



On-the-fly Routing for Zero-shot MoE Speaker Adaptation of Speech Foundation Models for Dysarthric Speech Recognition

Shujie Hu¹, Xurong Xie^{2*}, Mengzhe Geng³, Jiajun Deng¹, Huimeng Wang¹, Guinan Li¹, Chengxi Deng¹, Tianzi Wang¹, Mingyu Cui¹, Helen Meng¹, Xunying Liu^{1*}

¹The Chinese University of Hong Kong, Hong Kong SAR, China

²Institute of Software, Chinese Academy of Sciences, China

³National Research Council Canada, Canada

{sjhu, xyliu}@se.cuhk.edu.hk, xurong@iscas.ac.cn

Abstract

This paper proposes a novel MoE-based speaker adaptation framework for foundation models based dysarthric speech recognition. This approach enables zero-shot adaptation and real-time processing while incorporating domain knowledge. Speech impairment severity and gender conditioned adapter experts are dynamically combined using on-the-fly predicted speaker-dependent routing parameters. KL-divergence is used to further enforce diversity among experts and their generalization to unseen speakers. Experimental results on the UASpeech corpus suggest that on-the-fly MoE-based adaptation produces statistically significant WER reductions of up to 1.34% absolute (6.36% relative) over the unadapted baseline HuBERT/WavLM models. Consistent WER reductions of up to 2.55% absolute (11.44% relative) and RTF speedups of up to 7 times are obtained over batch-mode adaptation across varying speaker-level data quantities. The lowest published WER of 16.35% (46.77% on very low intelligibility) is obtained.

Index Terms: Speech Recognition, Speech Foundation Model, Speaker Adaptation, Dysarthric Speech, Mixture of Experts

1. Introduction

Despite the rapid progress of ASR technologies targeting normal and healthy users, their application to those suffering from speech disorders, such as dysarthria, remains a challenging task to date [1–10]. Dysarthric speech brings challenges on all fronts to current deep learning based ASR technologies predominantly targeting healthy users: **1) substantial mismatch** against typical voices; **2) data scarcity** [3, 9]; and **3) large speaker-level diversity** [11] among dysarthric talkers, including accent or gender, and speech pathology severity. Such heterogeneity among dysarthric speakers not only complicates the training or fine-tuning of speaker-independent (SI) ASR systems on this data but also hinders their effective personalization to individual users' voices. These challenges are further compounded in the fine-tuning of self-supervised learning (SSL) speech foundation models (SFM) [12–14] with their massive parameter counts. Despite the growing prominence of SFMs in ASR research, limited studies have investigated speaker adaptation of SFMs for dysarthric speech recognition [10, 15, 16].

Recent studies on SFM adaptation [17–20] have primarily focused on normal speech, with most approaches utilizing a single adapter for transfer learning. In contrast, Mixture of Experts (MoE) methods [21–23] demonstrate superior effectiveness addressing data heterogeneity such as speaker-level diversity. Individual experts develop specialized capabilities to handle specific data distributions, while their diversity enables comprehensive

coverage and generalization to unseen data. These MoE methods have been extensively applied in large language models (LLMs) [24–26] and widely adopted in speech recognition, where they have been integrated into end-to-end Transformer or Conformer [27–31] as well as pre-trained SFMs [32].

However, existing research on MoE approaches has predominantly focused on typical speech. When applying MoE to dysarthric speaker adaptation, three significant issues emerge: **a) Mobility issues** of dysarthric speakers hinder large-scale speech data collection, resulting in data sparsity and speaker bias. Such issues severely restrict the generalizability to unseen speakers; **b) Batch-mode unsupervised test-time adaptation** introduces substantial processing delays due to its two-stage process: pseudo-label generation followed by speaker-dependent (SD) parameter updates. The resulting high latency imposes a significant physical burden on dysarthric users during interactions; and **c) The pathological nature** of dysarthric speech necessitates the incorporation of domain knowledge to ensure both diverse expert specialization and comprehensive coverage of the MoE.

To address these challenges, we propose a novel MoE-based speaker adaptation framework for SFM based dysarthric speech recognition. This approach enables **zero-shot** adaptation and **real-time** processing while effectively incorporating **domain knowledge**. Specifically, feature-driven routing networks are designed to generate homogeneous SD routing parameters on the fly, enabling **a) zero-shot** adaptation and **b) real-time** processing. In addition, **c) domain knowledge**, such as severity and gender information, is incorporated by initializing each expert with severity and gender conditioned adapter parameters from adaptive training [10], allowing experts to focus on distinct severity and gender groups. These severity and gender labels are further utilized in classification tasks to better capture dysarthric speaker characteristics. Additionally, a Kullback-Leibler (KL) divergence loss is introduced during training to further enforce diversity among experts and their generalization to unseen speakers's data.

