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

Mahi Luthra[*,1], **Jiayi Shen**[*,1], **Maxime Poli**[*,2], **Angelo Ortiz Tandazo**[2], **Yosuke Higuchi**[1], **Youssef Benchekroun**[1], **Martin Gleize**[1], **Charles-Eric Saint-James**[1], **Dongyan Lin**[1], **Phillip Rust**[1], **Angel Villar**[1], **Surya Parimi**[1], **Vanessa Stark**[1], **Rashel Moritz**[1], **Juan Pino**[1], **Yann LeCun**[1], **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 of speech units 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 downstream spoken language modeling scores (sWUGGY, sBLIMP, tSC), surpassing in-domain toplines after training on less than 1h of target-language audio and delivering $100\times$ greater data efficiency than standard multi-task 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 specializing their perception of phonemic contrasts to the structures relevant to their native language (Eimas et al., 1971; Werker and Tees, 1984; Kuhl, 2004), all from continuous auditory input and with only 50 to 500 hours of speech exposure (Bergelson et al., 2019; Cychosz et al., 2021; Cristia, 2023).

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 (Lakhotia et al., 2021; Hassid et al., 2023) 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 (Kuhl, 1979, 2004; Saffran et al., 2007). 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. Conceptually, this episodic design draws loose inspiration from evolutionary processes, with a second-order op-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28705–28728
July 2-7, 2026

timization 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 approximates the second-order bi-level problem posed by MAdaPT. The inner loop trains the model to learn on unlabeled, under-resourced data, while the outer loop calibrates the meta-parameters using feedback from a gold-standard labeled set.
- **Interleaved light supervision**, which incorporates self-supervised training with occasional phone 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 monolingual 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 data-efficient speech representation learning, conceptually motivated by the efficiency of early human language acquisition.

2 Related Works

2.1 Self-supervised learning

Self-supervised learning has enabled speech models to learn rich representations from unlabeled audio, underpinning 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., 2023)—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, in particular their phonemic discriminability, directly impacts downstream spoken language performance (Poli et al., 2024; Hallap et al., 2023). Hence, in this 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 phone 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 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., 2025). 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., 2025). 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 address 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 $\mathcal{D}_\ell^u$ and, optionally, a small labeled corpus $\mathcal{D}_\ell^s$. 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 $\mathcal{D}_\ell^u$ 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 $\mathbf{D}_\ell^u$ 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 teacher-student architecture. We represent it as $\boldsymbol{\theta} = \{f(\cdot), E_s, E_t, \{\mathbf{W}^k\}, \{\mathbf{C}^k\}\}$, where $f(\cdot)$ is a convolutional downsampler, and E_s, E_t are Transformer encoders for the student and teacher, respectively. The teacher is an exponential moving average 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 $L-K \leq k \leq L$, with L the number of Transformer layers and K the number of codebooks).

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

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

where $\mathcal{L}_{\text{ssl}}$ denotes a self-supervised loss function, $\boldsymbol{\theta}$ represents all learnable parameters of the speech model SpidR, and $\boldsymbol{\theta}_\ell^*$ are the optimal model parameters specific to the language ℓ. We note that $\mathbf{D}_\ell^u$ 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

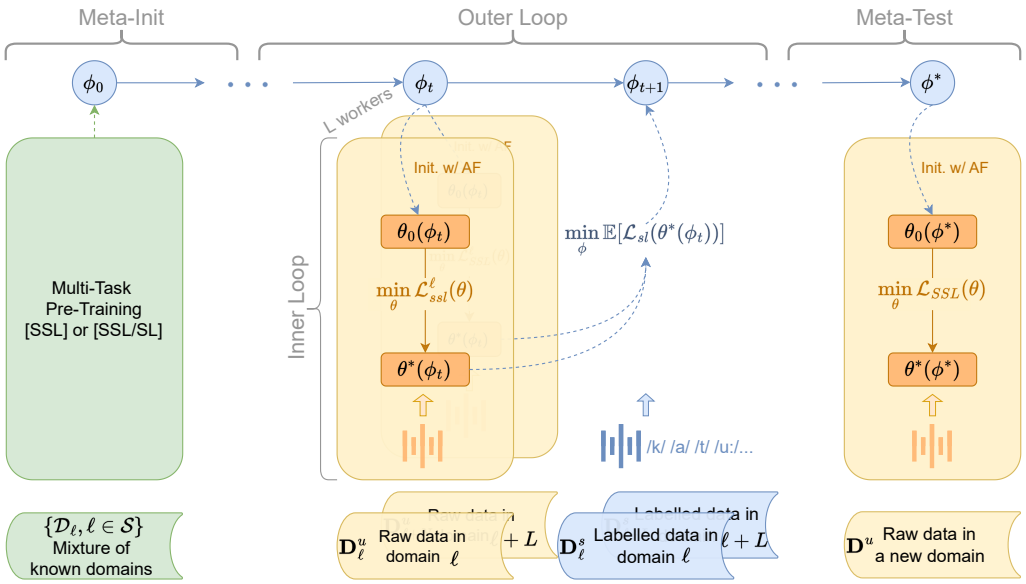

Figure 1: **Overview of SpidR-Adapt for few-shot speech adaptation.** It consists of three main phrases: (1) meta-initialization performs multi-task pre-training with interleaved supervision, learning a robust initialization ϕ_0 from a mixture of source domains. (2) meta-training through MAdaPT-FOBLO optimizes this initialization for fast adaption to $\mathcal{D}_\ell$. Each worker conducts inner loop adaptation with active forgetting (AF) on unlabeled data, followed by outer loop updates that refine 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.

an inner optimization; meanwhile, lightweight labeled data are deployed to supervise these adaptation processes at 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. Concretely, we interpret the learned meta-parameters ϕ as an adaptation prior: an initialization that can be efficiently reorganized toward the phonemic structure of a new language under limited unlabeled data.

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

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

where $\mathcal{L}_{\text{ssl}}$ denotes a self-supervised loss function in the inner level, performing adaptation from unlabeled speech data; and $\mathcal{L}_{\text{sl}}$ 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 to 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 $\boldsymbol{\phi}$ as default but reset all heads and codebooks, yielding the optimization with initialization $\boldsymbol{\theta}_{\text{AF}}(\boldsymbol{\phi})$ for each inner loop at both meta-training and meta-test stages:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{ssl}}(\boldsymbol{\theta}_{\text{AF}}(\boldsymbol{\phi}); \mathbf{D}^u_\ell), \text{ where} \tag{3}$$

$$\boldsymbol{\theta}_{\text{AF}}(\boldsymbol{\phi}) = \Big\{f(\boldsymbol{\phi}), E_s(\boldsymbol{\phi}), E_t(\boldsymbol{\phi}), \{\mathbf{W}^k_0\}, \{\mathbf{C}^k_0\}\Big\}.$$

Each codebook $\mathbf{C}^k_0$ is sampled from a normal distribution $\mathcal{N}(0, 1)$ and each head $\mathbf{W}^k_0$ is warmed up for 20 steps using the first batch of $\mathbf{D}^u_\ell$.

