

SpidR-Adapt: A Universal Speech Representation Model for Few-Shot Adaptation

Mahi Luthra^{*1}, Jiayi Shen^{*1}, Maxime Poli^{*2}, Angelo Ortiz^{1,2},
 Yosuke Higuchi¹, Youssef Benchekroun¹, Martin Gleize¹, Charles-Eric Saint-James¹,
 Dongyan Lin¹, Phillip Rust¹, Angel Villar¹, Surya Parimi¹, Vanessa Stark¹, Rashel Moritz¹,
 Juan Pino¹, Yann LeCun¹, Emmanuel Dupoux^{1,2}

^{*} Equal contribution, ¹ Meta AI, ² ENS-PSL, EHESS, CNRS

Correspondence: jiayishen@meta.com, mahiluthra@meta.com

Abstract

Human infants, with only a few hundred hours of speech exposure, acquire basic units of new languages, highlighting a striking efficiency gap compared to the data-hungry self-supervised speech models. To address this gap, this paper introduces **SpidR-Adapt** for rapid adaptation to new languages using minimal unlabeled data. We cast such low-resource speech representation learning as a meta-learning problem and construct a multi-task adaptive pre-training (MAdaPT) protocol which formulates the adaptation process as a bi-level optimization framework. To enable scalable meta-training under this framework, we propose a novel heuristic solution, first-order bi-level optimization (FOBLO), avoiding heavy computation costs. Finally, we stabilize meta-training by using a robust initialization through interleaved supervision which alternates self-supervised and supervised objectives. Empirically, SpidR-Adapt achieves rapid gains in phonemic discriminability (ABX) and spoken language modeling (sWUGGY, sBLIMP, tSC), improving over in-domain language models after training on less than 1h of target-language audio, over 100 \times more data-efficient than standard training. These findings highlight a practical, architecture-agnostic path toward biologically inspired, data-efficient representations. We open-source the training code and model checkpoints at <https://github.com/facebookresearch/spidr-adapt>.

1 Introduction

Human infants demonstrate a remarkable capacity for language acquisition: at under 6-months of age, they begin distinguishing phonemic contrasts and rapidly internalize the structure of their native language (Werker and Tees, 1984; Kuhl, 2004; Eimas et al., 1971), all from continuous auditory input and with only 100 to 500 hours of speech exposure (Bergelson et al., 2019; Cychosz et al., 2021).

In contrast, current self-supervised learning (SSL) models such as HuBERT (Hsu et al., 2021) and WavLM (Chen et al., 2022) require thousands of hours of training data to learn meaningful linguistic representations, and even then, their learned units are brittle—sensitive to acoustic and contextual variability (Gat et al., 2023; Hallap et al., 2023). When used as the basis for spoken language models (SLMs), these representations lead to limited language modeling performance compared to text-based systems (Hassid et al., 2023; Lakhotia et al., 2021) and far worse than the learning trajectories of human infants (Bergelson and Swingley, 2012).

A key reason for this discrepancy lies in inductive biases—infants begin with strong predispositions for speech perception, such as sensitivity to phones, rhythmic regularities, and speaker-invariance (Werner, 2007; Kuhl, 2004). These biases constrain learning to plausible linguistic structures, enabling rapid generalization from sparse input. By contrast, most machine learning systems are initialized from random weights and rely solely on statistical regularities of massive datasets. Without built-in inductive priors, they fail to discover linguistic abstractions of new languages efficiently.

To move toward the inductive efficiency of human learners, we propose a fast-adaptive self-supervised framework for speech representation learning including three broad components:

- **Multi-task Adaptive Pre-training (MAdaPT)**, a novel protocol that frames model learning as a bi-level optimization problem. The model is meta-optimized across several data-scarce adaptation episodes, each simulating a “lifetime” of low-resource language learning. Intuitively, this episodic design draws inspiration from evolutionary processes, with a second-order optimization occurring at an outer, population-like level that shapes the model’s inductive biases over generations. To further encourage cross-

lingual abstraction, we introduce controlled active forgetting between episodes, resetting key model components to simulate the onset of a new “lifetime,” thereby promoting robust, transferable representations.

- **First Order Bi-level Optimization (FOBLO)**, a meta-optimization heuristic that efficiently solves the second-order bi-level problem posed by MAdaPT. It trains the model to learn from unlabeled, under-resourced data in the inner-loop, with the outer-loop calibrating meta-parameters through feedback from a gold-standard labeled set.
- **Interleaved supervision**, which incorporates self-supervised training with occasional phoneme supervised steps, yielding an initialization that imitates human-robustness to contextual- and acoustic-variations of speech while being label-efficient.

Together, these mechanisms produce a model that achieves performance comparable to SSL systems trained on 6,000 hours of language data, despite seeing only 10 minutes to 100 hours of data in the target language. We further demonstrate that the resulting fast-adaptive model learns speech representations of an unseen language significantly faster than standard multi-task training.

We build on SpidR (Poli et al., 2025b), a speech SSL model that achieves state-of-the-art (SOTA) performance on phonemic discrimination and SLM metrics with efficient training. Our framework extends SpidR with the above fast-adaptive components, yielding **SpidR-Adapt**. Although our current implementation of MAdaPT-FOBLO uses SpidR as the backbone and focuses on speech representation, our framework is architecture-agnostic and broadly applicable to self-supervised models.

Our results demonstrate a step toward biologically inspired, data-efficient speech representation learning. Our paper makes three broad contributions: (1) Methodologically, we introduce MAdaPT, a general meta-training protocol that structures training as a series of episodes, each mirroring the low-resource language adaptation scenario.

The approach naturally formulates the adaptation process as a bi-level optimization problem. (2) Technically, we propose FOBLO, a novel heuristic solution to the bi-level optimization challenge formulated by MAdaPT. Additionally, we introduce interleaved supervision as a complementary

strategy to build stronger model initializations for meta-training. (3) Empirically, we conduct comprehensive experiments, including comparisons with alternative meta-learning heuristics (Reptile), demonstrating that the combination of MAdaPT and FOBLO consistently achieves superior performance, on par with in-domain language training.

2 Related Works

2.1 Self-supervised learning

Self-supervised learning (SSL) has enabled speech models to learn rich representations from unlabeled audio and now underpins a wide range of downstream applications—including ASR, emotion recognition, and spoken language modeling (SLM). Among these, SLM—where the objective is to capture linguistic structure directly from speech (Lakhotia et al., 2021; Dunbar et al., 2021; Borsos et al., 2022)—is particularly relevant for our work, given our motivation to build SSL models that enable human-like acquisition of spoken language. In the context of SLM, recent research has demonstrated that the semantic representativeness of learned units—especially their phonemic discriminability—directly impacts downstream spoken language performance (Poli et al., 2024; Hallap et al., 2023). Hence, in the current work, when evaluating the performance of speech SSL models, we employ measures of phonemic discriminability such as ABX (Schatz, 2016), PNMI (Hsu et al., 2021), and phoneme error rate.

Self-supervised models like HuBERT (Hsu et al., 2021) and WavLM (Chen et al., 2022) use masked prediction and clustering to build speech representations, but require extensive training time. SpidR (Poli et al., 2025b) improves on prior SSL models by combining self-distillation and online clustering, achieving SOTA SLM results with more efficient training. This efficiency makes SpidR an ideal backbone for current meta-learning approaches.

2.2 Meta-learning

Meta-learning aims to optimize models for rapid adaptation to new tasks, often in low-resource settings (Finn et al., 2017; Nichol et al., 2018). This is typically achieved by performing two loops of optimization—in the inner-loop, the model is repeatedly adapted to a new task and in the outer-loop, its meta-parameters are updated based on how well it adapts to that task. First-order model-agnostic meta-learning (FOMAML) and Reptile

(Nichol et al., 2018), in particular, use first-order outer-loop updates, making them computationally attractive heuristics for large-scale meta-learning.

Meta-learning has demonstrated significant effectiveness in improving out-of-domain (OoD) generalization. Recent studies have introduced risk-aware task selection frameworks that significantly improve adaptability and robustness without sacrificing training efficiency when facing distribution shifts (Wang et al., 2025; Qu et al., 2025) while others have proposed meta-learning for OoD detection and model selection (Qin et al., 2024). In this paper, we evaluate generalization capability by meta-testing on OoD languages that are not available during meta-training.

Recent work has also explored active forgetting as a complementary mechanism for improving model plasticity (Chen et al., 2023; Aggarwal et al., 2024). By periodically resetting parts of the model, such as embeddings or prediction layers, active forgetting encourages the formation of weights that can be reconfigured for new linguistic domains and prevents overfitting to unstable patterns. Here, we blend traditional meta-learning with active forgetting to amplify the adaptive benefits of both.

Despite their success in few-shot learning, meta-learning methods have seen limited application in speech models, where training typically relies on large, static corpora. Only a few studies explore meta-learning for speech classification or ASR (e.g., Chen et al., 2021; Hsu et al., 2020), and none target self-supervised speech representations. In contrast, we apply meta-learning at the level of SSL itself for the goal of spoken language modeling.

3 Methodology

Here we introduce **SpidR-Adapt**, a speech representation model tailored for rapid and robust adaptation to new languages with limited unlabeled audio data. First, we build a general multi-task training setup (**MAdaPT**; Sec. 3.1) that imitates fast-adaptation to new languages in low-resource scenarios, incorporating active forgetting to encourage stronger cross-lingual abstraction. This approach builds the adaptation process as a bi-level optimization problem. Then, to efficiently solve the nontrivial bi-level problem, we introduce an empirical solution called first-order bi-level optimization (**FOBLO**; Sec. 3.2), which avoids the heavy computational cost of second-order gradient steps in the outer-loop. Finally, to stabilize meta-

optimization, we propose initializing with a pre-trained model and design an interleaved supervised objective (**interleaved supervision**; Sec. 3.3).

3.1 Multi-task Adaptive Pre-Training (**MAdaPT**)

The goal of MAdaPT is to addresses the OoD generalization challenge: the model is pre-trained on source (seen) linguistic domains with sufficient data and subsequently adapted on target (new) linguistic domains for which only limited unlabeled data is available.