The main contributions of our work are summarized below:

1) Novelty: To the best of our knowledge, this paper is the first to investigate on-the-fly MoE-based speaker adaptation for dysarthric speech recognition, whereas prior efforts have primarily focused on typical speech [32]. Our method addresses the three major challenges outlined earlier: **a)** while previous methods lack the ability to adapt to unseen speakers, our **zero-shot** approach achieves greater generalization and extends applicability; **b)** our **on-the-fly** router achieves real-time speaker adaptation, which is more efficient compared to previous two-stage batch-mode methods [10,32]; and **c)** compared to previous approaches focusing on speaker identity only [32], we leverage **domain knowledge** to better model dysarthric speakers.

*Corresponding author.

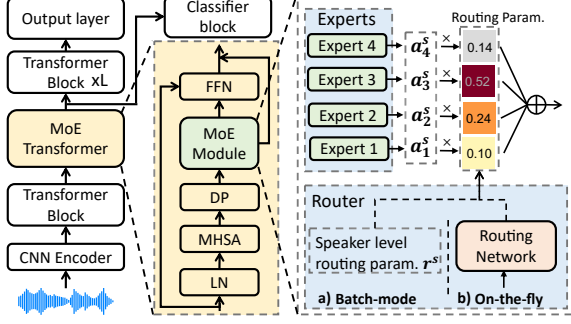


Figure 1: Architecture of MoE-based adaptation on SFM, where the routing parameters are derived from either **a)** speaker-dependent parameters in batch mode; or **b)** an on-the-fly routing network. “LN”, “MHSA”, “DP” and “FFN” are layer-norm, multi-head self-attention, dropout and feedforward.

2) Performance: Experimental results on the UASpeech [33] dysarthric corpus suggest that **i)** the proposed on-the-fly MoE-based adaptation approach produces statistically significant word error rate (WER) reductions of up to **1.34%** absolute (**6.36%** relative) over the baseline SI HuBERT and WavLM models. **ii)** Consistent WER reductions of up to **2.55%** absolute (**11.44%** relative) and **iii)** real-time factor (RTF) speedups of up to **7 times** are obtained over batch-mode adaptation across varying speaker-level data quantities. Further, **iv)** the lowest published WER of **16.35%** (**46.77%** on very low intelligibility) is obtained after cross-system multi-pass rescoring [9].

3) Analysis: Heatmap visualization intuitively reveals that the on-the-fly predicted SD routing parameters exhibit more **consistent** and **interpretable** speech impairment severity centric features than those obtained without domain knowledge.

2. Batch-Mode MoE Speaker Adaptation

Backbone Speech Foundation Models (SFMs): SSL speech foundation models such as Wav2vec2.0 [12], HuBERT [14], and WavLM [13] share similar Transformer-based backbones. For example, HuBERT contains three main components: 1) a multi-layer CNN-based feature encoder; 2) an L -layer transformer-based context network with a projection layer; and 3) a k -means quantization module. In this paper, we fine-tune the pre-trained HuBERT and WavLM with a CTC decoder.

MoE Architecture: As shown in Fig. 1, the MoE module is integrated into the 2^{nd} Transformer block, positioned between the feedforward layer and the dropout module. Residual Adapter Blocks (RAB) [10] act as expert network modules. All speakers share experts, while the router uses SD learnable parameters in batch mode. For speaker s , let N and $\mathbf{r}^s \in \mathbb{R}^N$ denote the number of experts and the SD routing parameters. The adapted hidden outputs of the MoE module are given as $\mathbf{h}^s = \sum_{i=1}^N r_i^s \mathbf{a}_i^s$, where \mathbf{a}_i^s denotes the outputs of the i -th expert for speaker s .

Multi-task Learning: To enforce diversity among experts and their generalization to unseen speakers’s data, **1)** a Kullback-Leibler (KL) divergence loss \mathcal{L}_{KL} is introduced to penalize similarity between the outputs of different experts:

$$\mathcal{L}_{KL} = - \sum_{(i,j) \in [N]^2: i \neq j} D_{KL}(\zeta(\mathbf{a}_i^s) || \zeta(\mathbf{a}_j^s)) \quad (1)$$

where D_{KL} is the KL divergence and $[N]^2$ denotes the Cartesian product of the set $\{1, \dots, N\}$ with itself. $\zeta(\cdot)$ is a Softmax function converting \mathbf{a}_i^s to a probability distribution¹. Further-

¹Alternative settings, such as $\zeta(\mathbf{a}_i^s) = \mathcal{N}(\mathbf{a}_i^s, 1)$ were found to degrade performance.

more, domain knowledge is incorporated by initializing the experts using adapter parameters from adaptive training [10], providing a robust foundation for specialized expert development. To better capture dysarthric speaker characteristics, **2)** an auxiliary classification task² with cross-entropy (CE) loss is used. As shown in Fig. 2(a), the combined batch-mode MoE-based adaptation cost function is $\mathcal{L}_B = \mathcal{L}_{CTC} + \alpha \mathcal{L}_{KL} + \beta \mathcal{L}_{CE}$, where α and β are empirically set to 5 and 0.1 in this paper.

