FONDAZIONE
BRUNO KESSLER

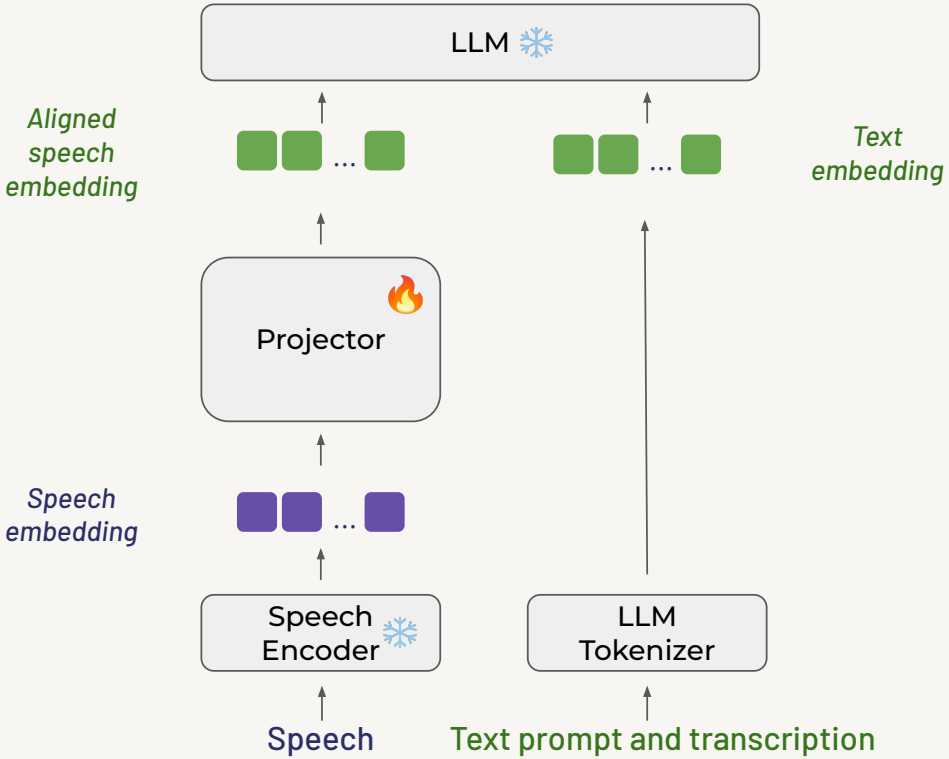
SPEECHTEK
SPEECH TECHNOLOGY LAB

Scaling Speech LLMs Responsibly: Languages, Privacy and Decentralization

What changes when the Speech-LLM is scaled across languages, tasks, and data owners?

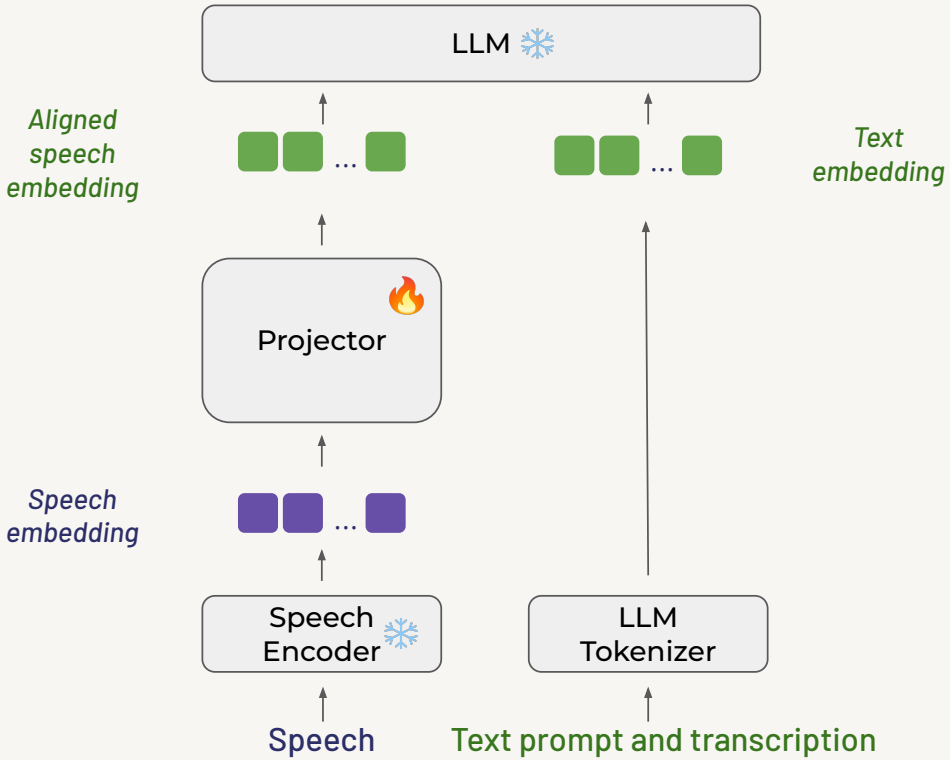
Alessio Brutti

SpeechLLM: an established architecture



Z. Ma, et al., “An Embarrassingly Simple Approach for LLM with Strong ASR Capacity,” 2024

SpeechLLM: an established architecture



Inclusivity

Privacy & Data Ownership

Z. Ma, et al., “An Embarrassingly Simple Approach for LLM with Strong ASR Capacity,” 2024



- Building and growing multilingual & multitask projectors
- Measuring and removing what they leak
- Training without centralizing speech










SPEECHTEK
SPEECH TECHNOLOGY LAB

Meetween

LLMs4EU

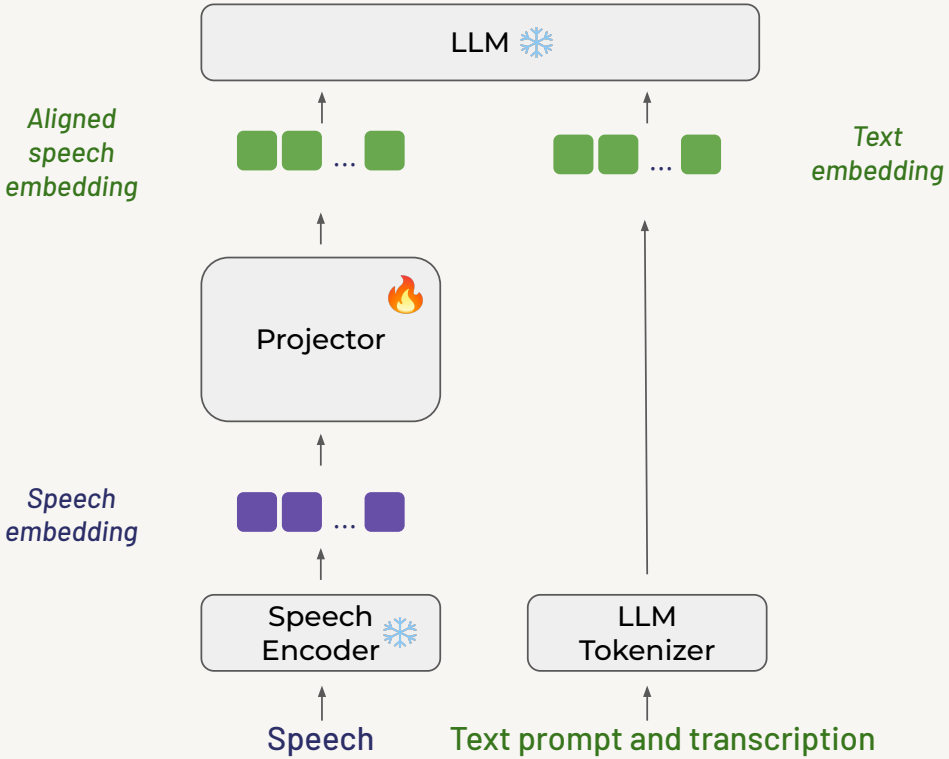
eloquence

Funded by
the European Union

alt-edic
Savez za jezične tehnologije | Alleanza per le tecnologie del linguaggio | Valodu tehnoloģiju alianse