Notation. Let \mathcal{S} denote the set of source languages available during training and \mathcal{T} represent the set of unseen target languages encountered during adaptation. For each source language $\ell \in \mathcal{S}$, we assume access to a sufficiently large unlabeled corpus \mathcal{D}_ℓ^u and, optionally, a small labeled corpus \mathcal{D}_ℓ^s . In contrast, for each target language in \mathcal{T} , only a limited unlabeled corpus is available.

Episodic multi-lingual setup. We cast the OoD challenge from seen to new languages as a meta-learning problem. To simulate fast adaptation to target languages with limited speech data, we partition the large unlabeled corpus \mathcal{D}_ℓ^u of each source language into multiple smaller data chunks $\{\mathbf{D}_\ell^u\}$. Thus, one task in this work corresponds to a specific language ℓ and one scarce data chunk \mathbf{D}_ℓ^u as the training set. During meta-training, the model is presented with a mini-batch of task-specific episodes and is optimized in the outer-loop based on learning performance of the inner-loops. At the meta-test stage, we fine-tune the learned model on data-scarce tasks derived from each target language, evaluating adaptation in low-resource scenarios.

SpidR as backbone speech model. In this work, we deploy the SOTA speech representation model SpidR (Poli et al., 2025b) as our backbone, which has a student-teacher architecture. Thus, we represent the speech model in detail: $\boldsymbol{\theta} = \{f(\cdot), E_s, E_t, \{\mathbf{W}^k\}, \{\mathbf{C}^k\}\}$, where $f(\cdot)$ is a convolutional downampler and E_s, E_t are Transformer encoders for the student and teacher, respectively. The teacher is an exponential moving average (EMA) of the student. \mathbf{W}^k is the prediction head of the student and \mathbf{C}^k is the target codebook of the teacher at the intermediate layer k (where $L - K \leq k \leq L$).

Given a language ℓ with its low-resource dataset \mathbf{D}_ℓ^u , we formalize the adaptation process as:

$$\boldsymbol{\theta}_\ell^* = \arg \min_{\boldsymbol{\theta}} \mathcal{L}_{\text{ssl}}(\boldsymbol{\theta}; \mathbf{D}_\ell^u), \quad (1)$$

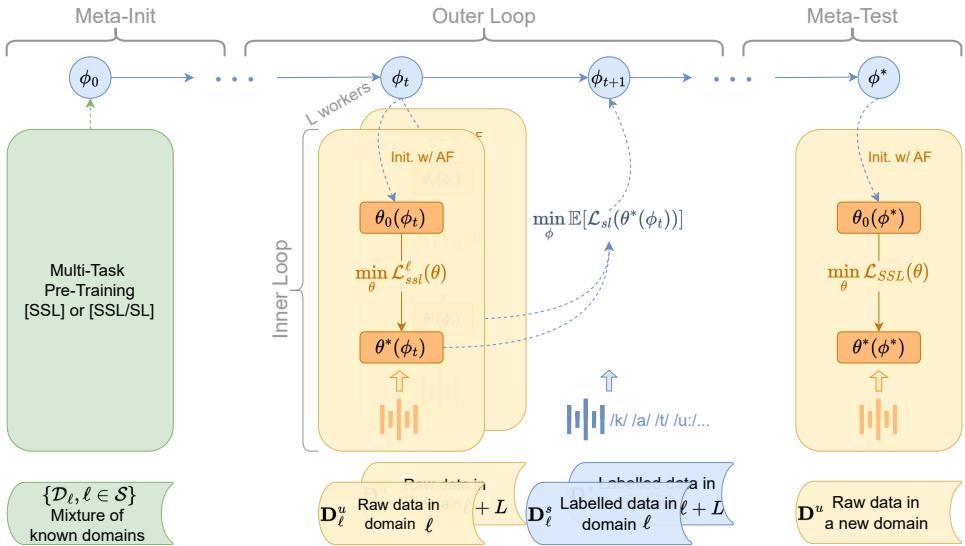


Figure 1: Overview of SpidR-Adapt for few-shot speech adaptation. It consists of three main phases: (1) meta-initialization performs multi-task pre-training with interleaved supervision, learning a robust initialization ϕ_0 from a mixture of source domains. (2) meta-training through MAdaPT-FOBLO optimizes this initialization for fast adaption to \mathcal{D}_ℓ . Each worker conducts inner-loop adaptation with active forgetting (AF) on unlabeled data, followed by outer-loop updates that refines ϕ by minimizing the expected task loss on labeled data. (3) at meta-test time, the learned ϕ^* is fast adapted to a new, unseen domain using only unlabelled data.

where \mathcal{L}_{ssl} denotes a self-supervised loss function, θ represents all learnable parameters of the speech model SpidR, and θ_ℓ^* are the optimal model parameters specific to the language ℓ . We note that \mathbf{D}_ℓ^u is not sufficient to train a specific speech model from scratch due to severe overfitting ([Dupoux, 2018](#)).

Bi-level optimization. To mitigate model’s overfitting to source languages, we propose a generic bi-level optimization framework which aims to learn meta-parameters from source languages that adapt rapidly to target languages. Within this framework, training with pure SSL in Equation (1) serves as an inner-optimization; meanwhile, light-weight labeled data are deployed to supervise these adaptation processes in the outer-level by shaping meta-parameters. Meta-parameters are shared across concurrent tasks and can be intuitively viewed as inductive biases for speech representation learning.

For clarity, we instantiate meta-parameters ϕ as the initial parameters of the backbone model in Equation (1). Thus, the expected bi-level objective for **MAdaPT** is:

$$\begin{aligned} & \min_{\phi} \mathbb{E}_{\ell \sim S} [\mathcal{L}_{sl}(\theta_\ell^*(\phi); \mathbf{D}_\ell^s)], \\ & \text{s.t. } \theta_\ell^*(\phi) = \arg \min_{\theta} \mathcal{L}_{ssl}(\theta, \phi; \mathbf{D}_\ell^u), \end{aligned} \quad (2)$$

where \mathcal{L}_{ssl} denotes a self-supervised loss function in the inner level, performing adaptation from un-

labeled speech data; and \mathcal{L}_{sl} denotes a supervised loss function in the outer level. In contrast to regular meta-learning frameworks designed for supervised learning (Finn et al., 2017), supervised information here is only used in the outer optimization while inner adaptations remain unsupervised. This preserves the assumption of low-resource, unlabeled data usage within the inner-loop, while leveraging supervised information in the outer-loop to resolve the ambiguities of pure self-supervision.

Active forgetting in task adaptation. To suppress unstable and language-specific learning from past episodes, we introduce an active forgetting mechanism. During meta-training, SpidR’s prediction heads and codebooks tend to be dominated by phonemic knowledge from source languages, hindering its generalization over new languages.

To this end, we reinitialize these components at the start of each inner loop. Concretely, we copy the student and teacher parameters from the shared meta-parameters ϕ as default but reset all heads and codebooks, yielding the optimization with initialization $\theta_{AF}(\phi)$ for each inner loop at both meta-training and meta-test stages:

$\min_{\theta} \mathcal{L}_{ssl}(\theta_{AF}(\phi); \mathbf{D}_\ell^u)$, where

$$\boldsymbol{\theta}_{AF}(\phi) = \{f(\phi), E_s(\phi), E_t(\phi), \{\mathbf{W}_0^k\}, \{\mathbf{C}_0^k\}\}. \quad (3)$$

Here each codebook \mathbf{C}_0^k is sampled from a normal distribution $\mathcal{N}(0, 1)$ and each head \mathbf{W}_0^k is warmed up for 20 steps using the first batch of \mathbf{D}_ℓ^u .

3.2 First-Order Bi-Level Optimization (FOBLO)

Solving the bi-level optimization in Equation (2) is non-trivial because both the inner- and outer-loops require multiple gradient steps. To make meta-training scalable, we introduce a first-order bi-level optimizer that yields a principled first-order approximation to the meta-gradient. In contrast to other first-order approximations (Finn et al., 2017; Nichol et al., 2018), our optimizer is intended for a more challenging case where the inner- and outer-loops are served by different loss functions.

Given a specific language ℓ , the updating of meta-parameters ϕ can be formulated as:

$$\phi \leftarrow \phi - \beta \nabla_\phi \mathcal{L}_{sl}(\theta_\ell^*(\phi); \mathbf{D}_\ell^s), \quad (4)$$

where β is a learning rate in the outer-loop to update the meta-parameter ϕ . Assume that the inner- and outer-loops perform M and N gradient steps, respectively. By applying chain rule to Equation (4) during backpropagation over M inner steps, we can reformulate the meta-update as:

$$\begin{aligned} \phi &\leftarrow \phi - \beta \nabla_\phi \mathcal{L}_{sl}(\theta_\ell^M(\phi); \mathbf{D}_\ell^s) \\ &\cdot \prod_{m=1}^M \left[\mathbf{I} - \alpha \nabla_{\phi_\ell^{m-1}} (\nabla_\phi \mathcal{L}_{ssl}(\theta_\ell^{m-1}(\phi))) \right], \end{aligned} \quad (5)$$

where α is the learning rate in the inner-loop update and the task-specific parameter θ_ℓ^m denotes the model's parameters after the m^{th} -inner step. To avoid the heavy computational cost in computing the Jacobian product of the second derivative in Equation (5), we adopt a first-order approximation by dropping the second-order term (i.e., we stop the gradient through the inner-loop).

The outer-loop typically performs N supervised steps on labeled speech corpora \mathbf{D}_ℓ^s . Following Reptile (Nichol et al., 2018), we approximate the outer-loop gradient by the parameter difference between the end of the inner-loop and the end of the outer-loop:

$$\nabla_\phi \mathcal{L}_{sl}(\theta_\ell^M(\phi); \mathbf{D}_\ell^s) = \theta_\ell^M - \theta_\ell^{M+N}, \quad (6)$$

where θ_ℓ^M is obtained after M self-supervised inner-steps starting from θ and θ_ℓ^{M+N} is obtained

by taking an additional N supervised steps from θ_ℓ^M .