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 update of meta-parameters $\boldsymbol{\phi}$ can be formulated as:

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

where β is a learning rate in the outer loop used to update the meta-parameters $\boldsymbol{\phi}$. 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:

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \beta \nabla_{\boldsymbol{\phi}} \mathcal{L}_{sl}(\boldsymbol{\theta}_\ell^M(\boldsymbol{\phi}); \mathbf{D}_\ell^s) \cdot \prod_{m=1}^{M} \left[\mathbf{I} - \alpha \nabla_{\boldsymbol{\phi}_\ell^{m-1}} \left(\nabla_{\boldsymbol{\phi}} \mathcal{L}_{ssl}(\boldsymbol{\theta}_\ell^{m-1}(\boldsymbol{\phi})) \right) \right], \tag{5}$$

where α is the learning rate in the inner loop update and the task-specific parameters $\boldsymbol{\theta}_\ell^m$ denote 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 $\mathbf{D}_\ell^s$. 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:

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

where $\boldsymbol{\theta}_\ell^M$ is obtained after M self-supervised inner steps starting from $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_\ell^{M+N}$ is obtained by taking an additional N supervised steps from $\boldsymbol{\theta}_\ell^M$. By substituting Equation (6) into Equation (5), **FOBLO** updates the meta-parameters as follows:

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \beta \mathbb{E}_{\ell \sim \mathcal{S}}[\boldsymbol{\theta}_\ell^M - \boldsymbol{\theta}_\ell^{M+N}]. \tag{7}$$

Both Reptile (Nichol et al., 2018) and the proposed FOBLO method use an outer loop learning rate β, but with different semantics: in Reptile, β controls how far the meta-parameters move toward the task-adapted parameters along the full adaptation trajectory; in Equation (7), β controls how much the meta-parameters move in the direction of the supervised correction defined in Equation (6), encouraging the model to meta-learn an initialization whose SSL adaptation aligns well with supervised targets. 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 performative meta-initialization, denoted as $\boldsymbol{\phi}_0$. 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:

$$\arg\min_{\boldsymbol{\phi}_0} \begin{cases} \lambda \, \mathcal{L}_{\text{ssl}}(\boldsymbol{\phi}_0; \{\mathcal{D}_\ell^u\}) \\ + (1-\lambda) \, \mathcal{L}_{\text{sl}}(\boldsymbol{\phi}_0; \{\mathcal{D}_\ell^s\}) \end{cases}, \tag{8}$$

where $\lambda \in \{0, 1\}$ is a binary hyperparameter. Here, $\mathcal{L}_{\text{ssl}}$ 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 $\mathcal{L}_{\text{sl}}$ is the supervised loss, applied to the union of labeled corpora $\{\mathcal{D}_\ell^s, \ell \sim \mathcal{S}\}$.

In the current work, we use two distinct meta-initializations: 1) `Multi-Task-PT [SSL]`: setting λ to 1 throughout pre-training, corresponding to standard SSL; and 2) `Multi-Task-PT [SSL/SL]` switching λ to 0 periodically, interleaving occasional supervised steps into the self-supervised training regime. The latter provides a stronger initialization for meta-training.

4 Experiments

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 lead to improvements when labeled data is unavailable during pretraining? (3) Can **SpidR-Adapt** produce improvements in downstream spoken language modeling? (4) Can **SpidR-Adapt** outperform 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 phone-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 `Common Voice` (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 random 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 with 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 ABX scores averaged across our three test languages and between within- and across-speaker ABX conditions (individual trends are consistent). Using SpidR as the backbone under two meta-initialization strategies, self- and interleaved-supervised initialization (Sec. 3.3), we compare: 1) **In-Domain Mono-Task-PT**: Standard in-domain pre-training on sufficient data (6k hours) from the target language, serving as topline. Since every small-scale evaluation subset is drawn from the in-domain training pool, we do not perform additional small-scale adaptation: In-Domain PT therefore 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. This serves as our primary OoD baseline. 3) **MAdaPT-FOBLO**: Our proposed approach, combining MAdaPT with its first-order approximation FOBLO. When used with SpidR as the backbone and with interleaved-supervised initialization, this constitutes our few-shot learning speech encoder, **SpidR-Adapt**.

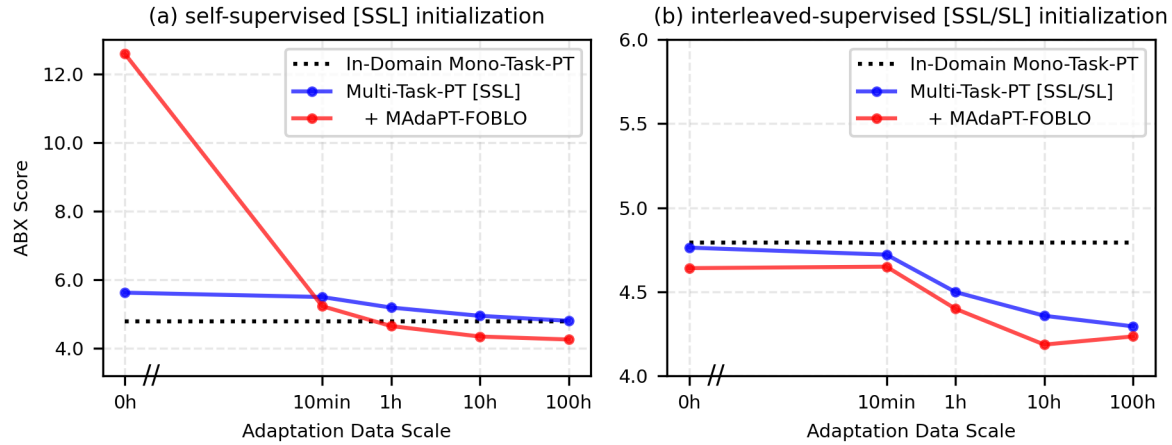

Figure 2: **Data-efficiency of SpidR-Adapt on new languages across different adaptation data scales**. We report ABX scores (lower is better) averaged across three test languages (French, German, English) for two initialization strategies (a) self-supervision `[SSL]` and (b) interleaved-supervision `[SSL/SL]`. Each sub-figure compares our approach with the following: `In-Domain Mono-Task-PT`, the topline method, pretrained on 6k hours of in-domain data and `Multi-Task-PT`, standard multi-task pretraining baseline 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.

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 pretraining does not guarantee better generalization.

