

# Federated Multilingual Speech-LLMs: Architecture and Aggregation Strategy Benchmarking

Jordi Luque <sup>1,\*\*</sup>, Aleix Sant <sup>1,2</sup>, Fernando López <sup>1,3</sup>

<sup>1</sup> Telefónica Innovación Digital, Scientific Research

<sup>2</sup> Universitat Politècnica de Catalunya

<sup>3</sup> Universidad Autónoma de Madrid

jordi.luque@telefonica.com

## Abstract

We present a comprehensive benchmark of Federated Learning (FL) for multilingual Automatic Speech Recognition (ASR), evaluating four Speech-LLM architectures on the Multilingual LibriSpeech dataset. We compare FedAvg and FedProx across frozen and unfrozen encoder configurations, demonstrating that optimized learning rates are critical for performance. Specifically, independently tuning the learning rates for the speech encoder, connector, and decoder yields the lowest error rates, with full three-component adaptation (LoRA for encoder and decoder, full training for the connector) producing the best FL results. We observe that FedProx efficacy is architecture-dependent, providing notable advantages in multilingual pre-trained architectures (e.g., EuroLLM over TinyLlama when keeping the encoder fixed); this indicates that LLM backbone capacity plays a key role in mediating resilience to heterogeneous data distributions. These findings offer concrete design guidance for deploying multilingual Speech-LLMs in privacy-sensitive, distributed environments.

**Index Terms:** federated learning, speech recognition, speech large language models, FedAvg, FedProx, multilingual ASR

## 1. Introduction

Speech large language models (LLMs) have emerged as a prominent paradigm for end-to-end automatic speech recognition (ASR), combining powerful acoustic encoders [1, 2] with pre-trained language decoders [3, 4] through a lightweight cross-modal connector [5]. This architecture leverages large-scale text pretraining to exploit rich language priors, enabling instruction-following speech understanding and achieving state-of-the-art results particularly in multilingual and low-resource settings. Privacy-sensitive deployments in healthcare, legal, and personal assistant applications motivate *federated learning* (FL) [6], where models are fine-tuned on distributed client data without centralising raw audio.

Federated speech learning predates Speech-LLMs and has mainly targeted ASR, where strongly heterogeneous, privacy-sensitive speech data have motivated methods for non-IID optimisation, personalisation, and communication-efficient training [7, 8, 9]. More recent work consolidates parameter-efficient fine-tuning for pretrained speech encoders in FL, showing that LoRA and adapter-based approaches can effectively adapt foundation models under tight communication and compute budgets [10, 11, 12]. However, these studies operate on monolithic speech encoder-only [11] or language-based architectures [8], and no previous available work has directly investigated federated training of Speech-LLMs, leaving open how encoder,

connector, and language-decoder components should be coordinated under multilingual, non-IID client distributions.

Motivated by these gaps, we focus explicitly on Speech-LLMs under the FL paradigm and benchmark four encoder-LLM families and two aggregation strategies in a realistic multilingual setting. Our study addresses the following questions:

- **(Q1) Architecture:** which encoder-LLM combination best handles multilingual non-IID client distributions. Does the encoder type (ASR-supervised vs. self-supervised) affect federated adaptability of Speech-LLMs?
- **(Q2) Aggregation:** does FedProx mitigate client language drift in this multilingual Speech-LLM setting?

These questions are particularly acute in multilingual settings, where clients differ simultaneously in language, accent, speaker population and recording conditions. Such strongly non-IID data cause *client drift* [13] that destabilises FL optimisation [14] and has motivated aggregation strategies such as FedProx [15] or SCAFFOLD [16].

Our FL experiments use a speaker-partitioned split of Multilingual LibriSpeech [17] (8 European languages, 316 clients, one speaker per client), inducing simultaneous acoustic, linguistic and domain-level heterogeneity. Local Speech-LLMs are fine-tuned with LoRA [18] under frozen and unfrozen speech encoders, with the connector and LLM always unfrozen. We analyse the interplay between pretrained encoders and LLM backbones, the effect of differential per-component learning rates, and the FL-ASR behaviour of FedAvg and FedProx in this heterogeneous setting.

