forum acusticum 2026
8-12 September · Graz, Austria

HYDRA VI-LORA: SPEAKER-AWARE MIXTURE-OF-EXPERTS FOR PERSONALIZED RECOGNITION OF ATYPICAL SPEECH

Jingjing Qiu[1], Niclas Pokel[2,4], Pehuen Moure[2,4], Yingqiang Gao[3,4], Roman Boehringer[2]

[1]*ETH Zurich, Zurich, Switzerland*
[2]*Institute of Neuroinformatics (INI), University of Zurich and ETH Zurich, Zurich, Switzerland*
[3]*Department of Computational Linguistics, University of Zurich, Zurich, Switzerland*
[4]*ETH AI Center, Zurich, Switzerland*

Abstract

Automatic Speech Recognition (ASR) for individuals with impaired speech remains a challenging task due to severe data scarcity and substantial acoustic variability. To better adapt to such variability, we incorporate Hydra-LoRA, a mixture-of-experts-style parameter-efficient fine-tuning method, in which routers dynamically allocate expert distributions. Since distinctive speech characteristics may appear at both the speaker level and the token/phoneme level, we investigate different routing strategies: global routing, which uses a single speaker-level representation to determine the expert distribution for all layers, and layer-wise routing, where each adapted module assigns experts based on layer-specific representations. To improve activation efficiency, we adopt a top-k routing mechanism and design a load-balancing loss to encourage more even expert utilization. We evaluate our method on two non-normative speech datasets, UASpeech and SAP. Experimental results show that our approach outperforms the standard LoRA baseline by 2–3 points in both WER and CER, while maintaining an equivalent number of activated parameters. We further analyze router distribution patterns using LibriSpeech, a normative speech dataset, and observe clear distributional differences between normative and non-normative speakers, demonstrating both the effectiveness and the interpretability potential of the proposed routing mechanism. Overall, dynamic expert routing is a promising direction for adapting ASR to highly variable impaired speech.

Keywords: *impaired speech recognition, parameter-efficient fine-tuning, mixture of experts, low-rank adaptation, expert routing*

***Corresponding author**: npokel@ethz.ch.*

1. Introduction

Automatic Speech Recognition (ASR) has made substantial progress with large-scale speech models, yet these systems perform substantially worse on impaired or otherwise non-normative speech. For affected individuals this gap is an accessibility issue: unreliable recognition restricts voice-based interfaces, communication tools, and assistive technologies [1], [2], [3], [4].

Impaired speech is difficult to recognize because its acoustic patterns differ from normative speech in highly speaker-dependent ways, including atypical articulation, unstable speaking rates, irregular prosody, and inconsistent phoneme pronunciation. Since impaired-speech data is also scarce, training robust models from scratch or fully fine-tuning large ASR models risks overfitting; effective adaptation must be data-efficient yet flexible enough to capture heterogeneous speech patterns [1], [2], [3]. Even modern multimodal audio-language models fail to accurately transcribe impaired speech [5].

Parameter-efficient fine-tuning (PEFT) methods such as LoRA adapt large pre-trained models with few trainable parameters [6], but a single adapter imposes a uniform adaptation path for all inputs, averaging over distinct patterns—some global and speaker-level, others tied to specific phonetic contexts—rather than learning specialized behaviors. This motivates a mixture-of-experts (MoE)-style architecture: multiple expert adapters combined with a routing mechanism that dynamically determines which experts are activated, so that experts can specialize in different speakers, impairment-related acoustic patterns, or phonetic difficulties, while sparse routing keeps the number of activated parameters comparable to a standard adapter [7], [8], [9], [10].

The main contributions of this project are: (i) we adopt a dynamically routed LoRA-based mixture-of-experts architecture for parameter-efficient adapta-

tion; (ii) we compare global and layer-wise routing strategies to study how routing granularity affects adaptation; and (iii) we analyze router distributions on normative and non-normative speech to explore the interpretability of the learned expert allocation.

2. Related Work

Parameter-efficient fine-tuning for ASR.
PEFT avoids the cost and overfitting risk of fully fine-tuning a large backbone by updating only a small set of additional parameters: LoRA parameterizes the update of a linear layer with a low-rank decomposition [6], and adapter-based PEFT has been benchmarked on multiple speech tasks [11], [12]. For impaired speech, Adapter Fusion supports target-speaker adaptation for dysarthric speech [13], hypernetworks enable utterance-level personalization with small parameter budgets [14], and VI-LoRA models LoRA parameters as variational distributions for uncertainty-aware personalization [15].

