SSVD: Structured SVD for Parameter-Efficient Fine-Tuning and Benchmarking under Domain Shift in ASR

Pu Wang
Department of Electrical Engineering
KU Leuven
Leuven, Belgium
pu.wang@esat.kuleuven.be

Shinji Watanabe
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
shinjiw@ieee.org

Hugo Van hamme
Department of Electrical Engineering
KU Leuven
Leuven, Belgium
hugo.vanhamme@esat.kuleuven.be

***Abstract*—Parameter-efficient fine-tuning (PEFT) has emerged as a scalable solution for adapting large foundation models. While low-rank adaptation (LoRA) is widely used in speech applications, its state-of-the-art variants, e.g., VeRA, DoRA, PiSSA, and SVFT, are developed mainly for language and vision tasks, with limited validation in speech. This work presents the first comprehensive integration and benchmarking of these PEFT methods within ESPnet. We further introduce structured SVD-guided (SSVD) fine-tuning, which selectively rotates input-associated right singular vectors while keeping output-associated vectors fixed to preserve semantic mappings. This design enables robust domain adaptation with minimal trainable parameters and improved efficiency. We evaluate all methods on domain-shifted speech recognition tasks, including child speech and dialectal variation, across model scales from 0.1B to 2B. All implementations are released in ESPnet to support reproducibility and future work.**

***Index Terms*—speech recognition, parameter-efficient fine-tuning (PEFT), domain adaptation, low-rank adaptation (LoRA), singular value decomposition (SVD), child and dialectal speech**

I. Introduction

Large-scale foundation models pretrained on multilingual speech corpora, such as OpenAI's Whisper [1], the open-source Whisper-style OWSM [2, 3], NVIDIA's Canary [4], etc., have demonstrated strong generalization and adaptability across diverse speech recognition tasks. Fine-tuning these models for downstream or personalized applications has therefore become the dominant paradigm. However, these models are typically trained on broad yet generic datasets dominated by standard accents and adult speech, limiting their effectiveness under domain shifts involving regional accents, dialectal variation, child speech, and other low-resource speech domains. In such scenarios, mismatches in acoustic features, linguistic structures, and articulatory behaviors can significantly degrade performance.

Recent studies on scaling laws of large-scale multilingual automatic speech recognition (ASR) models have shown that increasing model size substantially improves recognition accuracy, particularly for low-resource languages [5, 6, 7, 8]. Nonetheless, rapid scaling introduces practical challenges, as adapting large models to specific downstream or out-of-domain scenarios can become prohibitively expensive in terms of computational resources and storage requirements.

To address the limitations of full-model fine-tuning, parameter-efficient fine-tuning (PEFT) methods have been actively explored, particularly with large language models (LLMs) [9, 10, 11, 12]. Among them, low-rank adaptation (LoRA) has received considerable attention for significantly reducing the number of trainable parameters without modifying the original model architecture or introducing additional decoding latency [12, 13, 14, 15]. Specifically, LoRA freezes the pre-trained weight matrices and injects two trainable low-rank matrices to approximate weight updates. While originally proposed for LLMs, LoRA has also attracted increasing interest in speech tasks and has been successfully applied in various speech applications [15, 16, 17, 18, 19, 20]. Wang et al. [17] explore low-rankness in speech models and adapt it to spoken language understanding tasks. Liu et al. [18] adapt Whisper with LoRA to child speech ASR. Xu et al. [19] further investigate using LoRA to mitigate forgetting when fine-tuning Whisper to multiple languages. Song et al. [20] propose LoRA with a mixture of experts (MoE) to build an extensible multilingual LoRA-Whisper ASR model.

Although LoRA effectively reduces parameter count, it still incurs noticeable storage overhead as foundation models scale and suffers from convergence inefficiencies [21, 22]. To mitigate these issues, vector-based random matrix adaptation (VeRA) has been recently proposed as a more lightweight alternative [21]. VeRA randomly initializes and freezes the low-rank matrices of LoRA, sharing these matrices across layers while updating only two scaling vectors. Thus, further reducing the memory footprint with acceptable performance drops on natural language processing (NLP) and vision tasks. Additionally, recent studies in NLP have shown a persistent performance gap between LoRA and full fine-tuning. Liu et al. [23] analyzed the dynamic differences in weights updated between LoRA-tuned and fully fine-tuned models and proposed weight-decomposed LoRA (DoRA). DoRA reparameterizes LoRA into separate directional and magnitude components, demonstrating that tuning primarily one component is sufficient to match full fine-tuning performance in downstream NLP and vision tasks.

To further improve the training efficiency of LoRA variants, Meng et al. [22] and Lingam et al. [24] introduce singular value decomposition (SVD) to guide parameter tuning along more meaningful directions. Instead of relying on random initialization within LoRA, Meng et al. [22] proposed PiSSA, which initializes the low-rank matrices of LoRA using the principal singular vectors and values derived from the pre-trained weight matrix. It accelerates convergence and reduces the risk of becoming trapped in suboptimal local minima compared to standard LoRA in NLP tasks. Lingam et al. [24] extended this idea by proposing singular vector guided fine-tuning (SVFT), which further minimizes the number of trainable parameters by freezing the singular vectors entirely and updating only the singular values, significantly reducing parameter counts compared to PiSSA in NLP and vision domains.

Despite standard LoRA showing promising results in ASR, its advanced variants, as mentioned earlier, are primarily developed and studied in natural language and vision tasks, with their application to speech recognition remaining relatively underexplored.

Speech fundamentally differs from text in its dynamic, and highly variable nature, with significant acoustic diversity, pronunciation variability, dialectal shifts, etc. For instance, Flemish (Belgian Dutch) differs from standard Dutch, with a prominent distinction being the use of the *soft* $/g/$, compared to the *hard* $/g/$ commonly spoken in the

Netherlands. Even identical lexical content "Goed geregend, zeg!", such phonetic differences introduce substantial acoustic mismatch. Similarly, child speech presents unique acoustic characteristics, such as phonetic variations arising from developmental articulation patterns (e.g., pronouncing "wabbit" instead of "rabbit"). PEFT methods are thus expected to efficiently adapt to domain shifts in acoustic variation while maintaining consistent semantic recognition.

