Hypernetworks for Personalizing ASR to Atypical Speech

Max Müller-Eberstein[1*†] **Dianna Yee**[2*] **Karren Yang**[2]
Gautam Varma Mantena[2] **Colin Lea**[2]
[1]IT University of Copenhagen, Denmark
mamy@itu.dk
[2]Apple, USA
{dianna_yee, karren_yang, gmantena, colin_lea}@apple.com

Abstract

Parameter-efficient fine-tuning (PEFT) for personalizing automatic speech recognition (ASR) has recently shown promise for adapting general population models to atypical speech. However, these approaches assume a priori knowledge of the atypical speech disorder being adapted for—the diagnosis of which requires expert knowledge that is not always available. Even given this knowledge, data scarcity and high inter-/intra-speaker variability further limit the effectiveness of traditional fine-tuning. To circumvent these challenges, we first identify the minimal set of model parameters required for ASR adaptation. Our analysis of each individual parameter's effect on adaptation performance allows us to reduce Word Error Rate (WER) by half while adapting 0.03% of all weights. Alleviating the need for cohort-specific models, we next propose the novel use of a meta-learned hypernetwork to generate highly individualized, utterance-level adaptations on-the-fly for a diverse set of atypical speech characteristics. Evaluating adaptation at the global, cohort, and individual-level, we show that hypernetworks generalize better to out-of-distribution speakers, while maintaining an overall relative WER reduction of 75.2% using 0.1% of the full parameter budget.

1 Introduction

Large-scale automatic speech recognition (ASR) models are trained predominately on speech collected from the general population and historically have not been able to fully support speakers with atypical speech. Recent work has proposed parameter-efficient fine-tuning (PEFT) of large ASR models for adapting such general population models to work better for people with speech differences (Tomanek et al., 2021a,b; Qi and van Hamme, 2023). Such adaptations have focused either on fine-tuning using data from a group of individuals with common speech differences—referred to as cohort-level fine-tuning—or on fine-tuning on speech data at the level of an individual.

Individually personalized ASR models yield state-of-the-art transcription performance, however they require laborious data collection, which could be especially strenuous for people with severe speech disorders. Additionally, characteristics vary greatly, even for the same speaker over time, potentially leading to data drift and eventual performance degradation (Tomanek et al., 2023).

At the cohort level, PEFT of general population models has been shown to improve transcription performance for dysarthria (Tomanek et al., 2021b). While this approach reduces training data and compute requirements, it requires a priori knowledge of an individual's atypical speech category, which is not always available. As we later demonstrate, a precise diagnosis is crucial, as fine-tuning on a specific cohort does not transfer well to other types of atypical speech (Section 4.5). Furthermore, such solutions require discrete categorizations of individuals and do not share knowledge across cohorts, although sharing may be beneficial for individuals who express mixtures of speech differences, or between individuals with different severities of the same speech disorder.

In this work, we consolidate individual-level personalization with knowledge-sharing across cohorts by proposing the use of hypernetworks to generate adaptation parameters dynamically during inference, for individualized personalization amongst a heterogeneous cohort of speech disorders. As opposed to cohort and individual-level

[*]These authors contributed equally to this work.
[†]Research performed while at Apple.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1182–1196, 2024. https://doi.org/10.1162/tacl_a_00696
Action Editor: Liang Huang. Submission batch: 2/2024; Revision batch: 6/2024; Published 9/2024.

fine-tuning, which learn fixed adaptations that are difficult to transfer across etiologies, hypernetworks leverage a meta-learning procedure that instead learns to generate adaptation parameters, conditioned on the target speaker's speech characteristics. This approach enables flexibility with respect to the adaptations applied to the ASR model, as they can change depending on the individual utterance being adapted for. Simultaneously, the use of a single hypernetwork instead of multiple pre-trained adaptations for each cohort or individual reduces complexity and actively promotes sharing information that is useful across different types of atypical speech.

In our study, we include both phonological and fluency-related speech disorders, using speech from people with *stuttering*, *dysarthria* consistent with cerebral palsy, and *Parkinson* disease, for which a myriad of speech differences including dysarthria and stuttering may be exhibited. Stuttering includes dysfluencies, such as sound, word or phrase repetitions (''m-m-mall'', ''go go go''), prolongations (''baaall''), and audible pauses or blocks (Sander, 1963; Riley, 2009). Dysarthric speech may contain differences in pronunciation, pitch, intelligibility, strain, speaking rate and volume. It is particularly challenging as the expressed characteristics depend on the etiology and individual, varying even for one speaker (Rowe et al., 2022). For example, spastic dysarthria, commonly associated with cerebral palsy, is characterised by slow speaking rates, strained voice, and pitch breaks (Schölderle et al., 2016), whereas hypokinetic dysarthria, commonly associated with Parkinson disease, is characterized by monotonous speech, varying in volume, breathiness, hoarseness, rapid repetition of phones and imprecise consonant production (Duffy, 1995; Tjaden, 2008).

Towards improving ASR for these speech communities, we concretely contribute:

- To the best of our knowledge, the first study of adapting transformer-based ASR models to dysarthric, dysfluent, and Parkinson-influenced speech simultaneously;
- The highest resolution analysis to date, regarding which model parameters contribute most to adaptation (Section 4);
- A novel approach of using hypernetworks to generate individualized, zero-shot ASR adaptations dynamically across atypical speech types (Section 5);
- Experiments covering global, cohort, and individual adaptation, to compare hypernetworks with prior work and analyze factors important to its performance (Section 6).

2 Related Work

Personalized ASR adaptation for atypical speech is a broad yet under-explored topic. Prior work mainly focuses on large-scale ASR models trained on general population speech, which are subsequently fine-tuned on small datasets of atypical speech (Shor et al., 2019; Green et al., 2021). Such datasets are scarce, and even more so at the level of individual speakers, leading to overfitting and poor generalization. To overcome these challenges, past work has explored PEFT methods such as individually re-weighting transcription output probabilities (Morales and Cox, 2007), or using residual adapters (Rebuffi et al., 2017) to individually personalize ASR models (Tomanek et al., 2021b), while retaining the original model weights and only training the light-weight adapter modules.

Another approach leverages cohort-level transfer learning: Tomanek et al. (2021a) use a two stage fine-tuning process, where the ASR model is first fine-tuned on data from a cohort sharing atypical speech characteristics, before being further fine-tuned on data from an individual in the same cohort. Qi and van Hamme (2023) follow a similar approach, but make use of less resource-intensive adapter fusion (Pfeiffer et al., 2021), where a cohort-level adapted model is fused with multiple individual-level adapters to train personalized models for new target speakers.

These aforementioned studies either require sufficient data from the target speaker in order to train a personalized model, and/or knowledge of the cohort an individual belongs to—both of which may not be readily available. As we demonstrate in Section 4.5, the process of maintaining and selecting the correct cohort model is critical, however defining the cohort is nontrivial as assigning membership may not be limited to etiology but also severity thereof. Furthermore, prior approaches consider cohorts independently of each other and are thus unable to share knowledge that may be beneficial for better generalization performance across individuals.

In order to generate individualized adaptations while learning globally shared representations across cohorts, we reformulate ASR adaptation as a meta-learning problem. We propose to model this inductive bias via a light-weight hypernetwork meta-learner (Ha et al., 2017), which is tasked to generate the most effective adaptation weights for an individual based on their speaker characteristics as represented by a shared encoder.

