# Scalable Expansion of Multilingual Speech LLMs for ASR: A Continual Learning Approach

**Lorenzo Concina, Marco Matassoni, Alessio Brutti**
Center for Augmented Intelligence, Fondazione Bruno Kessler, Trento, Italy
{lconcina, matasso, brutti}@fbk.eu

**Abstract**

Speech Large Language Models have recently enabled the processing of spoken language by coupling powerful language models (LLMs) with pre-trained speech encoders. However, their multilingual scalability remains limited, particularly for low-resource and unseen languages, while naïve fine-tuning often triggers catastrophic forgetting of previously learned languages. This work investigates how Continual Learning (CL) can be used to sustainably expand multilingual Speech LLMs. We first demonstrate that multilingual projectors can be efficiently bootstrapped to new languages, even with extremely small datasets, but at the cost of severe degradation on the original supported languages. To address this, we adopt rehearsal-based CL strategies and show that interleaving even small amounts of replay data effectively stabilizes multilingual performance. Through extensive ablations, we quantify the minimum rehearsal budget required to prevent forgetting and identify fragile languages that require more targeted reinforcement. We further evaluate sequential acquisition of four linguistically diverse languages (Ukrainian, Japanese, Thai, and Vietnamese), revealing the trade-offs between buffer size and long-term stability. Finally, based on these empirical observations, we propose a Fragility-Based Sampling heuristic as a pathway to allocate rehearsal data more efficiently by tiering languages according to their stability thresholds. Our findings provide a practical roadmap for scalable, resource-efficient multilingual expansion of Speech LLMs, enabling inclusive ASR systems that can grow over time without sacrificing prior knowledge.

**Keywords:** continual learning, data replay, speech recognition, LLM, low-resource languages, multilinguality, fragility

## 1. Introduction

Large Language Models (LLMs) have transformed natural language processing, and their integration with audio modalities has birthed the era of Speech Large Language Models (Speech LLMs). By bridging pre-trained speech encoders with high-capacity language decoders, these models move beyond simple transcription to achieve a more context-aware understanding of spoken communication. Traditionally, speech understanding has relied on cascaded architectures that sequentially link ASR, language modeling, and TTS modules. However, this pipeline approach suffers from several flaws (1). First, it is prone to error propagation, where inaccuracies in transcription directly degrade the quality of the subsequent text generation. Second, these systems lack a shared context; once speech is converted to text, paralinguistic features, such as pitch and tone are lost. Consequently, the language model cannot take advantage of the rich information contained in the original audio signal. Furthermore, the multi-stage process introduces significant processing latency. Speech Language Models (SLMs) (2) seek to resolve these problems by processing audio directly, utilizing the inherent capabilities of LLMs to match or exceed the performance of specialized, task-specific models in a unified framework. Despite recent advances, there is still a gap in the multilingual scalability of these systems. Most state-of-the-art models are optimized primarily for high-resource languages such as English, often struggling with the phonetic and structural nuances of low-resource or underrepresented languages (3). In the context of Automatic Speech Recognition (ASR), extending Speech LLMs to a global scenario is not just a matter of data volume; it requires overcoming challenges such as catastrophic forgetting (4), where the acquisition of a new language degrades performance on previously learned ones.

Building on these challenges, the scope of this work focuses on the sustainable expansion of Speech LLM architectures through Continual Learning (CL). Rather than resorting to the computationally expensive and data-intensive process of retraining models from scratch to accommodate new linguistic domains, we investigate methods to incrementally integrate additional languages into an existing system. Using as a starting point the MEUSLI [1] projector (a multilingual interface connecting a frozen Whisper encoder to a 1.7B EuroLLM based on SLAM-ASR framework (5)), we examine how to efficiently bootstrap support for unseen and low-resource languages. We specifically explore rehearsal-based (6) (Data Replay) strategies to mitigate the catastrophic forgetting inherent in the SLAM-ASR framework. By analyzing the minimum rehearsal data required and proposing

[1] https://huggingface.co/SpeechTek/MEUSLI_projector_v2



*Proceedings of SPEAKABLE @ LREC 2026*, pages 59–65
11 May 2026. 

a Fragility-Based Sampling heuristic, we provide a pathway for ASR Speech-LLM systems that is both inclusive of underrepresented languages and resource-efficient.