## 2. System Architectures

*Speech-LLMs* couple three heterogeneous components: an acoustic encoder ( $\mathcal{E}$ ), a cross-modal connector ( $\mathcal{C}$ ), and a language decoder ( $\mathcal{L}$ ), producing a transcript as:

$$\hat{\mathbf{y}} = \mathcal{L}([\mathcal{C}(\mathcal{E}(\mathbf{x})); \mathbf{E}_{\text{text}}]). \quad (1)$$

All three components are trained jointly by minimising the standard autoregressive cross-entropy loss over the ground-truth transcript tokens  $\mathbf{y} = (y_1, \dots, y_S)$ :

$$\mathcal{L}_{\text{CE}}(\boldsymbol{\theta}) = - \sum_{s=1}^S \log p_{\boldsymbol{\theta}}(y_s | \mathcal{C}(\mathcal{E}(\mathbf{x})), \mathbf{E}_{\text{text}}, y_{<s}), \quad (2)$$

where  $\boldsymbol{\theta}$  collects all trainable parameters (LoRA adapters and connector projection), and  $y_{<s}$  denotes the preceding tokens supplied via teacher forcing. The loss is computed over transcript tokens; audio tokens and prompt embeddings appear as conditioning context. Figure 1 summarises the full pipeline.

\*\*indicates the corresponding author.

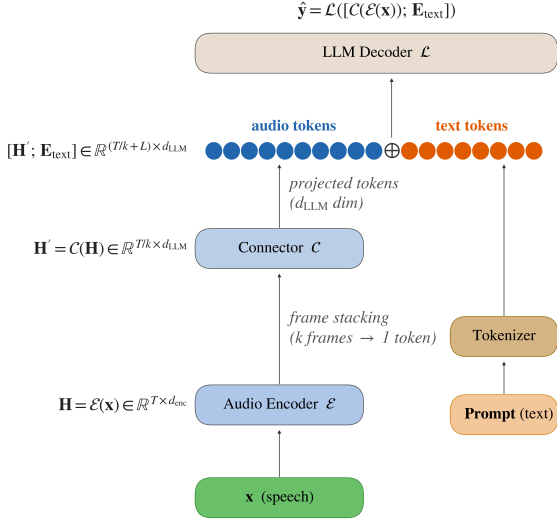


Figure 1: *Speech-LLM architecture: encoder  $\mathcal{E}$  maps speech  $\mathbf{x}$  to frame-level features; connector  $\mathcal{C}$  downsamples by  $k$ ; decoder  $\mathcal{L}$  generates  $\hat{\mathbf{y}}$  from the projected tokens and task prompt.*

### 2.1. Encoder-LLM Models

The combination of foundational acoustic encoders and pre-trained LLMs has become a prominent recipe for multilingual end-to-end speech recognition [3]. We evaluate four encoder-LLM pairings:

**Whisper + TinyLlama.** Whisper large-v3-turbo [1] encodes audio into 768-dimensional speech frame representations, projected to 2048-dimensional tokens to align with TinyLlama-1.1B [19] (22 layers). Whisper is a pretrained end-to-end multilingual ASR model, and its encoder (32 layers) already captures robust cross-lingual acoustic features.

**Whisper + EuroLLM.** Identical Whisper encoder, but the decoder is replaced by EuroLLM-1.7B-Instruct [4], a decoder-only model pretrained on multilingual European text corpora.

**WavLM + TinyLlama.** WavLM-Large [2] replaces Whisper as the encoder (1024-dimensional tokens, 24 layers); TinyLlama remains the decoder. WavLM is a self-supervised learning (SSL) model pretrained on masked speech prediction. Unlike Whisper it has *never* seen ASR supervision, making it a stronger test of whether FL can adapt an SSL encoder to the downstream ASR task.

**Voxtral-Mini** Voxtral [20] is an end-to-end multimodal Speech-LLM composed of a Whisper-large-v3-based audio encoder and a 30-layer Ministral-3B [21] text decoder, jointly pre-trained on audio understanding and ASR.