8ra CLOUD-EDGE CONTINUUM

SpeechLLM: an established architecture

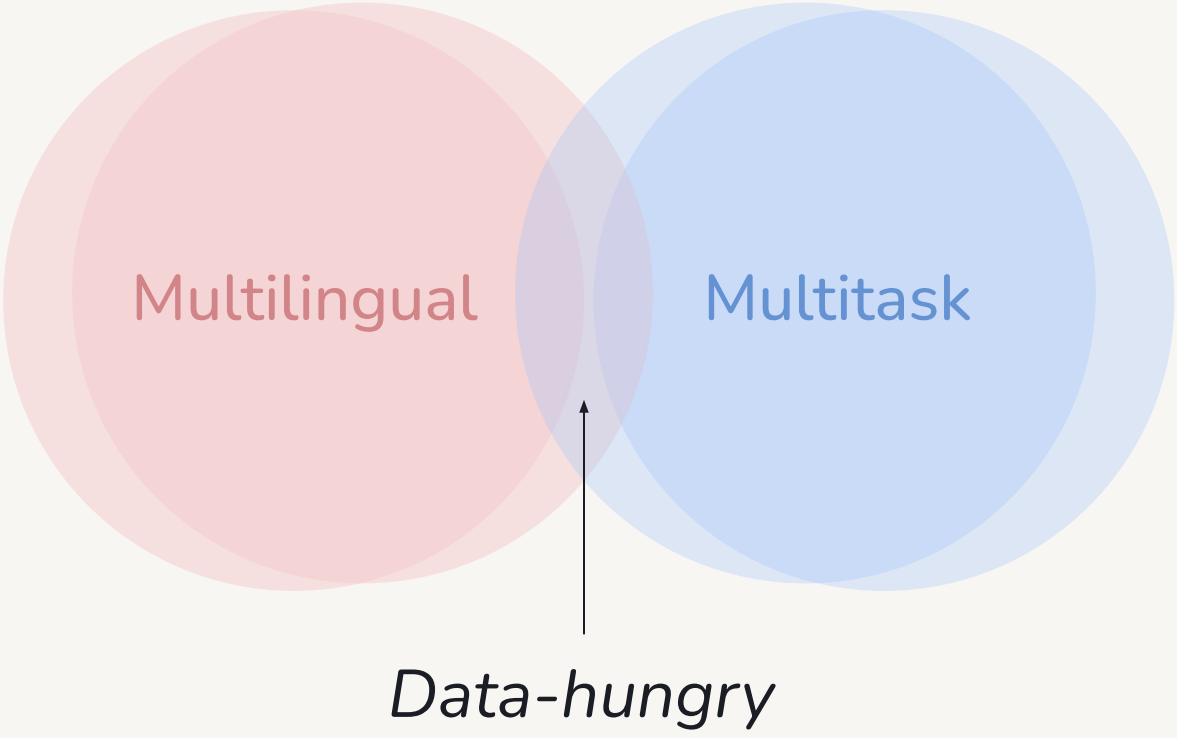


Inclusivity

Languages

Z. Ma, et al., “An Embarrassingly Simple Approach for LLM with Strong ASR Capacity,” 2024

SpeechLLMs are becoming..



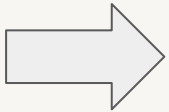
SpeechLLM: data hungry - what languages?

- High-resource ASR depends on massive datasets
- **~100/7,000** languages are supported by SOTA speech systems
- **Performance degrades for many speakers**: dialectal, accented, atypical, impaired speech

SpeechLLM: data hungry - what languages?

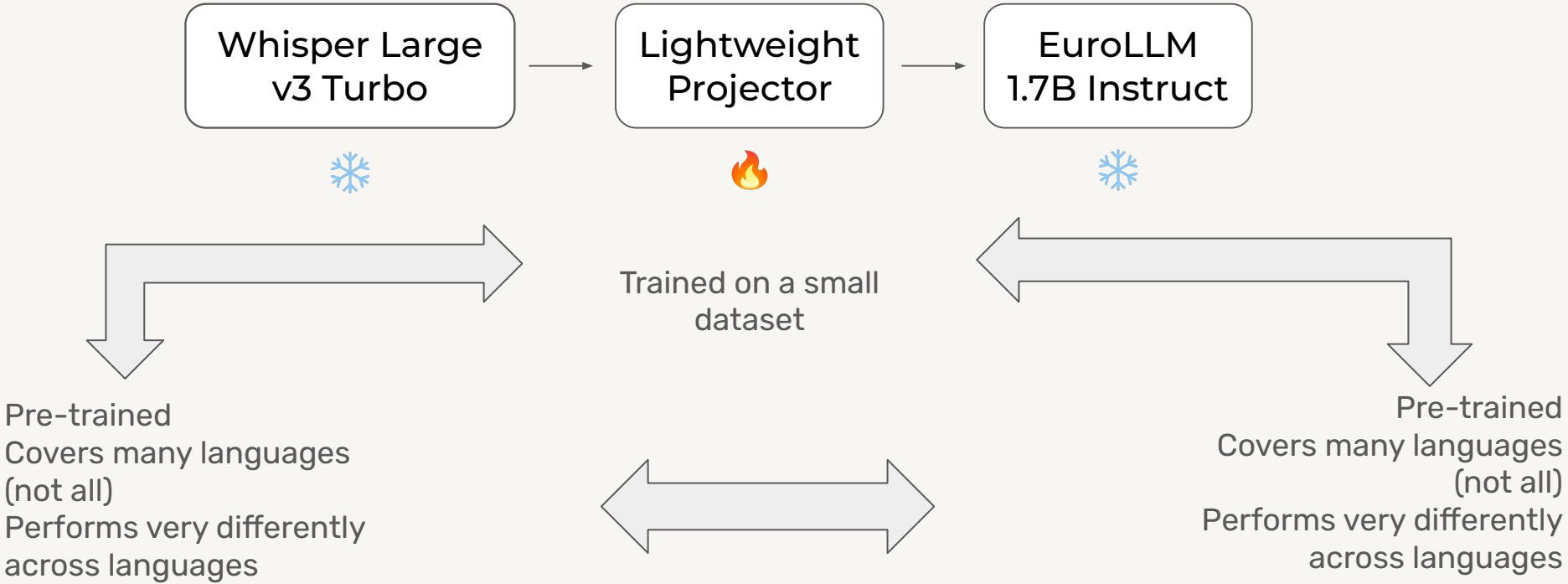
- High-resource ASR depends on massive datasets
- **~100/7,000** languages are supported by SOTA speech systems
- **Performance degrades for many speakers**: dialectal, accented, atypical, impaired speech

Towards inclusive, scalable, low-resource-friendly speech LLM



We still miss:

- Analysing behavior under data scarcity
- Understanding cross-lingual transfer dynamics
- Evaluations focused on low-resource settings (languages and domain)