In contrast, MAdaPT-FOBLO improves rapidly, demonstrating strong adaptability to OoD data. Notably, with just 1 hour of unlabeled target-language audio, MAdaPT-FOBLO reaches parity with In-Domain PT—a 100$\times$ improvement in data efficiency over Multi-Task-PT. This efficiency is critical for real-world scenarios where language corpora are scarce.

As shown in Figure 2 (b), 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.

We also note an important stability–plasticity tradeoff: MAdaPT-FOBLO has a weaker zero-shot performance, reflected in the high 0 hour ABX scores under SSL initialization. This is a caveat of the current formulation, since Equation (2) optimizes ϕ exclusively for post-adaptation performance without explicitly preserving zero-shot behavior. However, this also reveals that strong plasticity does not necessarily entail strong zero-shot performance. MAdaPT-FOBLO (with SSL initialization) yields a model that reorganizes more effectively once even minimal unlabeled data is available ($\geq$ 10 minutes), consistently surpassing Multi-Task-PT baselines.

Method	**Avg. ABX (w/o 0h)** $\downarrow$	
	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, 0 hours). Although Reptile under-performs FOBLO, it achieves better results than baseline Multi-Task-PT, demonstrating the effectiveness of MAdaPT.

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-

Method	0h	10m	1h	10h	100h	Avg. (w/o 0h) ↑
In-Domain Mono-Task-PT	6000 *hours in-domain training; Topline score is* 65.27.					
Multi-Task-PT [SSL]	63.49	63.80	64.63	64.51	65.20	64.54
+ MAdaPT-Reptile	64.42	64.42	64.47	64.59	64.72	64.55
+ MAdaPT-FOBLO	59.08	65.30	65.19	65.73	66.39	65.65
Multi-Task-PT [SSL/SL]	64.72	65.34	<u>65.77</u>	66.02	66.07	65.80
+ MAdaPT-Reptile	**65.73**	<u>65.44</u>	66.02	**66.77**	**67.00**	<u>66.31</u>
+ MAdaPT-FOBLO	<u>65.31</u>	**65.79**	**66.32**	<u>66.55</u>	<u>66.85</u>	**66.38**

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-initalizations, MAdaPT-FOBLO consistently outperforms the Multi-Task-PT baseline, surpassing the In-Domain topline. MAdaPT-Reptile comes a close second. The best results are shown in **bold**, and second-best are <u>underlined</u>. For detailed results, refer to Appendix C.2.

learning heuristic that approximates the meta-gradient assuming identical inner and outer loop objectives. Here, the meta-update is written as $\boldsymbol{\phi} \leftarrow (1-\beta)\boldsymbol{\phi} - \beta\mathbb{E}_{\ell\sim\mathcal{S}}[\boldsymbol{\theta}_\ell^{M+N}]$, where β trades off the previous meta-parameters and the task-specific solution after each episode. In contrast to our FOBLO update in Equation (7), $\boldsymbol{\theta}_\ell^{M+N}$ 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, used as a purely self-supervised instantiation of MAdaPT, outperforms the baseline Multi-Task-PT. These findings underscore the importance of a tailored multi-task framework for low-resource OoD adaptation.

4.3 Evaluating Downstream Spoken Language Models

We use OPT-1.3B (Zhang et al., 2022) as SLM backbones to evaluate the language modeling 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 SSL and SSL/SL initializations). MAdaPT-Reptile comes a close second, with especially strong zero-shot performance.

4.4 Evaluating on Phoneme Discovery

To further investigate the adaptability of the proposed methods, we compare them with performant speech SSL models, HuBERT (Hsu et al., 2021) and DinoSR (Liu et al., 2023) trained under the OoD Multi-Task-PT setup, on the DiscoPhon benchmark (Poli et al., 2026). This benchmark targets the automatic discovery of phoneme inventories from limited raw speech (10 hours maximum, in 6 development and 6 test languages). Models are evaluated by mapping the discrete speech units to their most frequently associated phonemes prior to evaluation (fixing the number of units either to 256 in the many-to-one setting or to the ground truth number of phonemes in the one-to-one setting). Evaluating our models in the many-to-one setting, we report the following metrics: 1) phone error rate (**PER**); 2) $\boldsymbol{R}$**-value** (Räsänen et al., 2009) and $\boldsymbol{F_1}$ segmentation scores; 3) phone-normalized mutual information (**PNMI**), measuring the uncertainty about a phone label eliminated by a predicted unit; 4) **ABX** on continuous representations (within- and

Method	Backbone Model	PER ↓	R-value ↑	F_1 ↑	PNMI ↑	ABX ↓
Multi-Task-PT [SSL]	HuBERT	98.02	24.37	61.16	57.56	5.49
Multi-Task-PT [SSL]	DinoSR	69.79	43.96	66.63	65.91	6.64
Multi-Task-PT [SSL]	SpidR	48.63	61.84	71.10	69.34	4.47
MAdaPT-Reptile	SpidR	<u>37.05</u>	**75.01**	**78.14**	<u>69.42</u>	<u>4.11</u>
MAdaPT-FOBLO	SpidR	**36.58**	<u>73.91</u>	<u>77.51</u>	**71.64**	**4.10**

Table 3: **Results on phoneme discovery.** We present the scores on the DiscoPhon benchmark after finetuning models on 10 hours of language data and mapping their 256 speech units to the ground truth phonemes. Models were individually finetuned on one of six test languages and aggregate results across languages is presented here. MAdaPT-FOBLO outperforms alternative speech SSL models (HuBERT, DinoSR) and is on par with MAdaPT-Reptile. Both MAdaPT-FOBLO and MAdaPT-Reptile are initialized using Multi-Task-PT [SSL/SL] here. Complete results are provided in Appendix C.3. Best scores (in %) are in **bold** and second best are <u>underlined</u>.

across-conditions). As shown in Table 3, **SpidR-Adapt** (i.e., MAdaPT-FOBLO) consistently outperforms the alternative speech SSL models (HuBERT, DinoSR) with the alternative meta-learning framework (Reptile) also demonstrating strong performance. 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. Improving over in-domain model performance with as little as 1 hour of target-language audio, **SpidR-Adapt** is over $100\times$ more data efficient than traditional multi-task methods and demonstrates the 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. Further hyperparameter exploration is needed to identify superior configurations of SpidR-Adapt—some initial explorations in this direction are summarized in Appendix E. 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

MP acknowledges PhD funding from Agence de l'Innovation de Défense and HPC resources from GENCI-IDRIS (Grant 2023-AD011014368). 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). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. YLC contributed to this work while at Meta.

References

Divyanshu Aggarwal, Ashutosh Sathe, and Sunayana Sitaram. 2025. Improving cross lingual transfer by pretraining with active forgetting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2378, Suzhou, China. Association for Computational Linguistics.