While this work, to the best of our knowledge, is the first to apply hypernetworks to ASR, recent studies have applied them to predict the task-specific adaptations of a text-based, pre-trained, large-scale Transformer architecture (Karimi Mahabadi et al., 2021; Phang et al., 2023). Additionally, language model adaptations generated by hypernetworks have also been shown to generalize to unseen task and language combinations (Ansell et al., 2021; Üstün et al., 2022a,b). Based on these results as well as recent successes in adapter fusion (Pfeiffer et al., 2021; Qi and van Hamme, 2023), we hypothesize that zero-shot personalization is possible by having the hypernetwork learn a mapping between speaker characteristics and ASR adaptation weights, effectively learning a manifold of personalized models. This procedure would require neither labelled audio data nor fine-tuning on the individual-level, leading to increased parameter and data efficiency.

3 Setup

3.1 Data

Our experiments use three datasets containing speech with phonological and fluency-related speech disorders, with content relating to common voice commands for digital assistants, as well as dictation. The first dataset X_D, as described in Yee et al. (2023), contains dysarthric speech with varying severities mostly consistent with cerebral palsy. All 33 participants read a common set of 51 phrases with at least 5 repetitions in multiple recording sessions with several microphone placements, across multiple days. The second dataset X_S, as described in Lea et al. (2023), contains speech from people who stutter with various degrees of fluency. All 91 participants were prompted from a common set of dictation and voice assistant tasks but had agency to personalize the commands. The speech of all participants within X_D and X_S was graded by a Speech-Language Pathologist as 'mild', 'moderate', or 'severe'. The third dataset X_P, is a subset of the Speech Accessibility Project (SAP), which contains speech from 113 individuals whose speech is consistent with Parkinson disease, saying a mixture of read and free-spoken prompts for common dictation and voice assistant commands, and has not been graded by severity. The full public benchmark, denoted by $X_{\mathbb{P}}$[1], contains a broader set of 253 participants, for which we run additional experiments to provide official benchmark results for future work. In each setup, we run a preliminary study with 3 random seeds where X_D, X_S, and X_P are split with no speaker overlap into 70% train, 10% validation and 20% test sets, and there is no known overlap of participants across datasets.

3.2 Models

We choose Whisper (Radford et al., 2023) as our base model architecture, a series of 10 encoder-decoder Transformer models, which vary with respect to model size and pre-training data. Trained on 680k hours of general population speech, they allow us to investigate how well the largest contemporary models for typical speech characteristics fare in low-resource adaptation scenarios.

To gain an understanding of how personalization affects the model, we ablate across multiple fine-tuning setups (Section 4). Going beyond previous work, we first run an extensive full fine-tuning sweep, additionally ablating across seven partial model components and layers. Next, we adapt sub-layer components, such as the attention and feed-forward layers, using Low-rank Adaptation (LoRA; Hu et al., 2021). We respectively denote these setups as $\{\textit{full}, \textit{partial}, \textit{LoRA}\}$. Based on these ablations, we identify which parameters contribute most to personalization, and then train hypernetworks to generate them dynamically (Section 5).

3.3 Evaluation

Since our proposed approach aims to generate dynamic personalizations, which generalize across a heterogeneous collection of speech disorders, the hypernetwork is trained using a concatenation of all three atypical speech types X_D, X_S, X_P. In our final experiments, we further include

[1]Planned for public release in 2024.

a hypernetwork solely trained on the $X_{\mathbb{P}}$ benchmark to enable future comparisons, as well as to investigate the effects of lower speaker diversity.

As prior work has mainly focused on cohort-level or individualized fine-tuning (Tomanek et al., 2021a,b; Qi and van Hamme, 2023), we define our baselines correspondingly, i.e., with access to speech disorder diagnoses. Additionally, we re-train these cohort-level baselines using the same concatenated training datasets as the hypernetwork to ensure a fair comparison. These dataset-specific and concatenated setups are denoted respectively by $\{cohort, global\}$. Finally, we evaluate personalization at the most granular, *individual* level, by continually fine-tuning/ adapting the matching cohort-level baselines on training data from the target speaker. Note that, while the baselines assume access to the speaker's cohort information and even individualized training data, our hypernetwork-based approach is the first to operate in a fully zero-shot manner, without additional speaker-specific training (meta-) data. All evaluations are computed using the same sets of utterances, and of speakers not observed during training, using the transcription Word and Match Error Rates (WER, MER; Morris et al., 2004). For the individual-level adaptation experiments, the baselines use part of a target speaker's utterances as training data and use the remainder as unseen test data, while the hypernetwork remains completely agnostic to the target speaker and is applied directly to the equivalent test data subset without additional training.

In our experiments, we observe that Whisper occasionally hallucinates, especially for stuttered speech, repeatedly decoding a stuttered syllable up until its maximum decoding length, even if the audio actually contains further content. While MER normalizes this high number of insertion errors into a [0, 100] range, for WER these hallucinations result in large, anomalous values, which hinder the comparison of results across setups. We therefore report performance in terms of median (P50) and interquartile range (IQR) WERs on the speaker level for robustness against such outliers.

4 Cohort-level Personalization

Generating an entire personalized model would be prohibitively expensive. As such we first follow the cohort-level personalization paradigm to identify which pre-trained models provide good initializations for adaptation (Section 4.1), and which components (Section 4.2) and individual parameters (Section 4.3) are crucial to fine-tune and adapt. This allows us to create a lighter-weight adaptation framework to which we later apply dynamic personalization (Section 5).

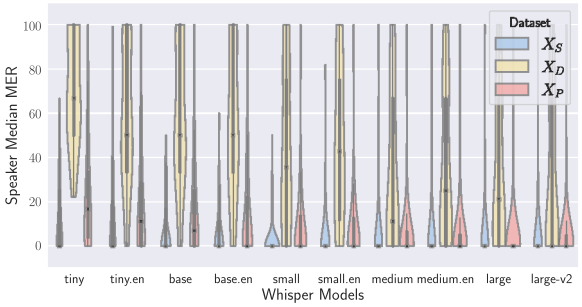


Figure 1: Speaker median MER of untuned pre-trained Whisper models on X_S (stuttering), X_D (dysarthria), and X_P (Parkinson).

4.1 Pre-trained Model Performance

Across the 10 different Whisper model sizes and pre-training paradigms, their transcription error rates on each dataset in Figure 1 show that performance tends to improve with model size. However, even the 1.6B parameter `large-v2` model is not performant enough, presuming a usable system has WER < 15% (MacDonald et al., 2021). We also observe that Whisper models tend to transcribe verbatim (e.g., stuttered repetitions), which may not be desirable for some downstream applications (Lea et al., 2023). Additionally, the most severe errors arise from infinite repetition loops in the decoding process, to which the monolingual variants appear to be slightly more robust. Our subsequent experiments focus on both ends of the model spectrum: the multilingual `tiny` and `large-v2` models, with 39M and 1.6B parameters respectively, plus the monolingual English `tiny.en`, also with 39M parameters.