Test-Time Adaptation: Unsupervised test-time adaptation is performed on speaker data without speech transcription or classification labels. The classification label³ \hat{C}^s for test speaker s is automatically predicted using the spectro-temporal feature based neural network classifiers in [4, 34]. The hypothesis supervision \hat{Y}^s for adaptation is generated by decoding the test data using the baseline unadapted SFMs. During unsupervised test-time adaptation, as shown in Fig. 2(c), the speaker-level routing parameters \mathbf{r}^s are re-initialized and optimized by:

$$\{\hat{\mathbf{r}}^s\} = \arg \min_{\{\mathbf{r}^s\}} \{\mathcal{L}_B(\hat{Y}^s, \hat{C}^s | \mathbf{X}^s; \mathbf{r}^s)\} \quad (2)$$

where \mathbf{X}^s is the input waveform for speaker s . All other parameters of the MoE-based SFM remain frozen.

Speaker Adaptive Training (SAT): During supervised training, SAT generates speaker-invariant canonical models that provide a more neutral and robust starting point for unsupervised test-time adaptation compared to standard non-SAT models. As shown in Fig. 2(a), the speaker-level routing parameters are jointly optimized with both the backbone SFM’s parameters Θ and those of the experts Θ_e during SAT, given as follows:

$$\{\hat{\Theta}, \hat{\Theta}_e, \hat{\theta}_S\} = \arg \min_{\{\Theta, \Theta_e, \theta_S\}} \sum_{s \in S} \{\mathcal{L}_B(\mathbf{Y}^s, \mathbf{C}^s | \mathbf{X}^s; \Theta, \Theta_e, \theta_S)\} \quad (3)$$

where $\theta_S = \{\mathbf{r}^s\}_{s \in S}$ is the SD parameter sets associated with training data. \mathbf{Y}^s and \mathbf{C}^s denote the ground truth transcription and classification label, respectively.

Adaptation Data: As shown in the line chart in Fig. 2(c), unsupervised test-time adaptation in batch mode is highly dependent on the utterance count. The routing parameters initially fluctuate significantly and require substantial data accumulation to converge, introducing notable processing delays.

3. On-the-fly MoE Speaker Adaptation

Routing Network Architecture: To achieve zero-shot speaker adaptation and reduce latency, a feature-driven routing network is designed to generate homogeneous SD routing parameters (shown in the line chart of Fig. 2(b)) on the fly, enabling efficient and real-time speaker adaptation. As depicted in Fig. 2(b), the routing network comprises two alternating feedforward layers and layernorm modules, an attentive pooling block for capturing intra-utterance speaker context, and a final linear layer.

Attentive Pooling: To capture both the internal contexts of utterances and their temporal dynamics, attentive statistics pooling [35] is integrated into the routing network. Let $\hat{\mathbf{h}}_t^{s,k}$ and $T^{s,k}$ respectively denote the normalized hidden outputs at time step t and the frame count for the k -th utterance of speaker s , respectively. The weighted mean and standard deviation are given as $\boldsymbol{\mu}^{s,k} = \sum_{t=1}^{T^{s,k}} \alpha_t^{s,k} \hat{\mathbf{h}}_t^{s,k}$ and $\boldsymbol{\sigma}^{s,k} = \sqrt{\sum_{t=1}^{T^{s,k}} \alpha_t^{s,k} \hat{\mathbf{h}}_t^{s,k} \odot \hat{\mathbf{h}}_t^{s,k} - \boldsymbol{\mu}^{s,k} \odot \boldsymbol{\mu}^{s,k}}$, where $\alpha_t^{s,k}$ is the normalized attention score at time step t , given as:

²Three levels of domain knowledge are used: severity, severity-gender, and speaker. Ablation studies are conducted in Sec. 4.3.

³Classification task is not performed for speaker-level domain knowledge in test-time adaptation.