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, Marseille, France. European Language Resources Association.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Can Balioglu, Alexander Erben, Martin Gleize, Artyom Kozhevnikov, Ilia Kulikov, 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 nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533.

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.

Alejandrina Cristia. 2023. A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations. *Developmental Science*, 26(1):e13265.

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.

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 *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 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 Lakhotia, 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. 1979. Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America*, 66(6):1668–1679.

Patricia K. Kuhl. 2004. Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5:831–843.

Kushal Lakhotia, 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.

Alexander H. Liu, Heng-Jui Chang, Michael Auli, Wei-Ning Hsu, and Jim Glass. 2023. DinoSR: Self-Distillation and Online Clustering for Self-supervised Speech Representation Learning. *Advances in Neural Information Processing Systems*, 36:58346–58362.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*, pages 498–502.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 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, Valencia, Spain. Association for Computational Linguistics.

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. *Preprint*, arXiv:2505.02692.

Maxime Poli, Manel Khentout, Angelo Ortiz Tandazo, Ewan Dunbar, Emmanuel Chemla, and Emmanuel Dupoux. 2026. DiscoPhon: Benchmarking the unsupervised discovery of phoneme inventories with discrete speech units. *Preprint*, arXiv:2603.18612.

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*.

Yuehan Qin, Yichi Zhang, Yi Nian, Xueying Ding, and Yue Zhao. 2025. MetaOOD: Automatic selection of OOD detection models. In *The Thirteenth International Conference on Learning Representations*.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 50865–50892. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Okko Johannes Räsänen, Unto Kalervo Laine, and Toomas Altosaar. 2009. An improved speech segmentation quality measure: the r-value. In *Interspeech 2009*, pages 1851–1854.

Jenny R. Saffran, Janet F. Werker, and Lynne A. Werner. 2007. *The Infant's Auditory World: Hearing, Speech, and the Beginnings of Language*, chapter 2. John Wiley & Sons, Ltd.

Thomas Schatz. 2016. *ABX-Discriminability Measures and Applications*. Theses, 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. *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.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, 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 unlabeled 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 are available in the accompanying GitHub repository.

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 `Common Voice` (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 by 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, which results in a labeled dataset of 372 hours.

To compute 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 the development languages, we use data from `Common Voice` (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	*5700 hours* VP 19 langs. (w/o target adaptation langs.)	*10 minutes, 1 hour, 10 hours* CV Swahili CV Tamil CV Thai CV Turkish CV Ukrainian
Test	*6000 hours* VP English VP French VP German	*5700 hours* VP 19 langs. (w/o target adaptation langs.)	*10 minutes, 1 hour, 10 hours, 100 hours* *(subset of In-Domain Training set)* VP English VP French VP German

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

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 monotask and the OoD multi-task models. In interleaved supervised pre-training (i.e., Multi-Task-PT [SSL/SL]), every tenth step is backpropagated using phone supervised loss (hence in equation 8, $\lambda = 0$ if step $\bmod 10 = 0$, else 1). For prediction of supervised labels, language-specific classifier heads (19 heads in total) are attached to the 8^{th} transformer layer of the SpidR model. Here, the 8^{th} 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, with 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 $\mathbf{D}_\ell^u$, we use a randomly chosen 10-hour data chunk from a randomly chosen source language. For the outer loop otimization, 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 6^{th} layer of the model; while, for the interleaved-supervised initialization, it is applied to the 8^{th} layer (staying consistent with the supervised layer during meta-initialization). The FOBLO supervised layers are selected based on best performing layers of the meta-initialization models computed on the development language set. When computing ABX scores, we thereby report results from the 6^{th} and 8^{th} layers for the self- and interleaved-supervised models, respectively. For Reptile, since no layer-specific optimization is employed, we identify the best-performing layers of models through exploration on the development set, finding 6^{th} and 8^{th} layers to be optimal for the self- and interleaved-supervised models, respectively, across all adaptation scales.

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 performance 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 5×10^{-5}. The detailed scheduler is 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 5×10^{-4} or 5×10^{-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), and hence are the toplines. 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 in-domain topline performance is not reported in the table. Across both tables, our proposed MAdaPT-FOBLO consistently outperforms the multi-task 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. A K-means model (with 256 units) is trained on the model embeddings to produce discrete tokens for language modeling. OPT-1.3B models (Zhang et al., 2022) are used as the SLMs, trained with fairseq2 (Balioglu et al., 2023) and following the architectural decisions made by previous works (Hassid et al., 2023; Poli et al., 2025b). The full 60k hour dataset of Libri-Light (Kahn et al., 2020) is used for training. We train on 16 GPUs, with a context length of 2048, and a batch of at most 40960 tokens, for 150000 steps. The learning rate is set at $1e-3$ with a 1000-step warmup period and with a cosine annealing schedule. Remaining hyperparameters follow OPT-1.3B defaults. We select the checkpoint with the lowest validation loss.

C.3 Detailed Results on Phoneme Discovery

DiscoPhon (Poli et al., 2026) is a benchmark specifically designed to investigate the abilities of speech representation models to encode phonemic information in a low resource setting. The benchmark includes 6 development languages (Swahili, Tamil, Thai, Ukrainian, Turkish, and German) and 6 test languages (French, English, Japanese, Mandarin, Wolof, and Basque) selected to span a diverse range of phonemic categories. Note that the development and test language sets in the benchmark differ from our previous experiments but are still disjoint from our training set.

In the current experiment, we apply our previously tuned Multi-Task-PT, MAdaPT-Reptile, and MAdaPT-FOBLO models with SpidR as backbone. Our models are trained on 19 `VoxPopuli` languages (Wang et al., 2021). We compare our approaches to OoD HuBERT (Hsu et al., 2021) and DinoSR (Liu

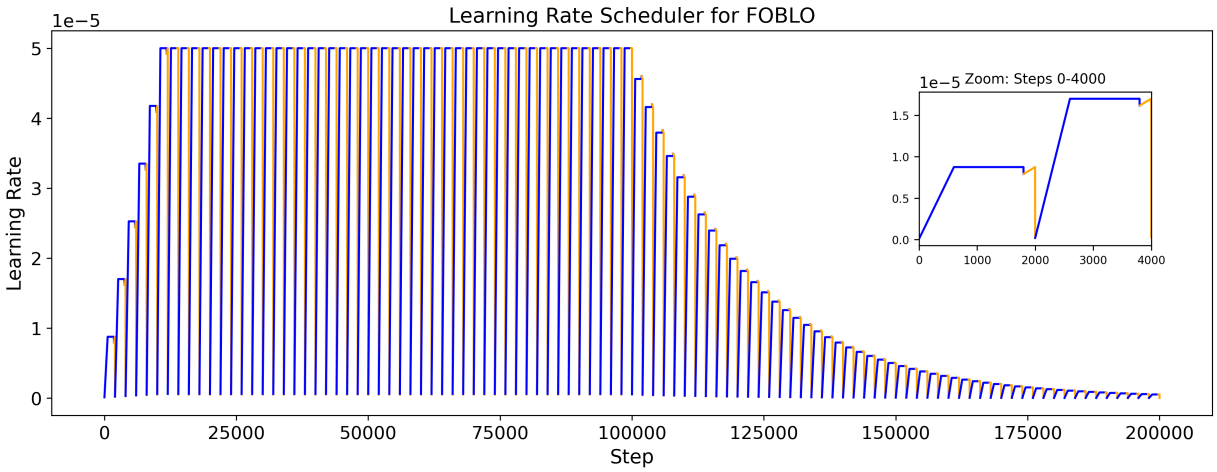

Figure 3: **Learning rate scheduler for FOBLO.** We use blue and orange to represent the learning rate for self-supervised inner steps and supervised outer steps, respectively. The overall training has 200,000 steps. The learning rate scheduler alternates between inner loop and outer loop steps within 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.