## 2. The Speech-LLM Paradigm for ASR

The emergence of Speech Large Language Models (SLMs) marks a shift toward unified, multimodal architectures that bridge the gap between raw acoustic signals and high-level linguistic reasoning. Unlike traditional cascaded systems, which reduce speech to plain text before processing, SLMs aim to maintain a continuous flow of information, preserving paralinguistic cues such as tone, pitch, and rhythm. These models typically leverage the vast pre-trained knowledge of Large Language Models (LLMs) to perform tasks like Automatic Speech Recognition (ASR) or Spoken Question Answering directly from audio input.

### 2.1. The SLAM-ASR Architecture

This study leverages the SLAM-ASR framework[2], which provides an efficient, modular approach to multimodal integration. As shown in Figure 1, the architecture is composed of three main components:

- **Speech Encoder**: Transforms raw audio into high-dimensional acoustic representations.
- **Projector**: A lightweight module that maps acoustic representations into the same embedding space as the LLM's token embeddings. This interface can range from simple linear layers to more complex neural architectures. In this work we leverage the MEUSLI linear projector.
- **Language Model**: A pre-trained LLM that processes the projected speech embeddings as if they were text tokens.

A significant advantage of this paradigm is the efficiency, during training in fact, both the encoder and the LLM are typically kept frozen, with the possibility of eventually use Low-Rank Adaptation (LoRA) (7) to further enhance LLM performance.

### 2.2. Baseline Model: The MEUSLI Projector

To evaluate the ability to extend a Speech-LLM based model to new languages by applying Continual Learning techniques, we utilize the MEUSLI (Multilingual EU Speech LInear projector) model as our experimental baseline. MEUSLI is an open-science initiative designed to connect a frozen **Whisper-large-v3-turbo** (8) encoder with the multilingual **EuroLLM 1.7B-Instruct** (9) backend.

The model was initially trained on 7,622 hours of open-source data from Common Voice, FLEURS, and VoxPopuli, covering 28 European languages. The projector itself is a linear layer with approximately 17.31M trainable parameters, supplemented by a LoRA configuration in the LLM adding 1.38M tunable parameters.

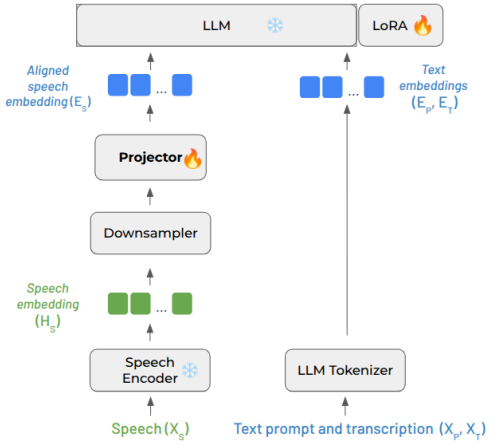


Figure 1: The proposed Speech LLM architecture used in MEUSLI (Multilingual EU Speech LInear projector) training pipeline (5; 3).

### 2.3. Multilingual Capabilities and Constraints

The current iteration of the projector natively supports 28 European languages, ranging from high-resource (Spanish, French, German) to medium and low-resource ones (Breton, Maltese, Welsh). To enhance performance, Low-Rank Adaptation (LoRA) is applied to the LLM, introducing 1.38M tunable parameters. Despite its broad coverage, the model faces some limitations:

- Language Gaps [1]: Performance remains uneven, with very low-resource languages (e.g., Irish or Breton) exhibiting significantly higher Word Error Rates (WER) compared to their high-resource counterparts.
- The model is only focused on European languages, but it does not cover them all.
- The model has been trained on open source data: Common Voice 17 (10), Fleurs (11) and VoxPopuli (12), therefore it is not suitable for other specific domains or accents.

## 3. Bootstrapping New Languages

To explore the model's capacity for expansion, we investigate the process of bootstrapping—fine-

[2] https://github.com/X-LANCE/SLAM-LLM

tuning the pre-trained MEUSLI projector on languages entirely absent from its initial training set. This stage is critical for assessing whether a multilingual foundation model can effectively transfer knowledge to low-resource scenarios with minimal train data.