#### 2.1.1. Cross-modal connector.

The connector  $\mathcal{C}$ , see Fig. 1, bridges the acoustic encoder and the LLM decoder in two steps. First, a *frame-stacking* operation with stride 2 concatenates each pair of consecutive encoder output frames. Second, a single trainable linear layer projects to the input LLM dimension, aligning acoustic representations with the LLM’s embedding space. This design is shared by Whisper+TinyLlama, Whisper+EuroLLM, and WavLM+TinyLlama, where only the single linear projection is trained from scratch. In contrast, Voxtral uses a different connector, downsampling the audio by a factor of 4, followed by a

**Algorithm 1** FedAvg [6].  $\theta$ : LoRA adapters + connector;  $K$ : clients;  $n_k$ : samples on client  $k$ ;  $C$ : client fraction/round;  $B$ : batch size;  $E=10$ : local epochs;  $\eta=10^{-4}$ .

**Server:**

```

initialise  $\theta_0$ 
for each round  $t = 1, 2, \dots, T$  do
   $S_t \leftarrow$  random set of  $\max(C \cdot K, 1)$  clients
  for each client  $k \in S_t$  in parallel do
     $\theta_t^k \leftarrow \text{ClientUpdate}(k, \theta_{t-1})$ 
   $\theta_t \leftarrow \sum_{k \in S_t} \frac{n_k}{\sum_{j \in S_t} n_j} \theta_t^k$ 

```

**ClientUpdate( $k, \theta$ ):**

```

for local epoch  $e = 1, \dots, E$  do
  for each batch  $b \subseteq \mathcal{P}_k, |b|=B$  do
     $\theta \leftarrow \text{AdamW}(\theta, \nabla \ell(\theta; b), \eta)$ 
return  $\theta$ 

```

2-layer MLP projector; a deeper, higher-compression connector in which both MLP layers are trained.

## 3. Federated Setup

### 3.1. Aggregation Strategies

Our federated setting uses synchronous client-server optimization where participating clients ( $C = 0.3$ ) perform  $E = 10$  local AdamW epochs before server aggregation via FedAvg or FedProx. We compute the loss gradient over all data held by these clients,  $C * K = 94$  clients for the global server; see Algorithm 1. Two aggregation strategies are considered:

**FedAvg:** Standard weighted averaging of client updates [6] or each round:

$$\theta^{(t+1)} = \sum_{k \in S_t} \frac{n_k}{\sum_{j \in S_t} n_j} \theta^{(t,k)}, \quad (3)$$

where  $n_k$  are the training samples on client  $k$  and  $S_t$  is the random fraction of clients sampled every round.

**FedProx:** Regularises the client loss, *ClientUpdate* in Algorithm 1, with the proximal term  $\mu$  to limit client drift [15], due to data heterogeneity, for each round:

$$\min_{\theta} \mathcal{L}_{\text{local}}(\theta) + \frac{\mu}{2} \|\theta - \theta^{(t)}\|_2^2, \quad (4)$$

where  $\mu > 0$  penalises deviation, e.g. due to narrow acoustic distribution in client data, from the current global model  $\theta^{(t)}$ .

### 3.2. Dataset and Partitioning

#### 3.2.1. Corpus

All experiments use the Multilingual LibriSpeech (MLS) corpus [17], an audiobook corpus covering 8 European languages derived from LibriVox recordings. We use the official MLS *train* splits as the federated training pool (685.7 h total), the MLS *dev* split for validation during training, and the MLS *test* split as the held-out evaluation benchmark (138 h, 19,492 samples). Table 1 reports training hours and speaker counts per language.

#### 3.2.2. Multilingual Partition (approximately IID)

The multilingual partition serves as a control to isolate the effect of data heterogeneity from architecture choice. A random

Table 1: *MLS training data by language under the stratified by speaker partition, 316 clients.*

| Language     | Hours        | Speakers/Clients |
|--------------|--------------|------------------|
| French       | 251.6        | 15               |
| German       | 160.9        | 19               |
| English      | 105.3        | 256              |
| Spanish      | 83.8         | 9                |
| Italian      | 27.3         | 7                |
| Polish       | 25.7         | 1                |
| Portuguese   | 18.5         | 5                |
| Dutch        | 12.7         | 4                |
| <b>Total</b> | <b>685.7</b> | <b>316</b>       |

multilingual mixture of utterances is assigned to each client irrespective of speaker identity, approximating the IID assumption. Because MLS is derived from LibriVox audiobooks, the same reader can appear across official train, dev, and test splits. The centralized training pools this full set and therefore carries a 3.5% speaker leakage. For FL experiments, partition (A) in Table 2, same identities overlap contaminating 59 clients out of  $K=316$  clients (18.7%). These clients contain utterances from test or dev sharing same speakers, totalling 60,825 of 1,726,583 training samples (i.e the same 3.5% speaker leakage). Note that these are different utterances of the same speaker, not duplicate audio.

