forum acusticum 2026
8-12 September · Graz, Austria

AN EVALUATION OF HYPERNETWORK-BASED PARAMETER-EFFICIENT FINE-TUNING FOR HUNGARIAN AND ENGLISH DYSARTHRIC ASR

Piroska Zsófia Barta[1, 2], Péter Mihajlik[1, 3]

[1] *Department of Telecommunications and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Hungary*
[2] *SpeechTex Ltd.*
[3] *ELTE Research Centre for Linguistics, Hungary*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

In this study, we evaluate a recently proposed hypernetwork-based, parameter-efficient fine-tuning approach for dysarthric speech recognition in Hungarian and English, using the Whisper ASR (Automatic Speech Recognition) model. We compare its performance to Low-Rank Adaptation (LoRA) demonstrating the promise of hypernetwork-driven personalization for atypical ASR for both languages. To better understand speaker-level variability, we analyze the encoder's latent-space representations and examine the presence of intra-cohort structure. Our results show that, even under severe data limitations, the hypernetwork-based method consistently improves recognition accuracy for previously unseen speakers while requiring substantially fewer trainable parameters than LoRA, yet achieving competitive performance.

Keywords: *Automatic Speech Recognition, Dysarthria, Hypernetwork*

1. Introduction

Dysarthria is a motor speech disorder [1] that affects speech production and can be attributed to a wide range of diseases, including stroke, brain injuries, and neurodegenerative diseases such as Parkinson's or Amyotrophic Lateral Sclerosis (ALS). People suffering from dysarthria may face difficulties in the course of everyday communication, as – beyond other possible complications – the intelligibility of their speech is impaired to varying degrees.

The reduction in speech clarity poses a significant challenge in the context of Automatic Speech Recognition (ASR) as well; the ASR systems fail to reliably transcribe dysarthric speech, which can be primary ascribed to the inter-personal variability of atypical speech patterns. The development of such models is constrained by severe data shortage, which can promote a shift toward personalized model approaches and the use of synthesized speech [2], [3], [4] in research. Although personalization is an effective method for enhancing model performance [5], many applications such as virtual assistants and customer service applications require models with strong generalization capabilities. Since individualization is prohibitively time-consuming and scales poorly, even in few-shot settings, it is impractical in the considered context.

To address the limitations of speaker-adapted models, in this research, building on [6], we examined whether utilizing the latent space representation produced by the encoder-decoder model's encoder would help to identify intra-cohort level similarities and differences in the speech of dysarthric speakers, enabling the formation of intra-cohort-level groups. Further exploiting the findings of [6], we experimented with integrating a hypernetwork [7] into a Whisper [8] model.

2. Related work

Owing to data scarcity, obtaining sufficiently large dysarthric speech datasets remains challenging even for well-resourced languages such as English, with this limitation being even more severe for lower-resourced languages such as Hungarian. This is reflected in both the length and the nature of the available datasets. Several dysarthric databases contain only isolated words, such as the UASpeech database [9], which consists of recordings from 16 speakers, and the Whitaker database of dysarthric (cerebral palsy) speech [10], which comprises nearly 20,000 utterances from 6 speakers. In contrast, the Nemours database [11] con-

***Corresponding author**: piroskazsofia.barta@edu.bme.hu.*

tains 74 short sentences recorded by 11 male speakers. The TORGO [12] database comprises data from 8 dysarthric individual and contains isolated words and short sentences that were fairly identical between speakers. The Chinese Dysarthria Speech Database [13] (CDSD) is the largest Chinese dataset, comprising recordings from 44 speakers and totaling 133 hours of speech. The Italian EasyCall corpus [14] includes a subcorpus consisting of recordings from 24 dysarthric speakers, featuring commands, isolated words, and sentences that are phonetically similar to commands. [15] describes a Dutch Dysarthric Speech Database comprising 6 hours of audio from 16 speakers. Recently, significant effort has been made to create larger scale atypical data sets, such as the Speech Accessibility Project [16] where approximately 1,500 hours of speech from 999 participants has been collected. The main limitation of this data is its monolingual nature (English).