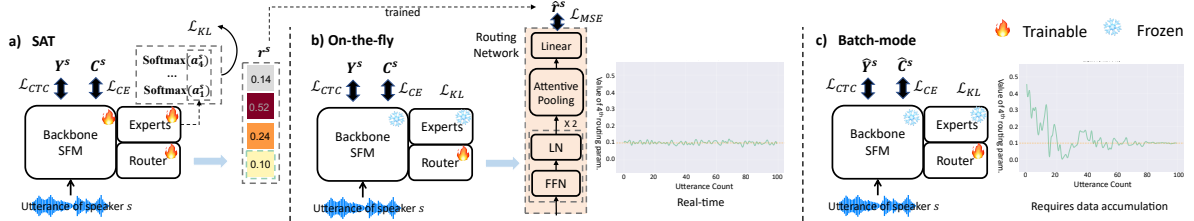


Figure 2: Examples of *on-the-fly* (a) & (b)) and *batch-mode* (a) & (c)) MoE-based speaker adaptation on the SFM. Routing parameters r^s in a) serve as SD parameters, while experts are shared by all speakers. The line charts in b) and c) illustrate the variation in a specific expert’s routing parameters as a function of utterance count.

$$\alpha_t^{s,k} = \frac{\exp(\mathbf{v}^T f(\mathbf{W}\hat{\mathbf{h}}_t^{s,k} + \mathbf{b}) + c)}{\sum_{t=1}^{T^{s,k}} \exp(\mathbf{v}^T f(\mathbf{W}\hat{\mathbf{h}}_t^{s,k} + \mathbf{b}) + c)} \quad (4)$$

where \mathbf{W} , \mathbf{b} , \mathbf{v} and c are the trainable parameters. $f(\cdot)$ denotes the $\text{Tanh}(\cdot)$ function. These statistics are then concatenated as $\mathbf{z}^{s,k} = [\boldsymbol{\mu}^{s,k}, \boldsymbol{\sigma}^{s,k}]$ and fed into the subsequent linear layer.

Multi-task Learning: The primary objective of the routing network is to minimize the mean squared error (MSE) between the predicted SD routing parameters and their corresponding training targets. The training targets \hat{r}^s can be obtained on the training data of speaker s in a supervised manner within the SAT framework (detailed in Sec. 2, and shown in Fig. 2(a)). Therefore, the overall on-the-fly MoE-based speaker adaptation learning cost function is given as $\mathcal{L}_O = \mathcal{L}_{CTC} + \alpha \mathcal{L}_{KL} + \beta \mathcal{L}_{CE} + \gamma \mathcal{L}_{MSE}$, where γ is empirically set as 0.5 in this paper.

As shown in Fig. 2(b), during on-the-fly speaker adaptation, the routing network Θ_P is optimized as:

$$\{\hat{\Theta}_P\} = \arg \min_{\{\Theta_P\}} \sum_{s \in S} \{\mathcal{L}_O(\mathbf{Y}^s, \mathbf{C}^s, \hat{\boldsymbol{\theta}}_S | \mathbf{X}^s; \Theta_P)\} \quad (5)$$

the parameters of the backbone SFM and experts are initialized using those obtained after SAT and frozen. During decoding, the SD routing parameters of each test utterance are predicted online and directly applied to the outputs of the experts for test-time on-the-fly adaptation.

4. Experiments

4.1. Task Description and Experimental Setup

UASpeech [33] is the largest publicly available dysarthric speech dataset containing 16 dysarthric and 13 control speakers. It includes 155 common and 300 uncommon words and is further divided into three subset blocks per speaker. The same 155 common words are used across all blocks, while the uncommon words vary. Data from Blocks 1 and 3 of all 29 speakers and Block 2 of 13 control speakers form the training set, while Block 2 of 16 dysarthric speakers is the test set. After silence stripping and speed perturbation based data augmentation [36], the training set comprises 173 hours of audio, with 9 hours for evaluation. The pre-trained models on UASpeech are HuBERT⁴ and WavLM⁵. The settings of RAB-based experts follow [10].

4.2. Main Results

Several trends can be observed from Table 1:

1) Comparison with the SI baseline, i-vector and x-vector adaptation: The proposed on-the-fly MoE-based speaker adaptation consistently outperforms these systems with statistically significant WER reductions of up to **1.34% absolute (6.36% relative)** on HuBERT and WavLM (Sys. 7 vs. 1,2,3 & Sys. 11 vs. 8). These WER reductions align with the more invariant

Table 1: Performance comparison between baseline, i-vector, x-vector, RAB-based speaker adaptation and the proposed MoE-based speaker adaptation on HuBERT and WavLM. † and * denote statistically significant (MAPSSWE [37], $\alpha = 0.05$) improvements obtained against the SI baseline ASR systems (Sys. 1, 8). “+” represents score interpolation, while “X→Y” denotes two-pass rescoreing of the N-best outputs from system X by system Y. “VL/L/M/H” denotes the speech intelligibility groups “very low”, “low”, “mid” and “high”.