et al., 2023) models trained on 20 `VoxPopuli` languages. Notably, supervised ASR encoders such as Whisper (Radford et al., 2023) exhibit poor phonemic discriminability despite strong recognition accuracy (Poli et al., 2025b), and are therefore not included in our comparisons. For HuBERT and DinoSR, the best performing layer for ABX as determined by the development languages is used. For discrete unit metrics (PNMI and PER) a K-means is trained on HuBERT embeddings while the model codebooks are used for DinoSR and SpidR (with 256 codewords across all models).

Test languages results are reported in Table 8 and development languages results in Table 9. As can be observed, on aggregate, our proposed MAdaPT-FOBLO achieves improved performance over alternate speech SSL models (HuBERT, DinoSR), with MAdaPT-Reptile also demonstrating strong performance.

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 subtables 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, reach-

Method	0h	10m	1h	10h	100h	Avg. (w/o 0h) ↓
			Within-Speaker ABX			
In-Domain Mono-Task-PT	6000 *hours training; Topline 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	3.78	3.69	4.01
Multi-Task-PT [SSL/SL]	4.10	4.08	3.94	3.78	3.71	3.88
+ MAdaPT-Reptile	**3.89**	**3.94**	**3.82**	**3.62**	**3.66**	**3.76**
+ MAdaPT-FOBLO	4.00	4.07	3.84	**3.62**	3.70	3.80
			Across-Speaker ABX			
In-Domain Mono-Task-PT	6000 *hours training; Topline 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	5.06
+ MAdaPT-Reptile	**5.19**	**5.16**	4.97	4.78	4.79	**4.93**
+ MAdaPT-FOBLO	5.28	5.23	**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 with SSL/SL regimes show superior performance, surpassing In-Domain Mono-Task-PT with limited data. The best scores are in **bold** and second best are underlined.

ing 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 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).

E Extended Ablation Studies

E.1 Impact of Interleaved Steps in the Multi-task Pretraining

We conduct an ablation over the number of interleaved supervision steps $\in \{2, 10, 50, 100\}$ (Table 15). We observe a trade-off between zero-shot and few-shot performance: more frequent interleaving improves zero-shot transfer by injecting stronger supervised structure, but degrades few-

Method	0h	10m	1h	10h	Avg. (w/o 0h) ↓
Within-Speaker ABX					
Multi-Task-PT [SSL]	8.84	8.34	7.33	6.12	7.26
+ MAdaPT-Reptile	7.97	7.10	6.61	5.79	6.50
+ MAdaPT-FOBLO	13.89	7.70	6.21	**5.29**	6.40
Multi-Task-PT [SSL/SL]	7.57	6.84	6.20	5.65	6.23
+ MAdaPT-Reptile	**7.05**	6.40	6.04	5.59	6.01
+ MAdaPT-FOBLO	7.73	**6.25**	**5.99**	5.56	**5.93**
Across-Speaker ABX					
Multi-Task-PT [SSL]	10.48	9.77	8.18	6.80	8.25
+ MAdaPT-Reptile	9.14	8.05	7.50	6.58	7.38
+ MAdaPT-FOBLO	16.28	8.61	6.78	**5.82**	7.07
Multi-Task-PT [SSL/SL]	8.23	7.40	6.50	6.06	6.65
+ MAdaPT-Reptile	**7.72**	6.93	6.40	6.05	6.46
+ MAdaPT-FOBLO	8.40	**6.82**	**6.37**	5.96	**6.38**

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.

shot adaptability. Crucially, few-shot performance is roughly stable across steps $\{2, 10\}$, and only degrades noticeably at steps ≥ 50. We select steps = 10 as it offers the best balance: competitive few-shot performance with substantially improved zero-shot transfer over steps = 2.

E.2 Analysis on Supervised Steps N during meta-training

To investigate the impact of supervised steps during meta-training, we systematically vary $N \in 0, 40, 200, 1000$ under two pre-training regimes: Multi-Task-PT [SSL/SL] and Multi-Task-PT [SSL]. As shown in Table 16, with SSL/SL pre-training, increasing N improves few-shot performance while slightly degrading zero-shot transfer, revealing a trade-off between initialization generality and adaptation depth. With SSL-only pre-training, $N = 0$ leads to codebook collapse, confirming that supervised inner loop steps are essential for producing informative meta-gradients. Note that we use the same hyperparameter configuration ($N = 200$) across all model variants, which may not be optimal for each setting; we leave this exploration to future work.

Method	0h	10m	1h	10h	100h	Avg. (w/o 0h) ↓
			sWUGGY (in-vocab and out-of-vocab)			
In-Domain Mono-Task-PT			6000 *hours training; Topline score is* 64.51.			
Multi-Task-PT [SSL]	63.74	64.07	65.28	64.85	65.94	65.04
+ MAdaPT-Reptile	64.17	63.99	64.53	64.54	64.88	64.49
+ MAdaPT-FOBLO	54.82	64.42	64.60	66.30	**67.46**	65.70
Multi-Task-PT [SSL/SL]	62.68	65.17	65.76	66.53	66.02	65.87
+ MAdaPT-Reptile	**65.65**	**66.37**	66.92	**67.79**	67.25	**67.08**
+ MAdaPT-FOBLO	64.32	66.04	**66.99**	67.26	67.26	66.89
			sBLIMP			
In-Domain Mono-Task-PT			6000 *hours training; Topline score is* 56.94.			
Multi-Task-PT [SSL]	55.73	56.07	57.12	56.46	57.94	56.90
+ MAdaPT-Reptile	56.23	56.88	56.75	57.54	56.84	57.00
+ MAdaPT-FOBLO	52.93	**57.60**	57.46	58.17	**58.11**	**57.84**
Multi-Task-PT [SSL/SL]	56.37	56.59	56.99	56.74	56.60	56.73
+ MAdaPT-Reptile	**56.58**	56.93	56.94	**58.47**	58.01	57.59
+ MAdaPT-FOBLO	55.48	57.38	**57.99**	57.54	57.87	57.69
			Spoken tSC			
In-Domain Mono-Task-PT			6000 *hours training; Topline score is* 74.36.			
Multi-Task-PT [SSL]	70.99	71.26	71.47	72.22	71.74	71.67
+ MAdaPT-Reptile	72.86	72.38	72.12	71.69	72.44	72.16
+ MAdaPT-FOBLO	69.50	73.88	73.50	72.70	73.61	73.42
Multi-Task-PT [SSL/SL]	75.11	**74.25**	**74.57**	74.79	75.59	**74.80**
+ MAdaPT-Reptile	74.95	73.02	74.20	74.04	**75.75**	74.25
+ MAdaPT-FOBLO	**76.12**	73.93	73.99	**74.84**	75.43	74.55