We conducted two primary bootstrapping experiments on languages that utilize different scripts or linguistic structures than those natively supported. For these two experiments, we used the training recipe of SLAM-ASR while loading the MEUSLI projector as starting checkpoint:

- Ukrainian: Fine-tuning on 30 hours of Common Voice data enabled the model to achieve a 16.3% Word Error Rate (WER), a notable improvement over the monolingual training of the same exact pipeline that only reaches 20.4% WER.
- Albanian: Even with an extremely limited dataset of only 46 minutes of Common Voice data, the model successfully bootstrapped to a 75.6% WER. In contrast, a monolingual projector trained from scratch on this same data failed to converge, reaching 389% WER.

These two experiments reported in Table 1 show promising results on how to extend this model to new languages. These results confirm that multilingual pre-training provides a superior initialization point for new language acquisition.

Table 1: WER for the bootstrapping experiment on Ukrainian and Albanian. The training data comes from Common Voice 17. "Mono" refers to a monolingual projector trained on the CV data, "→" indicates MEUSLI fine-tuned on CV data.

| Language | Training | Mono | MEUSLI→ |
|---|---|---|---|
| Ukrainian | 30h | 20.4% | **16.3%** |
| Albanian | 46min | 389% | **75.6%** |

### 3.1. The Emergence of Catastrophic Forgetting

While the model successfully acquires the new target language, this specialized fine-tuning induces a severe side effect known as catastrophic forgetting. This phenomenon occurs when the weights of the projector are adjusted to the distribution of the new task or in this case language, causing a sudden and near-total loss of performance on the original 28 supported European languages. Table 2 shows the side effects on some of the 28 supported languages after fine-tuning on Ukrainian. Specifically, for almost all these test languages in both test sets (Common Voice 17 and Fleurs), the model is no longer able to transcribe properly. The output transcription is instead full of hallucinations with Ukrainian characters.

Table 2: Comparison of WER showing the catastrophic forgetting effect after fine-tuning on Ukrainian. High WER values in the Fleurs and CV17 columns indicate a collapse of original multilingual capabilities compared to the base model performances reported on the third and fourth columns. "→" indicates MEUSLI fine-tuned on CV data

| Language | MEUSLI→ FL | MEUSLI→ CV | MEUSLI CV | MEUSLI FL |
|---|---|---|---|---|
| Spanish | 90.34 | – | 5.22 | 4.09 |
| German | 15.50 | 19.87 | 7.11 | 7.79 |
| French | 68.90 | – | 11.24 | 7.83 |
| Portuguese | 109.6 | 123.66 | 9.39 | 4.86 |
| English | 9.02 | 22.72 | 12.94 | 6.34 |
| Polish | 100.6 | – | 14.19 | 8.68 |
| Czech | 108 | – | 11.16 | 11.32 |
| Italian | 39.28 | 60.18 | 6.01 | 3.32 |
| Danish | 40.02 | 95.16 | 18.81 | 14.65 |
| Latvian | 68.38 | – | 27.12 | 17.23 |

## 4. Continual Learning for Multilingual ASR

To learn how to transcribe a new language and at the same time mitigate the catastrophic forgetting identified in Section 3.1, we adopt a **Continual Learning (CL)** framework based on **Data Rehearsal** (or Data Replay) and inspired by the framework proposed in CL-MASR (13). The core objective is to integrate new languages sequentially, without degrading the performance of the original 28 European languages already supported by the MEUSLI model.

Within this Continual Learning framework, we formulate the expansion of the Speech LLM as a **Task Incremental Learning (TIL)** problem. In this paradigm, the model is trained on a sequence of distinct tasks $T_1, T_2, \ldots, T_N$, where each task $T_i$ represents the acquisition of a new, previously unsupported language (e.g., $T_1$ = Ukrainian, $T_2$ = Japanese). The objective is to minimize the error on the new task $T_i$ while strictly bounding the performance degradation on all previously learned tasks $T_{<i}$, including the original 28 base languages.

To formally evaluate the global stability of the model across this sequence, we adopt the **Average Word Error Rate (AWER)** metric as in (13). Let $WER_{i,j}$ be the Word Error Rate evaluated on task $j$ after the model has finished training on task $i$. The AWER after training on task $i$ is defined as the arithmetic mean of the errors across all active tasks up to that point:

$$AWER_i = \frac{1}{i} \sum_{j=1}^{i} WER_{i,j} \quad (1)$$

This metric allows us to monitor the evolution of performances.