4.2 Full Component-level Fine-tuning

Full fine-tuning provides a theoretical upper-bound of ASR performance and compute requirements. In addition, we ablate training seven sub-architectures, including the full encoder/ decoder, the earlier$_{\downarrow}$/later$_{\uparrow}$ half of their respective layers, and the final decoder head. In terms of sub-architectures, Table 1 shows that tuning

Setup		Untuned	Full Tuning							LoRA		
		–	All	Enc	Enc↓	Enc↑	Dec	Dec↓	Dec↑	All	Enc	Dec
		0%	100%	28%	14%	14%	72%	36%	36%	9%	4%	6%
tiny	X_P	25.9 (39.4)	**11.7 (25.9)**	22.4 (55.5)	28.0 (42.7)	25.5 (37.8)	23.4 (36.8)	25.8 (36.9)	25.7 (37.7)	***24.1 (38.3)***	27.0 (39.5)	26.2 (37.8)
	X_S	16.8 (171.5)	**0.5 (9.2)**	2.6 (15.1)	9.7 (32.0)	2.3 (11.9)	6.8 (21.2)	8.6 (28.0)	2.3 (14.0)	***2.6 (14.7)***	5.8 (22.9)	8.7 (28.3)
	X_D	82.7 (96.8)	**2.6 (9.5)**	9.0 (16.7)	35.7 (62.1)	9.1 (16.9)	9.5 (20.8)	58.1 (81.6)	7.3 (14.7)	***19.7 (46.3)***	21.2 (46.5)	45.2 (81.1)
tiny.en	X_P	22.4 (34.1)	18.8 (31.3)	19.2 (32.4)	22.3 (34.0)	**17.7 (30.4)**	20.5 (31.5)	22.8 (34.0)	22.0 (31.7)	21.7 (35.2)	***19.1 (31.2)***	22.9 (34.5)
	X_S	97.3 (239.2)	**0.2 (7.6)**	2.0 (12.0)	5.1 (19.8)	1.5 (10.6)	2.7 (11.9)	5.0 (17.8)	2.8 (12.4)	***0.3 (8.8)***	2.2 (13.6)	3.6 (13.8)
	X_D	67.4 (91.8)	1.9 (9.5)	11.1 (18.8)	24.6 (48.5)	6.4 (12.6)	4.8 (11.9)	20.6 (42.4)	**0.0 (9.5)**	***3.6 (9.5)***	13.1 (21.5)	7.6(14.2)
large-v2	X_P	8.5 (17.7)	5.2 (9.4)	7.4 (15.5)	7.5 (15.1)	7.5 (15.6)	**5.0 (10.1)**	5.5 (13.4)	5.8 (11.5)	***4.9 (10.2)***	5.9 (13.6)	5.5 (11.1)
	X_S	19.3 (171.0)	**0.0 (1.0)**	0.0 (3.9)	0.0 (5.8)	0.0 (4.0)	0.0 (1.5)	0.1 (3.8)	0.0 (2.0)	***0.0 (2.7)***	1.5 (7.9)	0.2 (4.2)
	X_D	33.3 (63.8)	**0.0 (2.4)**	2.8 (6.3)	11.3 (16.3)	2.3 (6.0)	0.0 (5.3)	0.0 (9.4)	0.0 (4.0)	***0.0 (6.6)***	11.7 (19.2)	0.0 (9.5)

Table 1: Average speaker WER of untuned, fully tuned, low-rank adapted Whisper models on test splits of X_P (Parkinson), X_S (stuttering) and X_D (dysarthria), reported as 'P50 (IQR)'. Best fully tuned setups in **bold**, and best low-rank tuned setups in ***bolded-italics***. Setups vary as to whether all components (All), only the encoder (Enc) or decoder (Dec), or the earlier (↓) or later (↑) layers thereof were trained/adapted. Percentages indicate the amount of tuned parameters with respect to full fine-tuning.

all parameters in either the encoder or decoder leads to the lowest WER in most cases. Notably, `large-v2` is able to reach median speaker-wise WERs of close to 0, with tight interquartile ranges within the usability threshold of 15% WER. This indicates that it should theoretically be possible to achieve good coverage of most common voice commands across our examined speech disorders, given sufficient model capacity and compute. Tuning earlier or later parts of the model, such as the early encoder layers and the final decoding head, exhibited higher instability and worse performance. These patterns generalize across the three types of atypical speech as well as across model sizes and multi/monolinguality.

4.3 Parameter-efficient Sub-layer Adaptation

Given that the optimal sub-architectures for full fine-tuning generalize well, we aim to localize the optimal sub-layer components for adaptation using LoRA (Hu et al., 2021). In contrast to residual adapters, which adapt only the the multi-layer perceptron (MLP) component of each Transformer block, LoRA allows for more targeted adaptation, including the query, value, and attention matrices in each layer. To target these individual parameters, LoRA augments any pre-trained weight matrix W by adding a trainable low-rank matrix ΔW. The adapted weight W' is defined by

$$W' = W + \Delta W = W + BA, \qquad (1)$$

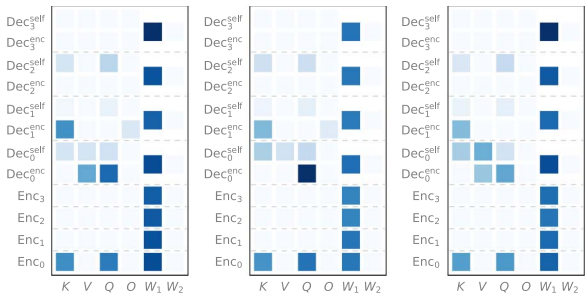
where ΔW is rank r and factorized by two low-rank matrices A and B.

We observe that adapting self-attention is unstable and rarely yields performance benefits. In contrast, the highest performance gains stem from adapting all components (Table 1), or the MLPs at the end of each layer. Applying LoRA ($r = 64$) to the MLPs alone, can thereby reduce training costs down to 4% of full fine-tuning, while retaining equivalent or better WER. These observations once again hold across all setups.

4.4 Parameter-level Adaptation Magnitudes

To understand the magnitude and localization of adaptations at a higher level of detail—specifically for individual parameters—we propose measuring the difference between each original weight W and its adapted matrix W' using Principal Subspace Angles (SSAs; Knyazev and Argentati, 2002). This measure keeps adaptation magnitudes comparable across different dimensionalities of W irrespectively of linear invariance by using the singular values of the transformation between the orthonormal bases of the two matrices to measure the ''energy'' required to map one to the other, expressed as an angle from 0° to 90° (similar/dissimilar).

We compute SSAs at the parameter level, plotting the resulting angles for `tiny.en` in Figure 2. We observe that the largest adaptation is concentrated in the first linear transformation W_1 of the MLP. Some adaptations are learned for the key K and query Q matrices of the early encoder and decoder layers, however these are



(a) X_D (b) X_S (c) X_P

Figure 2: Adaptation magnitudes of key K, value V, query Q, output projection O matrices of the self/cross-attention components, as well as the first W_1 and second W_2 layers of the MLP within each Whisper (`tiny.en`) encoder/decoder layer, measured in SSAs according to Section 4.4.

sparser and less pronounced. This pattern is consistent across all datasets and best-performing model configurations. To confirm these findings, we run experiments where only W_1 is adapted and compare the performance to when the entire MLP is adapted in Figure 3. We observe similar and even improved performance comparable to full fine-tuning in some cases. We thus conclude that W_1 is necessary and effective to adapt.

To pursue further parameter efficiency, we consider reducing the rank of the LoRA matrix ΔW, and focus on the decoder specifically, for which we observed larger improvements compared to tuning/adapting only the encoder (Table 1). Similar to observations in Hu et al. (2021), Figure 3 shows that adaptations are robust to the reduction in rank, down to even rank 2 and 1. By localizing the individual parameter type most relevant to adaptation, we are thus able to effectively halve WER while using 0.03% of the full parameter budget.