Mixture-of-experts routing and multi-LoRA.
Sparsely-gated MoE showed that conditional computation over multiple expert modules with a learned router can scale total parameters while activating only a few experts [7], an idea adopted in GShard, Switch Transformer, and GLaM [8], [16], [17]. Routing granularity is a key design choice: coarser task-level routing supports efficient task-specific inference [18], paralleling our comparison of global and per-layer routing. Load-balancing regularization [8] and alternative assignment strategies [19], [20] address expert collapse and training stability, and interaction-aware routing has been explored in multimodal settings [21].
Closest to our work, HydraLoRA shares the LoRA down-projection while routing among expert-specific up-projections [9]; Mixture of LoRA Experts [10] and mixture-of-LoRA strategies for low-resource multi-accent ASR [22] follow similar ideas. We adopt this line for impaired speech, asking whether expert allocation should be driven by global speaker characteristics or finer-grained layer-wise information.

ASR for non-normative and impaired speech.
Dysarthric and other non-normative speech differs substantially from normative speech: UA-Speech is a widely used dysarthric isolated-word benchmark [2], recognition difficulty correlates with reduced intelligibility [23], and surveys highlight data scarcity, speaker heterogeneity, and distribution mismatch as recurring challenges [3], [24]. Personalizing ASR with limited dysarthric or accented data substantially improves recognition [1]; recent work explores adaptation of modern backbones, data augmentation, hyperparameter adaptation, and uncertainty-aware methods [15], [25], [26], and the Speech Accessibility Project (SAP) provides a large-scale benchmark across diverse impairment conditions [27]. We instead investigate whether routed multi-expert low-rank adaptation captures heterogeneous non-normative speech better than a single LoRA adapter.

3. Methodology

We characterize Hydra-LoRA [9], an MoE-style PEFT framework, for impaired speech adaptation. It replaces the single LoRA adapter with multiple expert-specific branches and a router that dynamically determines expert allocation; the down-projection matrix is shared across experts while the up-projections are expert-specific, reducing parameter redundancy while enabling diverse adaptation behaviors.

3.1. Routing strategies

For a pretrained linear transformation $W \in \mathbb{R}^{d' \times d}$, standard LoRA [6] adds a low-rank update BA, with $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d' \times r}$. Hydra-LoRA [9] shares the down-projection A while introducing expert-specific up-projections $\{B_i\}_{i=1}^{N}$:

$$W' = W + \sum_{i=1}^{N} \alpha_i B_i A, \tag{1}$$

where N is the number of experts and α_i the routing weight of expert i. We investigate two routing strategies.

Global routing.
Routing decisions are made from a single speaker-level representation of the encoder output. Given encoder hidden states H_{enc}, a routing network predicts expert probabilities

$$\mathbf{p} = \text{Softmax}(W_r \, \text{Pool}(H_{\text{enc}}) + b_r), \tag{2}$$

where $\text{Pool}(\cdot)$ denotes temporal mean pooling and $\mathbf{p} \in \mathbb{R}^N$. The same distribution is shared across all LoRA modules and layers, assuming speaker characteristics remain consistent throughout the utterance.

Layer-wise routing.
Here, each LoRA target module contains an independent router, and routing is computed from layer-specific token representations. For layer l and token t,

$$\mathbf{p}_t^{(l)} = \text{Softmax}\left(W_r^{(l)} h_t^{(l)} + b_r^{(l)}\right), \tag{3}$$

where $h_t^{(l)}$ is the hidden representation at layer l. Different layers and tokens can thus activate different experts—useful for impaired speech, where pronunciation variability may depend on phonetic context and temporal position.

3.2. Expert selection choices

The original HydraLoRA uses a soft combination in which all expert branches are weighted by the routing probabilities, i.e., $\alpha_i = p_i$ in Eq. (1) [9]. Although fully differentiable, all experts then remain active for every input, so the method behaves more like a parameter-boosting ensemble, and averaging over all branches weakens expert specialization. We therefore adopt top-k expert selection, as widely used in sparse MoE

models [7], [8], [20]:

$$\tilde{p}_i = \begin{cases} p_i, & i \in \mathrm{TopK}(\mathbf{p}), \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

followed by probability renormalization; Eq. (1) is then applied with $\alpha_i = \tilde{p}_i$, so each input activates only a subset of specialized experts.