Table 7: **Detailed results of spoken language modeling metrics: sWUGGY, sBLIMP, and spoken tSC (in %).** MAdaPT-FOBLO shows consistently superior performance across tasks. The best scores are in **bold** and the second best are underlined.

Method	Backbone Model	0h	10m	1h	10h	Avg. (w/o 0h)
		PER ↓				
Multi-Task-PT [SSL]	HuBERT	127.02	114.12	101.03	98.02	104.39
Multi-Task-PT [SSL]	DinoSR	86.32	78.40	70.14	69.79	72.78
Multi-Task-PT [SSL]	SpidR	85.40	75.24	56.51	48.63	60.13
+ MAdaPT-Reptile	SpidR	153.77	66.00	60.03	54.42	60.15
+ MAdaPT-FOBLO	SpidR	87.40	66.94	41.93	**35.49**	48.12
Multi-Task-PT [SSL/SL]	SpidR	**50.82**	40.94	38.77	37.99	39.23
+ MAdaPT-Reptile	SpidR	110.71	40.77	**36.88**	37.05	38.23
+ MAdaPT-FOBLO	SpidR	51.06	**39.32**	37.23	36.58	**37.71**
		R-value ↑				
Multi-Task-PT [SSL]	HuBERT	9.39	13.82	22.81	24.37	20.34
Multi-Task-PT [SSL]	DinoSR	38.93	41.35	45.38	43.96	43.56
Multi-Task-PT [SSL]	SpidR	39.96	45.46	57.29	61.84	54.86
+ MAdaPT-Reptile	SpidR	-7.66	52.87	55.61	58.50	55.66
+ MAdaPT-FOBLO	SpidR	43.98	54.61	70.34	73.44	66.13
Multi-Task-PT [SSL/SL]	SpidR	**67.27**	71.21	72.03	72.78	72.01
+ MAdaPT-Reptile	SpidR	26.05	72.37	**75.28**	**75.01**	**74.22**
+ MAdaPT-FOBLO	SpidR	67.26	**72.47**	73.50	73.91	73.29
		F_1 ↑				
Multi-Task-PT [SSL]	HuBERT	58.18	58.98	60.65	61.16	60.26
Multi-Task-PT [SSL]	DinoSR	62.44	65.38	66.65	66.63	66.22
Multi-Task-PT [SSL]	SpidR	63.88	65.46	69.69	71.10	68.75
+ MAdaPT-Reptile	SpidR	53.18	66.01	66.94	66.75	66.57
+ MAdaPT-FOBLO	SpidR	29.55	69.54	76.06	77.05	74.22
Multi-Task-PT [SSL/SL]	SpidR	**74.50**	75.99	76.89	77.52	76.80
+ MAdaPT-Reptile	SpidR	60.46	**76.78**	**77.72**	**78.14**	**77.55**
+ MAdaPT-FOBLO	SpidR	74.24	76.62	77.06	77.51	77.06
		PNMI ↑				
Multi-Task-PT [SSL]	HuBERT	49.49	52.74	55.72	57.56	55.34
Multi-Task-PT [SSL]	DinoSR	54.82	61.06	64.85	65.91	63.94
Multi-Task-PT [SSL]	SpidR	58.25	61.89	67.16	69.34	66.13
+ MAdaPT-Reptile	SpidR	40.45	61.15	62.96	64.62	62.91
+ MAdaPT-FOBLO	SpidR	10.36	62.03	68.70	70.78	67.17
Multi-Task-PT [SSL/SL]	SpidR	**66.69**	**70.74**	**72.61**	**73.11**	**72.15**
+ MAdaPT-Reptile	SpidR	48.90	67.86	68.96	69.42	68.75
+ MAdaPT-FOBLO	SpidR	66.48	70.12	71.25	71.64	71.00
		ABX ↓				
Multi-Task-PT [SSL]	HuBERT	7.79	7.63	6.77	5.49	6.63
Multi-Task-PT [SSL]	DinoSR	8.48	7.73	7.20	6.64	7.19
Multi-Task-PT [SSL]	SpidR	6.45	6.05	5.23	4.47	5.25
+ MAdaPT-Reptile	SpidR	5.76	5.25	4.91	4.30	4.82
+ MAdaPT-FOBLO	SpidR	12.30	5.76	4.46	**3.88**	4.70
Multi-Task-PT [SSL/SL]	SpidR	5.54	4.82	4.28	4.13	4.41
+ MAdaPT-Reptile	SpidR	**5.07**	4.61	4.28	4.11	4.33
+ MAdaPT-FOBLO	SpidR	5.60	**4.56**	**4.26**	4.10	**4.31**

Table 8: **Detailed results (in %) on DiscoPhon on the 6 test languages** (Mandarin Chinese, English, Basque, French, Japanese, Wolof). Average ABX for within- and across-speaker conditions is reported. MAdaPT-FOBLO and MAdaPT-Reptile outperform alternate speech SSL models (HuBERT, DinoSR). The best scores are in **bold** and the second best are underlined.