To address this gap and facilitate deeper understanding and advancement in PEFT specifically for speech, we provide a thorough investigation of LoRA and its leading variants. We implement and comprehensively evaluate these methods on challenging speech recognition scenarios, including low-resource child speech, as well as dialectal variations between Dutch and Flemish, by fine-tuning on OWSM models at multiple scales (from 0.1B to 2B) using the widely adopted open-source ESPnet toolkit. All implemented methods are publicly released and fully integrated into ESPnet toolkit, facilitating reproducibility and enabling further research in speech-specific PEFT.

Meanwhile, this paper introduces a novel structured SVD-guided (SSVD) PEFT approach specifically designed to handle domain shifts in speech recognition tasks. The proposed SSVD method leverages SVD to first decompose pre-trained model weights into right singular vectors, representing the input (acoustic) feature space, and left singular vectors, capturing the output (semantic) feature space. During fine-tuning, our method selectively applies structured rotations to the input-associated right singular vectors, effectively adapting acoustic feature representations to align better with domain-shifted inputs. Meanwhile, the left singular vectors remain fixed, preserving stable semantic mappings. This innovative approach ensures efficient adaptation to acoustic domain shifts while maintaining robust linguistic outputs.

The main contributions of this paper are:

- Providing the first comprehensive benchmarking of SoTA PEFT methods, including LoRA, VeRA, DoRA, PiSSA, and SVFT, within the widely adopted ESPnet framework [25]. These methods are comprehensively evaluated on challenging speech recognition tasks involving low-resource child speech and dialectal variations (Dutch vs. Flemish) at multiple model scales, making all implementations publicly available and fully integrated into ESPnet for reproducibility and future research[1].
- Introducing a novel structured SVD-guided (SSVD) PEFT method that explicitly designed to efficiently adapt large-scale speech foundation models to challenging domain shifts by selectively adapting input-related singular vectors, while preserving semantic mappings through fixed output-related singular vector components.
- Empirically demonstrating that SSVD achieves comparable performance with significantly fewer trainable parameters and higher efficiency than LoRA and SoTA LoRA variants, approaching fully fine-tuned model performance.

II. Related Work for Benchmarking

In this section, we introduce the state-of-the-art (SoTA) LoRA variants evaluated in this work. Table I summarizes their corresponding formulations and trainable parameter counts. Trainable variables in this study are indicated by underlining.

LoRA. LoRA freezes the pre-trained weights of models and injects trainable low-rank matrices into each linear layer. For a pre-trained weight $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$, LoRA parameterizes its fine-tuning update as:

$$\mathbf{W}' = \mathbf{W}_0 + \underline{\mathbf{A}\mathbf{B}}^\top, \tag{1}$$

[1]All implementations including SSVD: https://github.com/wangpuup/SSVD

where $\mathbf{A} \in \mathbb{R}^{m\times r}$, $\mathbf{B} \in \mathbb{R}^{n\times r}$ are trainable low-rank matrices. LoRA significantly reduces the number of trainable parameters to $r \times (m+n)$ by choosing a rank $r \ll \min(m,n)$, but as model size scales, it can still accumulate a non-trivial number of learnable parameters.

VERA. Unlike LoRA, which introduces trainable low-rank matrices for each adapted layer, VeRA employs a single pair of frozen, randomly initialized low-rank matrices shared across all layers. Adaptation is achieved through the learning of small, trainable scaling vectors specific to each layer. For a pre-trained weight $\mathbf{W}_0$, VERA adapt it as:

$$\mathbf{W}' = \mathbf{W}_0 + (\underline{\mathbf{\Lambda}_d}\mathbf{A})(\underline{\mathbf{\Lambda}_b}\mathbf{B})^\top \tag{2}$$

where $\mathbf{A}$ and $\mathbf{B}$ are low-rank random matrices, generated once and kept frozen. $\mathbf{\Lambda}_d$ and $\mathbf{\Lambda}_b$ are diagonal matrices formed from small trainable scaling vectors $b \in \mathbb{R}^r$ and $d \in \mathbb{R}^m$, respectively. Since $\mathbf{A}$ and $\mathbf{B}$ are shared and reproducible from a fixed random seed, VeRA significantly lowers the memory footprint to $r + m$.

DoRA. While VeRA focuses on achieving decent performance with minimal trainable parameters, DoRA aims to bridge the performance gap between LoRA and full fine-tuning. For a pre-trained weight matrix $\mathbf{W}_0$, DoRA decomposes it into a magnitude vector $\mathbf{m} = \|\mathbf{W}_0\|_c$, where $\|.\|_c$ is the vector-wise (column-wise) norm, and a directional matrix $\dfrac{\mathbf{W}_0}{\|\mathbf{W}_0\|_c}$. By analyzing the change in the components during LoRA tuning versus full fine-tuning, Liu et al. [23] observed that LoRA struggles to simultaneously learn both direction and scale, while full fine-tuning tends to adapt primarily one of the two. DoRA hereby simplifies optimization and focuses solely on directional adaptation, injecting low-rank updates only into the directional component:

$$\mathbf{W}' = \underline{\mathbf{m}}\frac{\mathbf{W_0} + \underline{\mathbf{A}\mathbf{B}}^\top}{\|\mathbf{W_0} + \underline{\mathbf{A}\mathbf{B}}^\top\|_c} \tag{3}$$

Compared to LoRA, DoRA introduces an additional trainable magnitude vector $\mathbf{m} \in \mathbb{R}^m$, increasing the trainable parameter count to $r\times(m+n)+m$. However, this increase remains negligible compared to the size of the full model.