3.3. Routing load-balance design

Sparse routing may suffer from expert collapse, where a few experts dominate. Following Switch Transformer [8], let $r_{t,i}$ denote the routing probability of expert i for token t; the average routing probability and routing frequency are

$$p_i = \tfrac{1}{T}\textstyle\sum_t r_{t,i}, \quad f_i = \tfrac{1}{T}\textstyle\sum_t \mathbb{I}\big(i = \arg\max_j r_{t,j}\big), \tag{5}$$

where T is the number of routed tokens. The training objective combines the ASR loss with a load-balancing term,

$$L = L_{\mathrm{ASR}} + \lambda N \sum_{i=1}^{N} p_i f_i, \tag{6}$$

where λ controls the regularization strength, encouraging evenly distributed routing probabilities and assignments.

3.4. Backbone integration and adapter placement

Hydra-LoRA follows the same insertion principle as standard LoRA; in the per-layer setting each adapted module has its own branches and router. We apply it to the attention projections and feed-forward layers of the encoder-decoder backbone. For multimodal speech-language models such as Qwen3-ASR [28], where the acoustic encoder is connected to the language model through a linear projection rather than cross-attention, we explore encoder-and-projection-only, decoder-only, and joint adapter placements.

3.5. Exploratory VI-LoRA extension

We also explore combining Hydra-LoRA with Variational LoRA (VI-LoRA) [15], which models the LoRA matrices as random variables with an approximate posterior $q_\phi(A, B) = q_{\phi_A}(A)\, q_{\phi_B}(B)$, typically diagonal Gaussians learned via variational inference with

$$L_{\mathrm{VI}} = L_{\mathrm{ASR}} + \beta\, \mathrm{KL}[q_\phi(A, B) \,\|\, p(A, B)]\,, \tag{7}$$

where $p(A, B)$ is the prior and β the KL weight. Such uncertainty-aware regularization can reduce overfitting in low-resource personalization; we convert the point-estimated Hydra-LoRA experts into VI-LoRA-style probabilistic adapters.

4. Experiments

We evaluate Hydra-LoRA from both recognition performance and routing behavior perspectives, using UASpeech and SAP as non-normative datasets, LibriSpeech as a normative reference, and Whisper and Qwen3-ASR as backbones.

Table 1: Main ASR results on UASpeech and SAP. Hydra-G and Hydra-L denote global-routing and per-layer Hydra-LoRA (both $r = 32$, 4 experts, top-2). Zero-shot uses `whisper-large-v3` before adaptation; SAP is scored with the official SAP scorer. Lower is better.

Method	UASpeech		SAP	
	WER	CER	WER	CER
Whisper zero-shot	129.81	100.19	18.93	12.54
Pure LoRA, $r = 32$	20.19	17.73	10.47	7.15
Hydra-G	19.10	17.49	10.74	7.11
Hydra-L	**17.79**	**15.32**	**9.29**	**6.35**

4.1. Datasets and base models

We use the Hugging Face repackaging of UA-Speech [2] (low- and high-severity subsets): 57,396 isolated-word utterances from 12 dysarthric speakers and 449 word types, excluding the 10 control speakers, split by speaker–word–repetition groups into 70% training and 30% testing. SAP [27] provides sentence-level impaired speech across five impairment categories. We exclude utterances longer than Whisper's 30 s receptive context (about 2% of the data) and, due to the cost of repeated fine-tuning, use a 10% subset of the filtered training split (about 31,193 utterances from 834 speakers); runs on the full training set showed similar trends. As normative data, we use LibriSpeech [29] clean: `train.clean.100` (about 100 h, 28,539 utterances, 251 speakers) for optional mixed training and `test.clean` (2,620 utterances, 40 speakers) for evaluation. Whisper-large-v3 [30], a Transformer encoder-decoder with about 1.55B parameters, is the main backbone. Qwen3-ASR-1.7B [28], a multimodal speech-language model from the Qwen3-Omni family, couples an acoustic encoder to a large language model through a projection module rather than encoder-decoder cross-attention.