Based on these findings, our subsequent experiments for dynamic personalization via hypernetworks therefore focus on learning to adapt W_1 of each MLP in the decoder. Furthermore, each of these W_1 matrices will be adapted using LoRA following Equation 1 with rank $r = 2$.

4.5 Transferability across Etiologies

Despite the state-of-the-art parameter efficiency enabled by our previous analysis, adaptations are still cohort-specific, as in prior work. We next investigate the level of personalization required to adapt to different speaker cohorts. As shown in Figure 4, applying a model trained on one cohort to the same leads to the highest results, as expected. However, even within the same cohort, performance degrades for higher severities. While errors can be eliminated for mild and moderate cases, speakers with severe pathologies see the least benefit, even after full model fine-tuning. For dysarthric speech for instance, only fine-tuning or adapting the largest model yields error rates in a usable range.

Across datasets, we observe some transferability, as training on any type of atypical speech seems to improve performance on other types at least marginally. Training on X_P and X_S appears to transfer slightly better to each other and to X_D than vice-versa. This could be an effect of the mild and moderate cases of stuttering not differing as strongly phonetically from typical speech as dysarthria. Also, LoRA appears to allow for more stable transferability across different atypical speech types, while preserving original performance, as shown especially for the model adapted to X_D. This may be because X_D consists of a small vocabulary with repetitive utterances, making it prone to over-fitting when full-rank fine-tuning, whereas LoRA provides some regularization via the smaller number of adaptable parameters.

The detailed separation of severities across these transfer results also provides indication of these methods' performance on typical speech from the same domains: Mild stuttering ($X_{S,mild}$) contains only few dysfluencies and typical pronunciation compared to X_D or X_P. It is also the category with the consistently lowest WERs across training data regimens, including the untuned model (corresponding to Whisper's state-of-the-art transcription performance on typical speech at its time of publication; Radford et al., 2023). Nonetheless, our experiments demonstrate the need for finer-grained personalization, as populations with severe pathologies still see the least benefits from personalization, even within their own speech disorder cohort.

5 Dynamic Personalization

Our previous findings generalize across speech disorders to support higher parameter efficiency than prior work. However, in practice, cohort-level personalization still requires knowledge of the

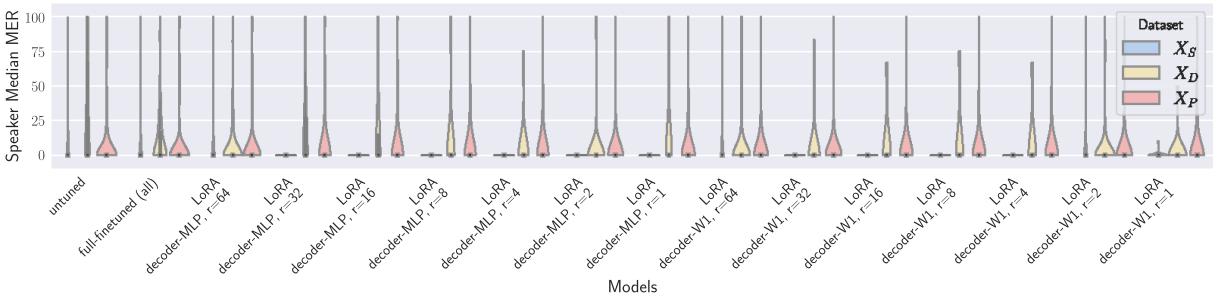
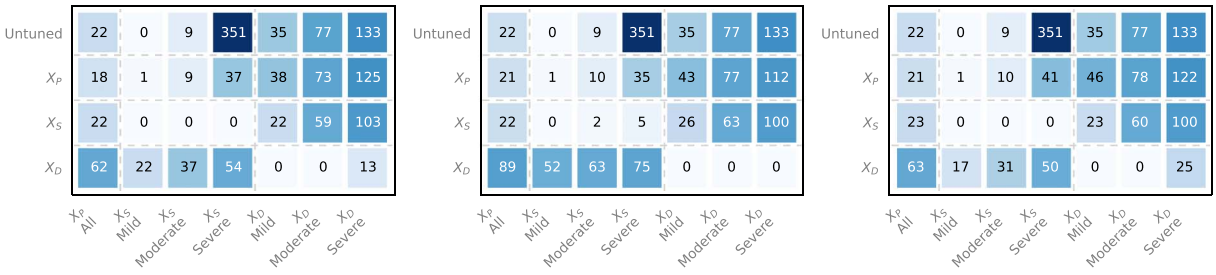


Figure 3: Speaker median MER on X_S (stuttering), X_D (dysarthria) and X_P (Parkinson) of Whisper (`large-v2`) untuned, fully tuned, and adapted using LoRA at both MLP layers or W_1, using $r \in [2; 64]$.



(a) Tuned (Full) (b) Tuned (Partial) (c) LoRA

Figure 4: Average speaker WER of Whisper (`tiny.en`) untuned and best fully/partially/LoRA configurations, across X_P (Parkinson), X_S (stuttering), and X_D (dysarthria), with various degrees of severity.

target etiology, as a model trained on data from one cohort does not transfer well to another. Additionally, severe cases of atypical speech, being rarer within the cohort, see less improvement than mild and moderate cases. Therefore, a data-centric design that is more cognizant of both speech disorder type and severity is necessary for successful personalization. Towards this goal, our second contribution is the design and use of light-weight hypernetworks (Ha et al., 2017) to dynamically generate personalized adaptation weights—essentially generating a new adapted model for each utterance at inference time.

5.1 Hypernetworks

We propose that the hypernetwork is a function $H(s, c; \theta)$ with trainable parameters θ and inputs consisting of a speaker characteristics vector s and the context of the generation c, such as the parameter type being adapted, as well as its location within the model. Both s and c can be manually defined (e.g., user self-identification, expert heuristics), based on external pre-trained models (e.g., speaker encoders), and/or acquired jointly during downstream meta-adaptation. The output of $H(s, c; \theta)$ are the vectorized LoRA-A and B matrices, which are reshaped and applied to the pre-trained weight matrix W following Equation 1. In our experiments, we explore two functional forms of $H(s, c; \theta)$, namely, a linear system, and an MLP with one hidden layer and ReLU activations.

Figure 5 shows the proposed architecture to adapt a parameter W, where the speaker characteristics vector s is computed using a speech encoder model. In our design, s is computed on the utterance-level, although it is possible to replace this with a coarser speaker-level characterisation. Additionally, $H(s, c; \theta)$ consists of separate output heads for predicting A and B, whose weights are respectively denoted by θ_A and θ_B, while the remainder of the hypernetwork is shared.