PiSSA. Unlike LoRA, VeRA, and DoRA, which rely on randomly initialized low-rank matrices and suffer from inefficient early training or suboptimal local minima, PiSSA initializes the low-rank components $\mathbf{A}$ and $\mathbf{B}$ using the top-k principal singular vectors and values from the SVD of the pre-trained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$.

Given the SVD, with $m \geq n$ (e.g., a feedforward layer):

$$\mathbf{W}_0 = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \tag{4}$$

where, $\mathbf{U} \in \mathbb{R}^{m\times n}$, $\mathbf{V} \in \mathbb{R}^{n\times n}$ contain the left and right singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ respectively, and $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ is a diagonal matrix of singular values σ_i. Fine-tuning is performed along the k principal directions, while the residual components with the $(n-k)$ smallest singular values are frozen:

$$\mathbf{W}' = \underline{\mathbf{A}\mathbf{B}}^\top + \sum_{i=k+1}^{n-k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \tag{5}$$

Here, $\mathbf{A}$ and $\mathbf{B}$ are initialized by $\sum_{i=1}^{k} \sigma_i^{\frac{1}{2}} \mathbf{u}_i$ resp. $\sum_{i=1}^{k} \sigma_i^{\frac{1}{2}} \mathbf{v}_i^{\mathrm{T}}$. This initialization enables more efficient optimization by aligning parameter updates with meaningful directions in the model's parameter space, leading to faster convergence and improved performance compared to standard LoRA.

SVFT. SVFT extends PiSSA by further reducing the number of trainable parameters, modifying only a sparse subset of singular

TABLE I: Summary of PEFT methods formulations and trainable parameter counts (underlined variables).

LoRA [12]	**VeRA** [21]	**DoRA** [23]	**PiSSA** [22]	**SVFT** [24]	**SSVD** (Ours)
$\mathbf{W}_0 + \underline{\mathbf{A}}\underline{\mathbf{B}}^\top$ (1)	$\mathbf{W}_0 + (\underline{\mathbf{\Lambda}}_d\mathbf{A})(\underline{\mathbf{\Lambda}}_b\mathbf{B})^\top$ (2)	$\underline{\mathbf{m}}\frac{\mathbf{W}_0 + \underline{\mathbf{A}}\underline{\mathbf{B}}^\top}{\|\mathbf{W}_0 + \underline{\mathbf{A}}\underline{\mathbf{B}}^\top\|_c}$ (3)	$\underline{\mathbf{A}}\underline{\mathbf{B}}^\top + \sum_{i=k+1}^{n-k}\sigma_i\mathbf{u}_i\mathbf{v}_i^\top$ (5)	$\mathbf{U}(\mathbf{\Sigma} + \underline{\mathbf{M}})\mathbf{V}^\top$ (6)	$\mathbf{U}(\Sigma + \underline{\mathbf{\Delta\Sigma}})\underline{\mathbf{G}}\mathbf{V}^\top$ (8)
$r \times (m+n)$	$r + m + 1$	$r \times (m+n) + m$	$r \times (m+n)$	$n \times q + (n-q)(q+1)$	$\frac{k(k+1)}{2}$

values and associated directions. Starting from the SVD of a pre-trained weight matrix (4), SVFT updates the weight as:

$$\mathbf{W}' = \mathbf{U}(\mathbf{\Sigma} + \underline{\mathbf{M}})\mathbf{V}^\top \tag{6}$$

where $\mathbf{M} \in \mathbb{R}^{n\times n}$ is a small, learnable matrix. Different structures of $\mathbf{M}$ lead to different SVFT variants: 1) $SVFT^p$ uses a diagonal matrix $\mathbf{M}$, adapting only singular values (akin to reweighting frozen singular directions); 2) $SVFT_d^B$ uses a banded matrix to introduce learnable off-diagonal interactions; 3) $SVFT_d^R$ samples $\mathbf{M}$ as a fixed random matrix; 4) $SVFT_d^T$ makes the top-k strong interactions between singular vector directions learnable. These variants enable SVFT to explicitly adjust the most critical directions in the parameter space. Among them, $SVFT_d^B$ demonstrated the best performance in [24], and has a trainable parameter count of $n \times q + (n-q)(q+1)$, where q denotes the bandwidth.

Although these methods have been extensively studied in text and vision domains, their effectiveness in speech remains largely unexplored. Moreover, none of them is fundamentally tailored to the unique characteristics of speech signals. This limitation is particularly evident in SVD-guided methods, which explicitly decompose the input and output feature spaces—an operation linked to the acoustic and semantic properties of speech.

In the following section, we highlight this intrinsic nature of SVD-based methods and introduce structured SVD-Guided fine-tuning (SSVD), which explicitly adapts the input feature space to enable efficient and robust domain adaptation in speech.

III. SSVD Method

As SVD is mathematically demonstrated in Section II - **PiSSA** (4), each right singular vector $\mathbf{v}_i$ is mapped to the corresponding left singular vector $\mathbf{u}_i$, scaled by the singular value $\mathbf{\Sigma}$.

An input $\mathbf{x} \in \mathbb{R}^n$, represented in the coordinate system spanned by the right singular vector $\mathbf{v}_i$, is mapped to an output $\mathbf{y} \in \mathrm{Col}\,\mathbf{O} \subset \mathbb{R}^m$ in the coordinate system spanned by the left singular vector $\mathbf{u}_i$.

$$\mathbf{y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{x} \tag{7}$$

Under domain shift in the speech input space, the input $\mathbf{x} \in \mathbb{R}^n$ is no longer aligned with the original right singular basis $\mathbf{v}_i$, but instead aligned with a shifted basis $\mathbf{v}_i'$. This constitutes an "inner" transform, whereas the output semantic space, governed by $\mathbf{u}_i$, remains unchanged (the "outer" transform). The shift in coordinate basis can be modeled through a rotation and scaling of the original input space. Accordingly, the transformation becomes:

$$\mathbf{y} = \mathbf{U}(\Sigma + \underline{\mathbf{\Delta\Sigma}})\underline{\mathbf{G}}\mathbf{V}^\top\mathbf{x} = \mathbf{W}'\mathbf{x} \tag{8}$$

where, the diagonal matrix $\mathbf{\Delta\Sigma}$ models axis-wise scaling (i.e., singular value shifts), and $\mathbf{G}$ is an orthogonal matrix representing a series of rotations in the right singular vector space.