### 4.1. Establishing a Continual Learning Baseline: The Case of Ukrainian

We first validate the effectiveness of data rehearsal by repeating the Ukrainian experiment as the initial step in our TIL sequence ($T_1 =$ Ukrainian). In this setup, we fine-tune the projector on the new task $T_1$ using the Common Voice (CV) training set, while simultaneously replaying a small subset of samples from the original 28 base languages. Initial results using a rehearsal budget of 1,000 samples per language—selected deterministically as the first 1,000 utterances from each training set—demonstrated that it is possible to acquire the new task (reaching 17.56% WER on $T_1$) while keeping performance on the base tasks $T_0$ stable. This result shown in Table 3 confirms that interleaving even a minimal amount of data from previous tasks acts as a powerful regularizer, anchoring the projector's weights and preventing the collapse of the shared multilingual embedding space.

Table 3: Comparison of Baseline (Meusli) vs. Sequential Sampling (First 1000) for Ukrainian acquisition. Metrics reported in WER (%).

| Exp | ES | DE | FR | IT | EN | DA | PL | CS | UK | Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| Meusli | 4.09 | 7.79 | 7.83 | 3.32 | 6.34 | 14.65 | 8.68 | 11.32 | 106.9 | **18.99** |
| first 1k | 4.82 | 9.54 | 7.96 | 5.94 | 5.20 | 18.67 | 10.95 | 15.21 | 17.56 | **10.65** |

### 4.2. Ablation Study: Rehearsal Buffer Sensitivity

To optimize computational efficiency and storage, we investigated the *minimum* rehearsal budget required to maintain stability. We tested buffer sizes ranging from 1000 down to a single sample per language. As shown in Figure 2 , the model demonstrates remarkable resilience:

- **High Stability**: As small as 5 samples per language are sufficient to preserve the performance of the original base tasks ($T_0$) after the acquisition of the first new task ($T_1$).
- **Random vs. Sequential Sampling**: We found that **Random Sampling** outperforms the "First $N$" sequential strategy used in the first Ukrainian CL experiment showed in section 4.1 when using a rehearsal buffer of 1000 or 500 samples per language. Randomization ensures a broader coverage of the acoustic and linguistic distribution of each language, which is vital for preventing drift in the projector and prevent overfitting on small amount of training data.
- **The Breakdown Point**: At only 1 sample per language, we observe the re-emergence of catastrophic forgetting in specific "fragile" languages such as Polish, Czech, Hungarian, and Estonian. In these cases, the model begins to hallucinate Ukrainian tokens when presented with base language audio, indicating that a single data point is insufficient for reinforcement.

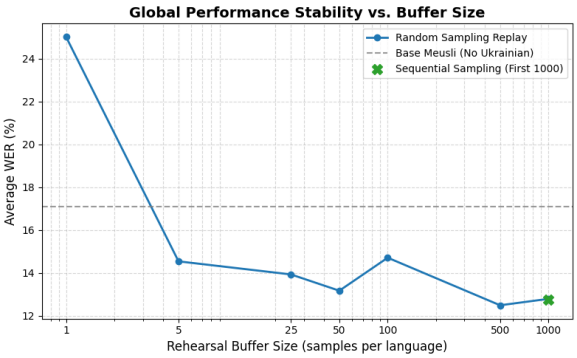


Figure 2: Ablation study of rehearsal buffer size on the global Average WER. The plot illustrates the stability plateau down to 5 samples per language and the collapse when the buffer is reduced to a single sample.