All entangled

SpeechLLM in low resource scenarios

Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

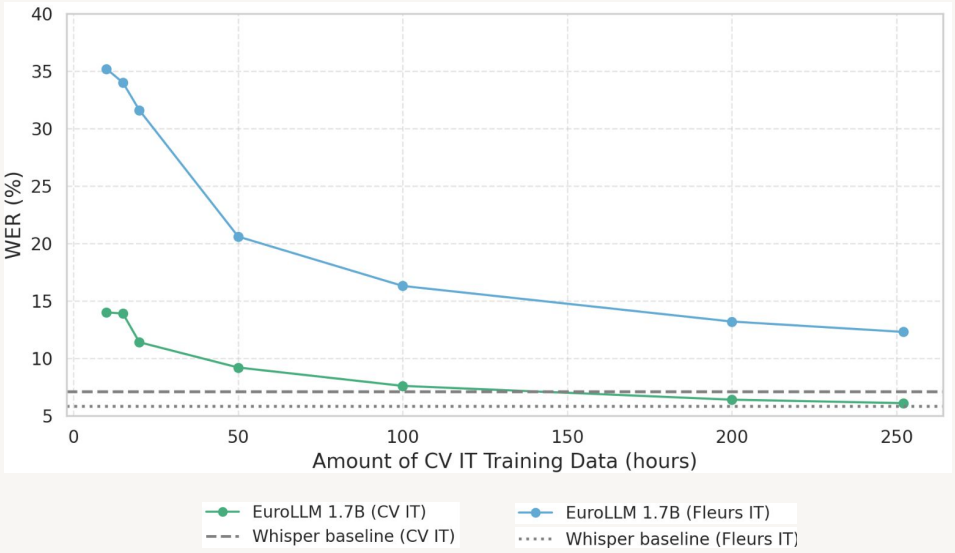
Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages

Seraphina Fong[1,2], *Marco Matassoni*[2], *Alessio Brutti*[2]

[1]Department of Information Engineering and Computer Science, University of Trento, Italy
[2]Center for Augmented Intelligence, Fondazione Bruno Kessler, Italy
meiyueseraphina.fong@unitn.it, matasso@fbk.eu, brutti@fbk.eu

ASR as a study case

- Encoder and LLM are pre-trained → can we use smaller data volume for the projector?



- **150-200 h to reach the baseline**
- Very *poor generalization*
- Prior knowledge in encoder and LLM does not really help….

SpeechLLM in low resource scenarios

Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

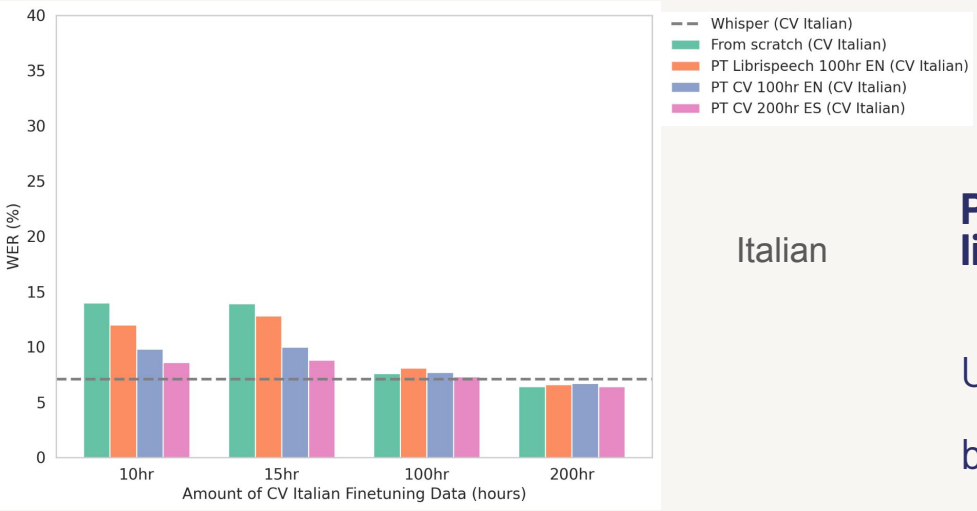
Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages

Seraphina Fong[1,2], *Marco Matassoni*[2], *Alessio Brutti*[2]

[1]Department of Information Engineering and Computer Science, University of Trento, Italy
[2]Center for Augmented Intelligence, Fondazione Bruno Kessler, Italy
meiyueseraphina.fong@unitn.it, matasso@fbk.eu, brutti@fbk.eu

ASR as a study case

- ~~Encoder and LLM are pre-trained → can we use smaller data volume for the projector?~~
- Can we leverage transfer learning for low-resource languages?



Italian

Pretrained (multilingual) projectors helps in limited data scenarios also on "unseen" data

Using *proper languages* for pre-training helps

but.... better whatever than nothing

SpeechLLM in low resource scenarios

Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

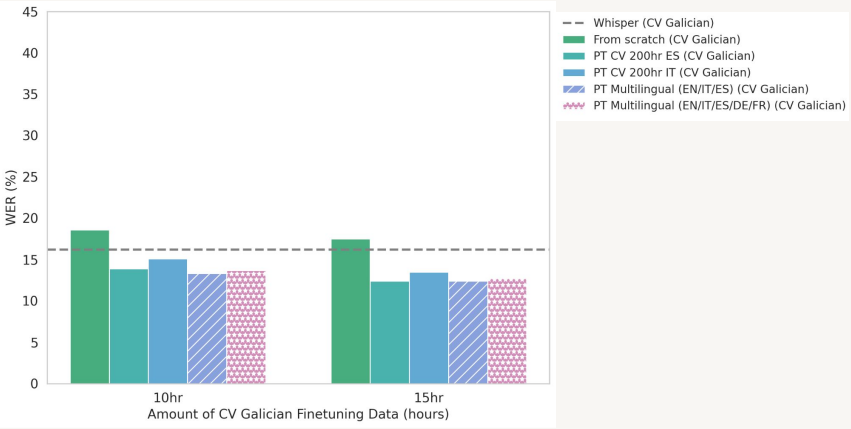
Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages

Seraphina Fong[1,2], *Marco Matassoni*[2], *Alessio Brutti*[2]

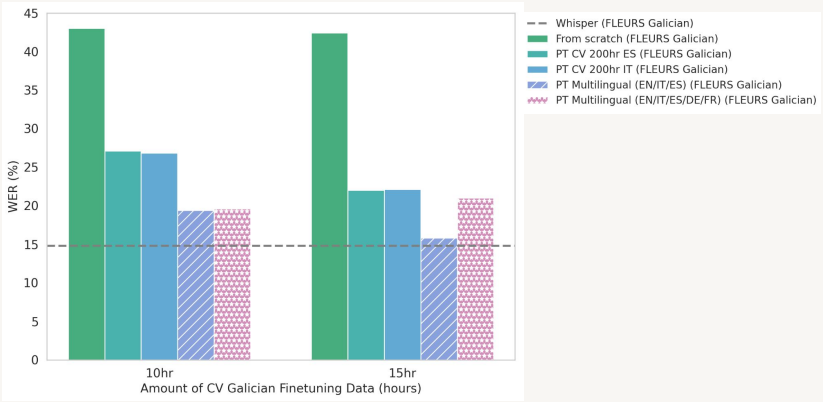
[1]Department of Information Engineering and Computer Science, University of Trento, Italy
[2]Center for Augmented Intelligence, Fondazione Bruno Kessler, Italy
meiyueseraphina.fong@unitn.it, matasso@fbk.eu, brutti@fbk.eu

Galician Common Voice

10-15 hours of training data



CV test



Fleurs test (cross-domain)

MEUSLI 🥣: Multilingual EU Speech LInear projector

MEUSLI: A MULTILINGUAL PROJECTOR FOR LLM-BASED ASR IN 28 EUROPEAN LANGUAGES (AND BEYOND)