Sys.	Model	Adapt. Method	# SD Param.	On The Fly	WER (%)				RTF	
					VL	L	M	H		All
1		SI	-	X	56.83	22.43	11.86	2.70	21.03	0.31
2		i-vector	-	✓	55.79†	21.87†	11.98	2.68	20.69†	0.36
3		x-vector	-	✓	55.20†	21.19†	11.53	2.58	20.26†	0.35
4	HuBERT	RAB	4M	X	54.90†	19.64†	9.33†	2.46	19.34†	2.59
5			-	✓	57.97	22.70	12.63	2.71	21.50	0.32
6		MoE	160	X	54.76†	20.93†	9.88†	2.48	19.75†	2.39
7			-	✓	54.42†	21.02†	9.88†	2.59	19.74†	0.39
8	WavLM	SI	-	X	56.49	22.76	11.57	2.89	21.06	0.42
9		RAB	4M	X	53.60*	20.13*	9.90*	2.35	19.26*	3.25
10		MoE	160	X	53.94*	20.65*	10.04*	2.82	19.65*	3.17
11		-	✓	53.92*	21.00*	10.02*	2.77	19.72*	0.45	
12	TDNN	LHUC-SAT	25K	X	61.62	24.56	15.82	6.50	24.64	-
13	(12→4) + (12→7) + (12→9) + (12→11)				46.77	16.62	6.53	2.64	16.35	-

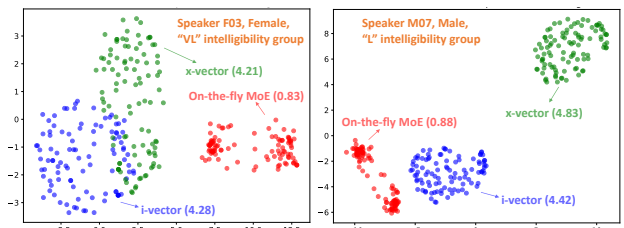


Figure 3: T-SNE visualization of the on-the-fly MoE-based, i-vector, and x-vector adaptation. The determinants of their covariance matrices are shown in each bracket.

speaker features produced by on-the-fly MoE in T-SNE visualization compared to i-vectors and x-vectors, as shown in Fig. 3.

2) Comparison with RAB-based⁶ methods: Both batch-mode and on-the-fly MoE-based adaptation achieve comparable average WERs compared to batch-mode RAB-based method [10], while producing lower WERs on the “VL” group (Sys. 6 & 7 vs. 4). Notably, the MoE-based methods require only **1/25000** of the SD parameters (Sys. 6 vs. 4) and operate at **1/7** of the RTF (Sys. 7 vs. 4). The parameter-heavy RAB method, when applied on-the-fly, underperforms against the SI baseline (Sys. 5 vs. 1). In contrast, the proposed on-the-fly MoE-based adaptation largely outperforms the on-the-fly RAB method with WER reductions of **1.76% absolute (8.19% relative, Sys. 7 vs. 5)**.

3) Comparison with batch-model MoE-based adaptation: The on-the-fly MoE-based speaker adaptation achieves comparable performance to offline batch-mode MoE-based method,

⁴huggingface.co/facebook/hubert-large-ls960-ft

⁵huggingface.co/microsoft/wavlm-large

⁶The RAB-based method can be considered as a special case of MoE, with a single expert and SD parameters as the expert parameters.

while operating approximately **7 times** faster in terms of RTF (Sys. 7 vs. 6 & Sys. 11 vs. 10).

4) Best performing system: By combining multiple adapted systems, including LHUC-SAT TDNN, RAB-based adapted SFMs, and the proposed MoE-based adapted SFMs via cross-system rescoring, the lowest published overall WER of **16.35%** (46.77% on very low intelligibility, Sys. 13) is obtained. Finally, the performance of our best system is contrasted against recently published state-of-the-art results in Table 2.

Table 2: WERs of published and our best system on UASpeech

System	On The Fly	VL	All
BUT-2022 Wav2vec2.0 + fMLLR + xvectors [15]	✓	57.72	22.83
Nagoya Univ.-2022 WavLM [38]	-	71.50	51.80
FAU-2022 Cross-lingual XLSR + Conformer [39]	-	62.00	26.10
JHU-2023 DuTa-VC (Diffusion) + Conformer [40]	-	63.70	27.90
CUHK-2024 HuBERT + sys. comb. [9]	✗	50.70	20.56
CUHK-2024 Wav2vec2/HuBERT + GAN Data Aug. + sys. comb. [41]	✗	46.47	16.53
CUHK-2024 DA + SVR adapt + sys. comb. [42]	✓	57.33	23.33
HuBERT/WavLM + MoE adapt. + sys. comb. (Sys. 13, Table 1, ours)	✓	46.77	16.35

Table 3: Performance (WER%) of HuBERT using batch-mode MoE-based speaker adaptation under different configurations of domain knowledge (“Know.”) integration and KL loss, as well as on-the-fly MoE-based speaker adaptation with and without attentive pooling (“Atten. Pool.”). Different expert types are investigated.