Method	Backbone Model	0h	10m	1h	10h	Avg. (w/o 0h)
		PER ↓				
Multi-Task-PT [SSL]	HuBERT	118.26	105.34	100.19	95.90	100.48
Multi-Task-PT [SSL]	DinoSR	79.75	77.12	70.97	65.18	71.09
Multi-Task-PT [SSL]	SpidR	81.41	76.02	58.22	51.47	61.90
+ MAdaPT-Reptile	SpidR	146.87	67.26	61.14	55.41	61.27
+ MAdaPT-FOBLO	SpidR	85.51	65.95	44.65	**37.80**	49.46
Multi-Task-PT [SSL/SL]	SpidR	**47.80**	45.63	40.37	39.73	41.91
+ MAdaPT-Reptile	SpidR	100.44	**42.89**	**38.69**	38.67	**40.08**
+ MAdaPT-FOBLO	SpidR	48.19	43.83	39.69	39.93	41.15
		R-value ↑				
Multi-Task-PT [SSL]	HuBERT	19.63	24.23	28.52	30.48	27.74
Multi-Task-PT [SSL]	DinoSR	47.38	46.73	48.35	51.37	48.82
Multi-Task-PT [SSL]	SpidR	46.18	48.69	59.44	62.64	56.92
+ MAdaPT-Reptile	SpidR	0.05	54.58	58.62	59.00	57.40
+ MAdaPT-FOBLO	SpidR	42.29	58.50	72.03	73.65	68.06
Multi-Task-PT [SSL/SL]	SpidR	**72.03**	71.17	74.05	74.51	73.24
+ MAdaPT-Reptile	SpidR	37.86	**73.39**	**75.86**	**75.87**	**75.04**
+ MAdaPT-FOBLO	SpidR	71.85	72.30	74.20	74.51	73.67
		F_1 ↑				
Multi-Task-PT [SSL]	HuBERT	58.87	59.88	60.41	61.45	60.58
Multi-Task-PT [SSL]	DinoSR	61.98	63.59	64.76	66.01	64.79
Multi-Task-PT [SSL]	SpidR	63.13	64.21	67.37	68.62	66.73
+ MAdaPT-Reptile	SpidR	53.20	63.96	66.13	64.53	64.87
+ MAdaPT-FOBLO	SpidR	26.25	68.76	74.13	74.42	72.44
Multi-Task-PT [SSL/SL]	SpidR	**73.83**	74.25	**75.77**	76.19	75.40
+ MAdaPT-Reptile	SpidR	60.42	**74.62**	75.51	**76.39**	**75.51**
+ MAdaPT-FOBLO	SpidR	73.48	74.22	75.62	76.14	75.33
		PNMI ↑				
Multi-Task-PT [SSL]	HuBERT	47.01	51.04	52.18	54.66	52.62
Multi-Task-PT [SSL]	DinoSR	52.46	57.36	60.60	62.70	60.22
Multi-Task-PT [SSL]	SpidR	55.09	58.22	63.20	65.83	62.42
+ MAdaPT-Reptile	SpidR	37.80	58.22	60.13	62.06	60.14
+ MAdaPT-FOBLO	SpidR	9.78	59.62	65.11	67.44	64.06
Multi-Task-PT [SSL/SL]	SpidR	**64.01**	**67.14**	**69.28**	**70.01**	**68.81**
+ MAdaPT-Reptile	SpidR	47.01	64.57	66.37	66.97	65.97
+ MAdaPT-FOBLO	SpidR	63.75	66.43	68.19	68.67	67.76
		ABX ↓				
Multi-Task-PT [SSL]	HuBERT	10.23	9.49	9.08	7.68	8.75
Multi-Task-PT [SSL]	DinoSR	11.44	10.27	9.94	9.02	9.74
Multi-Task-PT [SSL]	SpidR	8.79	8.26	7.12	6.02	7.13
+ MAdaPT-Reptile	SpidR	7.80	6.95	6.48	5.75	6.39
+ MAdaPT-FOBLO	SpidR	14.34	7.46	5.98	**5.15**	6.20
Multi-Task-PT [SSL/SL]	SpidR	7.19	6.54	5.84	5.40	5.93
+ MAdaPT-Reptile	SpidR	**6.73**	6.11	5.69	5.34	5.72
+ MAdaPT-FOBLO	SpidR	7.30	**6.01**	**5.67**	5.31	**5.66**

Table 9: **Detailed results (in %) on DiscoPhon on the 6 development languages** (German, Swahili, Tamil, Thai, Turkish, Ukrainian). Average ABX for within- and across-speaker conditions is reported. MAdaPT-FOBLO and MAdaPT-Reptile outperform alternate speech SSL models (HuBERT, DinoSR). The best scores are in **bold** and the second best are underlined.

Method	Active Forgetting	0h	10m	1h	10h	Avg. (w/o 0h) ↓
		Within-Speaker ABX				
Multi-Task-PT [SSL] + MAdaPT-FOBLO	✗	10.68	6.45	6.75	6.63	6.61
	✓	13.89	7.70	6.21	**5.29**	6.40
Multi-Task-PT [SSL/SL] + MAdaPT-FOBLO	✗	**7.45**	6.74	6.10	5.57	6.14
	✓	7.73	**6.25**	**5.99**	5.56	**5.93**
		Across-Speaker ABX				
Multi-Task-PT [SSL] + MAdaPT-FOBLO	✗	12.38	7.13	7.09	6.78	7.00
	✓	16.28	8.61	6.78	**5.82**	7.07
Multi-Task-PT [SSL/SL] + MAdaPT-FOBLO	✗	**8.12**	7.24	6.46	6.01	6.57
	✓	8.40	**6.82**	**6.37**	5.96	**6.38**

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) ↓
		Within-Speaker ABX					
In-Domain Mono-Task-PT	N.A.	6000 *hours training; Topline score is* 4.10.					
Multi-Task-PT [SSL] + MAdaPT-FOBLO	✗	21.89	4.40	4.54	4.37	4.20	4.38
	✓	10.05	4.51	4.05	3.78	3.69	4.01
Multi-Task-PT [SSL/SL] + MAdaPT-FOBLO	✗	**3.99**	**4.02**	3.87	3.71	**3.67**	3.82
	✓	4.00	4.07	**3.84**	**3.62**	3.70	**3.80**
		Across-Speaker ABX					
In-Domain Mono-Task-PT	N.A.	6000 *hours training; Topline score is* 5.47.					
Multi-Task-PT [SSL] + MAdaPT-FOBLO	✗	29.11	5.62	5.77	5.56	5.33	5.57
	✓	15.12	5.96	5.26	4.92	4.83	5.24
Multi-Task-PT [SSL/SL] + MAdaPT-FOBLO	✗	5.29	5.32	5.01	4.84	4.80	4.99
	✓	**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) ↓
		Within-Speaker ABX				
MAdaPT-FOBLO	Random	35.83	31.22	35.73	37.37	34.77
	Multi-Task-PT [SSL]	13.89	7.70	6.21	**5.29**	6.40
	Multi-Task-PT [SSL/SL]	**7.73**	**6.25**	**5.99**	5.56	**5.93**
		Across-Speaker ABX				
MAdaPT-FOBLO	Random	8.12	7.24	6.46	6.01	6.57
	Multi-Task-PT [SSL]	16.28	8.61	6.78	**5.82**	7.07
	Multi-Task-PT [SSL/SL]	**8.40**	**6.82**	**6.37**	5.96	**6.38**