By substituting Equation (6) into Equation (5), FOBLO updates the meta-parameters as follows:

$$\phi \leftarrow \phi - \beta \mathbb{E}_{\ell \sim \mathcal{S}} [\theta_\ell^M - \theta_\ell^{M+N}]. \quad (7)$$

Illustration of our work is provided in Figure 1. This work provides a principled and practical solution for few-shot self-supervised adaptation by nesting self-supervised inner-loops within supervised outer-loops.

3.3 Interleaved Supervision

In practice, we find that initializing the meta-parameters from random weights leads to unstable learning dynamics and poor convergence (see Appendix D.2). Thus, to facilitate effective bi-level optimization, it is necessary to perform a dedicated pre-training phase prior to the meta-training stage.

To this end, we introduce an interleaved pre-training objective to obtain the most performatve meta-initialization, denoted as ϕ_0 . During the dedicated pre-training phase, we alternate between self-supervised and supervised objectives in an interleaved manner. This mechanism leverages both unlabeled and labeled data, allowing the model to benefit from large-scale unsupervised corpora while grounding representations with supervised signals. This pre-training objective is defined as:

$$\arg \min_{\phi_0} \lambda \mathcal{L}_{ssl}(\phi_0; \{\mathcal{D}_\ell^u\}) + (1 - \lambda) \mathcal{L}_{sl}(\phi_0; \{\mathcal{D}_\ell^s\}), \quad (8)$$

where $\lambda \in \{0, 1\}$ is a binary hyperparameter. Here, \mathcal{L}_{ssl} denotes the self-supervised loss, applied to the union set of unlabeled corpora from all source languages, $\{\mathcal{D}_\ell^u, \ell \sim \mathcal{S}\}$; while \mathcal{L}_{sl} is the supervised loss, applied to the union of labeled corpora $\{\mathcal{D}_\ell^s, \ell \sim \mathcal{S}\}$.

In the current work, we utilize two distinct meta-initializations: 1) Multi-Task-PT [SSL]: setting λ to 1 throughout pre-training, yielding standard self-supervision. 2) Multi-Task-PT [SSL/SL] switching λ to 0 periodically, interleaving occasional supervised steps into the self-supervised training regime. This provides a more superior initialization for meta-training.

4 Experiments

In the experimental section, we seek to address the following key questions: (1) How data-efficient is

SpidR-Adapt in generalizing to the linguistic structure of new languages? (2) Can the MAdaPT framework produce improvements when labeled data is unavailable during pre-training? (3) Can SpidR-Adapt produce improvements in downstream spoken language modeling? (4) Can SpidR-Adapt achieve superior performance compared to existing speech models under the OoD setup?

Datasets. We collect data from 27 languages to evaluate adaptation capabilities of speech encoders under in-domain (ID) and out-of-domain (OoD) setups. We partition the languages as follows: 19 source languages for training; 5 target languages for development; and 3 target languages for testing. Importantly, there are no overlaps between source and target languages. Each source language is supported by a substantial unlabeled corpus (300 hours per language) collected from VoxPopuli (Wang et al., 2021) and a small phoneme-aligned corpus (maximum 50 hours per language) collected from VoxCommunis Corpus (Ahn and Chodroff, 2022) to serve as labels for the FOBLO outer-loop and for interleaved supervision.

Only small-scale unlabeled corpora are available for fast adaptation to target languages (mimicking infant learning settings). We construct four subsets per target language with durations 10 minutes, 1 hour, 10 hours, and 100 hours. To quantify the performance gap between ID and OoD training, we additionally collect large-scale in-domain training corpora from VoxPopuli for each test language. Each in-domain corpus comprises 6k hours—comparable in scale to the combined duration of the OoD corpora. The small-scale adaptation sets for these test languages are sampled from the same in-domain training pool; consequently, the OoD models are adapted using subsets of ID data. These choices are made to enable fair comparisons between ID and OoD models.

Small-scale adaptation corpora were also created for the meta-development languages, sourced from CommonVoice (Ardila et al., 2020) and used for model development. Further details on dataset construction are provided in Appendix A.

Training Setup. We perform multi-task pretraining of SpidR with self- or interleaved-supervised objectives (interleaving supervision every 10 steps; see Sec. 3.3). These models serve as initializations for meta-training wherein we train across 800 episodes, each episode consisting 1800 inner- and 200 outer-steps. In each inner-loop, the model is trained on a random 10 hour data chunk of a ran-

dom source language. Training is performed across 16 GPUs in a distributed fashion. Details regarding training can be found in Appendix B.

4.1 Data-Efficiency When Adapting on New Languages

To evaluate data efficiency, we adapt meta-trained models to new target languages using only limited unlabeled data. We benchmark our approach against baselines using ABX (lower is better), computed using the fastabx toolkit (Poli et al., 2025a).

ABX scores quantify how well model embeddings capture phone distinctions and correlate strongly with downstream SLM performance (Poli et al., 2025b), serving as an efficient zero-shot proxy. In the ABX task, embeddings are computed for three triphones: A , B , and X . Here, A and X are instances of the same triphone, while B differs in its central phone (e.g., /bag/ vs. /beg/). The model succeeds if X is closer to A than to B in embedding space. The within-speaker condition uses triphones from the same speaker, while the across-speaker condition uses A and B from one speaker and X from another, making the task more challenging.

Figure 2 shows results: the x-axis indicates adaptation data size, and the y-axis shows average ABX scores across our three test languages and across within- and across-speaker ABX conditions (individual trends are consistent). With SpidR as backbone and under two meta-initialization strategies (Sec. 3.3) self- and interleaved-supervised initialization, we compare: 1) **In-Domain Mono-Task-PT**: Standard in-domain pre-training with sufficient data from the target language. Because every small-scale evaluation subset is drawn from the ID training pool, we do not perform additional small-scale adaptation, so In-Domain PT appears as a horizontal line in Figure 2. 2) **Multi-Task-PT**: Standard OoD pre-training with ample unlabeled data from all source languages, using the same data-feeding protocol as In-Domain PT. 3) **MAdaPT-FOBLO**: Our proposed approach that combines MAdaPT with its trivial solution FOBLO. Used with SpidR as backbone and interleaved-supervised initialization, this method formulates our few-shot learning speech encoder, **SpidR-Adapt**.

Figure 2 (a) shows that Multi-Task-PT underperforms In-Domain PT especially when the adaptation budget is small (< 100 hours). This suggests that regular multi-task pre-training lacks the adaptation capacity needed for unseen targets, and simply mixing several source languages during pre-

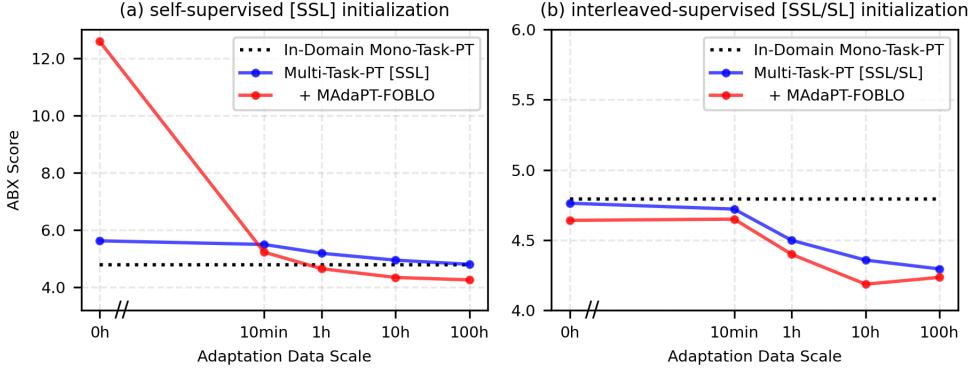


Figure 2: **Data-efficiency of SpidR-Adapt on new languages across different adaptation data scales.** We report ABX scores (lower is better) averaged across three test languages (French, German, English) for two initialization strategies (a) self-supervision [SSL] and (b) interleaved-supervision [SSL/SL]. Each sub-figure compares our approach with the baselines: In-Domain Mono-Task-PT, the oracle method pertained on 6k hours of in-domain data and **Multi-Task-PT**, standard multi-task pretraining using [SSL] or [SSL/SL] regimes. By integrating the proposed solution, **MAdaPT-FOBLO**, with Multi-Task-PT as meta-initialization, we achieve highly efficient adaptation to new languages. For detailed results, refer to Appendix C.1.

training does not guarantee better generalization.

In contrast, MAdaPT-FOBLO improves rapidly across data scales, indicating high effectiveness for adapting to OoD data. Notably, MAdaPT-FOBLO reaches parity with In-Domain PT after adapting with just 1 hour of unlabeled target-language audio, highlighting data efficiency improvements of 100×. Such efficiency is crucial for real-world scenarios where language corpora are scarce.

Finally, Figure 2 (b) indicates that the interleaved-supervised initialization (Multi-Task-PT [SSL/SL]) provides a better starting point (lower initial ABX) than self-supervised initialization (Multi-Task-PT [SSL]). However, regardless of initialization, the incorporation of MAdaPT-FOBLO delivers the largest gains in rapid adaptation to unseen languages. This suggests that while initialization can set a stronger baseline, the adaptation strategy is the primary driver of sustained performance improvements. Full results are provided in Appendix C.1 for detailed comparison.

4.2 MAdaPT for Pure Self-Supervision

Here we consider an extreme setting in which no supervised training data is available for source languages. In this regime, MAdaPT must be optimized using a purely self-supervised procedure.

To instantiate MAdaPT without labels, we adopt Reptile (Nichol et al., 2018), a first-order meta-learning heuristic that approximates the meta-gradient assuming identical inner- and outer-loop objectives. Here, the meta-update is written as $\phi \leftarrow (1 - \beta)\phi - \beta \mathbb{E}_{\ell \sim \mathcal{S}}[\theta_{\ell}^{M+N}]$, where β trades

Method	Avg. ABX (w/o 0h) ↓	
	Within-Speaker	Across-Speaker
Multi-Task-PT [SSL]	4.33	5.89
+ MAdaPT-Reptile	<u>4.19</u>	<u>5.59</u>
+ MAdaPT-FOBLO	4.01	5.24

Table 1: **Comparisons with MAdaPT-Reptile**, a purely SSL solution for MAdaPT. ABX scores averaged across 10 minutes to 100 hours training (excluding zero-shot, 0h). Although Reptile under-performs FOBLO, it achieves better results than baseline Multi-Task-PT, demonstrating the effectiveness of MAdaPT. The best scores are in **bold** and second best are underlined.

off the previous meta-parameters and the task-specific solution after each episode. In contrast to our FOBLO update in Equation (7), θ_{ℓ}^{M+N} in Reptile denotes parameters obtained by pure self-supervised training for a total of $M+N$ steps.