### 4.3. Sequential Language Acquisition

Building on the insights from the Ukrainian baseline, we extended the CL process to a multi-stage sequential integration of four new languages: Ukrainian (UK) $\rightarrow$ Japanese (JA) $\rightarrow$ Thai (TH) $\rightarrow$ Vietnamese (VI). This sequence was designed to test the model's ability to incorporate diverse scripts and distinct phonologies in a truly incremental fashion and with a challenging test using low resource languages very different from the ones natively supported by the base model. For this set of experiments, we employed the Ukrainian checkpoint obtained in section 4.1 with Common Voice data and then, for Japanese, Thai, and Vietnamese we opted to use the training set of the INTERSPEECH 2025 MLC-SLM challenge (14) to further diversify the train data domain. Figure 3 shows this pipeline. We evaluated two versions of this sequential loop to further probe the limits of memory retention introduced in section 4.2: one using a conservative buffer of 500 samples per language and a more aggressive setup with only 25 samples. A comparative analysis of these two experiments, shown in figure 4, reveals a significant trade-off between buffer size and long-term stability:

- **High-Fidelity Retention (500 Samples)**: With a 500-sample buffer, the model maintains strong retention of the original 28 European languages. For instance, high-resource languages such as Spanish (ES) remain stable, moving from a 4.09% baseline to 9.32% after four integration loops. However, we observe a steady performance degradation in

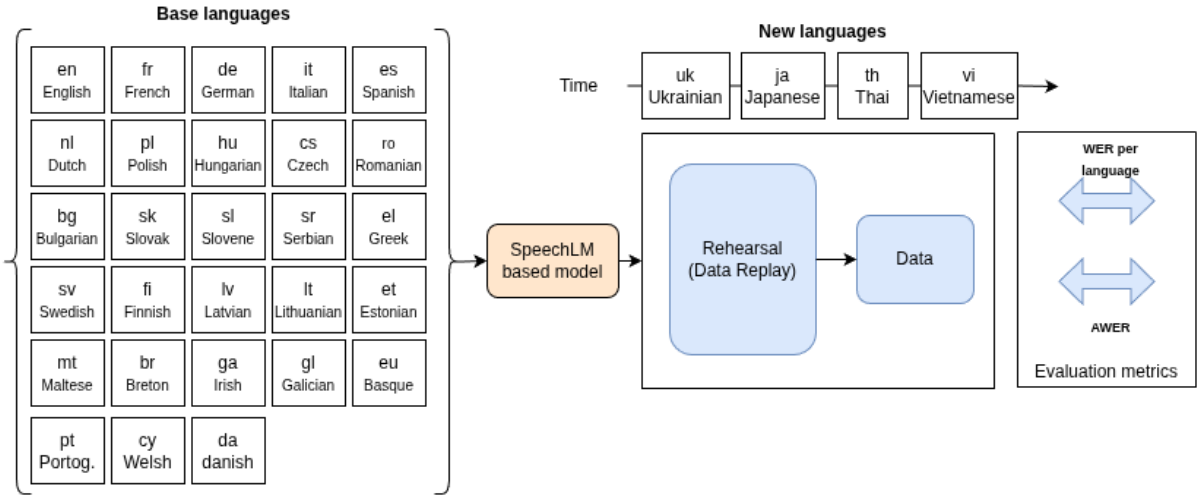


Figure 3: The proposed Continual Learning framework utilizing Data Rehearsal to mitigate catastrophic forgetting during the integration of new languages.

the newly acquired languages as the sequence progresses. While Ukrainian (UK) is initially learned at 17.10% WER, it drifts to 26.57% by the final loop. Similarly, Japanese (JA) worsens from an initial 27.28% to 40.45% following the addition of Thai and Vietnamese. This suggests that even a larger buffer provides only partial stability for newly integrated representations.

- **Accumulated Degradation (25 Samples):** In the aggressive 25-sample setup, we observe a "catastrophic drift" as the sequence progresses. The model reaches a breaking point during the Thai loop, where the **Avg WER** surges to **65.98%**. By the final stage, the AWER reaches 51.67%, more than double the error of the 500-sample experiment.

The divergence in stability is most evident when examining the **AWER** across all supported languages as shown in figure 4. In the 500-sample loop, the average error remains controlled, increasing from 15.74% at the base to 29.46% after the final loop. Conversely, the 25-sample loop experiences a "catastrophic drift," with the AWER nearly tripling to 56.06%. This global metric underscores that insufficient rehearsal does not just affect specific languages, but gradually destabilizes the entire multimodal alignment. These results, illustrated in our sequential evaluation logs, demonstrate that while Speech LLMs possess a high degree of "multilingual core" stability, a minimal rehearsal budget of 25 samples is insufficient to anchor representations over multiple sequential steps.