5.2 Hypernetwork Initialization

When training the hypernetwork, we recommend to not trivially initialize it randomly, since the generated adaptations will equate to random pertubations that are detrimental to any existing model capabilities. For language modeling, Phang et al. (2023) propose an additional hyper-pre-training

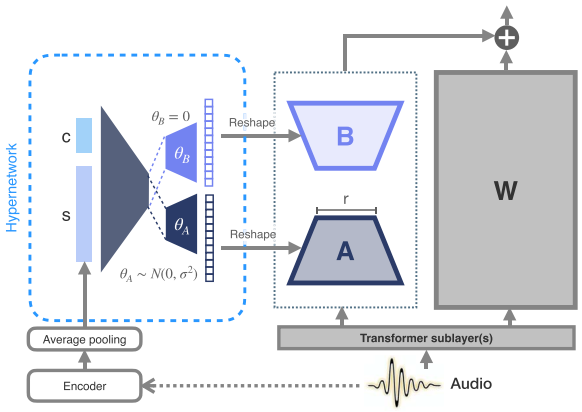


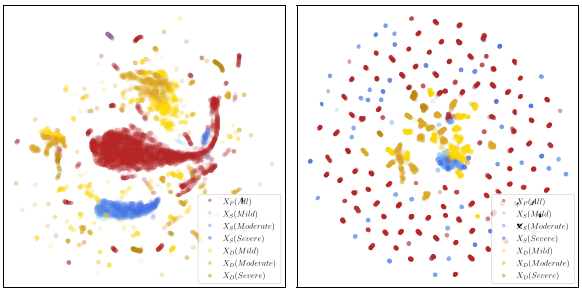
Figure 5: Hypernetwork for adapting W with LoRA weights A and B, generated by θ_A and θ_B, respectively initialized with $\mathcal{N}(0, \sigma^2)$ and zeroes. Generation is conditioned on speaker characteristics s from an audio encoder, and generation context c denoting the target parameter's location. All trainable parameters are within the hypernetwork.

phase to first learn a hypernetwork initialization matching the host model's parameter space. However, this approach is resource-intensive, requiring over 50k additional training steps, and cannot be trivially applied to ASR in a similar self-supervised manner.

Instead, we propose a simpler approach, which more closely follows the original LoRA design: specifically, initial adaptation weights, which leave the model unaugmented. For our hypernetworks, we propose implementing this design by initializing θ_B at zero, thereby nulling out any changes brought about by the initial ΔW. Simultaneously, θ_A is initialized close to zero, but randomly, ensuring gradient flow during backpropagation. We found this design choice to be crucial for training as it enables learning solutions that initially match the target model's parameter space without catastrophically deteriorating performance with random noise.

5.3 Speaker Characterization

As the speaker characteristics s must encode all necessary information for the hypernetwork to generate effective adaptation weights for personalization, we studied different audio-based encoding strategies to identify which factors are crucial to downstream performance. While it is possible to use manual features, such as flags to indicate speaker characteristics, we use automatic, pretrained speech encoder models which do not require expert annotations.



(a) Whisper (b) ECAPA-TDNN

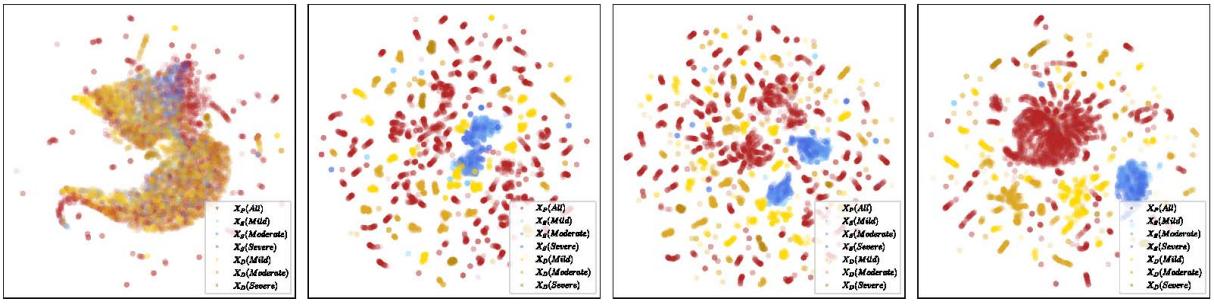
Figure 6: UMAP (Ghojogh et al., 2023) visualization of speaker characterization vectors s from the last encoder layer of Whisper (`large.v2`), and from the ECAPA-TDNN speaker verification model.

In our explorations, we ablate s from lower-level acoustic to higher-level concepts, such as speaker identity, leveraging speech encoder models either trained for ASR or speaker verification, respectively denoted as s_{ASR} and s_{SV}. For s_{ASR}, we use different layers from the encoder of Whisper `tiny`, `tiny.en` and `large-v2`, while for s_{SV}, we used a speaker verification model from the Speech Brain project (Ravanelli et al., 2021). The speaker verification model uses an Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network (ECAPA-TDNN; Desplanques et al., 2020), and is trained on the VoxCeleb datasets (Nagrani et al., 2017; Chung et al., 2018; Nagrani et al., 2020).

As shown in Figure 6, we observe that the s_{ASR} embeddings are localized by etiology more strongly than s_{SV}. The continuity of the embeddings manifold with regards to etiology also seems to be a benefit when training the hypernetwork as we had success with s_{ASR} but not with s_{SV}.

Additional to the learning task for which the speech encoder is trained on, we further explored the expressiveness of s_{ASR} when computed using earlier or later layers of the encoder. As shown in Figure 7, the clustering of etiology becomes more apparent in the later layers of the encoder.

Overall, we observe that effective, utterance-level adaptations require the hypernetwork to be conditioned on speaker characteristics s, which cover a continuous space with respect to a diverse set of features such as speaker characteristics and sufficient expressiveness of part-word acoustical units. The expressiveness of s dictates whether different parameters can be generated to accommodate various speech disorders and severities thereof. The task of speaker verification, while



(a) Layer 0 (b) Layer 7 (c) Layer 15 (d) Layer 23

Figure 7: UMAP (Ghojogh et al., 2023) visualization of Whisper (`large.v2`) encoder embeddings from various layer depths. Each point corresponds to an utterance from X_S (stuttering), X_D (dysarthria), and X_P (Parkinson), with color and hue corresponding to the dataset and severity of the speaker (severity annotations unavailable for X_P).

encoding individuals distinctly, does not appear to benefit sharing of lower-level acoustic characteristics. ASR on the other hand encodes acoustic properties more continuously, and s_{ASR} is most effective, when higher-level features correlated with etiology begin to be encoded as well (i.e., in deeper layers).

5.4 Final Adaptation Architecture

Based on our findings for improving parameter efficiency (Section 4) and hypernetwork construction (Section 5), our final dynamic adaptation architecture $H(s, c; \theta)$ is built as follows: A linear/MLP-based hypernetwork, which generates adaptations for the W_1 parameter type with rank $r = 2$. Since the hypernetwork is shared for all instances of this parameter, c provides context for the location of adaptation within the model in the form of a one-hot embedding lookup that is learned jointly during adaptation. The speaker characterization stems from the final encoder layer of Whisper `large-v2`, and is mean-pooled over time. For each forward pass, H generates all adaptations for a given utterance and inserts them into the model dynamically.

6 Results

We next compare our proposed approach to the *full*, *partial*, and *LoRA* baselines outlined in Section 3, and report the findings in Tables 2 and 3 in decreasing order of heterogeneity of the training data, namely at the *global*, *cohort*, and *individual* level. Results are reported for Whisper `large-v2`, which represents the upper bound in terms of performance.