In Table 1, we report ABX (in %) averaged over adaptation budgets from 10 minutes to 100 hours and three test languages. The results emphasize that MAdaPT-based optimization consistently improves over standard Multi-Task-PT, with FOBLO (which requires supervised labels) achieving the strongest performance. Notably, when all source languages lack supervision, Reptile that is used as a purely self-supervised instantiation of MAdaPT, outperforms baseline Multi-Task-PT. These findings underscore the importance of a tailored multi-task framework for low-resource OoD adaptation.

Method	0h	10m	1h	10h	100h	Avg. (w/o 0h) ↑
In-Domain Mono-Task-PT	6000 hours unlabeled data training; Oracle score is 61.85.					
Multi-Task-PT [SSL]	60.82	60.38	60.79	61.47	61.64	61.07
+ MAdaPT-Reptile	60.80	60.99	61.16	61.58	61.70	61.36
+ MAdaPT-FOBLO	56.92	62.19	62.12	62.41	63.60	62.58
Multi-Task-PT [SSL/SL]	61.36	62.13	<u>62.43</u>	62.68	62.70	62.49
+ MAdaPT-Reptile	62.38	<u>62.72</u>	61.89	<u>62.97</u>	<u>62.83</u>	<u>62.60</u>
+ MAdaPT-FOBLO	<u>61.65</u>	62.99	62.65	63.17	62.74	62.89

Table 2: **Spoken language modeling results (in %) of the English-adapted models.** We report the average SLM metrics, including sWUGGY, sBLIMP, and tSC (higher is better). Across both meta-initializations, FOBLO consistently outperforms Reptile and the oracle. The best results are shown in **bold**, and second-best are underlined. For detailed results, refer to Appendix C.2.

Method	Backbone Model	PNMI ↑	PER ↓	ABX ↓ Within-Speaker	ABX ↓ Across-Speaker
Multi-Task-PT [SSL]	HuBERT	0.58	76.01	6.62	7.77
Multi-Task-PT [SSL]	SpidR	0.66	60.17	4.83	5.72
MAdaPT-Reptile	SpidR	<u>0.69</u>	<u>38.27</u>	<u>4.12</u>	<u>4.57</u>
MAdaPT-FOBLO	SpidR	<u>0.71</u>	37.70	4.09	4.55

Table 3: **Comparisons on the Phoneme Discovery Benchmark.** MAdaPT-FOBLO outperforms alternate speech SSL model (HuBERT) and alternate meta-learning framework (Reptile). Here, MAdaPT-FOBLO and MAdaPT-Reptile are initialized using Multi-Task-PT [SSL/SL]. The best scores are in **bold** and second best are underlined.

4.3 Evaluating Downstream Spoken Language Models

We evaluate SLM performance of SSL models adapted on English test sets, using three complementary linguistic metrics. 1) **Lexical (sWUGGY)** (Nguyen et al., 2020) tests whether the model assigns higher probability to true words than to matched non-words. 2) **Syntax (sBLIMP)** requires the model to choose grammatical sentences from minimal pairs. 3) **Discourse/Narrative (Spoken Topic StoryCloze)** (Mostafazadeh et al., 2017) asks the model to select appropriate continuations for short stories. We report accuracy (in %) averaged across the three metrics in Table 2. Detailed per-task results are included in the Appendix C.2).

Table 2 shows that MAdaPT-FOBLO achieves rapid gains under the few-shot adaptation scenario (for both self- and interleaved-supervised initializations). MAdaPT-Reptile comes a close second, with especially strong zero-shot performance.

4.4 Comparisons on the Phoneme Discovery Benchmark

To further investigate the adaptability of the proposed methods, we compare them with a performant speech SSL model, HuBERT, trained under

the OoD Mutli-Task-PT setup using the Phoneme Discovery Benchmark (Poli et al., 2026). The metrics reported include: 1) **phone-normalized mutual information (PNMI)**, the uncertainty eliminated about a phone label by a predicted unit; 2) **phoneme error rate (PER)**, after mapping units to the most frequently associated phoneme 3) **ABX** (within- and across-conditions). Results reported in Table 3 broadly indicate superior performance of SpidR-Adapt over alternate speech SSL model (HuBERT) and alternate meta-learning framework (Reptile). Full results are reported in Appendix C.3.

5 Conclusion

We present **SpidR-Adapt**, a speech representation model that enables data-efficient adaptation to new languages by combining meta-adaptive pretraining, bi-level optimization, and interleaved supervision. Achieving superior performance with as little as 1 hour of target-language audio—100× less data than traditional mono- and multi-task methods—SpidR-Adapt demonstrates effectiveness of a tailored meta-learning framework for flexible representation learning in low-resource settings.

Limitations

This work offers promising data-efficiency in few-shot speech representation learning, but several limitations remain. Model performance is influenced by the choice of meta-initialization, suggesting that further research is needed into more robust meta-learning that can be trained without meta-initialization. Supervised information from source languages is still required at the outer-level, which limits scaling of source languages. Additionally, training of spoken language models has not been included into the meta-learning framework and hence is not data-efficient; future work could focus on applying meta-learning directly to SLM training to enhance efficiency and reduce data requirements.

Acknowledgements

ED in his EHESS role was supported in part by the Agence Nationale pour la Recherche (ANR-17-EURE0017 Frontcog, ANR10-IDEX-0001-02 PSL*) and an ERC grant (InfantSimulator).

References

- Divyanshu Aggarwal, Ashutosh Sathe, and Sunayana Sitaram. 2024. Exploring pretraining via active forgetting for improving cross lingual transfer for decoder language models. *arXiv preprint arXiv:2410.16168*.
- Emily Ahn and Eleanor Chodroff. 2022. VoxCommunis: A corpus for cross-linguistic phonetic analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5286–5294.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Can Balioglu, Martin Gleize, Artyom Kozhevnikov, Ilia Kulikov, Tuan Tran, and Julien Yao. 2023. fairseq2.
- Elika Bergelson, Andrei Amatuni, Shannon Dailey, Sharath Koorathota, and Shaelise Tor. 2019. Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science*, 22(1):e12715.
- Elika Bergelson and Daniel Swingley. 2012. At 6–9 months, human infants know the meanings of many common words. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258.
- Zalan Borsos, Matt Sharifi, Damien Vincent, Olivier Pietquin, Antoine Caillon, David Grangier, Andriy Zabolotskiy, Neil Zeghidour, and Marco Tagliasacchi. 2022. Audioltm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Ifeoluwa Adelani, Pontus Stenetorp, Sebastian Riedel, and Mikel Artetxe. 2023. Improving language plasticity via pretraining with active forgetting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 31543–31557.
- Yu Chen, Tom Ko, and Jian Wang. 2021. A meta-learning approach for user-defined spoken term classification with varying classes and examples. In *Proceedings of Interspeech 2021*, pages 4224–4228.
- Margaret Cychosz, Anele Villanueva, and Adriana Weisleder. 2021. Efficient estimation of children’s language exposure in two bilingual communities.

- Journal of Speech, Language, and Hearing Research*, 64(10):3843–3866.
- Ewan Dunbar, Mathieu Bernard, Nicolas Hamilakis, Tu Anh Nguyen, Maureen De Seyssel, Patricia Rozé, Morgane Rivière, Eugene Kharitonov, and Emmanuel Dupoux. 2021. **The Zero Resource Speech Challenge 2021: Spoken Language Modelling**. In *Interspeech 2021*, pages 1574–1578.
- Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330. IEEE.
- Emmanuel Dupoux. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.
- Peter D. Eimas, Eugene R. Siqueland, Peter Jusczyk, and James Vigorito. 1971. **Speech perception in infants**. *Science*, 171(3968):303–306.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.
- Itai Gat, Felix Kreuk, Tu Anh Nguyen, Ann Lee, Jade Copet, Gabriel Synnaeve, Emmanuel Dupoux, and Yossi Adi. 2023. **Augmentation invariant discrete representation for generative spoken language modeling**. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 465–477.
- Mark Hallap, Emmanuel Dupoux, and Ewan Dunbar. 2023. **Evaluating context-invariance in unsupervised speech representations**. In *INTERSPEECH 2023*, pages 2973–2977.
- Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, and 1 others. 2023. **Textually pretrained speech language models**. *Advances in Neural Information Processing Systems*, 36:63483–63501.
- Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee. 2020. **Meta-learning for end-to-end low-resource speech recognition**. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7844–7848.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhota, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. **HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units**. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.
- J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. **Libri-light: A benchmark for ASR with limited or no supervision**. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673.
- Patricia K. Kuhl. 2004. **Early language acquisition: cracking the speech code**. *Nature Reviews Neuroscience*, 5:831–843.
- Kushal Lakhota, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. **On Generative Spoken Language Modeling from Raw Audio**. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.
- Michael McAuliffe, Matthew Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldì. In *Proceedings of Interspeech 2017*, pages 498–502.
- Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. **LS-DSem 2017 shared task: The story cloze test**. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.
- Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. **The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling**. *Preprint*, arXiv:2011.11588.
- Alex Nichol, Joshua Achiam, and John Schulman. 2018. **On first-order meta-learning algorithms**. *Preprint*, arXiv:1803.02999.
- Angelo Ortiz Tandazo, Manel Khentout, Youssef Benchekroun, Thomas Hueber, and Emmanuel Dupoux. 2025. **MauBERT: Universal phonetic inductive biases for few-shot acoustic units discovery**. *Preprint*, arXiv:2512.19612.
- Maxime Poli, Emmanuel Chemla, and Emmanuel Dupoux. 2024. **Improving spoken language modeling with phoneme classification: A simple fine-tuning approach**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5284–5292.
- Maxime Poli, Emmanuel Chemla, and Emmanuel Dupoux. 2025a. **fastabx: A library for efficient computation of abx discriminability**. *arXiv preprint arXiv:2505.02692*.
- Maxime Poli, Manel Khentout, Angelo Ortiz Tandazo, Ewan Dunbar, and Emmanuel Dupoux. 2026. **The phoneme discovery benchmark**. Under review.