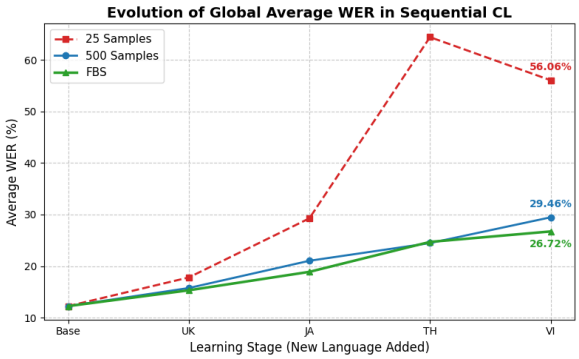


Figure 4: Evolution of Average WER across sequential learning loops (UK $\rightarrow$ JA $\rightarrow$ TH $\rightarrow$ VI). The 500-sample buffer (blue) maintains global stability, while the 25-sample buffer (red) exhibits a sharp catastrophic drift after the first loop, reaching a final Average WER of 51.67%. Finally the FBS based CL show the lowest AWER.

### 4.4. Fragility-Based Sampling

The varying levels of degradation observed in the uniform-buffer experiments indicate that not all languages require the same amount of rehearsal to prevent catastrophic forgetting. Motivated by this, we introduce a model agnostic **Fragility-Based Sampling (FBS)** heuristic. Instead of a uniform budget, FBS distributes the rehearsal samples based on the inherent stability (WER) of each language, keeping the overall system memory footprint constant.

For this experiment, we maintained the same total buffer capacity as the 500-sample uniform setup ($28$ languages $\times\ 500 = 14,000$ samples), since it is the buffer size that showed the best AWER. We categorized the base languages into three tiers and

Table 4: Sequential Continual Learning results: comparison between uniform rehearsal buffers (500 and 25 samples per language) and the Fragility-Based Sampling (FBS) strategy. Language codes: ES (Spanish), DE (German), FR (French), PT (Portuguese), EN (English), IT (Italian), PL (Polish), CS (Czech), DA (Danish), HU (Hungarian), BG (Bulgarian), RO (Romanian), ET (Estonian), LT (Lithuanian), EL (Greek), SL (Slovenian), FI (Finnish), UK (Ukrainian), JA (Japanese), TH (Thai), and VT (Vietnamese). The **AWER** column tracks the arithmetic mean across all active languages, illustrating the impact of different rehearsal strategies on global stability.