Sys.	Expert Init.	KL Loss	Class. Task	Domain Know.	# of Expert	Atten. Pool.	On The Fly	WER (%)			
								VL	L / M / H	All	
1	random	✗	✗	-	5	-	✗	57.97 / 21.96 / 11.57 / 2.61	21.07		
2		✗	✗				✗	57.01 / 21.33 / 10.98 / 2.73	20.63		
3		✗	✗				✗	55.72 / 20.74 / 10.16 / 2.82	20.08		
4	Init. from adapt. training RAB	✓	✓	Severity	5	✓	✗	55.61 / 20.54 / 10.29 / 2.58	19.94		
5								✓	✓	55.61 / 20.92 / 10.35 / 2.54	20.04
6								✗	✗	54.76 / 20.93 / 9.88 / 2.48	19.75
7								✓	✓	54.42 / 21.02 / 9.88 / 2.59	19.74
8				Severity, Gender	10	✓	✓	55.13 / 21.06 / 10.67 / 2.58	20.05		
9				Speaker	29	✓	✗	55.76 / 20.35 / 9.59 / 2.72	19.84		
10						✓	✓	54.12 / 21.03 / 10.06 / 2.71	19.75		

4.3. Ablation Studies

Table 3 presents the results of ablation studies on both batch-mode and on-the-fly MoE-based speaker adaptation on the HuBERT model. Several trends can be observed:

For batch-mode: **1)** Regarding the utilization of domain knowledge, initializing each expert with severity-specific adapter parameters from adaptive training (Sys. 2 vs. 1) and incorporating a severity classification task (Sys. 4 vs. 3) both lead to performance improvements; and **2)** the incorporation of KL loss produces large WER reductions (Sys. 3 vs. 2).

For on-the-fly: **1)** the on-the-fly MoE system with severity-gender experts outperforms the system with severity experts, while achieving comparable performance to the system with speaker-level experts (Sys. 7 vs. 5, 10); and **2)** the attentive pooling module produces better performance compared to the simple temporal average pooling (Sys. 7 vs. 8).

Table 4: Performance of on-the-fly MoE-based adaptation of WavLM with or without the speaker-level round-robin setting. * denotes statistically significant improvements against System 1.

Sys.	MoE Adapt.	Round-robin	VL / L / M / H	All
1	✗	✗	56.49 / 22.76 / 11.57 / 2.89	21.06
2	✓	✗	53.92* / 21.00* / 10.02* / 2.77	19.72*
3	✓	✓	54.63* / 21.80* / 11.51 / 3.02	20.21*

4.4. Analysis

Zero-shot Adaptation. To evaluate zero-shot performance of the on-the-fly MoE-based speaker adaptation, speaker-level round-robin experiments [43] are conducted. Specifically, for each test dysarthric speaker s , we exclude their data from the

training set before speaker adaptation. As shown in Table 4, the zero-shot on-the-fly MoE-based adaptation significantly outperforms the SI baseline (Sys. 3 vs. 1), even though the SI model is trained on data containing speakers from the test set.

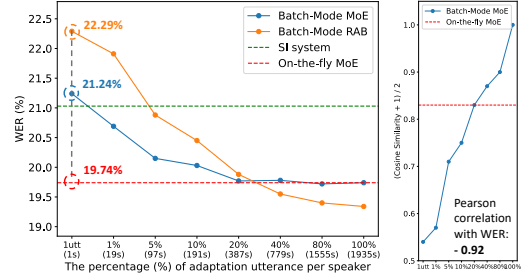


Figure 4: WER and cosine similarity on HuBERT systems with varying amounts of speaker adaptation data.

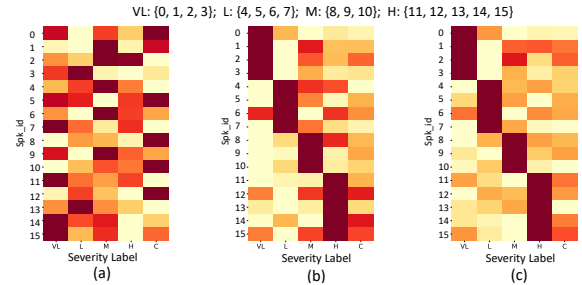


Figure 5: Heatmap visualization of routing parameters under varying settings on HuBERT: **a)** batch-mode without domain knowledge, **b)** batch-mode and **c)** on-the-fly with domain knowledge. “Severity Label: {Spk_ids}” is given at the top of Figure.

Low Processing Latency. The left part of Fig. 4 illustrates the WER variation of the HuBERT model across different speaker-level data quantities for various adaptation approaches. With just one utterance, the proposed on-the-fly MoE-based adaptation (red line) achieves considerable WER reductions of up to **2.55% absolute (11.44% relative)** compared to both batch-mode RAB and MoE approaches. The right sub-figure illustrates the cosine similarity between routing parameters obtained from partial versus complete data in the batch-mode MoE system. A strong negative correlation (Pearson coefficient = -0.92) between WER and cosine similarity is observed.