Maxime Poli, Mahi Luthra, Youssef Benchekroun, Yosuke Higuchi, Martin Gleize, Jiayi Shen, Robin Algayres, Yu-An Chung, Mido Assran, Juan Pino, and Emmanuel Dupoux. 2025b. *Spidr: Learning fast and stable linguistic units for spoken language models without supervision*. *Transactions on Machine Learning Research*. Accepted, in press.

Yuehan Qin, Yichi Zhang, Yi Nian, Xueying Ding, and Yue Zhao. 2024. Metaood: Automatic selection of ood detection models. *arXiv preprint arXiv:2410.03074*.

Yun Qu, Cheems Wang, Yixiu Mao, Yiqin Lv, and Xiangyang Ji. 2025. Fast and robust: Task sampling with posterior and diversity synergies for adaptive decision-makers in randomized environments. In *Forty-second International Conference on Machine Learning*.

Thomas Schatz. 2016. *ABX-discriminability measures and applications*. Ph.D. thesis, Université Paris 6 (UPMC).

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Qi Wang, Zehao Xiao, Yixiu Mao, Yun Qu, Jiayi Shen, Yiqin Lv, and Xiangyang Ji. 2025. Model predictive task sampling for efficient and robust adaptation. *arXiv preprint arXiv:2501.11039*.

Janet F. Werker and Richard C. Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development*, 7:49–63.

Lynne A. Werner. 2007. Infant hearing and perceptual development. In Robert F. Burkard, Manuel Don, and Jos J. Eggermont, editors, *Auditory Evoked Potentials: Basic Principles and Clinical Application*, pages 509–528. Lippincott Williams & Wilkins, Philadelphia, PA.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher De-wan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Appendix

A Details of Datasets

Table 4 summarizes the datasets used for meta-training, meta-development, and meta-testing. To prepare the unlabeled training data, we apply the Silero Voice Activity Detector to the audio files, segmenting them into smaller audio files ranging from 0.5 to 30 seconds in duration (with mean 14.6 seconds). This pre-processing step ensures that the model is exposed to realistic, variable-length speech segments during both training and evaluation. For reproducibility, the start and end timestamp metadata for all processed audio files used in training and evaluation will be made available in the accompanying GitHub codebase.

In addition to the unlabeled dataset, we also use a small supervised dataset, mainly sourced from VoxCommunis Corpus (Ahn and Chodroff, 2022). This corpus comprises phoneme alignments inferred on CommonVoice (Ardila et al., 2020) data using Montreal Forced Aligners (MFA; McAuliffe et al., 2017). While CommonVoice has data for 18 training languages, it does not contain data for one language—Croatian. To obtain a labeled set in Croatian, we use the transcribed set of Voxpopuli (Wang et al., 2021) and align phonemes using off-the-shelf MFA models. We clean the alignment data by applying similar filtering and phoneme mapping measures employed in (Ortiz Tandazo et al., 2025); this includes filtering out alignments with spn segments or with non-silent phones that are excessively long (which indicate alignment errors), fixing diacritics that were wrongly attached to adjacent phones, and replacing some MFA phones with their IPA equivalents ([g] becomes [g]). The amount of phoneme-aligned data available varied widely based on language—to avoid overfitting on any one language, we limit the maximum quantity to 50 hours per language, yielding a labeled dataset of total 372 hours.

For calculating ABX scores on test languages in experiments Sec. 4.1 and Sec. 4.2, we use phoneme alignments obtained from the test set of the Zero Resource 2017 Challenge (Dunbar et al., 2017). For computing the ABX score on the development languages, we use data from CommonVoice (Ardila et al., 2020) and alignments from VoxCommunis (Ahn and Chodroff, 2022).

Split	In-Domain Training	Out-of-Domain Training	
		Pre-Training	Fast Adaptation
Dev.	In-domain Training not performed	<u>5700 hours</u>	<u>10 minutes, 1 hour, 10 hours</u> CV Swahili CV Tamil CV Thai CV Turkish CV Ukrainian
		VP 19 langs. (w/o target langs.)	
Test	<u>6000 hours</u> VP English VP French VP German	<u>5700 hours</u>	<u>10 minutes, 1 hour, 10 hours, 100 hours</u> (subset of In-Domain Training set) VP English VP French VP German
		VP 19 langs. (w/o target langs.)	

Table 4: **Summary of unlabeled datasets utilized across training and evaluation.** Data was accumulated from the Voxpopuli (VP; Wang et al., 2021) corpus and the CommonVoice (CV; Ardila et al., 2020) corpus.

B Details of Training Setup

B.1 Pre-training

Models are trained using a distributed setup across 16 GPUs. Default SpidR hyperparameters (Poli et al., 2025b) are used for pre-training the ID mono-task and the OoD multi-task models. In interleaved supervised pre-training (i.e., Multi-Task-PT [SSL/SL]), every tenth step is backpropagated using phoneme supervised loss (hence in equation 8, $\lambda = 0$ if step mod 10 = 0, else 1). For prediction of supervised labels, language-specific classifier heads (19 heads in total) are attached to the 8th transformer layer of the SpidR model. Here, the 8th layer was used because exploration of hyperparameters indicated it as being optimal for few-shot performance on developmental languages. During supervised training steps, utterances are batched by language; while during self-supervised training steps, each batch consists of a mix of languages. In self-supervised pre-training (i.e., Multi-Task-PT [SSL]), standard SSL loss (as defined by the SpidR architecture) is used throughout. The OoD multi-task models trained under these schema are used as initialization weights for meta-training.

B.2 Meta-Training

Eight MAdaPT episodes are trained in parallel across 16 GPUs. During meta-training, each episode consists of 2,000 steps—1,800 steps for the inner-loop (self-supervised adaptation) and 200 steps for the outer-loop (supervised meta-optimization). For each inner-loop task D_ℓ^u , we utilize a randomly chosen 10-hour data chunk

from a randomly chosen source language. For the outer-loop optimization, the inner-loop language ℓ is retained, but data duration is not fixed at 10 hours. The overall training spans 200,000 steps, resulting in a total of 800 episodes (calculated as $200,000/2,000 \times 8 = 800$ episodes). This meta-training setup is chosen for both practicality of implementation (on limited compute and with limited time) and to closely mimic the low-resource adaptive fine-tuning scenario central to our research.

For the self-supervised initialization of FOBLO, the supervised outer-loop optimization is applied to the 6th layer of the model; while, for the interleaved-supervised initialization, it is applied to the 8th layer (staying consistent with the supervised layer during meta-initialization). The FOBLO supervised layers were selected based on best performing layers of the meta-initialization models used for meta-training. When computing ABX scores, we thereby report results from the 6th and 8th layers for the self- and interleaved-supervised models, respectively.

In SpidR, the teacher is trained as an exponential moving average of the student, with the decay of the teacher at the timestep t defined as $1 - (1 - \beta_0) \exp(-t/T)$. We find that some meta-training configurations (specifically, trainings initialized using interleaved supervision or meta-trained using FOBLO) perform better when trained with $\beta_0 = 1$, effectively producing a frozen teacher. Hence, we select the best performing value of β_0 (from 1.0 and the default 0.999) for each meta-training variant (i.e., Reptile or FOBLO with SSL or SSL/SL initializations) based on few-shot per-

formance on the development language set.

Within each meta-training inner-loop, we use a constant learning rate adding a small warmup for 600 timesteps at the beginning of each loop. The learning rate within each episode is identified through a tri-stage learning rate scheduler with maximum learning rate of $5e - 5$. The detailed scheduler has been illustrated in Figure 3.

B.3 Fast Adaptation Training

For fast adaptive fine-tuning to the OoD target languages, we use a single GPU. For each model variant (i.e., Multi-Task-PT, MAdaPT-Reptile, or MAdaPT-FOBLO with SSL or SSL/SL initializations) and each adaptation dataset size (10 minutes to 100 hours), we conduct a hyperparameter exploration on the development language set to identify optimal training timesteps (varied between 4,000 and 24,000), learning rate (constant learning rate of $5e-4$ or $5e-5$), and β_0 for the teacher decay (1.0 or default 0.999). The best checkpoint for each adaptation run is selected based on the lowest validation loss, ensuring optimal model performance for downstream evaluations.

C Detailed Experimental Results of the Main Manuscript

C.1 Detailed Results of ABX scores

Here, we present detailed ABX scores for both within-speaker and across-speaker setups as illustrated in Figure 2. As shown in Table 5, the In-Domain Mono-Task-PT [SSL] models are trained with sufficient in-domain data (6k hours per language), resulting in oracle-level performance. Moreover, we evaluate all methods on the five development languages, with their ABX scores reported in Table 6. Due to the lack of unlabeled corpora for these five development languages, the oracle performance is not reported in the table. Across both tables, our proposed MAdaPT-FOBLO consistently outperforms the Oracle baseline and achieves performance comparable to the MAdaPT-Reptile method. Notably, when self-supervised initialization is applied, our approach rapidly improves performance as adaptation time increases, highlighting its data efficiency and overall effectiveness.

C.2 Detailed Results of Spoken Language Modeling

Detailed results for downstream spoken language modeling are provided under Table 7. As described

in Experiment 4.3, we used sWuggy, sBlimp, and spoken tSC to estimate performance of the spoken language models. For all tasks, candidates are scored by length-normalized log-likelihood (log-likelihood divided by token count) for comparability across strings, and decisions are made by selecting the higher-scoring alternative.