Seraphina Fong, Lorenzo Concina, Marco Matassoni, Alessio Brutti

Center for Augmented Intelligence, Fondazione Bruno Kessler, Trento, Italy

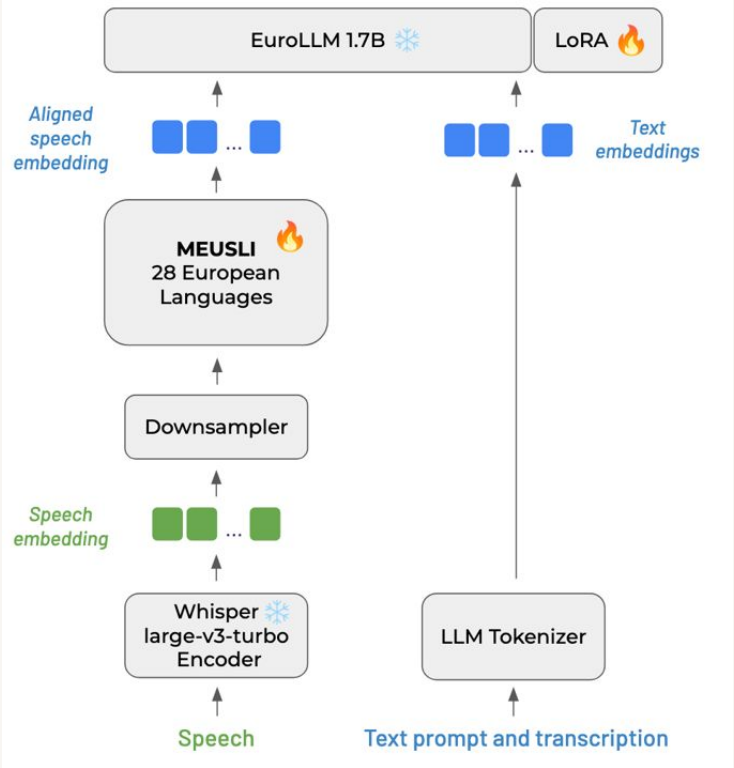
Multilingual projector trained on **28 languages**:
CV, Fleurs, VoxPopuli → 7622 hours

- ➢ Automatic Speech recognition
- ➢ Extension to other languages or further fine tuning
- ➢ Extension to other tasks

LLMs covered

- EuroLLM 1.7B & 9B
- Salamandra 7B
- Apertus 8B
- Villanova 2B

https://huggingface.co/SpeechTek/MEUSLI_projector_v2



eloquence
Funded by the European Union

MEUSLI 🥣: Multilingual EU Speech LInear projector

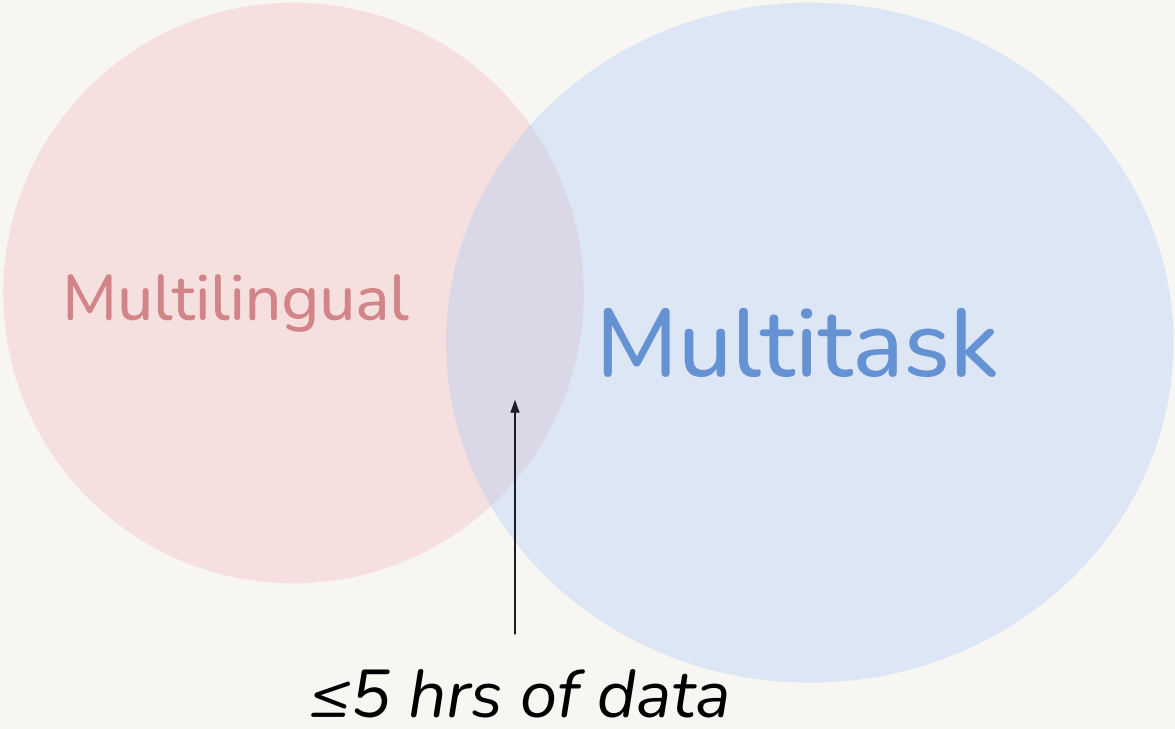
Solid starting point for low-resource languages

Strong adaptation to *unseen languages* (e.g., Albanian)

Lang.	CV data	Monolingual projector WER ↓	MEUSLI WER ↓	Finetuned MEUSLI WER ↓
Breton*	2.5 hrs	84.3%	95.7%	**79.7%**
Maltese	18 hrs	45.8%	58.8%	**39.4%**
Ukrainian†	30 hrs	20.4%	107%	**16.3%**
Albanian*†	46 mins	389%	166%	**75.6%**

**unsupported by EuroLLM* *†unseen during training*

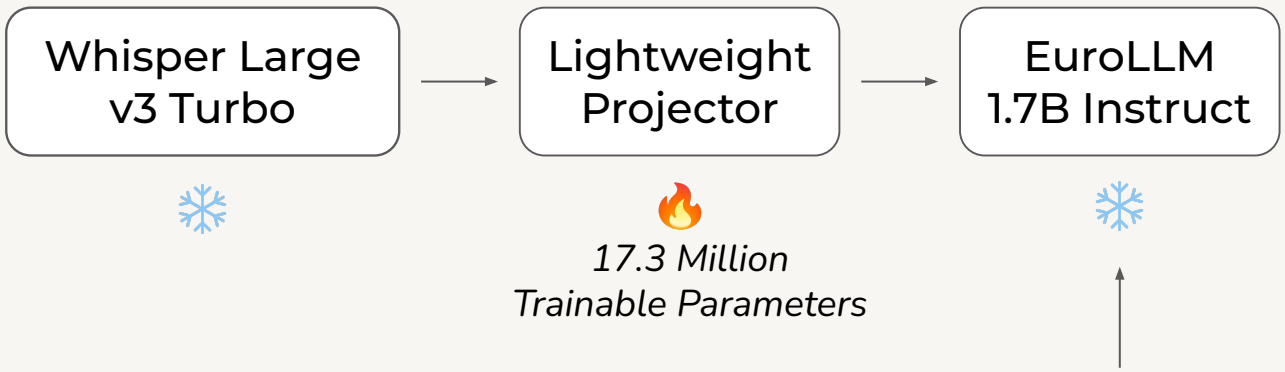
What happens when we need both, but labeled data are scarce?
Can prior knowledge & transfer help?



Multilingual Multitask SpeechLLM

Italian, Spanish,
Galician, Czech, Finnish

SIB-Fleurs speech (3-5 hr per task language)



What can we learn with only ≤5 hours of data per task/language?