Since the principal singular values and vectors for the best matrix approximation (Eckart-Young theorem) [26], we use

$$\mathbf{\Delta\Sigma} = \begin{bmatrix} \mathbf{\Delta\Sigma}_k & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{G} = \begin{bmatrix} \mathbf{G}_k & 0 \\ 0 & I \end{bmatrix} \tag{9}$$

to apply adaptations to the top-k components. In Section V-A, we discuss different choices for the proportion of adapted components.

To implement rotation within the inner transformation, we explore three parameterizations of the transformation matrix $\mathbf{G}_k$, which trade off orthogonality constraints for computational efficiency.

A. Strict orthogonal constraint

We enforce strict orthogonality by defining $\mathbf{G}_k$ using the Cayley transform [27]:

$$\mathbf{G}_k = (\mathbf{I} - \mathbf{K})(\mathbf{I} + \mathbf{K})^{-1}, \tag{10}$$

where $\mathbf{K} \in \mathbb{R}^{k\times k}$ is a skew-symmetric matrix (i.e, $\mathbf{K}^T = -\mathbf{K}$). By learning a skew-symmetric matrix $\mathbf{K}$, the number of trainable parameters in SSVD is reduced to $(\frac{k(k-1)}{2} + k)$, which is significantly fewer than the $(k^2 + k)$ parameters required to directly learn a full matrix $\mathbf{G}_k$ ('$+k$' accounts for singular values scaling update).

However, the matrix inversion in the Cayley transform incurs a computational cost of approximately $\mathcal{O}(k^3)$ cost, which becomes prohibitive for large dimensions k. To address this, we introduce an approximate orthogonality constraint using a first–order Cayley approximation in III-B.

B. Approximate orthogonal constraint

When $\|\mathbf{K}\| \ll 1$, we can approximate the inverse in the Cayley transform via a truncated Neumann series:

$$(\mathbf{I} + \mathbf{K})^{-1} = \mathbf{I} - \mathbf{K} + \mathcal{O}(\mathbf{K}^2) \tag{11}$$

leading to a first-order approximation:

$$\mathbf{G}_k \;\approx\; (\mathbf{I} - \mathbf{K})(\mathbf{I} - \mathbf{K}) = \mathbf{I} - 2\mathbf{K} + \mathcal{O}(\mathbf{K}^2). \tag{12}$$

By keeping only the linear term, we obtain the simplified form:

$$\mathbf{G}_k \approx \mathbf{I} - 2\mathbf{K} \tag{13}$$

which avoids matrix inversion and reduces the cost to $\mathcal{O}(k^2)$. This approximation introduces an orthogonality error of order $\|\mathbf{K}\|^2$, which is often acceptable for small k, a common setting in PEFT. We use **approximate orthogonality constraint** by default in this study. In Section V-C, we compare the strict and approximate orthogonality settings and demonstrate that the resulting error is trivial.

C. Unconstrained rotations

As a further relaxation, we drop the orthogonality constraint altogether, allowing $\mathbf{G}_k \in \mathbb{R}^{k\times k}$ to be a freely parameterized matrix. This maximally reduces computational overhead but sacrifices the beneficial geometric properties of orthogonal transformations, such as preserving norms and angles. It also increases the number of trainable parameters to $(k^2 + k)$. The performance of this unconstrained rotation is also evaluated in Section V-C.

IV. Experiments

In this study, we choose the open Whisper-style speech models OWSM [2, 3, 5] as our initial models, as they support a wide range of model scales, from 0.1B to 18B parameters. Moreover, the OWSM series is fully open-sourced, providing researchers with detailed information about the corpora used in pre-trainig. This transparency

ensures that there is no domain overlap in our evaluation setup, allowing for a reliable study of domain-shift issues in ASR.

All PEFT methods are implemented within the ESPnet framework, we evaluate them by fine-tuning OWSM models at 0.1B, 1B, and 2B (OWLS) scales on two domain-shifted speech datasets: MyST [28] (child speech) and CGN [29] (dialectal Flemish and Dutch).

A. Dataset

TABLE II: Summary of corpora.

Corpus	Duration (Hours)	Train Utts	Dev Utts	Test Utts
MyST	179	55,703	9,047	10,328
CGN	341	179,440	25,766	51,615

The MyST corpus [28] consists of English dialogues between elementary school students and virtual science tutors, spanning eight topics. The transcriptions provide verbatim orthographic annotations that capture hesitations, repetitions, and disfluencies. Following the protocol of [30], we filter out utterances with a word error rate (WER) higher than 50% on Whisper-large-v2 to ensure transcription quality.

The CGN (Spoken Dutch Corpus) [29] is a manually annotated speech database containing approximately 900 hours of Dutch speech, including 270 hours of Flemish Dutch. It consists of 15 components covering various speaking styles, such as read speech, interviews, spontaneous conversations, and telephone dialogues. In this study, we use a subset containing both Flemish and Dutch dialects, excluding component `c` (spontaneous interview), component `d` (discussions), and component `f` (spontaneous telephone dialogues), resulting in a total of approximately 341 hours.

The dataset statistics are summarized in Table II.

B. Implementation details

TABLE III: Architectural configurations of models.