| Loop Stage | Buffer | ES | DE | FR | PT | EN | IT | PL | CS | DA | HU | BG | RO | ET | LT | EL | SL | FI | UK | JA | TH | VT | Avg WER |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Meusli (Base) | – | 4.09 | 7.79 | 7.83 | 4.86 | 6.34 | 3.32 | 8.68 | 11.32 | 14.65 | 16.87 | 15.20 | 9.65 | 19.93 | 24.30 | 18.35 | 19.41 | 15.29 | – | – | – | – | **12.23** |
| **+ Ukrainian** | 25 | 4.95 | 10.96 | 7.60 | 6.38 | 8.97 | 6.67 | 10.98 | 17.23 | 22.22 | 25.68 | 18.62 | 15.62 | 30.87 | 32.20 | 33.34 | 27.33 | 23.35 | 17.16 | – | – | – | **17.79** |
| | 500 | 4.56 | 8.54 | 7.86 | 7.09 | 7.20 | 6.64 | 11.21 | 14.21 | 19.97 | 22.22 | 17.47 | 18.50 | 24.75 | 30.80 | 23.33 | 21.90 | 19.97 | 17.10 | – | – | – | **15.74** |
| | FBS | 4.78 | 9.01 | 7.92 | 7.31 | 7.18 | 6.80 | 11.26 | 14.94 | 18.38 | 22.66 | 16.78 | 14.35 | 25.03 | 28.90 | 22.84 | 22.30 | 18.30 | 16.62 | – | – | – | **15.30** |
| **+ Japanese** | 25 | 15.91 | 13.23 | 15.27 | 14.99 | 15.67 | 11.53 | 18.69 | 25.56 | 28.53 | 43.03 | 30.26 | 29.85 | 56.14 | 57.52 | 41.33 | 43.11 | 36.92 | 28.48 | 29.90 | – | – | **29.26** |
| | 500 | 7.06 | 10.58 | 10.51 | 9.65 | 8.40 | 7.67 | 13.54 | 18.86 | 22.83 | 34.98 | 22.49 | 20.78 | 32.79 | 39.43 | 29.78 | 34.41 | 24.36 | 24.17 | 27.28 | – | – | **21.03** |
| | FBS | 6.68 | 9.36 | 10.04 | 9.22 | 8.97 | 7.44 | 14.04 | 17.62 | 19.73 | 26.72 | 21.32 | 16.07 | 29.23 | 36.96 | 27.31 | 27.97 | 23.22 | 22.89 | 23.87 | – | – | **18.88** |
| **+ Thai** | 25 | 34.42 | 36.10 | 44.56 | 25.71 | 49.28 | 34.53 | 40.86 | 58.61 | 63.48 | 122.11 | 67.50 | 60.84 | 98.33 | 92.43 | 92.75 | 86.67 | 42.20 | 40.57 | 278.0 | 19.45 | – | **64.42** |
| | 500 | 8.23 | 11.73 | 11.49 | 11.20 | 10.93 | 8.89 | 16.57 | 23.41 | 25.34 | 41.67 | 26.83 | 26.95 | 34.69 | 41.27 | 38.23 | 35.39 | 30.48 | 28.15 | 38.62 | 19.21 | – | **24.46** |
| | FBS | 8.25 | 12.80 | 13.19 | 10.24 | 12.52 | 8.72 | 16.68 | 23.04 | 25.51 | 36.16 | 33.58 | 26.11 | 37.18 | 44.91 | 37.85 | 37.71 | 29.70 | 24.53 | 35.35 | 19.86 | – | **24.69** |
| **+ Vietnamese** | 25 | 23.47 | 32.43 | 34.72 | 22.75 | 23.81 | 30.71 | 40.04 | 58.64 | 52.25 | 90.36 | 55.09 | 55.20 | 82.19 | 80.51 | 74.73 | 67.59 | 64.85 | 24.99 | 142.44 | 106.45 | 13.95 | **56.06** |
| (Final Loop) | 500 | 9.32 | 16.73 | 12.73 | 11.60 | 11.16 | 11.59 | 18.53 | 26.05 | 27.22 | 42.49 | 27.88 | 27.46 | 42.50 | 48.93 | 37.12 | 38.18 | 34.59 | 26.57 | 40.45 | 91.62 | 15.98 | **29.46** |
| | FBS | 9.44 | 19.04 | 13.42 | 13.59 | 11.78 | 14.14 | 18.55 | 23.60 | 28.14 | 38.36 | 30.49 | 22.37 | 39.54 | 46.97 | 36.61 | 35.18 | 31.51 | 26.45 | 47.47 | 39.44 | 15.01 | **26.72** |

allocated samples inversely proportional to their baseline robustness:

- **Tier A (Robust):** 10 high-performing languages (e.g., ES, DE, FR) allocated $250$ samples each ($2,500$ total).
- **Tier B (Intermediate):** 8 stable languages (e.g., PL, CS, UK) allocated the standard $500$ samples each ($4,000$ total).
- **Tier C (Fragile):** 10 vulnerable languages (e.g., ET, LT, EL) allocated high-density reinforcement of $750$ samples each ($7,500$ total).

As shown in Table 4, this reallocation strategy effectively balances the multilingual embedding space. By the final sequential loop (VT), fragile Tier C languages show improved retention compared to the uniform 500-sample setup; for instance, Estonian (ET) degrades to 39.54% (compared to 42.50% previously). Crucially, halving the buffer size for Tier A languages does not trigger catastrophic drift—Spanish (ES) finishes at a stable 9.44%, nearly identical to the 9.32% achieved with double the rehearsal data. Also the AWER, as shown in figure 4, results to be lower with respect to the uniform 500-sample setup. FBS achieves a significantly fairer and more equitable performance distribution across high- and low-resource languages.

## 5. Conclusions and Future Work

In this paper, we investigated the sustainable expansion of multilingual Speech LLMs through Continual Learning. Using the MEUSLI projector as a foundation, we demonstrated that while the SLAM-ASR paradigm is highly efficient for bootstrapping unseen languages like Ukrainian and Albanian, it is inherently vulnerable to catastrophic forgetting. Without intervention, fine-tuning on a single new language causes a near-total collapse of the model's original multilingual knowledge, often resulting in linguistic hallucinations. Our experiments with Data Rehearsal provide a clear roadmap for mitigating this collapse.