6.1 Global Adaptation

From Table 2, we observe that global adaptation, i.e., training on data from all cohorts simultaneously, works well for people with mild to moderate speech differences. Indeed, the speaker-wise median WER of 0 reflects our initial observation from Section 4 in that the majority of common voice commands are covered well using most adaptation approaches. As mentioned in Section 4.5, $X_{S,mild}$ further indicates that all approaches would likely perform comparably to the untuned model on typical speech. Even on X_P, which has the most diverse set of utterances, the WER and IQR typically fall within the 15% usability threshold. This global approach further circumvents the need of managing cohort-specific models, however some cohorts are more represented than others, leading to model bias against rarely observed cohort characteristics. For example, we generally see higher WERs for higher severities, which are more rarely observed, and notably poorer performance for severe dysarthria $X_{D,sev}$ with only two speakers seen during training. These effects are most prominent when the tunable parameter budget is low, as with LoRA. Given the ability to tune larger parts of the model, i.e., full and partial fine-tuning, the globally tuned model can still be applied to a broader range of atypical speech types, however this requires tuning 72%–100% of the 1.6B parameters. Our proposed approach of using hypernetworks to dynamically generate personalized adaptations appears to most effectively leverage global data

Setup	Untuned	Full Tuning		Partial Tuning		LoRA		Hyper	Hyper$_{\mathbb{P}}$
	–	Global	Cohort	Global	Cohort	Global	Cohort	Global	Cohort
	0%	100%	100%	72%	72%	3%	3%	0.1%	0.1%
X_P	8.5 (17.7)	**4.4 (9.4)**	5.2 (9.4)	5.0 (9.8)	5.0 (10.1)	6.9 (14.9)	7.1 (16.3)	***6.0 (13.8)***	–
X_S	19.3 (171.0)	0.0 (1.9)	**0.0 (1.0)**	0.0 (2.5)	0.0 (1.5)	0.2 (6.2)	***0.0 (4.0)***	***0.0 (4.0)***	0.0 (8.9)
$X_{S,mild}$	0.0 (0.9)	**0.0 (0.0)**	**0.0 (0.0)**	**0.0 (0.0)**	**0.0 (0.0)**	0.0 (22.9)	***0.0 (0.0)***	***0.0 (0.0)***	0.0 (0.0)
$X_{S,mod}$	1.5 (12.4)	0.0 (1.6)	**0.0 (0.7)**	0.0 (2.0)	**0.0 (0.7)**	0.4 (4.4)	0.0 (3.6)	***0.0 (2.9)***	0.0 (6.4)
$X_{S,sev}$	70.6 (626.6)	0.5 (4.4)	**0.0 (2.6)**	0.0 (6.0)	0.0 (4.5)	0.0 (15.7)	***0.0 (8.8)***	0.0 (10.0)	0.0 (22.3)
X_D	33.3 (63.8)	0.0 (9.5)	**0.0 (2.4)**	0.0 (9.5)	0.0 (5.3)	15.3 (35.1)	***7.1 (14.7)***	8.3 (8.9)	24.3 (54.6)
$X_{D,mild}$	0.0 (38.9)	**0.0 (0.0)**	**0.0 (0.0)**	**0.0 (0.0)**	**0.0 (0.0)**	0.0 (6.5)	***0.0 (0.0)***	***0.0 (0.0)***	0.0 (27.8)
$X_{D,mod}$	44.4 (87.4)	**0.0 (0.0)**	**0.0 (0.0)**	**0.0 (0.0)**	**0.0 (0.0)**	7.87 (49.4)	0.0 (10.8)	***0.0 (0.0)***	25.5 (74.1)
$X_{D,sev}$	100.0 (68.1)	0.0 (66.7)	**0.0 (16.7)**	0.0 (66.7)	0.0 (37.2)	83.3 (77.8)	***50.0 (70.6)***	58.3 (62.5)	93.3 (76.4)
X_{P+S+D}	16.2 (83.7)	**2.1 (6.5)**	2.4 (5.2)	2.4 (6.9)	2.3 (6.1)	5.5 (14.41)	4.4 (11.3)	***4.0 (9.3)***	–
$X_{\mathbb{P}}$	11.8 (33.6)	–	**1.4 (7.7)**	–	2.5 (10.0)	–	***3.5 (16.3)***	–	4.1 (17.1)

Table 2: Average speaker WER of untuned, fully/partially tuned and LoRA/hypernetwork-adapted Whisper (`large-v2`) on test splits of X_P (Parkinson), X_S (stuttering), and X_D (dysarthria), reported as 'P50 (IQR)'. Best fully tuned setups in **bold**, and best PEFT setups in ***bolded-italics***. Models were trained on the *global* concatenation of datasets X_{P+S+D}, or on each individual target *cohort*. On the public benchmark $X_{\mathbb{P}}$, we further report cohort-level adaptation results, including for Hyper$_{\mathbb{P}}$ solely trained on $X_{\mathbb{P}}$. Note that models trained on X_P are not evaluated on $X_{\mathbb{P}}$ (and vice-versa), due to speaker overlap. Percentages indicate the amount of trainable parameters with respect to full fine-tuning.

Setup	Full	Partial	LoRA	Hyper
	100%	72%	3%	0.1%
X_P	4.7 (9.3)	**4.4 (9.3)**	8.1 (16.8)	***7.1 (23.9)***
X_S	**0.0 (1.5)**	0.0 (1.9)	1.4 (12.2)	***0.0 (5.1)***
X_D	**0.0 (1.1)**	0.0 (4.4)	***0.0 (6.7)***	7.0 (7.9)
X_{P+S+D}	2.2 (5.1)	**2.1 (5.8)**	4.4 (13.7)	***4.3 (14.3)***

Table 3: Average speaker WER of individual-level personalized Whisper models on test splits of X_P, X_S, and X_D, reported as 'P50 (IQR)'. Best fully tuned setups in **bold**, and best PEFT setups in ***bolded-italics***. Full/partial/low-rank adaptations are trained using 70% of an individual's data, while the hypernetwork generates zero-shot adaptations without any data from the individual. Fine-tuned setups are initialized using the corresponding cohort model. Percentages indicate the amount of trainable parameters with respect to full tuning.

sharing, outperforming standard LoRA and even full fine-tuning on $X_{S,sev}$, for an overall WER of 4.0, while using 0.1% of the full parameter budget. It further maintains the base model's original performance on close-to-typical speech ($X_{S,mild}$) best, as indicated by its substantially lower IQR compared to LoRA.

6.2 Cohort-level Adaptation

When knowledge of a speaker's cohort-membership is available, we observe from Table 2 that fine-tuning on cohort-specific data provides better performance than global adaptation regardless of the fine-tuning technique applied, alluding to the need for a higher degree of personalization. However, this increased granularity comes at the cost of necessitating one model per atypical speech type.

In contrast, we find that despite having to share a single model across all cohorts, lacking explicit knowledge of a target speaker's etiology, and representing a magnitudes smaller architecture, hypernetworks are able to generate adaptations which are competitive to cohort-level full fine-tuning, and LoRA. This improvement in performance could be attributed to the relatively flexible inductive bias imposed on the hypernetwork, allowing it to share representations that may be beneficial across heterogeneous cohorts. The cross-cohort transfer of Hyper$_{\mathbb{P}}$ to X_S and X_D reflects this capability in particular, as it has substantially lower WER than the untuned model, despite only being trained on the $X_{\mathbb{P}}$ cohort. As the global hypernetwork, trained on X_{P+S+D}, nonetheless outperforms the cohort variant, training on a diverse set of speech characteristics appears crucial for learning sharable

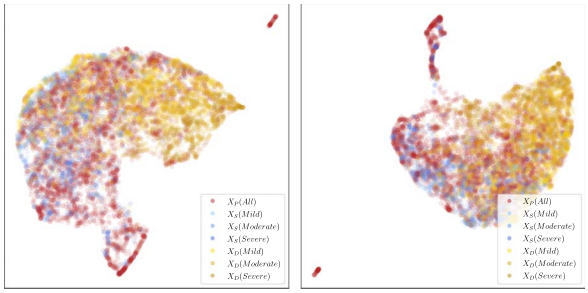
representations. We further examine this hypothesis in Section 6.4.