4.2. Evaluation metrics

We report word and character error rates (WER/CER), computed on normalized text from the Levenshtein alignment as substitutions plus deletions plus insertions over the reference length. For SAP, we follow the official protocol: each utterance provides two normalized references, with and without disfluencies; the hypothesis is scored against both and the lower error rate is used.

4.3. Main results on UASpeech and SAP

All methods fine-tune Whisper-large-v3 for 5 epochs under identical hyperparameters: LoRA uses rank $r = 32$ and scaling $\alpha = 64$; Hydra-LoRA uses 4 experts with top-2 routing and load-balancing weight 0.03. As shown in Table 1, per-layer Hydra-LoRA consistently outperforms the LoRA baseline, reducing WER/CER from 20.19/17.73 to 17.79/15.32 on UASpeech and from 10.47/7.15 to 9.29/6.35 on SAP, whereas global routing brings only a modest improve-

Table 2: Effect of expert number in per-layer Hydra-LoRA on SAP. All models use top-k routing with rank $r = 8$. Lower is better.

Method	WER	CER
Pure LoRA, $r = 8$	12.12	9.01
Hydra-L, $r = 8$, 4 experts	11.11	7.52
Hydra-L, $r = 8$, 8 experts	10.09	7.65
Hydra-L, $r = 8$, 16 experts	10.10	7.04
Hydra-L, $r = 8$, 32 experts	**10.09**	**6.88**

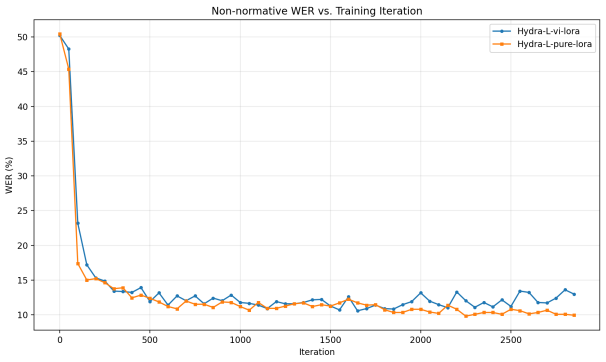


Figure 1: Non-normative WER over training iterations on SAP for Hydra-VI-LoRA and standard Hydra-LoRA.

ment on UASpeech and does not improve SAP WER. Routing thus benefits from layer-wise local information: speaker-level routing can still help isolated-word recognition, but for sentence-level data, per-layer routing adapts expert allocation to layer- and token-level representations, better capturing pronunciation variability and linguistic context.

4.4. Effect of expert scale

We vary the number of experts under per-layer top-k routing on SAP, with hyperparameters as in Section 4.3 except that the rank is reduced to $r = 8$ due to GPU memory limits. As shown in Table 2, reducing the rank clearly weakens standard LoRA (12.12 vs. 10.47 WER at $r = 32$), while enlarging the Hydra-LoRA expert pool compensates for the lost capacity: with top-2 routing the activated parameter count stays fixed, yet WER improves from 11.11 to 10.09 and CER from 7.52 to 6.88 as the pool grows from 4 to 32 experts. Even at $r = 8$, Hydra-LoRA clearly outperforms rank-matched LoRA, showing that expert specialization improves adaptation efficiency under a limited rank budget, although gains saturate beyond 8 experts.

4.5. Exploratory study of Hydra-VI-LoRA

We evaluate the Hydra-VI-LoRA variant of Section 3.5 against standard per-layer Hydra-LoRA, tracking non-normative WER on SAP over training iterations.

As shown in Fig. 1, both methods improve quickly at first, but standard Hydra-LoRA keeps a smoother trajectory and converges around 10.0 WER, while Hydra-VI-LoRA fluctuates and rises to about 13.0 WER; CER shows a similar trend. Hydra-LoRA already introduces adaptive diversity through routing, and adding parameter uncertainty inside each expert may complicate optimization under limited data, while the KL regularization may conflict with sparse routing and load balancing; a successful combination may require a routing-aware variational objective or a staged training schedule.