Regarding ASR results, while emphasizing that personalized fine-tuning remains essential in the process of adapting ASR models to dysarthric speech, in [17] using the TORGO dataset, the authors demonstrate that using model that were pretrained on a multi-speaker dysarthric dataset prior to personalization can yield better performance compared to relying solely on speaker-specific data. Within the study [6], a hypernetwork-based approach [7] was utilized in order to allow knowledge sharing and personalization to occur in parallel. They found that this method outperformed LoRA [18] by only adapting the first linear sub-layer of each Whisper model layer. They further established that, for severe pathological cases, using the data of speakers with mild impairments did not improve performance.

In this work, we place particular emphasis on the separate evaluation of ASR performance improvements for severe and non-severe dysarthric cases by comparing two parameter-efficient fine-tuning approaches, namely LoRA [18] and a hypernetwork-based method [6], using Hungarian and English dysarthric data.

3. Data

For our experiments we used two significantly different databases, namely the Hungarian Dysarthria Database and the English TORGO dataset.

3.1. HDD

The Hungarian Dysarthria Database (HDD) [1] consisting of the speech data of approximately 50 native Hungarian speakers. The dataset comprised of two distinct modules: short command-based utterances read by the speakers (e. g. "Nyisd ki az ablakot a konyhában" ["Open the window in the kitchen"]), and a narrative component, namely the "North Wind and the Sun" by Aesop, which was likewise read aloud by the speakers. We utilized the command-based part of the database by categorizing the speakers into four dysarthria severity groups (not perceived as dysarthric or normal, mild, moderate, severe) based on the intelligibility of their speech evaluated by non-experts, with the final classification determined through a voting procedure.

Based on the classification, we created three, speaker independent subsets – train, development (dev) and evaluation (eval) – from the original database, with each subset maintaining the representation of all dysarthric groups. Since the database contained only phonetic transcriptions, we manually corrected them to obtain a standard orthographic reference text.

The metadata for the subsets of HDD is presented in Table 1.

	HDD – subset		
	train	dev	eval
# of spk. [m|f]	19|8	2|2	3|3
normal	3	0	0
mild dys.	12	1	2
moderate dys.	8	1	2
severe dys.	3	2	2
# of segments	7119	1052	1668
# of words	41600	6304	9641
duration [h]	7.65	1.07	1.41

Table 1: Metadata of the Hungarian HDD subsets. Abbreviations: spk. – speakers, dys. – dysarthria, m|f – male|female, h – hours.

3.2. TORGO

Since the original TORGO dataset does not provide predefined subsets; we partitioned it into three speaker-independent subsets ensuring that there was no contextual overlap between the splits, and the dev and eval sets are balanced with respect to the number of segments belonging to each speaker. The metadata corresponding to the utilized subsets are presented in Table 2.

	TORGO – subset		
	train	dev	eval
# of speakers [m|f]	3|1	1|1	1|1
# of segments	1204	55	81
# of words	2900	155	207
duration [h]	1.24	0.06	0.09

Table 2: Metadata of the English TORGO subsets. Abbreviations: m|f – male|female, h – hours.

4. Encoder output representations

Figure 1 presents the latent space representations corresponding to HDD, whereas Figure 2 illustrates those corresponding to TORGO.

For each dataset's train subset, we examined the output of the last encoder layer of whisper-large-v3. We used Uniform Manifold Approximation and Projection

[1] http://lsa.tmit.bme.hu/databases/dysarthria-hun.html

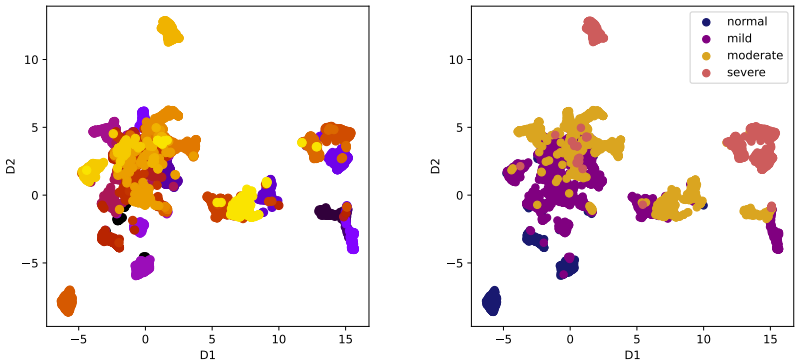

Figure 1: 2-dimensional UMAP representations of the encoder output belonging to the train samples of HDD. The two plots are organized as follows: by speaker (left) and by dysarthria severity (right).