Model	# Params	Encoders	Decoders	# dim	# head
OWSM-0.1B	101 M	6 layers	6 layers	384	6
OWSM-1B	1.01 B	18 layers	18 layers	1024	16
OWLS-2B	2.30 B	16 layers	16 layers	2048	64

OWSM uses an E-Branchformer [31] encoder with a Transformer [32] decoder, while OWLS is a standard Transformer encoder–decoder. Table III lists their full configurations. OWSM and OWLS models are trained on 180k hours of public speech data, as documented in [5, 3], which do not include child speech or Flemish Dutch, although standard Dutch data are included. Therefore, domain-shifted scenarios are expected.

In our study, PEFT methods are applied to all linear layers in each model, including query, key, and value projections. For different low-rank configurations, we denote them as $LoRA_{r=rank}$, $VeRA_{r=rank}$, $DoRA_{r=rank}$, and $PiSSA_{r=rank}$. For SVFT, we follow [24] and use the best-performing banded matrix $\mathbf{M}$, denoted as $SVFT^{B}_{d=band\ size}$. For SSVD, with different choices of k in (9), rotation adaptations are applied to $p\%$ of the right singular vectors and values, denoted as $SSVD_{p=portion}$. All experiments are conducted on a single NVIDIA H100 80GB GPU.

V. Results

A. ASR performance

In Table IV, we summarize the WERs on the MyST test data after fine-tuning the OWSM-0.1B, OWSM-1B and OWLS-2B models with different PEFT methods using various configurations. For each model the best-performing method is highlighted in **bold**, and the second-best is marked with a superscript †.

TABLE IV: PEFT methods across OWSM models on MyST.

Model	PEFT Method	# Params	WER (%) ↓
OWSM-0.1B	Zero-shot	–	25.0
	Full fine-tuning	101 M	14.9
	$LoRA_{r=8}$	2.07 M	19.2
	$LoRA_{r=16}$	4.13 M	17.3
	$LoRA_{r=32}$	8.26 M	16.4
	$VeRA_{r=32}$	1.74 M	22.4
	$VeRA_{r=64}$	3.36 M	21.1
	$VeRA_{r=128}$	6.59 M	20.2
	$DoRA_{r=8}$	2.19 M	21.2
	$DoRA_{r=16}$	4.26 M	19.1
	$DoRA_{r=32}$	8.39 M	17.8
	$PiSSA_{r=8}$	2.07 M	19.1
	$PiSSA_{r=16}$	4.13 M	16.8
	$PiSSA_{r=32}$	**8.26** M	**16.0**
	$SVFT^{B}_{d=8}$	1.23 M	23.9
	$SVFT^{B}_{d=16}$	2.39 M	20.4
	$SVFT^{B}_{d=32}$	4.67 M	18.2
	$SVFT^{B}_{d=64}$	9.03 M	16.8
	$SSVD_{p=40\%}$	1.51 M	18.4
	$SSVD_{p=60\%}$	3.67 M	17.0
	$SSVD_{p=80\%}$	6.57 M†	16.2†
OWSM-1B	Zero-shot	–	19.3
	Full fine-tuning	1.01 B	12.4
	$LoRA_{r=8}$	10.56 M	17.6
	$LoRA_{r=16}$	21.13 M	16.1
	$LoRA_{r=32}$	42.26 M	15.1
	$VeRA_{r=256}$	13.82 M	18.7
	$VeRA_{r=384}$	20.40 M	17.7
	$DoRA_{r=8}$	11.22 M	16.7
	$DoRA_{r=16}$	21.76 M	15.3
	$DoRA_{r=32}$	42.92 M†	14.2†
	$PiSSA_{r=8}$	10.56 M	16.8
	$PiSSA_{r=16}$	21.13 M	15.4
	$PiSSA_{r=32}$	42.26 M	14.3
	$SVFT^{B}_{d=8}$	7.00 M	19.0
	$SVFT^{B}_{d=16}$	13.55 M	16.3
	$SVFT^{B}_{d=32}$	26.52 M	15.1
	$SVFT^{B}_{d=64}$	51.88 M†	14.2†
	$SSVD_{p=22\%}$	9.83 M	15.3
	$SSVD_{p=25\%}$	12.50 M	14.5
	$SSVD_{p=29\%}$	16.26 M	14.1
	$SSVD_{p=33\%}$	22.16 M	14.1
	$SSVD_{p=40\%}$	**31.86** M	**13.8**
OWLS-2B	Zero-shot	–	20.3
	Full fine-tuning	2.30 B	13.1
	$LoRA_{r=8}$	12.69 M	18.3
	$LoRA_{r=16}$	25.39 M	16.9
	$LoRA_{r=32}$	50.78 M	15.6
	$VeRA_{r=256}$	14.16 M	19.1
	$VeRA_{r=384}$	20.86 M	18.0
	$DoRA_{r=8}$	13.47 M	17.1
	$DoRA_{r=16}$	26.16 M	16.1
	$DoRA_{r=32}$	51.55 M	15.2
	$PiSSA_{r=8}$	12.69 M	17.4
	$PiSSA_{r=16}$	25.39 M	16.1
	$PiSSA_{r=32}$	50.78 M†	15.0†
	$SVFT^{B}_{d=8}$	9.38 M	19.9
	$SVFT^{B}_{d=16}$	18.20 M	17.3
	$SVFT^{B}_{d=32}$	35.74 M	15.3
	$SSVD_{p=17\%}$	15.04 M	15.1
	$SSVD_{p=20\%}$	21.63 M†	14.7†
	$SSVD_{p=25\%}$	**33.88** M	**14.6**