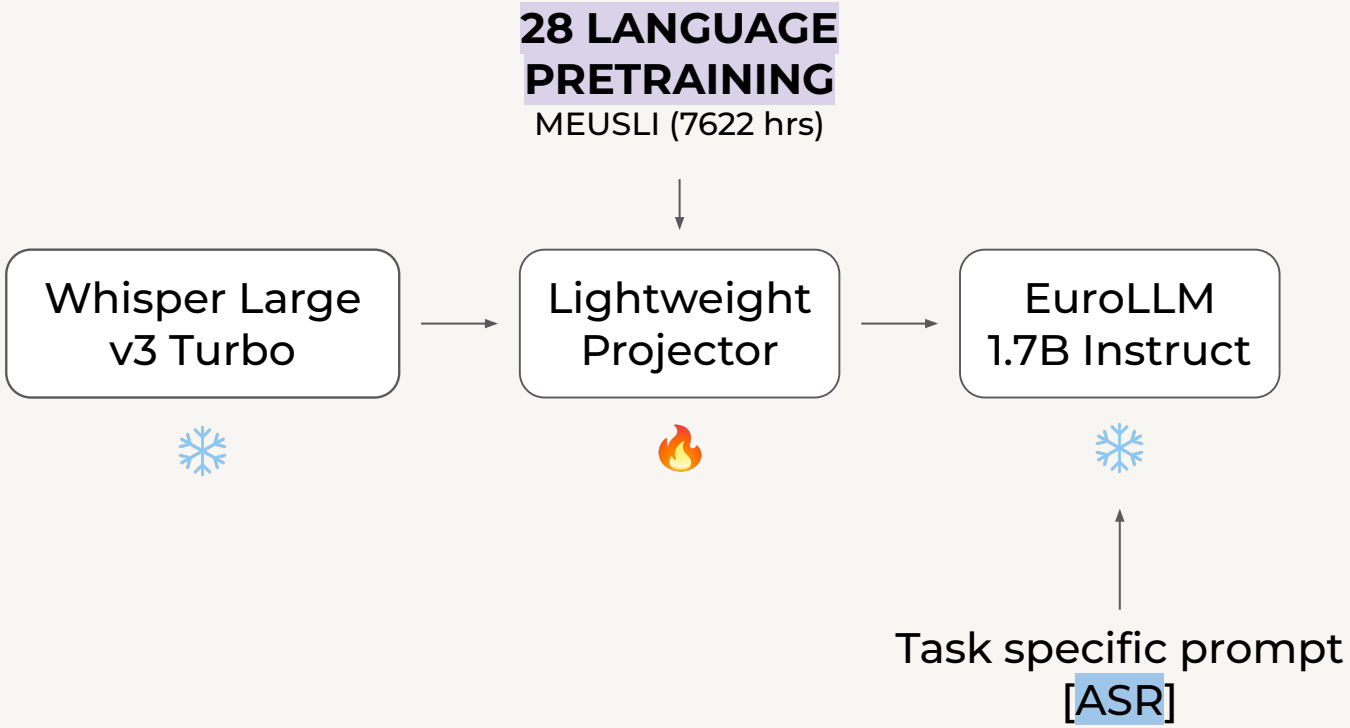
Not much.

Monolingual

	IT	ES	GL	CS	FIN
ASR (WER % ↓)	129.0	138.3	161.9	158.4	155.7
ST (BLEU % ↑)	0.8	0.4	0.2	0.1	0.1
TID (ACC % ↑)	71.8	70.1	67.5	63.9	57.8

Can ASR Provide a Better Starting Point?

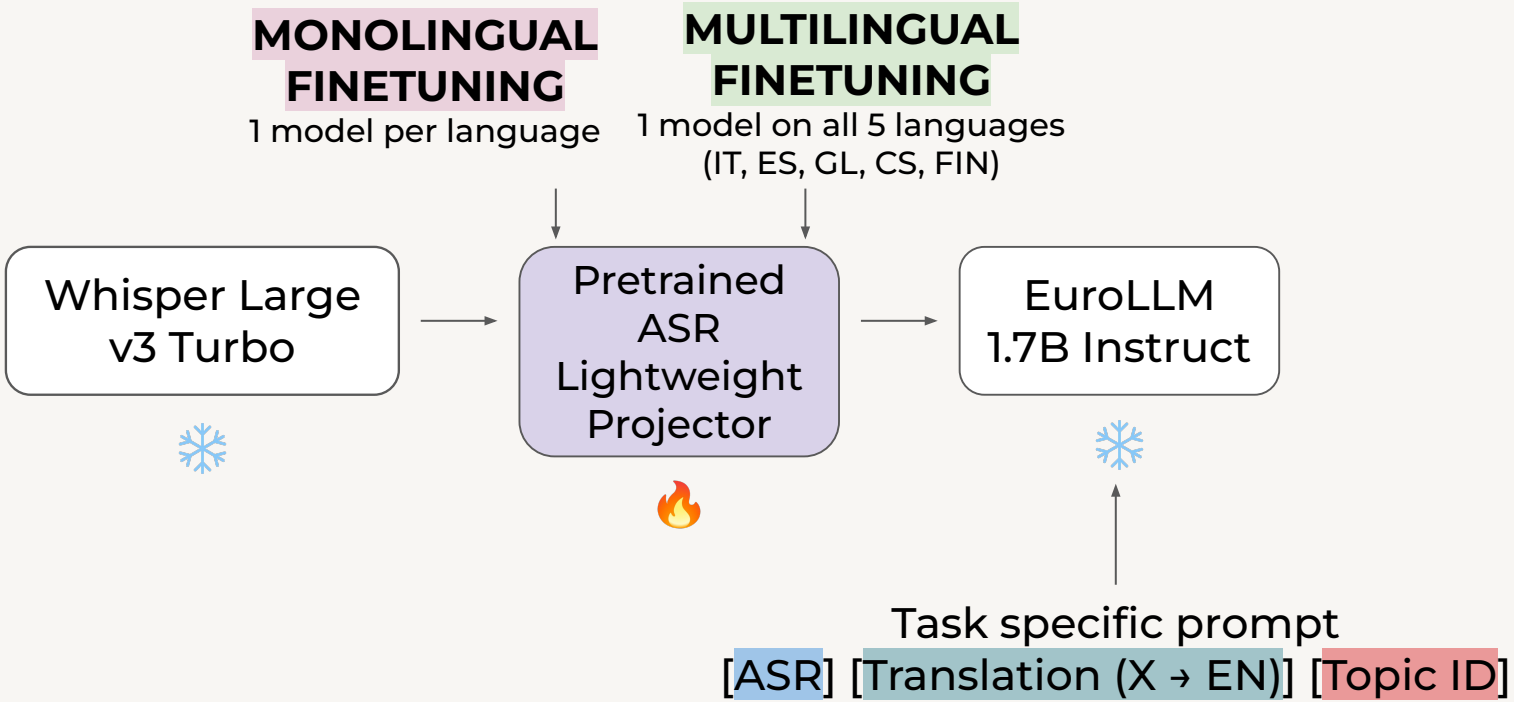
1. *Learn alignment with ASR (large ASR corpus → ASR-pretrained projector)*



Can ASR Provide a Better Starting Point?

Yes!

2. *Finetune on tasks with SIB-Fleurs data (3-5 hr per task language)*
   Pretrained ASR projector → ASR + ST + TID



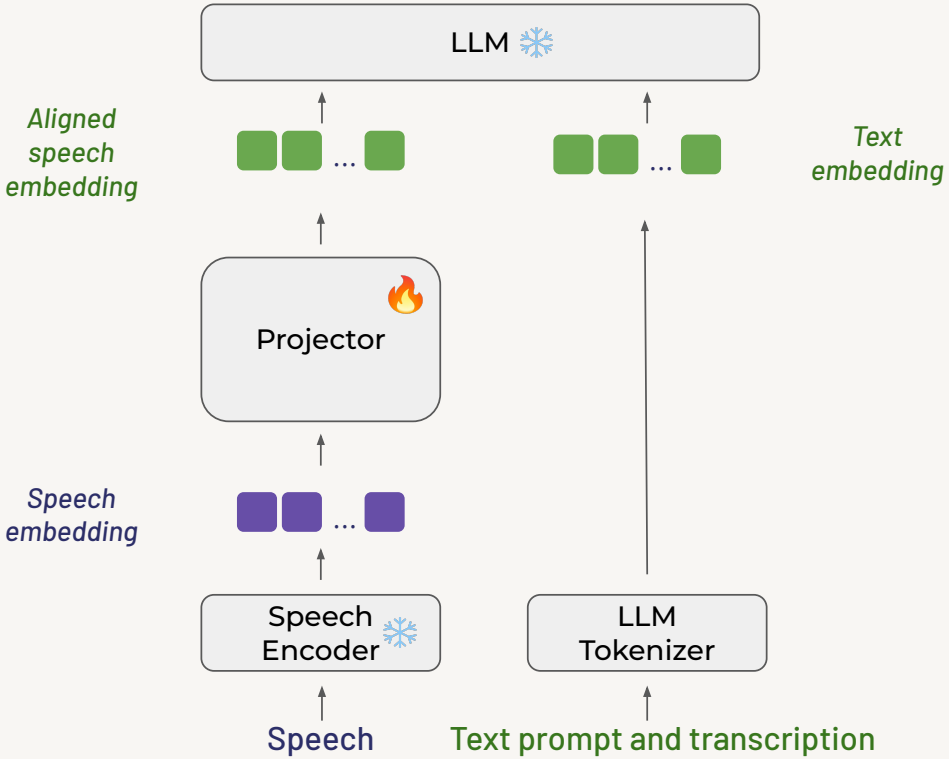
From MEUSLI ASR to Multilingual Multitask