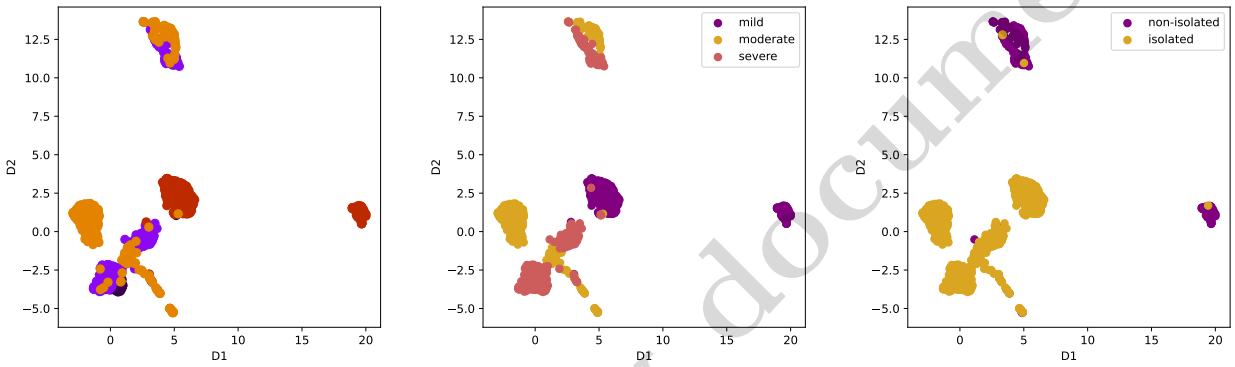

Figure 2: 2-dimensional UMAP representations of the encoder output belonging to the train samples of TORGO. The three plots are organized as follows (left to right): by speaker, by dysarthria severity, and by data points corresponding to isolated and non-isolated word utterances.

(UMAP) [19] dimension reduction technique to obtain 2-dimensional representations from the model output. Our aim was to determine whether intra-cohort group structures are present in the dimensionally reduced latent space.

By examining the subset belonging to HDD, we found that in both cases, speakers within the same severity category tend to form groups at the intra-cohort level. Due to the limited number of samples in the TORGO dataset, we focused on the differences between isolated words and longer segments. We observed that, for each speaker, the segments are separated into two distinct groups depending on whether they correspond to isolated words or longer segments.

5. Experimental setup

For the speech recognition experiments, we modified the decoder part of the whisper-large-v3 model according to the architecture described in [6]. First, we examined the extent of the parameter changes introduced by LoRA across the different sublayer components of three model versions; tiny, medium and large-v3. Our findings aligned with those reported in the original study: the hypernetwork should be integrated into the first linear sublayer of the decoder layers, as this component exhibited the greatest magnitude of change during LoRA fine-tuning. We also found, consistent with the original paper, that the optimal rank for the low-rank matrices generated by the hypernetwork is r=2. For our experiments, we modified the HuggingFace's Transformers [20] training script. As in [6] during fine-tuning we consistently used approximately 0.03% of the trainable parameters ($\approx$ 0.5 M), which is roughly an order of magnitude smaller than the amount we used during LORA fine-tuning ($\approx$ 7.5 M). Across all training setups, we used the default AdamW optimizer and the models were fine-tuned for a maximum of 10 epochs. We evaluated the model using two metrics: Word Error Rate (WER) and Character Error Rate (CER).