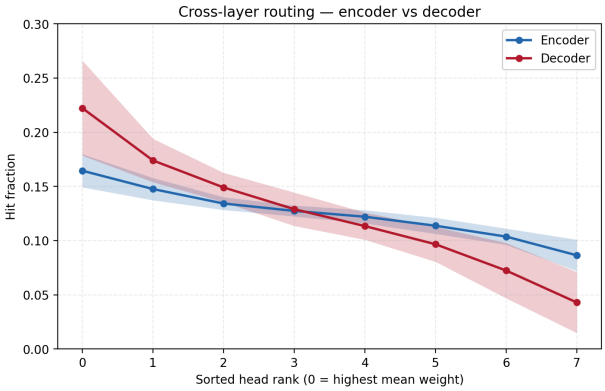


Figure 2: Cross-layer routing concentration for a representative speaker. Experts are sorted per layer by average routing weight; curves show the mean over layers at each sorted rank, shaded regions the standard error of the mean.

4.6. Routing pattern analysis

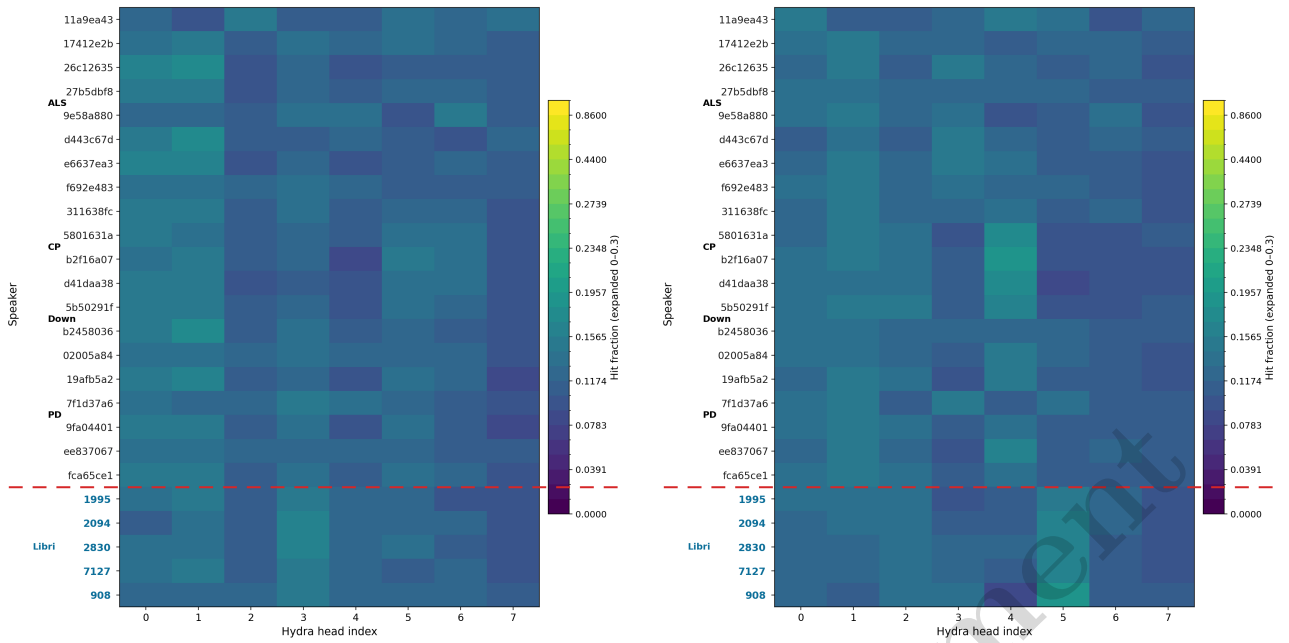
Beyond accuracy, we analyze the routing behavior of per-layer Hydra-LoRA with 8 experts and top-2 routing from two perspectives: cross-layer routing concentration for a representative speaker, and speaker-level heatmaps of the *hit fraction*, i.e., the fraction of routed tokens for which an expert is selected among the top-2.

In Fig. 2, the encoder curve is relatively flat near 0.125, the uniform value 1/8 for 8 experts, so encoder routing stays close to average expert usage across layers, whereas the decoder curve varies strongly across sorted expert ranks, indicating more selective, concentrated routing.