Domain Knowledge Benefits. Heatmap visualization of routing parameters under different settings on HuBERT in Fig. 5 shows that both on-the-fly and batch-mode MoE-based adaptation incorporating domain knowledge exhibit more consistent and interpretable speech impairment severity-centric features than those without domain knowledge (Fig. 5(b), (c) vs. (a)).

5. Conclusion

This paper presents a novel on-the-fly MoE-based speaker adaptation for SSL pre-trained SFMs on dysarthric speech. Feature-driven routing networks are designed to produce homogeneous SD routing parameters on the fly, thereby facilitating zero-shot and real-time speaker adaptation. Incorporating domain knowledge, such as severity and gender, ensures diverse expert specialization and comprehensive MoE coverage. Experiments on UASpeech shows that the proposed on-the-fly MoE-based adaptation approaches produce up to 1.34% absolute (6.36% relative) WER reductions over unadapted SFMs. Comparable WER performance and RTF speedup ratios of 7 times are also obtained over batch-mode adaptation. Heatmap visualization further demonstrates the interpretability of the proposed methods.

6. Acknowledgements

This research is supported by Hong Kong RGC GRF grant No. 14200220, 14200021, 14200324, TRS T45-407/19N, Innovation Technology Fund grant No. ITS/218/21, the project of China Disabled Persons Federation (CDPF2023KF00002), Basic Research Project of ISCAS (ISCAS-JCMS-202306), Youth Innovation Promotion Association CAS Grant (2023119), and Guangzhou CASTF project (2022MZK02).