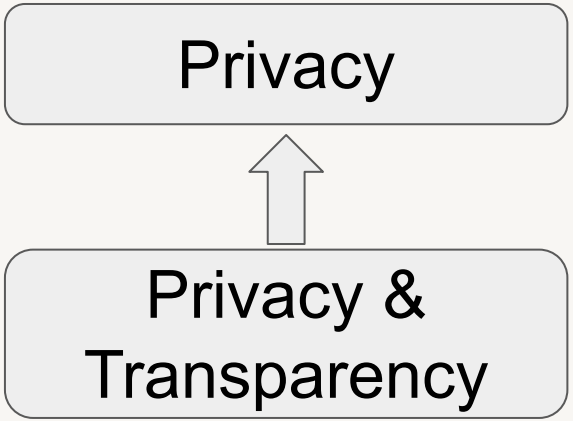
Language	ASR ↓ WER (%)		ST ↑ BLEU (%)		Topic ID ↑ Acc (%)		Topic ID ↑ $F1_{macro}$ (%)	
	Multi	**Mono**	**Multi**	**Mono**	**Multi**	**Mono**	**Multi**	**Mono**
Italian	3.7	3.4	54.5	54.6	86.8	84.3	86.4	82.3
Spanish	4.7	4.2	55.5	56.2	86.4	84.7	85.1	83.3
Galician	11.1	10.2	47.5	46.8	86.1	83.6	85.0	81.6
Czech	12.9	11.0	39.3	40.7	86.7	77.4	86.1	76.9
Finnish	18.6	17.3	26.6	27	81.7	78.0	79.2	74.5

Take-home messages

- Data scarcity does not have to mean capability scarcity
- ASR-pretrained multilingual projectors support multilingual, multitask SpeechLLM with low data volume
- No transfer learning across tasks (!)

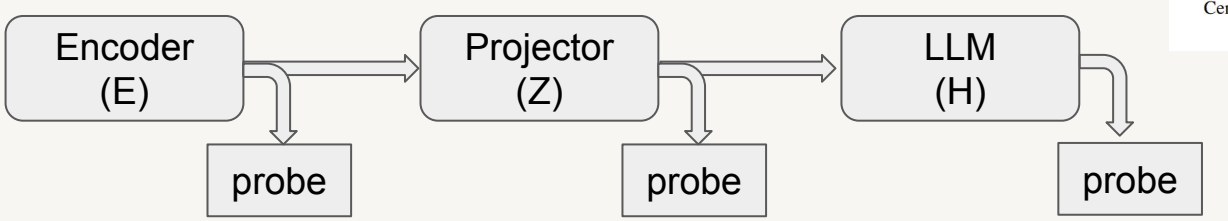
SpeechLLM: an established architecture





Z. Ma, et al., "An Embarrassingly Simple Approach for LLM with Strong ASR Capacity," 2024

Speaker attributes stay recoverable at every stage



Privacy Leakage in Speech-LLMs: Detection and Mitigation

Siavosh Sepanta, Marco Matassoni, Alessio Brutti
Center for Augmented Intelligence, Fondazione Bruno Kessler, Trento, Italy
{ssepanta, matasso, brutti}@fbk.eu

Sensitive Speaker Attribute Leakage in Speech–LLM Pipelines

Siavosh Sepanta, Alessio Brutti

Center for Augmented Intelligence, Fondazione Bruno Kessler, Trento, Italy
ssepanta@fbk.eu, brutti@fbk.eu

Probe accuracy	Speaker ID	Gender	Age	Accent	Emotion
Encoder (E)	51%	96%	86%	92%	91%
Projector (Z)	23%	81%	72%	78%	89%
LLM (H)	31%	72%	78%	71%	87%

Classification accuracy using a linear probe

CV +
Ravdess, Savee and Tess

71 speakers, 3 age groups, 5 accents, 7 emotions

Privacy-Aware Projector (PAP)

- A closed-form linear concept eraser (LEACE) in the projector
  → It removes concepts (e.g. gender, speaker id) from representations

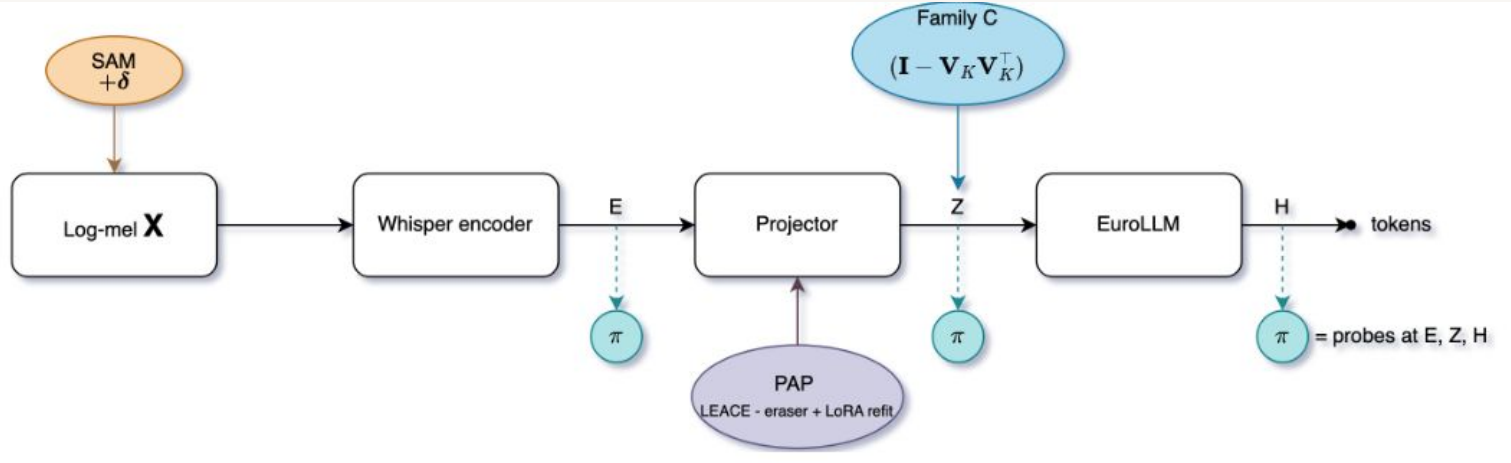
- *Idea:* Find the smallest (least-squares) edit to the features X to linearly uncorrelate them with the target labels.
- *Selective*: it removes a target attribute, leaving the others “intact”
- Designed for *linear classifiers*

Applied to the projector!