Fig. 3 gives a more detailed view. In the encoder layer, the near-uniform distribution makes group differences weak, though non-normative speakers activate heads 0 and 1 slightly more often, while normative LibriSpeech speakers use head 3 more strongly. The decoder layer shows clearer group-level differences: normative speakers activate head 5 more strongly, and condition-related patterns appear among non-normative speakers, e.g., speakers with cerebral palsy show noticeably stronger activation of head 4. Although qualitative, this suggests the router captures meaningful variation related to normative/non-normative differences and impairment categories: encoder layers reflect broader acoustic patterns, while decoder layers show stronger speaker- or context-dependent specialization.

4.7. Preliminary experiments on Qwen3-ASR

As preliminary exploration, we evaluate Qwen3-ASR-1.7B only on SAP, whose sentence-level utterances are closer to realistic ASR scenarios. We compare



(a) Encoder layer **(b)** Decoder layer

Figure 3: Speaker-level expert hit fractions in representative encoder and decoder layers. Non-normative speakers are shown above the dashed line, and normative LibriSpeech speakers are shown below it.

Table 3: Qwen3-ASR-1.7B on SAP under different fine-tuning placements. Hydra-LoRA uses 4 experts, top-2 routing. Lower is better.

Placement	Method	WER	CER
–	Qwen3-ASR zero-shot	14.28	9.43
Enc.+Proj	Pure LoRA	12.27	8.02
Enc.+Proj	Hydra-LoRA	12.07	7.88
LM-Dec	Pure LoRA	10.76	7.39
LM-Dec	Hydra-LoRA	**10.64**	7.23
All	Pure LoRA	10.94	**7.14**
All	Hydra-LoRA	10.87	7.28

three adapter placements—encoder-and-projection-only, LM-decoder-only, and joint adaptation—each with pure LoRA and per-layer Hydra-LoRA (4 experts, top-2 routing).

As shown in Table 3, Qwen3-ASR achieves strong zero-shot performance (14.28 WER, 9.43 CER), substantially better than Whisper's zero-shot result in Table 1. After fine-tuning, however, the best result (10.64 WER, 7.14 CER) is only comparable to the fine-tuned Whisper results, and Hydra-LoRA brings only marginal gains over pure LoRA in most placements. Placement matters more: LM-decoder fine-tuning (10.76 WER with pure LoRA) clearly outperforms encoder-and-projection fine-tuning (12.27 WER), with a similar trend for Hydra-LoRA, while joint fine-tuning does not consistently beat decoder-only adaptation.

Notably, tuning only the encoder and projection often produces hallucination-like repetitive outputs, even though encoder and LM decoder have comparable trainable parameter counts. Since the two are coupled through a projection module rather than cross-attention, shifting to non-normative speech may require more data or targeted techniques to maintain acoustic-to-language alignment: modifying only the acoustic side may disturb the representations passed to the language model, causing unstable generation. Multimodal models may thus require architecture-specific fine-tuning strategies.

5. Discussion and Conclusion

Per-layer Hydra-LoRA improves over pure LoRA on both isolated-word and sentence-level impaired speech, supporting our hypothesis that non-normative speech contains heterogeneous patterns that a single shared adapter—or a single utterance-level routing decision—cannot capture well. Enlarging the expert pool partly compensates for a smaller rank, although the benefit eventually saturates, and the routing visualizations indicate that the learned router is not arbitrary, capturing differences between encoder and decoder behavior and between normative and non-normative speech.

Three exploratory directions were less successful: global routing underperformed per-layer routing, especially on SAP, indicating that speaker-level information alone is insufficient for sentence-level speech; naively combining Hydra-LoRA with VI-LoRA was unstable, possibly because pairing routing diversity with per-expert parameter uncertainty complicates optimization; and Qwen3-ASR gained little from LoRA or Hydra-LoRA fine-tuning, suggesting multimodal speech-language models need architecture-specific strategies.

Limitations include the 10% SAP training subset used in many experiments, the qualitative routing analysis based on representative layers and speakers, non-exhaustive hyperparameter tuning (expert number, top-k, rank, load-balancing weight, target modules), and the preliminary Qwen3-ASR study. Future work includes more datasets and seeds with speaker- or condition-wise analysis, expert-diversity and speaker-aware routing objectives, and better acoustic-to-language alignment and cross-modal routing for multimodal ASR. Overall, MoE-style parameter-efficient fine-tuning is a promising direction for accessible ASR, but routed adaptation is sensitive to expert design and backbone architecture; future progress will likely require routing strategies matched to the to modern multimodal ASR architectures.