7. References

- [1] S. Sehgal *et al.*, “Model adaptation and adaptive training for the recognition of dysarthric speech,” in *SLPAT*, 2015.
- [2] F. Xiong *et al.*, “Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition,” in *ITG-Symposium*. VDE, 2018, pp. 1–5.
- [3] S. Liu *et al.*, “Recent Progress in the CUHK Dysarthric Speech Recognition System,” *TASLP*, vol. 29, pp. 2267–2281, 2021.
- [4] M. Geng *et al.*, “Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition,” *TASLP*, vol. 30, pp. 2597–2611, 2022.
- [5] S. Hu *et al.*, “Exploring Self-supervised Pre-trained ASR Models For Dysarthric and Elderly Speech Recognition,” in *ICASSP*. IEEE, 2023, pp. 1–5.
- [6] Z. Yue *et al.*, “Acoustic Modelling From Raw Source and Filter Components for Dysarthric Speech Recognition,” *TASLP*, vol. 30, pp. 2968–2980, 2022.
- [7] S. Hu *et al.*, “Exploiting Cross-Domain And Cross-Lingual Ultrasound Tongue Imaging Features For Elderly And Dysarthric Speech Recognition,” in *INTERSPEECH*, 2023, pp. 2313–2317.
- [8] S. Hu, S. Liu *et al.*, “Exploiting Cross Domain Acoustic-to-articulatory Inverted Features for Disordered Speech Recognition,” in *ICASSP*. IEEE, 2022, pp. 6747–6751.
- [9] S. Hu *et al.*, “Self-Supervised ASR Models and Features for Dysarthric and Elderly Speech Recognition,” *TASLP*, vol. 32, pp. 3561–3575, 2024.
- [10] S. Hu, X. Xie *et al.*, “Structured Speaker-Deficiency Adaptation of Foundation Models for Dysarthric and Elderly Speech Recognition,” *arXiv preprint arXiv:2412.18832*, 2024.
- [11] B. L. Smith *et al.*, “Temporal characteristics of the speech of normal elderly adults,” *JSLHR*, vol. 30, pp. 522–529, 1987.
- [12] A. Baevski *et al.*, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in NeuralIPS*, 2020, pp. 12 449–12 460.
- [13] S. Chen *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *JSTSP*, vol. 16, pp. 1505–1518, 2022.
- [14] W.-N. Hsu *et al.*, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *TASLP*, vol. 29, pp. 3451–3460, 2021.
- [15] M. K. Baskar *et al.*, “Speaker adaptation for Wav2vec2 based dysarthric ASR,” in *INTERSPEECH*, 2022, pp. 3403–3407.
- [16] Y. Jiang, T. Wang *et al.*, “Perceiver-Prompt: Flexible Speaker Adaptation in Whisper for Chinese Disordered Speech Recognition,” in *INTERSPEECH*, 2024, pp. 2025–2029.
- [17] B. Thomas *et al.*, “Efficient adapter transfer of self-supervised speech models for automatic speech recognition,” in *ICASSP*. IEEE, 2022, pp. 7102–7106.
- [18] Z.-C. Chen *et al.*, “Exploring Efficient-Tuning Methods in Self-Supervised Speech Models,” in *SLT*, 2023, pp. 1120–1127.
- [19] B. Li *et al.*, “Efficient Domain Adaptation for Speech Foundation Models,” in *ICASSP*, 2023, pp. 1–5.
- [20] Z.-C. Chen *et al.*, “Chapter: Exploiting Convolutional Neural Network Adapters for Self-Supervised Speech Models,” in *ICASSPW*, 2023, pp. 1–5.
- [21] R. A. Jacobs *et al.*, “Adaptive mixtures of local experts,” *Neural computation*, vol. 3, pp. 79–87, 1991.
- [22] M. I. Jordan *et al.*, “Hierarchical mixtures of experts and the EM algorithm,” *Neural computation*, vol. 6, pp. 181–214, 1994.
- [23] N. Shazeer *et al.*, “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” *arXiv preprint arXiv:1701.06538*, 2017.
- [24] A. Q. Jiang *et al.*, “Mixtral of experts,” *arXiv preprint arXiv:2401.04088*, 2024.
- [25] D. Dai *et al.*, “Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models,” *arXiv preprint arXiv:2401.06066*, 2024.
- [26] X. Sun *et al.*, “Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent,” *arXiv preprint arXiv:2411.02265*, 2024.
- [27] Z. You *et al.*, “SpeechMoE: Scaling to Large Acoustic Models with Dynamic Routing Mixture of Experts,” in *INTERSPEECH*, 2021, pp. 2077–2081.
- [28] Y. Bai *et al.*, “Parameter-Efficient Conformers via Sharing Sparsely-Gated Experts for End-to-End Speech Recognition,” in *INTERSPEECH*, 2022, pp. 1676–1680.
- [29] P. Chen *et al.*, “BA-MoE: Boundary-Aware Mixture-of-Experts Adapter for Code-Switching Speech Recognition,” in *ASRU*, 2023, pp. 1–7.
- [30] Y. Kwon *et al.*, “MoLE: Mixture Of Language Experts For Multi-Lingual Automatic Speech Recognition,” in *ICASSP*, 2023, pp. 1–5.
- [31] M. Perez *et al.*, “Aphasic Speech Recognition Using a Mixture of Speech Intelligibility Experts,” in *INTERSPEECH*, 2020, pp. 4986–4990.
- [32] Q. Zhao *et al.*, “SAML: Speaker Adaptive Mixture of LoRA Experts for End-to-End ASR,” in *INTERSPEECH*, 2024, pp. 777–781.
- [33] H. Kim *et al.*, “Dysarthric speech database for universal access research,” in *INTERSPEECH*, 2008, pp. 1741–1744.
- [34] M. Geng *et al.*, “Use of Speech Impairment Severity for Dysarthric Speech Recognition,” in *INTERSPEECH*, 2023, pp. 2328–2332.
- [35] K. Okabe *et al.*, “Attentive statistics pooling for deep speaker embedding,” in *INTERSPEECH*, 2018, pp. 2252–2256.
- [36] M. Geng *et al.*, “Investigation of Data Augmentation Techniques for Disordered Speech Recognition,” in *INTERSPEECH*, 2020, pp. 696–700.
- [37] L. Gillick and S. J. Cox, “Some statistical issues in the comparison of speech recognition algorithms,” in *ICASSP*. IEEE, 1989, pp. 532–535.
- [38] L. P. Violeta, W. C. Huang, and T. Toda, “Investigating Self-supervised Pretraining Frameworks for Pathological Speech Recognition,” in *INTERSPEECH*, 2022, pp. 41–45.
- [39] A. Hernandez *et al.*, “Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition,” in *INTERSPEECH*, 2022, pp. 51–55.
- [40] H. Wang *et al.*, “DuTa-VC: A Duration-aware Typical-to-atypical Voice Conversion Approach with Diffusion Probabilistic Model,” in *INTERSPEECH*, 2023, pp. 1548–1552.
- [41] H. Wang, X. Xie *et al.*, “Enhancing Pre-Trained ASR System Fine-Tuning for Dysarthric Speech Recognition Using Adversarial Data Augmentation,” in *ICASSP*, 2024, pp. 12 311–12 315.
- [42] M. Geng *et al.*, “Homogeneous Speaker Features for On-the-Fly Dysarthric and Elderly Speaker Adaptation,” *arXiv preprint arXiv:2407.06310*, 2024.
- [43] N. M. Joy and S. Umesh, “Improving Acoustic Models in TORGO Dysarthric Speech Database,” *IEEE TNSRE*, vol. 26, no. 3, pp. 637–645, 2018.