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) ↓
		Within-Speaker ABX					
In-Domain Mono-Task-PT	N.A.	6000 *hours training; Topline score is* 4.10.					
MAdaPT-FOBLO	Random	32.68	25.12	24.61	23.75	24.52	24.50
	Multi-Task-PT [SSL]	10.05	4.51	4.05	3.78	**3.69**	4.01
	Multi-Task-PT [SSL/SL]	**4.00**	**4.07**	**3.84**	**3.62**	3.70	**3.80**
		Across-Speaker ABX					
In-Domain Mono-Task-PT	N.A.	6000 *hours training; Topline score is* 5.47.					
MAdaPT-FOBLO	Random	38.75	33.25	32.81	32.31	32.78	32.79
	Multi-Task-PT [SSL]	15.12	5.96	5.26	4.92	4.83	5.24
	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) ↓
	Within-Speaker ABX				
0.001	**7.59**	6.33	6.06	5.59	6.00
0.01	7.73	**6.25**	5.99	**5.56**	**5.93**
0.1	11.63	6.31	**5.97**	5.61	5.96
1	8.23	6.63	6.09	5.58	6.10
	Across-Speaker ABX				
0.001	**8.22**	6.92	6.42	6.05	6.46
0.01	8.42	**6.82**	**6.38**	**5.96**	**6.39**
0.1	12.66	6.84	6.41	6.04	6.43
1	9.10	7.15	6.51	6.05	6.57

β	0h	10m	1h	10h	100h	Avg. (w/o 0h) ↓
	Within-Speaker ABX					
0.001	4.10	4.02	3.87	3.66	3.69	3.81
0.01	**4.00**	4.07	3.84	3.62	3.70	3.80
0.1	6.20	**3.91**	3.81	3.57	3.64	**3.73**
1	5.89	4.04	**3.72**	**3.53**	**3.62**	**3.73**
	Across-Speaker ABX					
0.001	5.39	5.25	4.97	4.79	4.77	4.95
0.01	**5.28**	5.23	4.96	4.76	4.77	4.93
0.1	8.56	**5.16**	5.00	4.76	4.75	4.92
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.

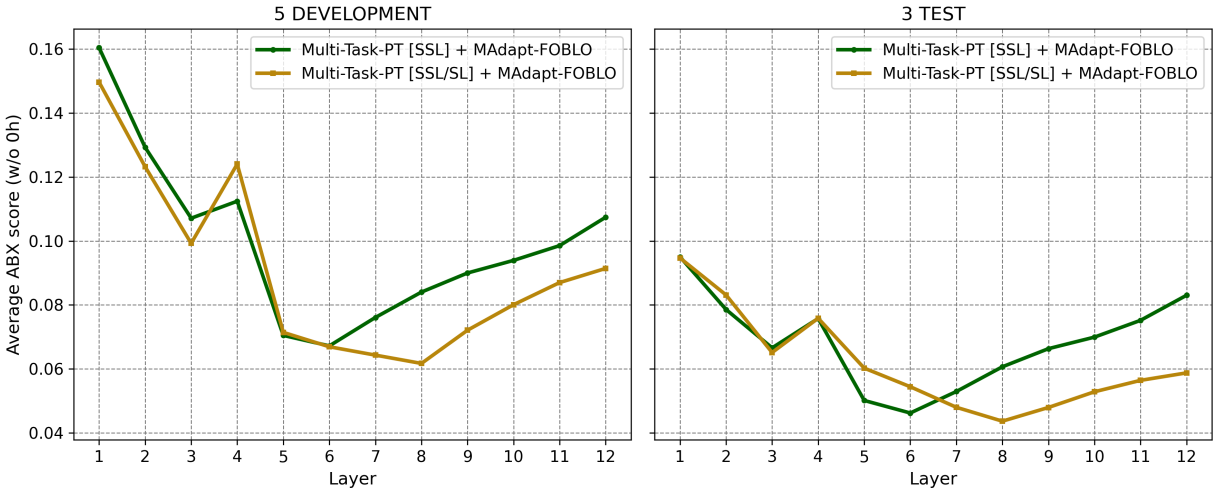

Figure 4: **Layer-wise analysis on the model's discriminability over phonemes.** We present the ABX scores averaged over the target languages, and across the two within- and across-speaker conditions for two evaluation sets: (a) 5 development and (b) 3 test language sets. We report results for our proposed 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 Multi-Task-PT[SSL] initialization and 8 for Multi-Task-PT[SSL/SL] initialization.

Interleaved Steps in Multi-Task-PT [SSL/SL]	**0h**	**10m**	**1h**	**10h**	**Avg. (w/o 0h) ↓**
Within-Speaker ABX					
2	9.49	6.66	6.01	5.56	6.08
10	7.57	6.84	6.20	5.65	6.23
50	7.13	6.65	6.36	6.14	6.38
100	7.01	6.81	6.41	6.14	6.45
Across-Speaker ABX					
2	10.17	7.25	6.47	5.93	6.55
10	8.23	7.40	6.50	6.06	6.65
50	7.88	7.30	6.83	6.52	6.88
100	7.71	7.42	6.92	6.70	7.01

Table 15: **Impact of interleaved steps in the multi-task pretraining on 5 DEVELOPMENT languages**. Overall performance is stable across higher frequency interleaving with smaller steps {2, 10} and degrades at steps larger than 50, revealing a trade-off between zero-shot transfer and few-shot adaptability.

Method	Supervised Steps N	0h	10m	1h	10h	Avg. (w/o 0h) ↓
		Within-Speaker ABX				
Multi-Task-PT [SSL] + MAdaPT-FOBLO	0	*codebook collapse*				
	40	23.16	7.81	6.83	5.62	6.76
	200	13.89	7.70	6.21	5.29	6.40
	1000	8.02	6.62	5.80	5.16	5.86
Multi-Task-PT [SSL/SL] + MAdaPT-FOBLO	0	7.62	6.53	6.16	5.64	6.11
	40	7.68	6.23	6.10	5.62	5.98
	200	7.73	6.25	5.99	5.56	5.93
	1000	7.89	6.31	5.99	5.54	5.95
		Across-Speaker ABX				
Multi-Task-PT [SSL] + MAdaPT-FOBLO	0	*codebook collapse*				
	40	27.31	8.78	7.45	6.36	7.53
	200	16.28	8.61	6.78	5.82	7.07
	1000	9.28	7.42	6.45	5.60	6.49
Multi-Task-PT [SSL/SL] + MAdaPT-FOBLO	0	8.35	6.98	6.48	6.03	6.50
	40	8.34	6.77	6.36	5.96	6.36
	200	8.40	6.82	6.37	5.96	6.38
	1000	8.51	6.77	6.28	5.91	6.32

Table 16: **Impact of supervised steps during meta-training on 5 DEVELOPMENT languages**. Multi-Task-PT [SSL/SL] makes the meta-model more robust to variation in the number of supervised steps compared to Multi-Task-PT [SSL].