We use SSL models finetuned on the English adaptation sets (0 hours to 100 hours) as encoders for the downstream SLM. OPT-125M models (Zhang et al., 2022) are utilized as the SLMs, trained using fairseq2 (Balioglu et al., 2023) and following the architectural decisions made by previous works (Hassid et al., 2023; Poli et al., 2025b). The 6k hour subset of Libri-Light (Kahn et al., 2020) is used as the training dataset. We train on 8 GPUs, with a context length of 2048, and a batch of at most 81920 tokens, for 25000 steps. The learning rate is set at $1e - 2$ with a 1000-step warmup period and with a cosine annealing schedule. Remaining hyperparameters follow OPT-125M defaults. We select the checkpoint with the lowest validation loss.

C.3 Detailed Results of Phoneme Discovery Benchmark

The Phoneme Discovery Benchmark (Poli et al., 2026) is specifically designed to investigate the abilities of speech representation models to encode phonemic information in a low resource setting. It employs metrics such as PNMI, PER, and ABX (which are described in the main paper). The benchmark includes 6 development languages—Swahili, Tamil, Thai, Ukrainian, Turkish, and German—and 6 test languages—French, English, Japanese, Mandarin, Wolof, and Basque. Note that the development and test language set in the benchmark differs from our previous experiments but is disjoint from our training set.

In the current work, we applied our previously tuned MAdaPT-Reptile and MAdaPT-FOBLO models to the tasks. Our models are trained on 19 Voxpopuli languages (Wang et al., 2021). We compare our approaches to an OoD HuBERT trained on 20 Voxpopuli languages. Test language results are reported in Table 8 and development language results are reported in Table 9. As can be observed, on aggregate, our proposed MAdaPT-FOBLO achieves improved performance over alternate meta-learning heuristics (Reptile) and alternate speech SSL models (HuBERT) for both test and development languages.

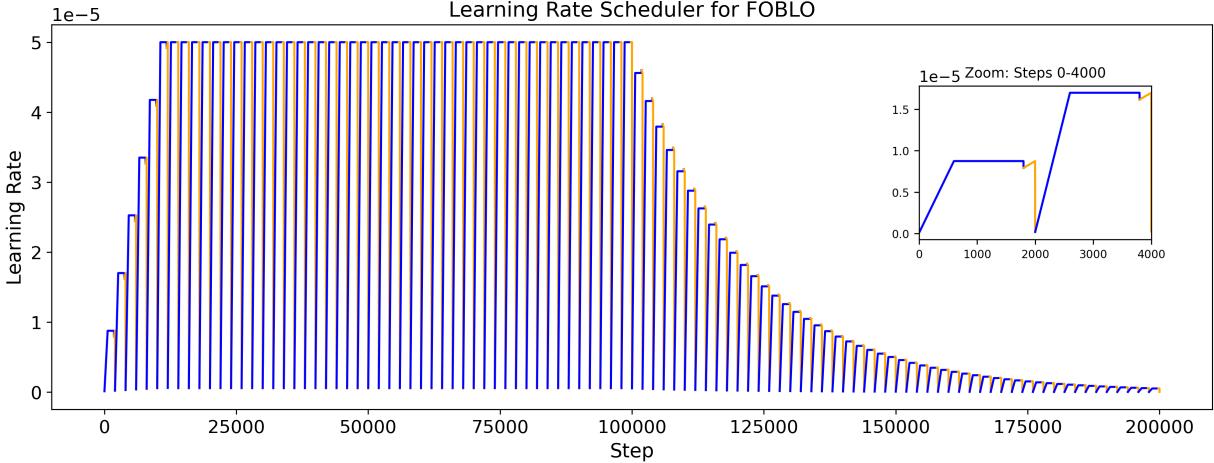


Figure 3: Learning rate scheduler for FOBLO. We use blue and orange to represent the learning rate for self-supervised inner-steps and supervised outer-steps, respectively. The overall training has 200,000 steps. The learning rate scheduler alternates between inner-loop and outer-loop steps within each episode, with resets every 2,000 steps. The inner-loop uses a constant rate after a warmup, while the outer-loop follows a tri-stage schedule.

D Ablation Studies

D.1 Impact of Active Forgetting

To investigate the impact of active forgetting in our approach we conduct ablation studies by removing the active forgetting mechanism from the inner-loop on the 5 development and 3 test languages. As shown in Table 10 and Table 11, incorporating active forgetting consistently outperforms the variant without this mechanism. This demonstrates that resetting the prediction heads and codebooks helps the model alleviate overfitting to previous episodes, thereby improving overall performance.

D.2 Impact of Meta-Initialization

To explore the influence of meta-initialization, we meta-train our model from three types of initialization. Multi-Task-PT [SSL] and Multi-Task-PT [SSL/SL] have been introduced in the main manuscript, both obtained via multi-task pre-training. Here we attempt random initialization, wherein the backbone is initialized by random sampling from the default parameter distribution. Table 12 and Table 13 present ablation studies with different meta-initializations on 5 development and 3 test languages, respectively.

We find that random meta-initialization does not work for meta-training. Without a meaningful starting point, meta-training may fail to converge or require significantly more data and iterations to achieve competitive performance for self-supervised speech models. Thus, the success of meta-learning for speech representation learning is

tightly coupled with the quality and relevance of the initial representations encoded in the backbone.

D.3 Analysis of meta-learning rate

To systematically investigate the impact of the meta-learning rate β in our approach, we conduct a series of experiments with SpidR-Adapt, utilizing interleaved supervised meta-initialization and varying β across 0.001, 0.01, 0.1 and 1. Table 14 presents the results, with the left and right sub-tables corresponding to the 5 development and 3 test language sets, respectively. Our analysis reveals a clear trend on the development set as β increases: ABX scores initially become lower, reaching its peak at $\beta = 0.01$ before declining at higher values. This suggests that a moderate meta-learning rate strikes the best balance between adaptation and stability, while excessively high rates may lead to suboptimal generalization.

To ensure robust hyperparameter selection and prevent overfitting to the 3 test languages, we rely on the 5 development results to identify the optimal β . Consequently, all reported results in the paper are based on $\beta = 0.01$, which consistently yields the strongest performance across our evaluation.

D.4 Layer-wise analysis on the model’s discriminability.

To investigate how layer-specific embeddings affect the model’s ability to discriminate between phonemes, we present ABX scores for each student layer in Figure 4. The scores are averaged across (a) 5 development or (b) 3 test languages

Method	0h	10m	1h	10h	100h	Avg. (w/o 0h) ↓
<i>Within-Speaker ABX</i>						
In-Domain Mono-Task-PT	6000 hours training; Oracle score is 4.10.					
Multi-Task-PT [SSL]	4.65	4.56	4.40	4.23	4.13	4.33
+ MAdaPT-Reptile	4.50	4.34	4.29	4.10	4.03	4.19
+ MAdaPT-FOBLO	10.05	4.51	4.05	<u>3.78</u>	<u>3.69</u>	4.01
Multi-Task-PT [SSL/SL]	4.10	4.08	3.94	<u>3.78</u>	3.71	3.88
+ MAdaPT-Reptile	3.89	3.94	3.82	3.62	3.66	3.76
+ MAdaPT-FOBLO	<u>4.00</u>	<u>4.07</u>	<u>3.84</u>	3.62	3.70	<u>3.80</u>
<i>Across-Speaker ABX</i>						
In-Domain Mono-Task-PT	6000 hours training; Oracle score is 5.47.					
Multi-Task-PT [SSL]	6.60	6.44	5.99	5.68	5.48	5.89
+ MAdaPT-Reptile	5.97	5.82	5.72	5.45	5.38	5.59
+ MAdaPT-FOBLO	15.12	5.96	5.26	4.92	4.83	5.24
Multi-Task-PT [SSL/SL]	5.42	5.36	5.06	4.93	4.88	<u>5.06</u>
+ MAdaPT-Reptile	5.19	5.16	<u>4.97</u>	<u>4.78</u>	<u>4.79</u>	4.93
+ MAdaPT-FOBLO	<u>5.28</u>	<u>5.23</u>	4.96	4.76	4.77	4.93

Table 5: **Detailed Within-Speaker and Across-Speaker ABX scores (in %) on 3 TEST languages.** MAdaPT-FOBLO and MAdaPT-Reptile show superior performance, surpassing In-Domain Mono-Task-PT with limited data. The best scores are in **bold** and second best are underlined.

and averaged across 10 minute to 100 hour adaptation data scales. Our analysis reveals distinct trends for different meta-initialization strategies applied to MAdaPT-FOBLO: 1) with Multi-Task-PT[SSL], the phone discriminability improves with increasing layer depth, peaking at layer 6. Beyond this point, performance declines, suggesting that intermediate layers capture the most relevant phonetic representations, while deeper layers may become overly specialized or abstracted for the ABX task. 2) with Multi-Task-PT[SSL/SL], the optimal performance is observed at layer 8.

These results suggest that the best performing layer is consistent with the layer at which supervision is applied during the outer-loop of FOBLO. For SSL meta-initialization, the a supervision head is attached to the 6th encoder layer for outer-loop supervision while for SSL/SL meta-initialization, it is attached to the 8th layer (see Appendix B for more details here).

Method	0h	10m	1h	10h	Avg. (w/o 0h) ↓
<i>Within-Speaker ABX</i>					
Multi-Task-PT [SSL]	8.84	8.34	7.30	6.14	7.26
+ MAdaPT-Reptile	7.94	7.08	6.64	5.78	6.50
+ MAdaPT-FOBLO	13.89	7.69	6.21	5.29	6.40
Multi-Task-PT [SSL/SL]	<u>7.59</u>	6.85	6.20	5.67	6.24
+ MAdaPT-Reptile	<u>7.07</u>	<u>6.42</u>	<u>6.04</u>	5.60	<u>6.02</u>
+ MAdaPT-FOBLO	7.76	6.26	6.01	<u>5.58</u>	5.95
<i>Across-Speaker ABX</i>					
Multi-Task-PT [SSL]	10.49	9.79	8.19	6.81	8.27
+ MAdaPT-Reptile	9.15	8.05	7.51	6.58	7.38
+ MAdaPT-FOBLO	16.31	8.60	6.79	5.73	7.04
Multi-Task-PT [SSL/SL]	<u>8.24</u>	7.40	6.51	6.06	6.66
+ MAdaPT-Reptile	<u>7.74</u>	<u>6.93</u>	<u>6.40</u>	6.05	<u>6.46</u>
+ MAdaPT-FOBLO	8.42	6.82	6.38	<u>5.96</u>	6.39

Table 6: **Detailed Within-Speaker and Across-Speaker ABX scores (in %) on 5 DEVELOPMENT languages.** MAdaPT-FOBLO outperforms alternate methods in phoneme representation. Hyperparameters are tuned using results from the development language set. The best scores are in **bold** and second best are underlined.