N. Belrose, et al, “LEACE: Perfect linear concept erasure in closed form”, Neurips

Privacy-Aware Projector (PAP)

- Comparative methods: selective attribute masking (SAM), subspace projection (Family C)
- **Adaptive** linear probe on the protected representations



Privacy-Aware Projector (PAP)

Experimental example:

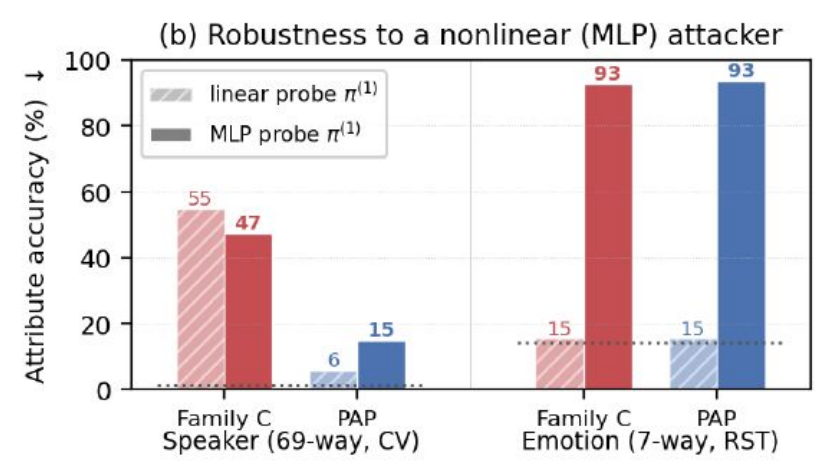
- *Remove speaker ID*
- Preserve the other attributes and the utility (WER)
- Linear probe after the projector adapted to the protected representations

Method	Speaker	Gender	Age	Accent	WER
SAM	75.4→98.7	92.6→92.3	82.7→85.7	88.2→90.8	7.7→7.9
Family C (K=20)	75.4→54.6	92.6→82.6	82.7→67.3	88.2→64.6	7.9→9.1
PAP	**75.4→5.5**	**92.6→77.8**	**82.7→63.3**	**88.2→53.1**	**7.9→9.3**



Privacy-Aware Projector (PAP)... but...

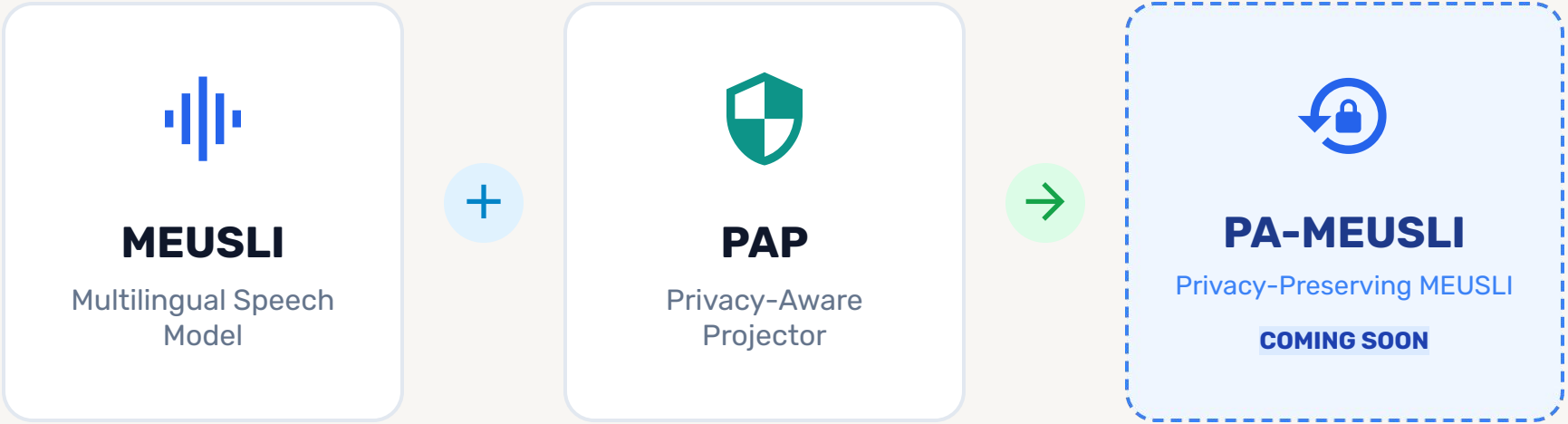
- A non linear attacher (MLP) recovers a lot of information



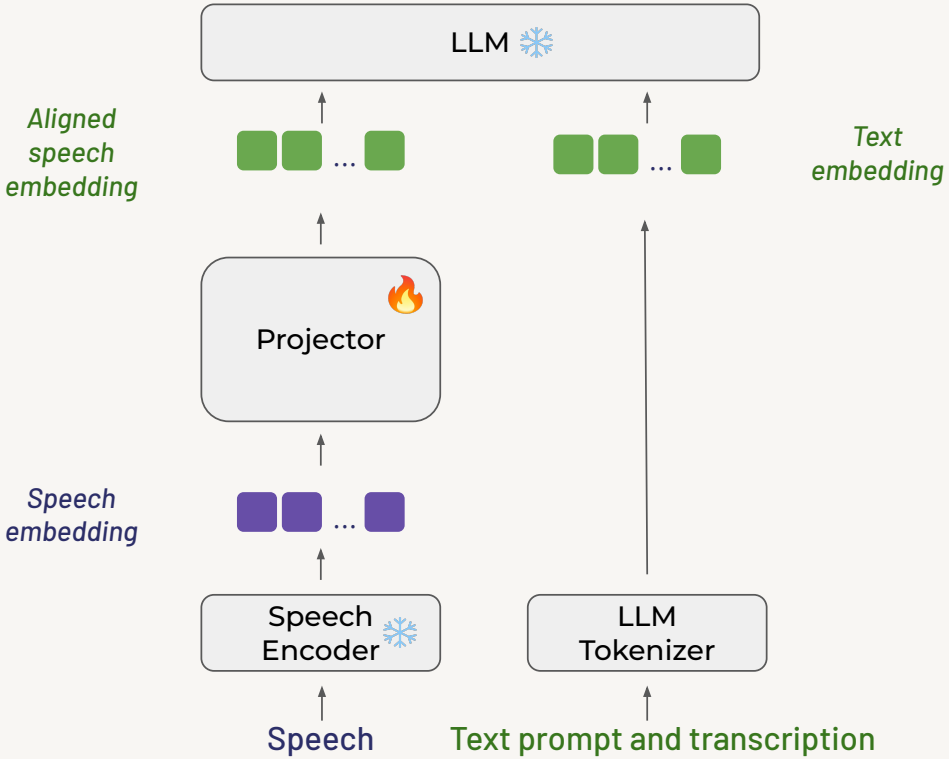
Limitations:

- *Non-linear probe breaks the protection on emotions*
- *Limited configurations and datasets*
- *Privacy preservation benchmarks*

PA-MEUSLI - Coming soon....