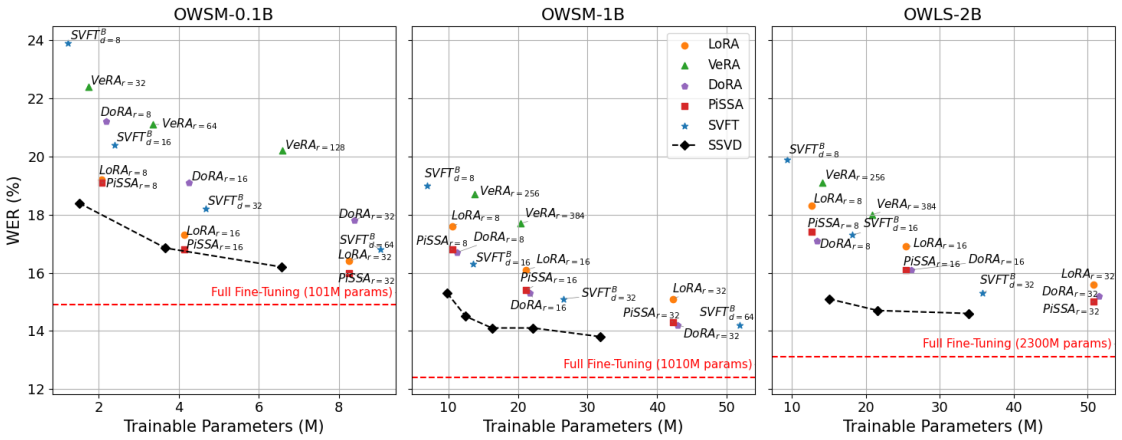

Fig. 1: WER (%) versus trainable parameters for PEFT methods fine-tuning OWSM-0.1B, OWSM-1B, OWLS-2B on MyST.

Across all three models, the zero-shot WERs remain around 20% or higher. After fine-tuning, for the smaller model OWSM-0.1B, all PEFT methods with similar parameter scales yield comparable performance, with $PiSSA_{r=32}$ slightly outperforming others. Notably, $SSVD_{p=80\%}$ achieves the second-best performance, with only a 0.2% higher WER than $PiSSA_{r=32}$ while requiring nearly **2M** fewer trainable parameters (6.57M vs. 8.26M).

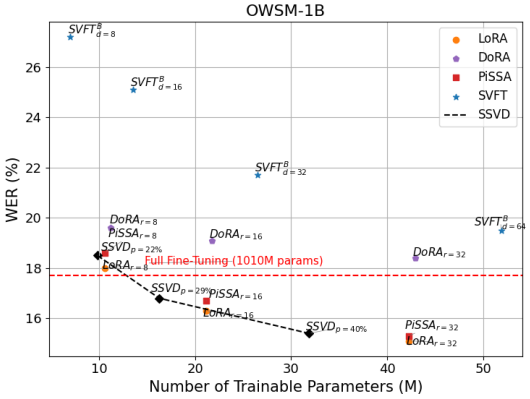

Fig. 2: WER (%) versus trainable parameters for PEFT methods fine-tuning OWSM-1B on CGN.

For the larger-scale OWSM-1B and OWLS-2B models, clearer performance trends show among PEFT methods. SSVD consistently outperforms other approaches while using significantly fewer trainable parameters. For example, when fine-tuning OWSM-1B, $SSVD_{p=40\%}$ achieves a WER of **13.8%**, approaching the 12.4% obtained by full fine-tuning, using only around **32M** parameters, compared to 14.2% WER from the second best ($DoRA_{r=32}$ and $SVFT^{B}_{d=64}$) with approximately 52M and 43M parameters. Similarly, for the OWLS-2B model, SSVD achieves **14.7%** WER with only **21.6M** trainable parameters, outperforming LoRA, DoRA, and PiSSA, each of which requires around 51M parameters, demonstrating SSVD's high efficiency in a domain-shift scenario.

To better illustrate the trade-off between ASR performance and parameter efficiency, Figure 1 presents WERs as a function of the number of trainable parameters for each PEFT method. Across all model scales, SSVD (depicted by black diamonds and dashed lines) consistently achieves lower WERs compared to other methods under similar trainable parameters (i.e., vertically). The performance gap becomes increasingly pronounced for larger models: in OWSM-1B and OWLS-2B, the SSVD curve lies significantly below those of other methods, indicating its strong ability to balance performance and efficiency, particularly in larger-scale ASR settings.

TABLE V: PEFT methods across OWSM models on CGN.

Model	PEFT Method	# Params	WER (%) ↓
OWSM-0.1B	Zero-shot	–	65.7
	Full fine-tuning	101 M	17.8
	$LoRA_{r=8}$	2.07 M	25.2
	$LoRA_{r=16}$	4.13 M	21.9
	$LoRA_{r=32}$	**8.26** M	**19.7**
	$DoRA_{r=8}$	2.19 M	24.2
	$DoRA_{r=16}$	4.26 M	21.9
	$DoRA_{r=32}$	8.39 M	20.7
	$PiSSA_{r=8}$	2.07 M	25.5
	$PiSSA_{r=16}$	4.13 M	22.1
	$PiSSA_{r=32}$	8.26 M†	20.2†
	$SVFT^{B}_{d=8}$	1.23 M	36.7
	$SVFT^{B}_{d=16}$	2.39 M	32.6
	$SVFT^{B}_{d=32}$	4.67 M	28.6
	$SVFT^{B}_{d=64}$	9.03 M	25.0
	$SSVD_{p=40\%}$	1.51 M	25.8
	$SSVD_{p=60\%}$	3.67 M	22.9
	$SSVD_{p=80\%}$	6.57 M	20.9
OWSM-1B	Zero-shot	–	46.3
	Full fine-tuning	1.01 B	17.7
	$LoRA_{r=8}$	10.56 M	18.0
	$LoRA_{r=16}$	21.13 M	16.3
	$LoRA_{r=32}$	**42.26** M	**15.1**
	$DoRA_{r=8}$	11.22 M	19.6
	$DoRA_{r=16}$	21.76 M	19.1
	$DoRA_{r=32}$	42.92 M	18.4
	$PiSSA_{r=8}$	10.56 M	18.6
	$PiSSA_{r=16}$	21.13 M	16.7
	$PiSSA_{r=32}$	42.26 M†	15.3†
	$SVFT^{B}_{d=8}$	7.00 M	27.2
	$SVFT^{B}_{d=16}$	13.55 M	25.1
	$SVFT^{B}_{d=32}$	26.52 M	21.7
	$SVFT^{B}_{d=64}$	51.88 M	19.5
	$SSVD_{p=22\%}$	9.83 M	18.5
	$SSVD_{p=29\%}$	16.26 M	16.8
	$SSVD_{p=40\%}$	31.86 M†	15.4†