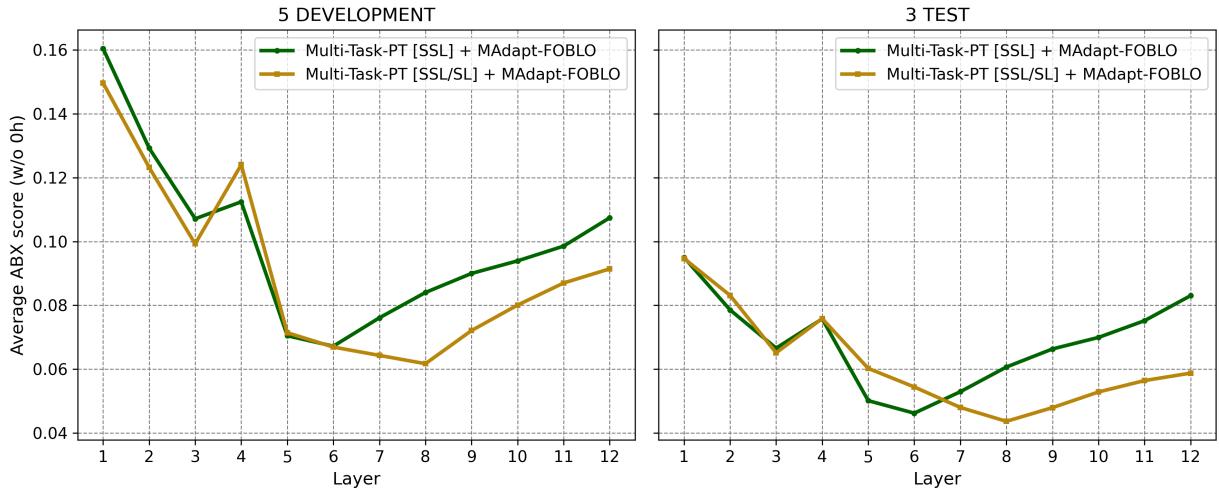


Figure 4: **Layer-wise analysis on the model’s discriminability over phonemes.** We present the ABX scores averaged over the corresponding new languages, and across the two within- and across-speaker conditions: (a) 5 development and (b) 3 test languages. We report results for our proposed MAdaPT-FOBLO method with two types of meta-initialization, Multi-Task-PT[SSL] and Multi-Task-PT[SSL/SL]. The optimal layer for ABX performance remains consistent across both ABX conditions, but varies depending on the meta-initialization. Specifically, the optimal layer is 6 for initialization (a) and 8 for initialization (b), respectively.

Method	0h	10m	1h	10h	100h	Avg. (w/o 0h) ↓
<i>sWuggy (in-vocab and out-of-vocab)</i>						
In-Domain Mono-Task-PT	6000 hours training; Oracle score is 62.34.					
Multi-Task-PT [SSL]	61.05	61.21	62.48	63.22	63.91	62.70
+ MAdaPT-Reptile	61.77	62.26	62.53	62.80	64.28	62.97
+ MAdaPT-FOBLO	<u>54.06</u>	61.87	63.12	64.00	<u>65.09</u>	63.52
Multi-Task-PT [SSL/SL]	<u>61.00</u>	62.91	63.58	64.87	64.17	63.88
+ MAdaPT-Reptile	63.72	<u>64.38</u>	64.45	<u>65.69</u>	65.60	65.03
+ MAdaPT-FOBLO	60.84	64.80	<u>64.37</u>	65.80	65.04	<u>65.00</u>
<i>sBlimp</i>						
In-Domain Mono-Task-PT	6000 hours training; Oracle score is 53.60.					
Multi-Task-PT [SSL]	52.39	53.15	53.56	54.00	53.86	53.64
+ MAdaPT-Reptile	54.18	53.71	52.78	53.67	54.22	53.59
+ MAdaPT-FOBLO	50.62	54.14	<u>54.39</u>	<u>54.15</u>	54.60	<u>54.32</u>
Multi-Task-PT [SSL/SL]	53.05	54.13	53.52	53.57	<u>54.22</u>	53.86
+ MAdaPT-Reptile	53.66	<u>54.27</u>	53.38	53.57	53.50	53.68
+ MAdaPT-FOBLO	<u>53.37</u>	55.11	54.46	54.86	54.00	54.61
<i>Spoken tSC</i>						
In-Domain Mono-Task-PT	6000 hours training; Oracle score is 69.60.					
Multi-Task-PT [SSL]	69.02	66.77	66.35	67.20	67.15	66.87
+ MAdaPT-Reptile	66.45	66.99	68.16	68.27	66.60	67.51
+ MAdaPT-FOBLO	66.08	70.57	68.86	69.07	71.10	69.90
Multi-Task-PT [SSL/SL]	<u>70.03</u>	69.34	70.19	<u>69.61</u>	<u>69.71</u>	<u>69.71</u>
+ MAdaPT-Reptile	69.76	<u>69.50</u>	67.84	69.66	69.39	69.10
+ MAdaPT-FOBLO	70.73	69.07	<u>69.12</u>	68.86	69.18	69.06

Table 7: **Detailed results of spoken language modeling metrics—sWuggy, sBlimp, and spoken tSC (in %).** MAdaPT-FOBLO and MAdaPT-Reptile show superior performance, surpassing In-Domain Mono-Task-PT with limited data. The best scores are in **bold** and the second best are underlined.

Method	Backbone Model	0h	10m	1h	10h	Avg. (w/o 0h)
<i>PNMI \uparrow</i>						
Multi-Task-PT [SSL]	HuBERT	0.50	0.53	0.59	0.62	0.58
Multi-Task-PT [SSL]	SpidR	0.58	0.62	0.67	0.69	0.66
+ MAdaPT-Reptile	SpidR	0.40	0.61	0.63	0.65	0.63
+ MAdaPT-FOBLO	SpidR	0.10	0.64	0.70	<u>0.72</u>	0.69
Multi-Task-PT [SSL/SL]	SpidR	0.67	0.71	0.73	0.73	0.72
+ MAdaPT-Reptile	SpidR	0.49	0.68	0.69	0.69	0.69
+ MAdaPT-FOBLO	SpidR	<u>0.66</u>	<u>0.70</u>	<u>0.71</u>	<u>0.72</u>	<u>0.71</u>
<i>PER \downarrow</i>						
Multi-Task-PT [SSL]	HuBERT	126.05	88.68	69.98	69.37	76.01
Multi-Task-PT [SSL]	SpidR	85.41	75.30	56.54	48.67	60.17
+ MAdaPT-Reptile	SpidR	153.92	66.03	60.10	54.48	60.20
+ MAdaPT-FOBLO	SpidR	87.41	64.77	39.84	34.96	46.52
Multi-Task-PT [SSL/SL]	SpidR	50.86	40.95	38.80	38.02	39.26
+ MAdaPT-Reptile	SpidR	110.77	<u>40.80</u>	36.92	37.08	<u>38.27</u>
+ MAdaPT-FOBLO	SpidR	<u>51.01</u>	39.35	<u>37.12</u>	<u>36.61</u>	37.70
<i>ABX (Within-Speaker) \downarrow</i>						
Multi-Task-PT [SSL]	HubERT	6.77	7.80	6.46	5.61	6.62
Multi-Task-PT [SSL]	SpidR	5.73	5.51	4.82	4.16	4.83
+ MAdaPT-Reptile	SpidR	5.32	4.84	4.53	4.02	4.46
+ MAdaPT-FOBLO	SpidR	10.59	5.28	4.03	3.69	4.33
Multi-Task-PT [SSL/SL]	SpidR	<u>5.16</u>	4.63	<u>4.04</u>	3.94	4.20
+ MAdaPT-Reptile	SpidR	4.73	<u>4.38</u>	4.09	<u>3.88</u>	4.12
+ MAdaPT-FOBLO	SpidR	5.22	4.32	4.08	<u>3.88</u>	4.09
<i>ABX (Across-Speaker) \downarrow</i>						
Multi-Task-PT [SSL]	HuBERT	8.84	9.10	7.61	6.60	7.77
Multi-Task-PT [SSL]	SpidR	7.20	6.63	5.70	4.81	5.72
+ MAdaPT-Reptile	SpidR	6.26	5.68	5.33	4.62	5.21
+ MAdaPT-FOBLO	SpidR	14.04	6.26	4.84	4.14	5.08
Multi-Task-PT [SSL/SL]	SpidR	<u>5.94</u>	5.04	4.55	4.36	4.65
+ MAdaPT-Reptile	SpidR	5.44	<u>4.85</u>	<u>4.50</u>	<u>4.34</u>	<u>4.57</u>
+ MAdaPT-FOBLO	SpidR	6.01	4.80	4.49	4.36	4.55

Table 8: **Detailed results of phoneme discovery benchmark—PNMI, PER (in %), and ABX (in %) on test languages.** MAdaPT-FOBLO outperforms alternate meta-training framework (Reptile) and alternate speech SSL model (HuBERT). The best scores are in **bold** and the second best are underlined.