6.3 Individualized Adaptation

Moving to the highest level of personalization, we next compare our zero-shot hypernetworks to full/partial/LoRA-tuned, individually personalized ASR models in Table 3. For the baselines, 70% of the data for each test subject in Table 2 is used to continually fine-tune their relevant cohort-level model, while for the hypernetwork, we evaluate on the same 30% remainder of the test data, but do not train on any target speaker data. Similarly to prior work utilizing target speaker data—either via continual fine-tuning (Tomanek et al., 2021a) or via the fusion of multiple individualized models based on their similarity to the target speaker (Qi and van Hamme, 2023)—this individual-level personalization generally improves the baselines' performance. However these approaches require training, retaining and selecting an even higher number of models than at the cohort-level. Conversely, the adaptations generated by our single hypernetwork remain competitive, especially to individualized LoRA, despite not observing any training data from the target speaker and having no information regarding their cohort-membership. Relative to the performance upper-bound of full and partial fine-tuning, the hypernetwork provides 2% higher WER, however remaining far below the usability threshold of 15% WER for most speakers, while requiring up to three orders of magnitude fewer tuned parameters.

6.4 Analysis of Parameter Space

To gain a better understanding of how hypernetworks balance global knowledge sharing while maintaining effective individualized personalization, we perform an analysis of the generated parameter manifold with respect to the different atypical speech cohorts being represented. Figure 8 shows a subsample of generated parameters for 10k utterances in the test set. Regardless of the hypernetwork's functional form, there are regions where generated adaptations overlap across all three datasets. We also observe the general trend that there is some overlap between dysfluent utterances X_S and those associated with Parkinson X_P, while there is far less sharing between X_P and X_D. This aligns with the previous observations from Table 2, where the hypernetwork trained solely on Parkinson data transfered well to stuttering, but not to speech consistent with Cerebal Palsy. As these overlaps do not occur in the original speech characterizations s (Figures 6 and 7), we stipulate that the hypernetwork was able to learn common adaptations that generalize well cross different speech disorders, while simultaneously generating adaptations that are unique to an etiology as seen in the non-overlapping regions.



(a) Linear (b) MLP

Figure 8: UMAP (Ghojogh et al., 2023) visualization of adaptions W' for 10k utterances from X_S (stuttering), X_D (dysarthria), and X_P (Parkinson) generated by a linear or MLP-based hypernetwork, colored by dataset and severity.

In terms of the functional form of H, we observe only minor performance differences between the linear and MLP variant, with the latter exhibiting slightly lower WER overall. Together with the similar overlaps exhibited by the generated parameters, this points towards the general meta-learned individualization and knowledge sharing approach being more crucial than the exact form of the architecture employed to achieve this goal.

6.5 Alternative Backbone Architectures

So far, we have found hypernetworks to consistently generate effective adaptations compared to more training-intensive approaches, across global, cohort, and individual-level personalization as heterogeneity of atypical speech characteristics increase (Tables 2 and 3). To explore how generalizable our findings are with respect to alternative backbone architectures, we next apply hypernetworks to host models with different architectures and/or pre-training data, namely, Whisper `tiny` and `tiny.en` respectively. Both consist of 4 instead of `large-v2`'s 32 encoder/decoder layers, having 0.02% the size of the larger architecture, with `tiny.en` having further been trained on

Setup	tiny		tiny.en	
	LoRA 0.08%	Hyper 1.47%	LoRA 0.08%	Hyper 1.47%
X_P	21.7 (36.0)	**20.4 (35.9)**	20.1 (31.7)	**20.1 (31.2)**
X_S	42.0 (69.5)	**28.4 (49.5)**	24.5 (56.5)	**19.7 (36.4)**
X_D	7.0 (24.7)	**5.4 (22.0)**	6.7 (21.5)	**5.1 (21.8)**
X_{P+S+D}	18.9 (36.4)	**15.7 (32.4)**	15.5 (31.3)	**14.2 (28.3)**

Table 4: Average speaker WER of LoRA and hypernetwork trained on X_{P+S+D}, applied to Whisper `tiny`/`tiny.en`, and evaluated on test splits of X_P, X_S, and X_D, reported as 'P50 (IQR)', with best setups in **bold**. Percentages indicate the amount of tuned parameters with respect to full fine-tuning.

English speech alone. Based on the findings from Figure 2, which indicate that W_1 accounts for the largest adaptations, we focus on adapting only the W_1 matrices of the encoder and decoder of the `tiny/tiny.en` models.

Table 4 shows that hypernetworks continue to consistently outperform LoRA across all datasets, despite being trained just once overall, instead of once per-cohort, and having a substantially smaller parameter budget. Similarly to our initial experiments in Section 4, we observe slightly higher decoding stability and fewer hallucinations for the monolingual `tiny.en` model. In general, the smaller backbones have a lower base performance compared to `large-v2`, however, relative to untuned `tiny.en`, we are nonetheless able to reduce WERs from 22.4 $\rightarrow$ 20.1 for X_P, 97.3 $\rightarrow$ 19.7 for X_S, and 67.4 $\rightarrow$ 5.1 for X_D, using hypernetworks-generated LoRA. With an average WER of 14.2 across datasets, hypernetworks further bring us into the range of practical usability, despite the much more parameter-constrained environment. The overall approach of using hypernetworks to dynamically generate adaptation parameters therefore seems robust to not only different types of atypical speech (Sections 6.1, 6.2 and 6.3), and choice of parameter generator (Section 6.4), but also with respect to the host model architecture.

7 Conclusion

Due to data scarcity and the highly variable nature of atypical speech, prior work relied on cohort-level fine-tuning and PEFT, followed by individualized adaptation. While this approach substantially reduces WER, it necessitates training as many models as there are speakers, and requires expert knowledge of the cohort a speaker belongs to. In Section 4.5, we demonstrate that this knowledge is critical, as a model trained on one cohort does not transfer to others. In addition, higher severity speakers and those with multiple pathologies still do not benefit from the improvements for mild and moderate cases, as they are least represented in the data.

Combining meta-level knowledge sharing and highly individualized personalization, we therefore proposed using hypernetworks to dynamically generate adaptations during inference (Section 5). As generating an entire model is computationally infeasible, we first conducted a study regarding which individual model parameters have the highest influence on adaptation performance (Section 4). Analyzing model components at increasing levels of detail using a novel combination of LoRA and SSAs, we identified a single parameter type W_1 which contributes most to adaptation performance. As this effect was consistent across different model sizes, pre-training strategies and LoRA ranks, we were able to halve WER while using 0.03% of the parameter budget required for full fine-tuning.

Based on these findings, we were able to scale our hypernetwork-based ASR adaptation approach to an even smaller parameter budget of 0.01%. Despite sharing a single model across cohorts and having access to neither cohort or individual speaker information, hypernetworks reach a WER of 4.0, consistently outperforming LoRA, and performing competitively to full fine-tuning (Section 6.1). These results hold, even when the latter two approaches are specifically fine-tuned using in-cohort data (Section 6.2) and/or training data from a target individual (Section 6.3). Further ablating the hypernetworks' parameter generators (Section 6.4) and host model architectures (Section 6.5), we find our general meta-learning approach to generalize well across both datasets and models.