### 3.2.3. Speaker Partition (non-IID)

The speaker partition, (B) in Table 2, assigns all utterances of a single MLS speaker to one client. With  $K=316$  clients, this creates the strongest possible non-IID distribution: each client’s data is drawn from a single acoustic identity, language, and recording environment, producing simultaneous *linguistic* (each client speaks at most one language) and *acoustic* (microphone, room, speaking rate) heterogeneity. Due to the LibriVox origin of MLS, 8 of 316 clients (2.5%) correspond to speakers also present in the MLS test split, accounting for 4,747 of 169,586 training samples (2.8%), while 3.5% of centralized training data and 18.7% of clients in partition (A) share speakers with the test set. Because these overlap rates are inherited from the standard MLS splits, the centralized and IID upper bounds benefit from acoustic speaker familiarity. Consequently, the actual degradation caused by FL non-IID conditions is slightly less severe than the raw distance to the upper bounds suggests.

### 3.3. Training Configuration

All FL experiments use the Flower simulation framework with Ray as the backend [22]. Experiments on the non-IID speaker partition (B) train for  $T = 40$  global rounds. For the multilingual partition (A), results are reported at  $T = 9$  global rounds ( $\ddagger$  in Table 2). This is because the multilingual mixture in partition (A) exhibits rapid convergence leading to early WER optimization collapse. All hyper-parameter choices—encoder learning-rate multiplier  $\lambda$ , FedProx coefficient  $\mu$ , and local epochs  $E$ —were selected on the MLS dev set, and the test split is used only to evaluate each selected checkpoint. Local clients perform  $E=10$  local epochs per round, using AdamW optimiser with maximum learning rate  $\eta=10^{-4}$ , cosine decay, batch size 16, and half precision bf16. Centralized (non-FL) training uses the same AdamW optimiser with cosine decay

Table 2: *WER on MLS test set. B: non-IID speaker partition. A: approximately IID multilingual partition. Gap: WER relative to corresponding FedAvg baseline; negative = improvement.  $\dagger$ : best encoder learning rate multiplier ( $0.02 \times \text{LLM lr}$ ).  $\ddagger$ : reported at round 9.*

| System                                 | Encoder Setting        | WER           | Gap            |
|----------------------------------------|------------------------|---------------|----------------|
| <i>Whisper + TinyLlama-1.1B</i>        |                        |               |                |
| Centralized                            | frozen                 | 0.0719        | —              |
| FedAvg                                 | frozen (B)             | 0.1415        | —              |
| FedAvg                                 | frozen (A) $\ddagger$  | 0.1284        | −0.0131        |
| FedAvg                                 | unfrozen (B) $\dagger$ | 0.1418        | +0.0003        |
| FedProx $\mu=0.001$                    | frozen (B)             | 0.1499        | +0.0084        |
| FedProx $\mu=0.001$                    | frozen (A) $\ddagger$  | 0.1386        | +0.0102        |
| FedProx $\mu=0.05$                     | frozen (B)             | 0.1615        | +0.0200        |
| FedProx $\mu=0.1$                      | frozen (B)             | 0.1843        | +0.0428        |
| <i>Whisper + EuroLLM-1.7B-Instruct</i> |                        |               |                |
| Centralized                            | frozen                 | 0.0660        | —              |
| FedAvg                                 | frozen (B)             | 0.1330        | —              |
| FedAvg                                 | frozen (A) $\ddagger$  | 0.1224        | −0.0106        |
| FedAvg                                 | unfrozen (B)           | 0.1778        | +0.0448        |
| FedAvg                                 | unfrozen $\dagger$ (B) | <b>0.1170</b> | <b>−0.0160</b> |
| FedProx $\mu=0.001$                    | frozen (B)             | 0.1217        | −0.0113        |
| <i>WavLM-Large + TinyLlama-1.1B</i>    |                        |               |                |
| Centralized                            | frozen                 | 0.2409        | —              |
| FedAvg                                 | frozen (B)             | 0.5559        | —              |
| FedAvg                                 | unfrozen (B)           | 0.5338        | −0.0221        |
| <i>Voxtral-Mini-3B-2507</i>            |                        |               |                |
| Centralized                            | frozen                 | 0.1238        | —              |
| FedAvg                                 | frozen (B)             | 0.1442        | —              |
| FedAvg                                 | unfrozen (B)           | 0.1362        | −0.0080        |
| FedProx $\mu=0.001$                    | frozen (B)             | 0.1492        | +0.0050        |
| FedProx $\mu=0.001$                    | unfrozen (B)           | 0.1468        | +0.0106        |

