transcription and Translation for Extended Reality

Multimodal technology with two use-cases in mind:

- Customer service
- **Meeting assistant**

Unbabel

NAVER LABS Europe