6. Results

For evaluation, the dev and eval sets of the two databases were partitioned into two subsets to analyze performance with respect to speaker intelligibility levels. The normal, mild, and moderate categories

were combined into a single non-severe group.
Concerning the results on the HDD dataset reported in 3, a substantial improvement in the target metrics can be observed with both parameter-efficient fine-tuning approaches on both the severe and non-severe speaker subsets. Notably, the error rates of the hypernetwork approach are consistently lower for both the severe and non-severe subset in terms of WER and CER, too.
Regarding the experiments on TORGO presented in Table 4, although the evaluation metrics indicate that the fine-tuned model remains far from real-world applicability, in the overall and results, a consistent improvement can be observed across both the dev and eval datasets when employing the fine-tuned hypernetwork and LoRA. On the eval subset, the inherently low metric values of the non-severe speaker did not improve or slightly degrade after fine-tuning, which may be due to the speaker's speech being relatively close to typical speech.
Based on the results obtained on the two datasets, the performance differences between the two parameter efficient fine-tuning approaches appear marginal. Nevertheless, the hypernetwork-based method is substantially more parameter-efficient than the LoRA approach, as it was fine-tuned using an order of magnitude fewer (trainable) parameters.

	HDD			
	dev		eval	
	WER	CER	WER	CER
overall (zs)	67.07	39.36	73.95	41.28
overall (LoRA)	9.66	6.39	10.69	6.33
overall (HNW)	9.11	5.78	9.72	5.50
severe (zs)	88.54	56.01	85.79	53.28
severe (LoRA)	15.35	11.16	12.38	7.57
severe (HNW)	14.0	9.45	10.99	6.26
non-severe (zs)	67.07	39.36	73.95	41.28
non-severe (LoRA)	9.66	5.48	9.59	5.51
non-severe (HNW)	5.51	3.03	8.89	5.0

Table 3: ASR Results [%] for whisper-large-v3 with and without an integrated fine-tuned hypernetwork component on the dev and eval subsets of HDD. Abbreviations: HNW – fine-tuned with hypernetwork, LoRA – fine-tuned with LoRA, zs – zero-shot.

7. Conclusion

We compared hypernetwork and LoRA-based parameter-efficient fine-tuning approaches on Hungarian and English datasets. On the Hungarian dataset, a radical improvement was observed with both techniques; however, the limited lexical diversity of the dataset represents a potential confounding factor, and its impact should be further investigated. Regarding the English dataset, the results exhibited a modest improvement with a minimal training set. This result further supports the findings of [6], indicating that the hypernetwork-based approach can be an effective form of parameter-efficient fine-tuning, requiring less than one-tenth of the number of parameters used by

	TORGO			
	dev		eval	
	WER	CER	WER	CER
overall (zs)	47.10	29.03	71.98	49.95
overall (LoRA)	37.42	23.39	49.76	32.29
overall (HNW)	36.13	21.37	49.76	33.52
severe (zs)	64.94	40.27	105.07	73.65
severe (LoRA)	54.55	34.05	71.74	47.15
severe (HNW)	50.65	29.46	71.74	48.86
non-severe (zs)	29.49	17.91	5.80	2.56
non-severe (LoRA)	20.51	12.83	5.80	2.56
non-severe (HNW)	21.79	13.37	5.80	2.85

Table 4: ASR Results [%] for whisper-large-v3 with and without an integrated fine-tuned hypernetwork component on the dev and eval subsets of TORGO. Abbreviations: HNW – fine-tuned with hypernetwork, LoRA – fine-tuned with LoRA, zs – zero-shot.

LoRA, and has the potential to improve performance on speakers not included in the training set. As a future work we will also explore the integration of hypernetworks into other model architectures. Furthermore, since the decoder adaptation targets the internal language model, it may have the tendency of overfitting to the lexical characteristics of the training data, thereby limiting the diversity of the generated output. Consequently, future work will investigate whether decoder or encoder adaptation offers greater potential for dysarthric speech recognition.

8. Limitations

Due to the limited availability of dysarthric speech data, our experiments were conducted on two relatively small datasets, subset of HDD and a subset of TORGO, comprising 10 and 1.5 hours of Hungarian and English dysarthric speech, respectively.
In case of the TORGO data set, it consists of mainly short words and restricted sentences which limits the validity of conclusions for general dysarthric speech. Similarly, although the given subset of the HDD corpus is significantly larger, it is also characterized by a restricted vocabulary, confined to a limited number of short commands and the repetitive nature of the dataset may have biased the results. In a control experiment, independent Hungarian dysarthric speech featuring more diverse lexical content was transcribed as command-like text by the fine-tuned models, which may indicate a loss of generalization ability if fine-tuning is performed on small, textually homogeneous data.