over a maximum of 10 epochs; the best-validation-WER checkpoint is selected (typically around epoch 4). Evaluation uses the MLS test split with overall WER reported across all 8 languages combined. All models are fine-tuned with LoRA [18] adapters (rank  $r=8$ ,  $\alpha=16$ , dropout 0.05; Voxtral uses  $\alpha=32$ ) applied to the attention projections of the speech encoder ( $q, k, v$ ) and LLM decoder ( $q, v$ ), together with the fully-trained connector; only these parameters are trainable and are transmitted between clients and the FL server, reducing communication cost over full fine-tuning. Unlike the pretrained encoders and LLM decoders, we initialise the connector from scratch (except for Voxtral), so it must learn to bridge the modalities entirely from the federated fine-tuning data. Unless stated otherwise, WER values are reported as proportions (e.g., 0.1415 corresponds to 14.15%).

## 4. Results and Discussion

### 4.1. Centralized and IID Upper Bounds

Centralized training (non-FL) on the pooled partition sets the per-architecture performance ceiling (Table 2); all centralized models converge within 4–7 epochs. Whisper+EuroLLM reaches WER 0.066 versus 0.072 for Whisper+TinyLlama, confirming the multilingual LLM advantage even without non-IID pressure; WavLM+TinyLlama reaches 0.24 (3.3 $\times$  higher than the Whisper variant), showing that ASR pretraining of the encoder decisively beats SSL pretraining for ASR.

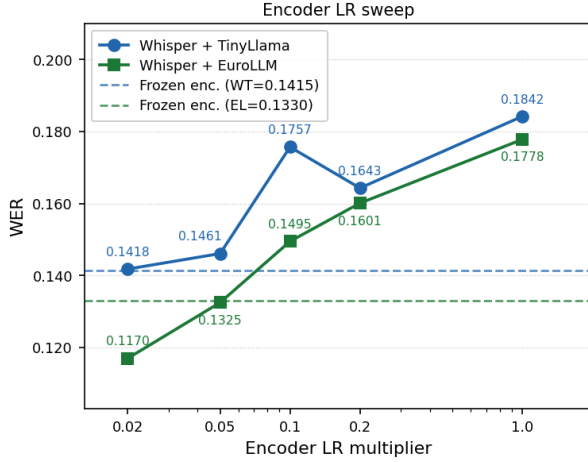


Figure 2: WER vs. encoder learning-rate multiplier (relative to LLM LR). Whisper+TinyLlama (blue,  $E=5$ ) stays flat at its frozen baseline while Whisper+EuroLLM (green,  $E=10$ ) improves monotonically; dashed: frozen-encoder baselines.

#### 4.2. Aggregation Strategy Comparison

With a frozen encoder, FedAvg reaches WER 0.1415 (96.8% relative degradation from centralized) for Whisper+TinyLlama; replacing only the decoder with EuroLLM reduces it to 0.133 (−6.0% relative), and Voxtral (0.144) underperforms the EuroLLM pipeline. WavLM reaches 0.56, confirming that ASR-pretrained encoders are decisive under FL non-IID conditions. FedProx regularisation degrades Whisper+TinyLlama monotonically with  $\mu$ , and also hurts Voxtral (frozen 0.144  $\rightarrow$  0.149, unfrozen 0.1362  $\rightarrow$  0.1468), contradicting its non-IID motivation. Strikingly, FedProx  $\mu=0.001$  improves Whisper+EuroLLM (WER **0.1217**, −8.5% relative), the best frozen-encoder result.