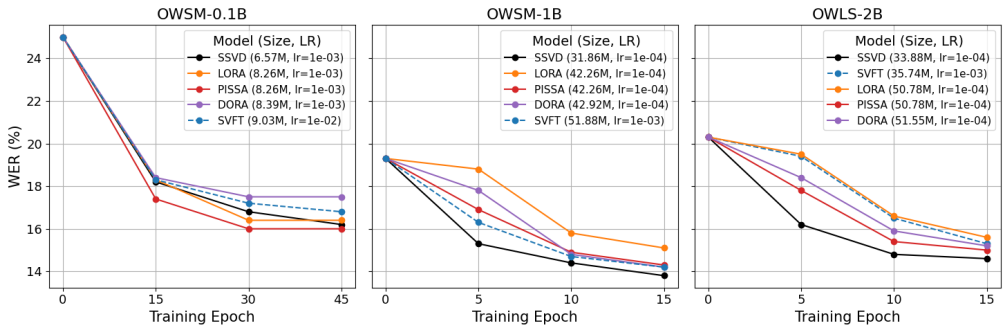

Fig. 3: WER (%) versus training epoch for PEFT methods fine-tuning OWSM-0.1B, OWSM-1B, OWLS-2B.

Moreover, as observed vertically in Figure 1, SVD-guided methods and the directionally tuned DoRA, consistently outperform LoRA when using a similar number of trainable parameters. This trend suggests that structured initialization (as in SVD-guided methods) and directional constraints (as in DoRA) are more effective than LoRA's random initialization with unconstrained updates, a finding that aligns with prior observations in NLP and vision tasks [23, 22, 24].

Table V reports the WERs on the CGN test set after fine-tuning. Since VeRA performs poorly in this scenario, we exclude it from the table. The results for OWSM-0.1B follow a similar trend as on MyST, except SVFT, which will be discussed further in Section V-B. For the OWSM-1B model, full fine-tuning becomes less effective, and methods such as DoRA, which mimic full fine-tuning behavior, also yield suboptimal results.

SSVD achieves comparable WERs (**15.4%** vs. 15.1% from LoRA and 15.3% from PiSSA) with **10M** fewer parameters (32M vs. 42M). Figure 2 visualizes the WER versus trainable parameter trade-off for OWSM-1B. On CGN, LoRA outperforms DoRA and SVFT at similar scales. One explanation is that CGN data require larger update deviations, and LoRA, unlike structured methods, offers more flexibility due to its unconstrained parameter updates. Nevertheless, SSVD shows a similar trend to LoRA, particularly when a larger portion of components (e.g., at 40%) are made trainable.

B. *Efficiency analysis*

TABLE VI: Comparing SSVD constraint methods on the MyST data.

Model	p	**# Params**	**Strict**	**Approx.**	**# Params**	**None**
OWSM-0.1B	40%	1.51 M	18.6	18.4	3.02 M	17.0
	60%	3.67 M	17.2	17.0	7.34 M	16.9
	80%	6.57 M	16.3	16.2	13.13 M	15.6
OWSM-1B	22%	9.83 M	15.1	15.3	19.67 M	14.9
	25%	12.50 M	14.9	14.5	25.00 M	14.6
	29%	16.26 M	14.5	14.1	32.51 M	13.8
	33%	22.16 M	14.2	14.1	44.32 M	13.7
	40%	31.86 M	13.8	13.8	63.72 M	13.3
OWLS-2B	17%	15.04 M	15.2	15.1	30.09 M	14.9
	20%	21.63 M	14.7	14.7	43.26 M	14.6
	25%	33.88 M	14.4	14.6	67.77 M	14.3

To further evaluate the training efficiency, Figure 3 presents accuracy versus epoch. Since methods with close model size have similar computing time, epoch here serves as a proxy for training time. Solid lines indicate training with a uniform learning rate, e.g., $1e{-}4$, while dashed lines correspond to a tenfold higher learning rate, e.g., $1e{-}3$. For larger-scale models, a clear convergence hierarchy shows: SSVD $>$ PiSSA $>$ DoRA $>$ LoRA. The convergence speed of SVFT, however, is highly sensitive to the chosen band size. Since SVFT updates only a fixed band of singular values, it requires a larger band size to achieve convergence comparable to PiSSA, resulting in increased trainable parameters. Additionally, SVFT benefits more from higher learning rates, as it only updates singular values while keeping singular vectors fixed, limiting the flexibility of optimization under smaller learning rates.

C. *Ablation study*

In Tables IV and V, SSVD is implemented in the "Approx." variant as described in Section III-B, which introduces minor orthogonality errors. As further evaluated in Table VI, we compare this variant against the "Strict" implementation (Section III-A) and a "None" constraint baseline (Section III-C). The results show that "Strict" and "Approx." yield nearly identical WERs, suggesting that the orthogonality deviations in the approximate implementation are relatively minor and do not significantly affect performance in the PEFT setting. The "None" constraint implementation, as explained in Section III-C, doubles the number of trainable parameters. While increasing parameters can improve performance, the "None" constraint performs worse than "Strict" and "Approx." when the number of trainable parameters is close. For instance, with OWLS-2B, it achieves a WER of 14.9% using 30.09M parameters, higher than the 14.7% WER achieved with only 21.63M parameters.