Method	Backbone Model	0h	10m	1h	10h	Avg. (w/o 0h)
<i>PNMI \uparrow</i>						
Multi-Task-PT [SSL]	HuBERT	0.47	0.46	0.52	0.58	0.52
Multi-Task-PT [SSL]	SpidR	<u>0.54</u>	0.57	0.62	0.65	0.62
+ MAdaPT-Reptile	SpidR	0.37	0.57	0.59	0.61	0.59
+ MAdaPT-FOBLO	SpidR	0.10	0.60	0.66	<u>0.68</u>	0.65
Multi-Task-PT [SSL/SL]	SpidR	0.63	0.66	0.68	0.69	0.68
+ MAdaPT-Reptile	SpidR	0.46	0.64	0.65	0.66	0.65
+ MAdaPT-FOBLO	SpidR	0.63	<u>0.65</u>	<u>0.67</u>	<u>0.68</u>	<u>0.67</u>
<i>PER \downarrow</i>						
Multi-Task-PT [SSL]	HuBERT	118.33	96.54	77.23	70.48	81.42
Multi-Task-PT [SSL]	SpidR	82.33	76.99	58.99	52.50	62.83
+ MAdaPT-Reptile	SpidR	147.89	68.19	61.98	56.32	62.16
+ MAdaPT-FOBLO	SpidR	84.73	65.85	43.88	37.40	49.05
Multi-Task-PT [SSL/SL]	SpidR	48.69	46.58	41.30	<u>40.65</u>	42.84
+ MAdaPT-Reptile	SpidR	101.44	43.80	39.61	39.60	41.01
+ MAdaPT-FOBLO	SpidR	<u>49.08</u>	<u>44.76</u>	<u>40.61</u>	<u>40.65</u>	<u>42.01</u>
<i>ABX (Within-Speaker) \downarrow</i>						
Multi-Task-PT [SSL]	HubERT	9.48	10.92	9.36	7.93	9.40
Multi-Task-PT [SSL]	SpidR	8.31	7.90	6.99	6.00	6.96
+ MAdaPT-Reptile	SpidR	7.52	6.79	6.40	5.68	6.29
+ MAdaPT-FOBLO	SpidR	13.35	7.31	5.99	5.19	6.17
Multi-Task-PT [SSL/SL]	SpidR	<u>7.17</u>	6.56	5.96	5.52	6.01
+ MAdaPT-Reptile	SpidR	6.73	<u>6.17</u>	<u>5.82</u>	<u>5.42</u>	<u>5.80</u>
+ MAdaPT-FOBLO	SpidR	7.29	6.05	5.79	5.43	5.75
<i>ABX (Across-Speaker) \downarrow</i>						
Multi-Task-PT [SSL]	HuBERT	11.83	12.69	10.80	9.09	10.86
Multi-Task-PT [SSL]	SpidR	10.18	9.57	8.15	6.96	8.22
+ MAdaPT-Reptile	SpidR	8.94	7.99	7.50	6.70	7.40
+ MAdaPT-FOBLO	SpidR	16.11	8.46	6.87	5.94	7.09
Multi-Task-PT [SSL/SL]	SpidR	<u>8.13</u>	7.44	6.64	6.23	6.77
+ MAdaPT-Reptile	SpidR	7.68	<u>7.00</u>	<u>6.51</u>	6.19	<u>6.57</u>
+ MAdaPT-FOBLO	SpidR	8.26	6.91	6.49	<u>6.12</u>	6.51

Table 9: **Detailed results of phoneme discovery benchmark—PNMI, PER (in %), and ABX (in %) on development languages.** MAdaPT-FOBLO outperforms alternate meta-training framework (Reptile) and alternate speech SSL model (HuBERT). The best scores are in **bold** and the second best are underlined.

Method	Active Forgetting	0h	10m	1h	10h	Avg. (w/o 0h) ↓
<i>Within-Speaker ABX</i>						
Multi-Task-PT [SSL] +	✗	10.69	<u>6.46</u>	6.77	6.64	6.62
MAdaPT-FOBLO	✓	13.89	7.69	6.21	5.29	6.40
<i>Across-Speaker ABX</i>						
Multi-Task-PT [SSL] +	✗	12.38	7.13	7.09	6.78	7.00
MAdaPT-FOBLO	✓	16.31	8.60	6.79	5.73	7.04
Multi-Task-PT [SSL/SL] +	✗	8.12	<u>7.24</u>	<u>6.47</u>	6.01	<u>6.57</u>
MAdaPT-FOBLO	✓	<u>8.42</u>	6.82	6.38	<u>5.96</u>	6.39

Table 10: **Impact of active forgetting on 5 DEVELOPMENT languages.** ✓ and ✗ denote whether we deploy the active forgetting mechanism in the inner-loop or not, respectively. Broadly, active forgetting improves adaptation performance, preventing overfitting to training languages. The best scores are in **bold** and second best are underlined.

Method	Active Forgetting	0h	10m	1h	10h	100h	Avg. (w/o 0h) ↓
<i>Within-Speaker ABX</i>							
In-Domain Mono-Task-PT	N.A.	6000 hours training; Oracle score is 4.10.					
Multi-Task-PT [SSL] +	✗	21.89	4.40	4.54	4.37	4.20	4.38
MAdaPT-FOBLO	✓	10.05	4.51	4.05	3.78	<u>3.69</u>	4.01
Multi-Task-PT [SSL/SL] +	✗	3.99	4.02	<u>3.87</u>	<u>3.71</u>	3.67	<u>3.82</u>
MAdaPT-FOBLO	✓	4.00	<u>4.07</u>	3.84	3.62	3.70	3.80
<i>Across-Speaker ABX</i>							
In-Domain Mono-Task-PT	N.A.	6000 hours training; Oracle score is 5.47.					
Multi-Task-PT [SSL] +	✗	29.11	5.62	5.77	5.56	5.33	5.57
MAdaPT-FOBLO	✓	15.12	5.96	5.26	4.92	4.83	5.24
Multi-Task-PT [SSL/SL] +	✗	<u>5.29</u>	<u>5.32</u>	<u>5.01</u>	<u>4.84</u>	<u>4.80</u>	4.99
MAdaPT-FOBLO	✓	5.28	5.23	4.96	4.76	4.77	4.93

Table 11: **Impact of active forgetting on 3 TEST languages.** ✓ and ✗ denote whether we deploy the active forgetting mechanism in the inner-loop or not, respectively. Similar to results in development languages, active forgetting here improves adaptation performance, preventing overfitting to training languages. The best scores are in **bold** and second best are underlined.

Method	Meta-Initialization	0h	10m	1h	10h	Avg. (w/o 0h) ↓
<i>Within-Speaker ABX</i>						
MAdaPT-FOBLO	Random	35.66	31.21	35.74	37.42	34.79
	Multi-Task-PT [SSL]	<u>13.89</u>	<u>7.69</u>	<u>6.21</u>	5.29	<u>6.40</u>
	Multi-Task-PT [SSL/SL]	7.76	6.26	6.01	<u>5.58</u>	5.95
<i>Across-Speaker ABX</i>						
MAdaPT-FOBLO	Random	39.13	35.73	38.04	39.15	37.64
	Multi-Task-PT [SSL]	<u>16.31</u>	<u>8.60</u>	<u>6.79</u>	5.73	<u>7.04</u>
	Multi-Task-PT [SSL/SL]	8.42	6.82	6.38	<u>5.96</u>	6.39

Table 12: **Impact of meta-initialization on 5 DEVELOPMENT languages.** Random initialization produces unstable model training. The best scores are in **bold** and second best are underlined.

Method	Meta-Initialization	0h	10m	1h	10h	100h	Avg. (w/o 0h) ↓
<i>Within-Speaker ABX</i>							
In-Domain Mono-Task-PT	N.A.	6000 hours training; Oracle score is 4.10.					
MAdaPT-FOBLO	Random	32.68	25.12	24.61	23.75	24.52	24.50
	Multi-Task-PT [SSL]	<u>10.05</u>	<u>4.51</u>	<u>4.05</u>	<u>3.78</u>	3.69	<u>4.01</u>
	Multi-Task-PT [SSL/SL]	4.00	4.07	3.84	3.62	<u>3.70</u>	3.80
<i>Across-Speaker ABX</i>							
In-Domain Mono-Task-PT	N.A.	6000 hours training; Oracle score is 5.47.					
MAdaPT-FOBLO	Random	38.75	33.25	32.81	32.31	32.78	32.79
	Multi-Task-PT [SSL]	<u>15.12</u>	<u>5.96</u>	<u>5.26</u>	<u>4.92</u>	<u>4.83</u>	<u>5.24</u>
	Multi-Task-PT [SSL/SL]	5.28	5.23	4.96	4.76	4.77	4.93

Table 13: **Impact of meta-initialization on 3 TEST languages.** Random initialization produces unstable model training. The best scores are in **bold** and second best are underlined.

β	0h	10m	1h	10h	Avg. (w/o 0h) ↓	β	0h	10m	1h	10h	100h	Avg. (w/o 0h) ↓
<i>Within-Speaker ABX</i>												
0.001	7.60	6.34	6.07	<u>5.59</u>	6.00	0.001	<u>4.10</u>	<u>4.02</u>	3.87	3.66	3.69	3.81
0.01	<u>7.76</u>	6.26	<u>6.01</u>	5.58	5.95	0.01	4.00	4.07	3.84	3.62	3.70	<u>3.80</u>
0.1	11.64	<u>6.32</u>	5.99	5.61	<u>5.98</u>	0.1	6.20	3.91	<u>3.81</u>	<u>3.57</u>	<u>3.64</u>	3.73
1	8.25	6.64	6.11	5.58	6.11	1	5.89	4.04	3.72	3.53	3.62	3.73
<i>Across-Speaker ABX</i>												
0.001	8.23	6.92	6.42	6.05	6.46	0.001	<u>5.39</u>	5.25	4.97	4.79	4.77	4.95
0.01	<u>8.42</u>	6.82	6.38	5.96	6.39	0.01	5.28	<u>5.23</u>	<u>4.96</u>	<u>4.76</u>	4.77	4.93
0.1	12.66	<u>6.85</u>	<u>6.41</u>	<u>6.04</u>	<u>6.43</u>	0.1	8.56	5.16	5.00	<u>4.76</u>	<u>4.75</u>	<u>4.92</u>
1	9.13	7.16	6.52	6.06	6.58	1	7.87	5.29	4.86	4.68	4.72	4.89

Table 14: **Impact of meta-learning rate β on 5 DEVELOPMENT and 3 TEST languages.** Best performing β is 0.01 for MAdaPT-FOBLO [SSL/SL] on development languages and is retained for test language inference. The best scores are in **bold** and second best are underlined.