SpeechLLM: an established architecture



Privacy & Transparency

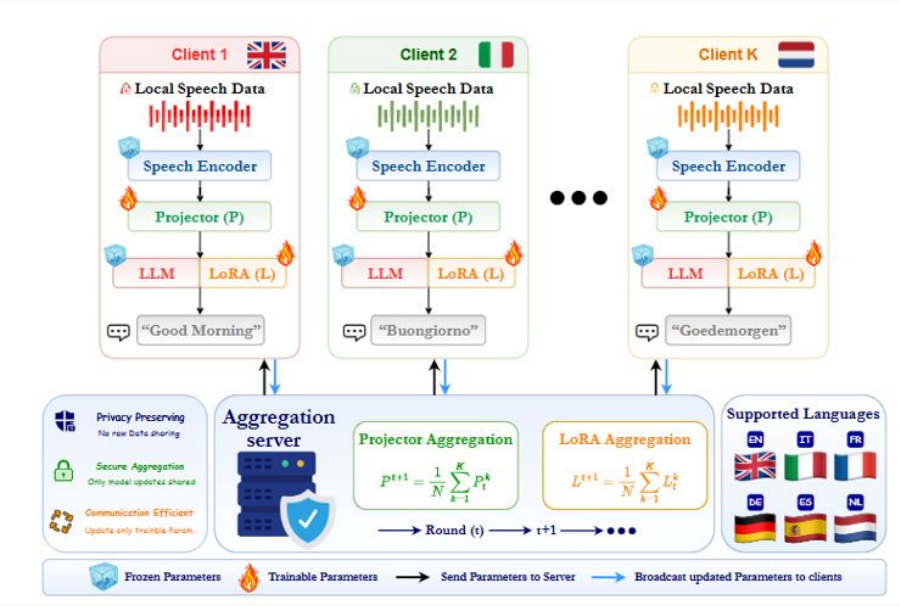
FedL

Z. Ma, et al., “An Embarrassingly Simple Approach for LLM with Strong ASR Capacity,” 2024

SpeechLLM Meets Federated Learning:
Keeping data local

Realistic settings:

- Multilingual data
- Multiple monolingual clients
- Clients and data distribution uneven across languages



Method:

- Communication efficient federation of projectors
- Effective client balance
- Released in Flower.ai

Fed-SpeechLLM: Federated Learning Speech Language Models for Multilingual ASR

Mohamed Nabih Ali, Daniele Falavigna, Alessio Brutti

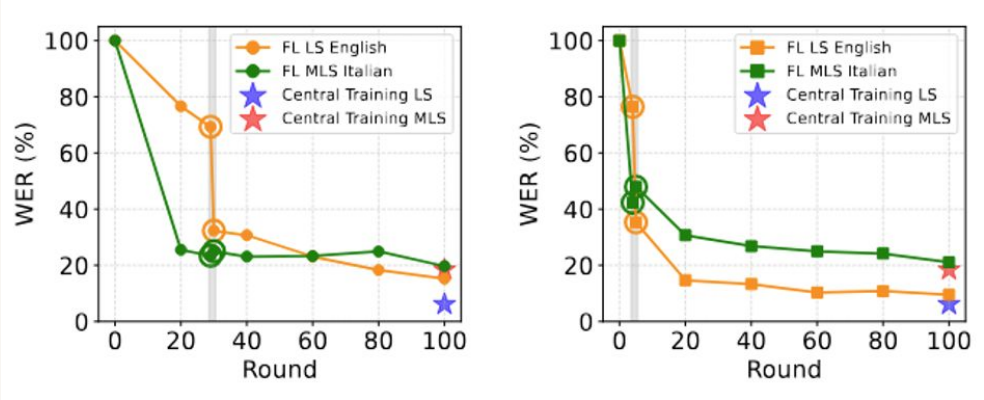
SpeechLLM Meets Federated Learning: bilingual case

	LibriSpeech		MLS	
	hours	# Spks.	hours	# Spks.
Train	100	251	247.38	65
Test	5.4	40	5.27	10

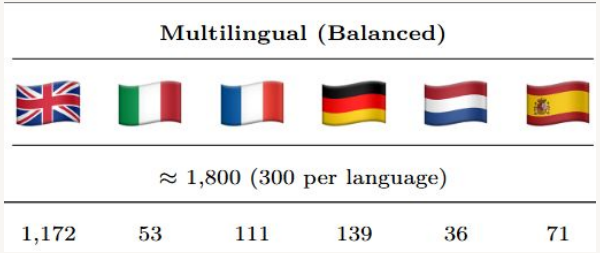
1 client == 1 speaker

More of small & asap

Dataset	Training Scenario	LS	MLS
Central Training	Monolingual	6.1	20.1
	Multilingual	6.1	18.4
	Monolingual FL	6.4	22.6
FL Random Selection	Client-Balanced	33.8	22.7
	Full-Clients	16.8	19.7
	Full-Clients + FT	12.1	21.1
FL Cluster Selection	Sample Clustering	9.7	19.6
	Language Clustering (Late Biasing)	15.2	19.7
	Language Clustering (Early Biasing)	9.5	21.4



SpeechLLM Meets Federated Learning: multilingual



Data from LS and MLS

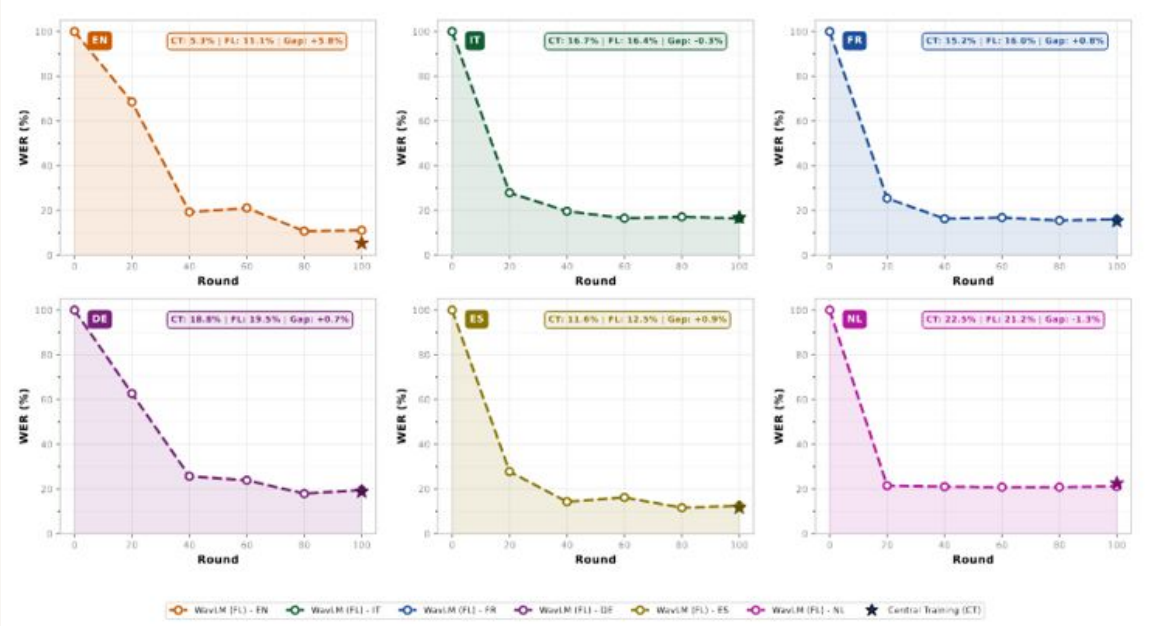
Same amount of hours but different number of clients

How do we decide where to focus at each iteration?

WER-based selection **FAILED**

Round-robin focus: each language gets priority for a while **WORKED**

SpeechLLM Meets Federated Learning: multilingual



Limitations: limited number of languages, 2 datasets, 1 task

Future directions

- Fine tuning MEUSLI on new languages and tasks → Catastrophic forgetting Continual learning on new tasks and languages (SLT paper)

- Further extension of inclusivity targeting pathological speech

- FedL: further scaling towards more languages and tasks

Thanks

Questions?

MEUSLI - https://arxiv.org/abs/2607.22100

Low-resourced/Multi-task:
https://www.isca-archive.org/interspeech_2025/fong25_interspeech.html;
https://www.isca-archive.org/interspeech_2026/fong26b_interspeech.pdf

Privacy leakage and mitigation:
https://www.isca-archive.org/odyssey_2026/sepanta26_odyssey.pdf

Multilingua federated learning: https://arxiv.org/abs/2607.25716

Hugging Face Model
SpeechTek/MEUSLI_projector_v2

Fed-SpeechLLM Repository
SpeechTechLab/Fed-SpeechLLM

Flower Framework App
flower.ai/apps/mnabih/speech_llm_fl

LLMs4EU

Meetween

Funded by the European Union

eloquence

alt-edic

CLOUD-EDGE CONTINUUM