VI. Conclusion

In this work, we presented a comprehensive evaluation of PEFT methods for adapting large-scale speech foundation models under domain-shifted conditions, such as child speech and dialectal variation. We integrated and benchmarked SoTA PEFT techniques, including LoRA, DoRA, VeRA, PiSSA, SVFT, and our proposed SSVD, within ESPnet across model sizes from 0.1B to 2B. Our results highlight that while several PEFT methods perform comparably on small models, performance gaps widen at larger scales, especially under constrained compute budgets. Among all methods, SSVD consistently achieves the best trade-off between performance and parameter efficiency, closely approaching full fine-tuning performance with significantly fewer trainable parameters.

Acknowledgment

Experiments of this work used the Bridges2 system at PSC and Delta system at NCSA through allocations CIS210014 and IRI120008P from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, supported by National Science Foundation grants #2138259,#:2138286, #:2138307, #:2137603, and #:2138296. This research was supported by the Flemish Government under "Onderzoeksprogramma AI Vlaanderen", the FWO-SBO grant S004923N: NELF, and the FWO grant V401325N.

References

[1] Alec Radford et al. “Robust speech recognition via large-scale weak supervision”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2023, pp. 28492–28518.

[2] Yifan Peng et al. “Reproducing whisper-style training using an open-source toolkit and publicly available data”. In: *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2023, pp. 1–8.

[3] Yifan Peng et al. “OWSM v3.1: Better and faster open Whisper-style speech models based on E-Branchformer”. In: *Proceedings of Interspeech*. 2024, pp. 352–356.

[4] Krishna C Puvvada et al. “Less is more: Accurate speech recognition & translation without web-scale data”. In: *Proceedings of Interspeech*. 2024, pp. 3964–3968.

[5] William Chen et al. “OWLS: Scaling laws for multilingual speech recognition and translation models”. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2025).

[6] Ruchao Fan, Natarajan Balaji Shankar, and Abeer Alwan. “Benchmarking children’s ASR with supervised and self-supervised speech foundation models”. In: *Proceedings of Interspeech*. 2024, pp. 5173–5177.

[7] Pu Wang and Hugo Van hamme. “Benefits of pre-trained mono-and cross-lingual speech representations for spoken language understanding of Dutch dysarthric speech”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2023.1 (2023), p. 15.

[8] Yaroslav Getman et al. “Exploring adaptation techniques of large speech foundation models for low-resource ASR: a case study on Northern Sámi”. In: *Proceedings of Interspeech*. 2024, pp. 2539–2543.

[9] Neil Houlsby et al. “Parameter-efficient transfer learning for NLP”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2019, pp. 2790–2799.

[10] Brian Lester, Rami Al-Rfou, and Noah Constant. “The power of scale for parameter-efficient prompt tuning”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021, pp. 3045–3059.

[11] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. “BitFit: Simple parameter-efficient fine-tuning for Transformer-based masked language-models”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 2: Short Papers)*. 2022, pp. 1–9.

[12] Edward J. Hu et al. “LoRA: Low-rank adaptation of large language models”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2022.

[13] Yuhui Gu et al. “QA-LoRA: Quantization-aware low-rank adaptation of large language models”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2023.

[14] Qingru Zhang et al. “Adaptive budget allocation for parameter-efficient fine-tuning”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2023.

[15] Pranay Dighe et al. “Leveraging large language models for exploiting asr uncertainty”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 12231–12235.

[16] Arun Baby, George Joseph, and Shatrughan Singh. “Robust speaker personalisation using generalized low-rank adaptation for automatic speech recognition”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 11381–11385.

[17] Pu Wang and Hugo Van hamme. “Bottleneck low-rank Transformers for low-resource spoken language understanding”. In: *Proceedings of Interspeech*. 2022, pp. 1248–1252.

[18] Wei Liu et al. “Sparsely shared lora on whisper for child speech recognition”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 11751–11755.

[19] Tianyi Xu et al. “Towards rehearsal-free multilingual ASR: A LoRA-based case study on Whisper”. In: *Proceedings of Interspeech*. 2024, pp. 2534–2538.

[20] Zheshu Song et al. “LoRA-Whisper: Parameter-efficient and extensible multilingual ASR”. In: *Proceedings of Interspeech*. 2024, pp. 3934–3938.

[21] Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. “VeRA: Vector-based random matrix adaptation”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2024.

[22] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. “PiSSA: Principal singular values and singular vectors adaptation of large language models”. In: *Advances in Neural Information Processing Systems (NeurIPS)* 37 (2024), pp. 121038–121072.

[23] Shih-Yang Liu et al. “DoRA: Weight-decomposed low-rank adaptation”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2024.

[24] Vijay Chandra Lingam et al. “SVFT: Parameter-efficient fine-tuning with singular vectors”. In: *Advances in Neural Information Processing Systems (NeurIPS)* 37 (2024), pp. 41425–41446.

[25] Shinji Watanabe et al. “ESPnet: End-to-end speech processing toolkit”. In: *Proceedings of Interspeech* (2018).

[26] Carl Eckart and Gale Young. “The approximation of one matrix by another of lower rank”. In: *Psychometrika* 1.3 (1936), pp. 211–218.

[27] Asher Trockman and J Zico Kolter. “Orthogonalizing convolutional layers with the Cayley Transform”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2021.

[28] Sameer Pradhan and et al. “My Science Tutor (MyST): A large corpus of children’s conversational speech”. In: *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. May 2024, pp. 12040–12045.

[29] Nelleke Oostdijk and et al. “The Spoken Dutch Corpus: Overview and first evaluation”. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*. Athens, Greece, 2000, pp. 887–894.

[30] Ahmed Adel Attia and et al. “Kid-Whisper: Bridging the performance gap in automatic speech recognition for children”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. Vol. 7. 2024, pp. 74–80.

[31] Kwangyoun Kim et al. “E-branchformer: Branchformer with enhanced merging for speech recognition”. In: *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2023, pp. 84–91.

[32] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).