References

[1] J. Shor et al., "Personalizing asr for dysarthric and accented speech with limited data," *arXiv preprint arXiv:1907.13511*, 2019.

[2] H. Kim et al., "Dysarthric speech database for universal access research," in *Proceedings of Interspeech 2008*, 2008, pp. 1741–1744.

[3] Z. Qian and K. Xiao, "A survey of automatic speech recognition for dysarthric speech," *Electronics*, vol. 12, no. 20, p. 4278, 2023.

[4] N. Pokel, P. Moure, R. Boehringer, and Y. Gao, "Adapting foundation speech recognition models to impaired speech: A semantic re-chaining approach for personalization of german speech," in *Proc. DiSS 2025*, 2025, pp. 82–86.

[5] P. Moure et al., *When audio-language models fail to leverage multimodal context for dysarthric speech recognition*, 2026. arXiv: 2605.02782 [cs.AI].

[6] E. J. Hu et al., "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[7] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[8] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *arXiv preprint arXiv:2101.03961*, 2021.

[9] C. Tian, Z. Shi, Z. Guo, L. Li, and C. Xu, "Hydralora: An asymmetric lora architecture for efficient fine-tuning," *arXiv preprint arXiv:2404.19245*, 2024.

[10] X. Wu, S. Huang, and F. Wei, "Mixture of lora experts," *arXiv preprint arXiv:2404.13628*, 2024.

[11] N. Inoue et al., "Parameter efficient adapter tuning for various speech processing tasks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[12] A. M. Samin et al., "Investigating adapters for parameter-efficient low-resource speech recognition," in *Proceedings of the 2025 Workshop on Representation Learning for NLP*, 2025.

[13] J. Qi and H. Van hamme, "Parameter-efficient dysarthric speech recognition using adapter fusion and householder transformation," in *Proceedings of Interspeech 2023*, 2023.

[14] M. Müller-Eberstein, D. Yee, K. Yang, G. V. Mantena, and C. Lea, "Hypernetworks for personalizing asr to atypical speech," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 1182–1196, 2024.

[15] N. Pokel, P. Moure, R. Boehringer, S.-C. Liu, and Y. Gao, "Variational low-rank adaptation for personalized impaired speech recognition," *arXiv preprint arXiv:2509.20397*, 2025.

[16] D. Lepikhin et al., "Gshard: Scaling giant models with conditional computation and automatic sharding," *arXiv preprint arXiv:2006.16668*, 2020.

[17] N. Du et al., "Glam: Efficient scaling of language models with mixture-of-experts," in *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[18] S. Kudugunta et al., "Beyond distillation: Task-level mixture-of-experts for efficient inference," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 3577–3599.

[19] M. Lewis, S. Bhosale, T. Dettmers, N. Goyal, and L. Zettlemoyer, "Base layers: Simplifying training of large, sparse models," in *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[20] Y. Zhou et al., "Mixture-of-experts with expert choice routing," in *Advances in Neural Information Processing Systems*, 2022.

[21] J. Xin et al., "I2moe: Interpretable multimodal interaction-aware mixture-of-experts," *arXiv preprint arXiv:2505.19190*, 2025.

[22] Y. Liu et al., "Mixture of lora experts for low-resourced multi-accent automatic speech recognition," *arXiv preprint arXiv:2505.20006*, 2025.

[23] M. Tu, A. Wisler, V. Berisha, and J. M. Liss, "The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, EL416–EL422, 2016.

[24] Z. Qian and K. Xiao, "A survey of technologies for automatic dysarthric speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, 2023.

[25] T. Wang et al., "Hyper-parameter adaptation of conformer asr systems for elderly and dysarthric speech recognition," *arXiv preprint arXiv:2306.15265*, 2023.

[26] H. Wang et al., "Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation," *arXiv preprint arXiv:2401.00662*, 2024.

[27] X. Zheng et al., "The interspeech 2025 speech accessibility project challenge," *arXiv preprint arXiv:2507.22047*, 2025.

[28] X. Shi et al., "Qwen3-asr technical report," *arXiv preprint arXiv:2601.21337*, 2026.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 2023, pp. 28 492–28 518.