We established that stability can be maintained with surprisingly low rehearsal budgets—down to 5 samples per language—provided that random sampling is used rather than sequential selection. However, we also identified a "fragility threshold": when the budget is reduced to a single sample, performance on specific languages like Polish, Czech, and Estonian degrades significantly. Furthermore, our sequential integration loops (UK $\rightarrow$ JA $\rightarrow$ TH $\rightarrow$ VI) revealed that while a conservative buffer of 500 samples ensures high-fidelity retention across multiple stages, a minimal buffer of 25 samples leads to an accumulated "catastrophic drift," where the global Average WER nearly triples by the final stage. To address these non-uniform stability trade-offs, we introduced and evaluated a **Fragility-Based Sampling (FBS)** strategy. By categorizing languages into tiers, we optimized the rehearsal budget to distribute samples based on observed robustness rather than applying a uniform allocation. Our results demonstrate that this performance-aware reallocation significantly improves the retention of vulnerable linguistic representations without destabilizing robust languages. Crucially, FBS maintains a constant total memory budget while achieving a lower global Average WER compared to the uniform-buffer baseline, ensuring a much more equitable performance distribution across high- and low-resource languages. Future work will focus on automating this tiering process through dynamic sensitivity analysis during the training phase. Moreover, we will run experiments to better understand how the language diversity and order of these languages impact the final WER. By adaptively prioritizing vulnerable linguistic representations on the fly, we move closer to universal, resource-efficient Speech-LLM systems that can seamlessly grow their linguistic capabilities without sacrificing prior knowledge.

## 6. Acknowledgements

This paper was partially funded by the European Union's Horizon 2020 project ELOQUENCE (grant 101070558).

## 7. Bibliographical References

[1] C. Wenqian, Y. Dianzhi, J. Xiaoqi, M. Ziqiao, Z. Guangyan, W. Qichao, G. Yiwen, and K. Irwin, "Recent advances in speech language models: A survey," *arXiv:2410.03751*, 2024.

[2] P. Jing, W. Yucheng, F. Yangui, X. Yu, L. Xu, Z. Xizhuo, and Y. Kai, "A survey on speech large language models," *arXiv:2410.18908*, 2024.

[3] S. Fong, M. Matassoni, and A. Brutti, "Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages," in *Interspeech 2025*, 2025, pp. 2003–2007.

[4] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," *arXiv:1312.6211*, 2015.

[5] Z. Ma, G. Yang, Y. Yang, Z. Gao, J. Wang, Z. Du, F. Yu, Q. Chen, S. Zheng, S. Zhang *et al.*, "An embarrassingly simple approach for LLM with strong ASR capacity," *arXiv preprint arXiv:2402.08846*, 2024.

[6] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, "Experience replay for continual learning," in *Advances in Neural Information Processing Systems*, 2019.

[7] 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:2106.09685*, 2021.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv:2212.04356*, 2022.

[9] P. H. Martins, P. Fernandes, J. Alves, N. M. Guerreiro, R. Rei, D. M. Alves, J. Pombal, A. Farajian, M. Faysse, M. Klimaszewski, P. Colombo, B. Haddow, J. G. C. de Souza, A. Birch, and A. F. T. Martins, "Eurollm: Multilingual language models for europe," *arXiv:2409.16235*, 2024.

[10] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC 2020*, 2020, pp. 4211–4215.

[11] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," *arXiv preprint arXiv:2205.12446*, 2022. [Online]. Available: https://arxiv.org/abs/2205.12446

[12] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, 2021.

[13] L. D. Libera, P. Mousavi, S. Zaiem, C. Subakan, and M. Ravanelli, "Cl-masr: A continual learning benchmark for multilingual asr," *arXiv:2310.16931*, 2023.

[14] B. Mu, P. Guo, Z. Sun, S. Wang, H. Liu, M. Shao, L. Xie, E. S. Chng, L. Xiao, Q. Feng, and D. Wang, "Summary on the multilingual conversational speech language model challenge: Datasets, tasks, baselines, and methods," *arXiv preprint arXiv:2509.13785*, 2025.