#### 4.3. Encoder LR Tuning under FL

Unfreezing the Whisper encoder under FedAvg yields WER 0.1418, comparable to the frozen baseline (0.1415), even after sweeping the encoder learning rate  $\in \{0.01, 0.02, 0.05, 0.10\}$  and for different values of local epochs  $E \in \{3, 5, 10\}$ , see Fig. 2. We hypothesise that this stems from encoder-update cancellation, in which each client adapts the encoder to its own speaker’s acoustics, producing divergent updates that average to near-zero under FedAvg, while LLM updates on the shared linguistic task survive averaging. WavLM confirms the pattern. The centralized gap (3.2 $\times$ ) grows to 3.9 $\times$  under FL (0.24  $\rightarrow$  0.56), suggesting SSL pretraining misalignment is amplified by heterogeneous data. Unfreezing the WavLM encoder provides only marginal gain. For EuroLLM, after sweeping the learning rate multiplier for the encoder we can observe a monotonic improvement from 0.178  $\rightarrow$  **0.117**, see Fig. 2, a 12% relative below the encoder-frozen EuroLLM variant. Voxtral also benefits modestly from unfreezing (0.144  $\rightarrow$  0.136). These results suggest that, for multicomponent Speech-LLMs, separately adjusting the encoder and decoder is crucial when the encoder is well pre-trained, in order to prevent update cancellation during federated aggregation.

#### 4.4. Per-Language Analysis

Table 3 compares FedAvg TinyLlama and EuroLLM against the FedProx with  $\mu=0.001$  which reaches the lower WER 0.122 between frozen-encoder FL results. For the FedAvg algorithm,

Table 3: Per-language WER on MLS test set for FL encoder frozen variants trained on speaker (non-IID) partition. Cent. = Centralized Whisper+TinyLlama. FedAvg<sub>WT</sub> = FedAvg Whisper+TinyLlama. FedAvg<sub>EL</sub> = FedAvg Whisper+EuroLLM. FP<sub>EL</sub> = FedProx  $\mu=0.001$  Whisper+EuroLLM. Bold: best FL result per language.

| Lang.          | Cent.        | FedAvg <sub>WT</sub> | FedAvg <sub>EL</sub> | FP <sub>EL</sub> |
|----------------|--------------|----------------------|----------------------|------------------|
| Dutch          | 0.110        | 0.270                | 0.265                | <b>0.195</b>     |
| English        | 0.049        | 0.100                | 0.098                | <b>0.090</b>     |
| French         | 0.061        | 0.075                | 0.078                | <b>0.072</b>     |
| German         | 0.071        | 0.104                | <b>0.098</b>         | 0.125            |
| Italian        | 0.121        | 0.233                | <b>0.186</b>         | 0.189            |
| Polish         | 0.110        | 0.270                | 0.237                | <b>0.228</b>     |
| Portuguese     | 0.093        | 0.225                | <b>0.160</b>         | 0.177            |
| Spanish        | 0.046        | 0.085                | 0.075                | <b>0.068</b>     |
| <b>Overall</b> | <b>0.072</b> | 0.142                | 0.133                | <b>0.122</b>     |

EuroLLM also outperforms TinyLlama on six of eight languages; with the largest gains are on most of low-resource languages, see Table 1, where multilingual priors matter most: Portuguese (−6.5% abs.), Italian (−4.8%), Polish (−3.3%). FedProx EuroLLM improves over FedAvg on five of eight languages: Dutch (0.265  $\rightarrow$  0.195, −7.0% abs.), Polish, English, French, and Spanish. Conversely, it degrades German (+2.6% abs.), Italian, and Portuguese. Note that benefit is strongest for the least-represented languages ( $\leq 4$  clients), while moderate-resource languages show mixed results.

Because the client distribution is strongly skewed toward English (256 of 316 clients) and the overall WER is dominated by high-resource evaluation words, we also report macro-averaged WER (unweighted mean of the per-language values in Table 3): FedAvg Whisper+TinyLlama is 0.170 versus 0.150 for Whisper+EuroLLM. This macro gap (−2.0 points) is larger than the word-weighted gap (−0.9 points), confirming that the EuroLLM advantage is driven particularly by the low-resource languages rather than being masked by them.