9. Acknowledgments

This work was supported by the Hungarian NRDI Fund through projects NKFIH K143075 and by KIFÜ Kommondor.

References

[1] J. R. Duffy, *Motor Speech Disorders E-Book: Motor Speech Disorders E-Book*. Elsevier Health Sciences, 2019.

[2] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, *Synthesizing dysarthric speech using multi-talker tts for dysarthric speech recognition*, 2022. arXiv: 2201.11571 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2201.11571.

[3] W.-Z. Leung, M. Cross, A. Ragni, and S. Goetze, *Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis*, 2024. arXiv: 2406.08568 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2406.08568.

[4] E. Hermann and M. Magimai-Doss, "Few-shot dysarthric speech recognition with text-to-speech data augmentation," Aug. 2023, pp. 156–160. DOI: 10.21437/Interspeech.2023-2481.

[5] P. Mihajlik, É. Székely, P. Barta, M. S. Kádár, G. Dobsinszki, and L. Tóth, *Improved dysarthric speech to text conversion via tts personalization*, 2025. arXiv: 2508.06391 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2508.06391.

[6] M. Müller-Eberstein, D. Yee, K. Yang, G. V. Mantena, and C. Lea, *Hypernetworks for personalizing asr to atypical speech*, 2024. arXiv: 2406.04240 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2406.04240.

[7] D. Ha, A. Dai, and Q. V. Le, *Hypernetworks*, 2016. arXiv: 1609.09106 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1609.09106.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022. arXiv: 2212.04356 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2212.04356.

[9] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," Sep. 2008, pp. 1741–1744. DOI: 10.21437/Interspeech.2008-480.

[10] J. R. Deller, M. S. Liu, L. J. Ferrier, and P. Robichaud, "The whitaker database of dysarthric (cerebral palsy) speech.," *The Journal of the Acoustical Society of America*, vol. 93 6, pp. 3516–8, 1993. [Online]. Available: https://api.semanticscholar.org/CorpusID:39962202.

[11] X. Menendez-Pidal, J. Polikoff, S. Peters, J. Leonzio, and H. Bunnell, "The nemours database of dysarthric speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, vol. 3, 1996, 1962–1965 vol.3. DOI: 10.1109/ICSLP.1996.608020.

[12] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language resources and evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[13] Y. Wan et al., "Cdsd: Chinese dysarthria speech database," in *Interspeech 2024*, ser. interspeech$_2$024, ISCA, Sep. 2024, pp. 4109–4113. DOI: 10.21437/interspeech.2024-1597. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2024-1597.

[14] R. Turrisi et al., "Easycall corpus: A dysarthric speech dataset," in *Interspeech 2021*, ser. interspeech$_2$021, ISCA, Aug. 2021, pp. 41–45. DOI: 10.21437/interspeech.2021-549. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2021-549.

[15] E. Yilmaz, M. Ganzeboom, L. Beijer, C. Cucchiarini, and H. Strik, "A Dutch dysarthric speech database for individualized speech therapy research," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, N. Calzolari et al., Eds., Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 792–795. [Online]. Available: https://aclanthology.org/L16-1127/.

[16] M. Hasegawa-Johnson et al., "Community-supported shared infrastructure in support of speech accessibility," *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 11, pp. 4162–4175, 2024. DOI: 10.1044/2024_JSLHR-24-00122. eprint: https://pubs.asha.org/doi/pdf/10.1044/2024_JSLHR-24-00122. [Online]. Available: https://pubs.asha.org/doi/abs/10.1044/2024_JSLHR-24-00122.

[17] V. Raja, A. V. Ganesan, A. Syamkumar, R. Banerjee, and H. A. Schwartz, *Idiosyncratic versus normative modeling of atypical speech recognition: Dysarthric case studies*, 2025. arXiv: 2509.16718 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2509.16718.

[18] E. J. Hu et al., *Lora: Low-rank adaptation of large language models*, 2021. arXiv: 2106.09685 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2106.09685.

[19] L. McInnes, J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction*, 2020. arXiv: 1802.03426 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1802.03426.

[20] T. Wolf et al., *Huggingface's transformers: State-of-the-art natural language processing*, 2020. arXiv: 1910.03771 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1910.03771.