Our analyses in Sections 5.2, 5.3, and 6.4 further surface factors critical to hypernetwork performance: Firstly, improving upon prior approaches, we find that nulling out initial adaptations is crucial for not deteriorating existing model performance. Second, for speaker characterization, continuous coverage across acoustic

features appears more important than discrete speaker featurizations. Third, the hypernetwork meta-learns adaptation weights which exhibit overlaps, matching etiological intuitions beyond the information present in the input embeddings. Improving upon prior work, we demonstrate that this general approach holds across multiple types of atypical speech, enabling zero-shot dynamic personalization without explicit knowledge of cohort membership, nor training data from the target speaker.

In this work, we target a diverse set of dysfluency and phonology-related speech disorders. However, future work may investigate other categorizations of atypical speech that are not included in our datasets. Additionally, we believe exploring a broader range of ASR scenarios could be fruitful, especially since standard adapters have been shown to work well for heavy-accented speech, and multi-lingual settings (Le et al., 2021; Tomanek et al., 2021b). Applying hypernetworks to these, as well as a myriad of other downstream scenarios, could therefore form an interesting extension to this work. Lastly, our experiments centered around multiple variants of Whisper, confirming our findings across different model sizes and pre-training data regimens. We believe future work could follow our general approach for identifying relevant adaptation parameters, and subsequently generate them using meta-learned hypernetworks, in order to better optimize parameter efficiency when adapting alternative backbone architectures in low-resource scenarios.

Acknowledgments

Thanks to Leah Findlater and Jeff Bigham for supporting this project, as well as to the TACL action editor and the anonymous reviewers for their helpful comments and insightful discussions.

References

Speech Accessibility Project. https://speechaccessibilityproject.beckman.illinois.edu. Accessed: 2023-08-22.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.410

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Proceedings of Interspeech 2018*, pages 1086–1090. https://doi.org/10.21437/Interspeech.2018-1929

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proceedings of Interspeech 2020*. https://doi.org/10.21437/Interspeech.2020-2650

Joseph R. Duffy. 1995. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. St. Louis. Mosby.

Benyamin Ghojogh, Mark Crowley, Fakhri Karray, and Ali Ghodsi. 2023. *Uniform Manifold Approximation and Projection (UMAP)*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-10602-6_17

Jordan R. Green, Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, and Katrin Tomanek. 2021. Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. In *Proceedings of Interspeech 2021*, pages 4778–4782. https://doi.org/10.21437/Interspeech.2021-1384

David Ha, Andrew M. Dai, and Quoc V. Le. 2017. Hypernetworks. In *International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Computing Research Repository (CoRR), arXiv e-prints, 2106.09685*.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. 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 565–576, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.47

Andrew V. Knyazev and Merico E. Argentati. 2002. Principal angles between subspaces in an A-based scalar product: Algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040. https://doi.org/10.1137/S1064827500377332

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 817–824. https://doi.org/10.18653/v1/2021.acl-short.103

Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P. Bigham, and Leah Findlater. 2023. From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581224

Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, Jordan R. Green, and Katrin Tomanek. 2021. Disordered speech data collection: Lessons learned at 1 million utterances from Project Euphonia. In *Proceedings of Interspeech 2021*, pages 4833–4837. https://doi.org/10.21437/Interspeech.2021-697

Omar Caballero Morales and Stephen Cox. 2007. Modelling confusion matrices to improve speech recognition accuracy, with an application to dysarthric speech. In *Proceedings of Interspeech 2007*, pages 1565–1568. https://doi.org/10.21437/Interspeech.2007-126

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In *Proceedings Interspeech 2004*, pages 2765–2768. https://doi.org/10.21437/Interspeech.2004-668

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*, 60(C). https://doi.org/10.1016/j.csl.2019.101027

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A large-scale speaker identification dataset. In *Proceedings of Interspeech 2017*, pages 2616–2620. https://doi.org/10.21437/Interspeech.2017-950

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. https://doi.org/10.18653/v1/2021.eacl-main.39

Jason Phang, Yi Mao, Pengcheng He, and Weizhu Chen. 2023. HyperTuning: Toward adapting large language models without backpropagation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27854–27875. PMLR.

Jinzi Qi and Hugo van Hamme. 2023. Parameter-efficient dysarthric speech recognition using adapter fusion and householder transformation. In *Proceedings of INTERSPEECH 2023*, pages 151–155. https://doi.org/10.21437/Interspeech.2023-1627

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.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad,

Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. *Computing Research Repository*, arxiv:2106.04624. Version 1.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Glyndon D. Riley. 2009. *SSI-4: Stuttering Severity Instrument Fourth Edition*. Pro-Ed Inc.

Hannah P. Rowe, Sarah E. Gutz, Marc F. Maffei, Katrin Tomanek, and Jordan R. Green. 2022. Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective. *Frontiers in Computer Science*, 4. https://doi.org/10.3389/fcomp.2022.770210, PubMed: 37860708

Eric K. Sander. 1963. Frequency of syllable repetition and 'stutterer' judgments. *Journal of Speech and Hearing Disorders*, 28(1):19–30. https://doi.org/10.1044/jshd.2801.19, PubMed: 13976192

Theresa Schölderle, Anja Staiger, Renée Lampe, Katrin Strecker, and Wolfram Ziegler. 2016. Dysarthria in adults with cerebral palsy: Clinical presentation and impacts on communication. *Journal of Speech, Language, and Hearing Research*, 59(2):216–229. https://doi.org/10.1044/2015_JSLHR-S-15-0086, PubMed: 27057824

Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, Avinatan Hassidim, and Yossi Matias. 2019. Personalizing ASR for Dysarthric and Accented Speech with Limited Data. In *Proceedings of Interspeech 2019*, pages 784–788. https://doi.org/10.21437/Interspeech.2019-1427

Kris Tjaden. 2008. Speech and swallowing in parkinson's disease. *Topics in Geriatric Rehabilitation*, 24(2):115–126. https://doi.org/10.1097/01.TGR.0000318899.87690.44, PubMed: 19946386

Katrin Tomanek, Françoise Beaufays, Julie Cattiau, Angad Chandorkar, and Khe Chai Sim. 2021a. On-device personalization of automatic speech recognition models for disordered speech. *Computing Research Repository*, arxiv:2106.10259. Version 1.

Katrin Tomanek, Katie Seaver, Pan-Pan Jiang, Richard Cave, Lauren Harrell, and Jordan R. Green. 2023. An analysis of degenerating speech due to progressive dysarthria on asr performance. In *International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, pages 1–5. https://doi.org/10.1109/ICASSP49357.2023.10097195

Katrin Tomanek, Vicky Zayats, Dirk Padfield, Kara Vaillancourt, and Fadi Biadsy. 2021b. Residual adapters for parameter-efficient ASR adaptation to atypical and accented speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6751–6760, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.541

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2022a. UDapter: Typology-based language adapters for multilingual dependency parsing and sequence labeling. *Computational Linguistics*, 48(3):555–592. https://doi.org/10.1162/coli_a_00443

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord, and Sebastian Ruder. 2022b. Hyper-X: A unified hypernetwork for multi-task multilingual transfer. https://doi.org/10.18653/v1/2022.emnlp-main.541

Dianna Yee, Colin Lea, Jaya Narain, Zifang Huang, Lauren Tooley, Jeffrey P. Bigham, and Leah Findlater. 2023. Latent phrase matching for dysarthric speech. In *Proceedings INTERSPEECH 2023*, pages 161–165. https://doi.org/10.21437/Interspeech.2023-1921