## 5. Conclusion

This work benchmarks federated training for Speech-LLMs across four architectures and two aggregation strategies using a partition of 316 clients from the Multilingual LibriSpeech dataset. The results show that multilingual pre-training is a primary driver of FL performance, with EuroLLM outperforming TinyLlama, by 6% relative WER, and Voxtral pipelines. Furthermore, the efficacy of the FedProx algorithm is architecture-dependent, improving Whisper+EuroLLM while degrading Whisper+TinyLlama and Voxtral configurations. Controlled paired comparisons suggest that stronger multilingual LLM capacity helps absorb client drift under proximal regularization, though broader multi-component variations also introduce secondary effects from connector depth and encoder pre-training. We also report that differential learning rate tuning (low-rate encoder vs. high-rate decoder) provides superior results compared to frozen encoders, since the components of a Speech-LLM exhibit distinct gradient scales and adaptation requirements that preclude treating the model as monolithic during FL optimisation. These results offer a road map for practitioners deploying robust, cross-lingual Speech-LLM systems in federated settings.

## 6. Acknowledgements

This work has received funding from the European Union's Horizon Europe research and innovation programme under the project ELOQUENCE (Grant Agreement No. 101135916). This work was supported by computational resources from the EuroHPC Joint Undertaking under the EuroHPC AI Factory grant EHPC-AIF-2026LS01-004.

## 7. References

- [1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.
- [2] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," in *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, 2022, pp. 1505–1518.
- [3] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities," *arXiv preprint arXiv:2305.11000*, 2023.
- [4] P. H. Martins, P. Fernandes, M. Faysse *et al.*, "EuroLLM: Multilingual language models for europe," *arXiv preprint arXiv:2409.16235*, 2024.
- [5] C. Tang, W. Yu, G. Sun, X. Chen, W. Tan, L. Li, Z. Lu, S. Watanabe, and C. Zhang, "SALMONN: Towards generic hearing abilities for large language models," in *Proc. ICLR*, 2024.
- [6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, 2017, pp. 1273–1282.
- [7] H. Zhu, J. Wang, G. Cheng, P. Zhang, and Y. Yan, "Decoupled federated learning for asr with non-iid data," in *Interspeech 2022*, 2022, pp. 2628–2632.
- [8] D. Guliani, F. Beaufays, and G. Motta, "Federated learning for ASR based on wav2vec 2.0," in *Proc. ICASSP*, 2021, pp. 3040–3044.
- [9] Y. Gao, T. Parcollet, S. Zaiem, J. Fernandez-Marques, P. P. B. de Gusmao, D. J. Beutel, and N. D. Lane, "End-to-end speech recognition from federated acoustic models," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7227–7231.
- [10] Y. Du, Z. Zhang, L. Yue, X. Huang, Y. Zhang, T. Xu, L. Xu, and E. Chen, "Communication-efficient personalized federated learning for speech-to-text tasks," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10 001–10 005.
- [11] X. Kan, Y. Xiao, T.-J. Yang, N. Chen, and R. Mathews, "Parameter-efficient transfer learning under federated learning for automatic speech recognition," *CoRR*, vol. abs/2408.11873, 2024. [Online]. Available: <https://arxiv.org/abs/2408.11873>
- [12] M. N. Ali, D. Falavigna, and A. Brutti, "EFL-PEFT: A communication-efficient federated learning framework using peft sparsification for asr," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.
- [13] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. ICML*, 2020, pp. 5132–5143.
- [14] P. Kairouz, H. B. McMahan, B. Avent *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
- [15] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. MLSys*, 2020.
- [16] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," *arXiv preprint arXiv:1806.00582*, 2018.
- [17] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Proc. Interspeech*, 2020, pp. 2757–2761.
- [18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2022.
- [19] P. Zhang, G. Zeng, T. Wang, and W. Lu, "TinyLlama: An open-source small language model," *arXiv preprint arXiv:2401.02385*, 2024.
- [20] H. Liu *et al.*, "Voxtral," *arXiv preprint arXiv:2507.13264*, 2025. [Online]. Available: <https://arxiv.org/abs/2507.13264>
- [21] —, "Ministral 3," 2026. [Online]. Available: <https://arxiv.org/abs/2601.08584>
- [22] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmao *et al.*, "Flower: A friendly federated learning research framework," in *Proc. ICLR Workshop on Distributed and Private Machine Learning*, 2020.