University of Zurich and ETH Zurich
Institute of Neuroinformatics
Swiss Data Science Center

Master's Thesis Opportunity
Personalizing Automatic Speech Recognition Models for Non-normative Articulation via Active Learning

Supervisors

Prof. Dr. Benjamin Grewe	Primary PI	
Roman Boehringer	Supervisor	roman@ini.ethz.ch
Niclas Pokel	Supervisor	npokel@ethz.ch
Dr. Yingqiang Gao	Supervisor	yingqiang.gao@sdsc.ethz.ch

Abstract

Automatic Speech Recognition (ASR) for individuals with impaired speech remains a significant challenge due to data scarcity and high acoustic variability. This project builds upon a Bayesian personalization framework developed by our team, which has recently been accepted for publication [14, 16, 15] and got merged into a web-app application [17]. Further work involved experiments with Language Model Judges for Reinforcement Learning, Generation of synthetic training data and MoE Architectures with varying success.

Together, these components are the building blocks of a personalization loop that could keep improving the model long after the initial enrollment, as the user's speech, vocabulary and recording conditions change. So far, however, each building block has been studied in isolation, and every improvement step still relies on the user. This thesis investigates how a personalized ASR model can improve *continuously* while keeping the burden on the user minimal: which data is most informative to learn from next, which supervision can be obtained without the user, and how the user's own feedback can be made cheaper or partly automated. The project is research-oriented, leaves room for the student to set priorities, and aims at a publication at an international conference.

Research Goals and Problem Statement

Personalization does not end after enrollment. A user's speech changes over time, new vocabulary appears, and the model encounters conditions it was never trained on. Keeping it up to date requires new training data, yet labelled speech from the user is the scarcest resource in this setting, as recording and correcting are tiring, especially with co-occurring motor impairments [17]. Classical active learning reduces labelling effort by querying the most informative samples, but assumes a single, reliable oracle of uniform cost [19]. Here, several sources of supervision are available that differ in both cost and reliability: user corrections, synthetic speech and automatic judgements by

language models [2]. Automatic sources carry their own risks. Whisper is prone to hallucinations, particularly for speakers with aphasia [9], and models repeatedly trained on their own outputs can degrade [20], together with the uncertainty estimates the loop relies on [13]. The central question of the thesis is therefore:

> *How can a personalized ASR model for impaired speech improve continuously over time while minimizing the effort required from the user?*

Proposed Work

The following directions are starting points; the student is expected to select and combine them as the research develops.

- **Informative data selection.** The Phoneme Difficulty Score [16] and the epistemic uncertainty of VI-LoRA [15] show where the model struggles. Combined with Semantic Re-chaining [14], they can guide the search for the phrases that are most informative to add to the fine-tuning repertoire, building on Bayesian active learning [5, 3, 8].
- **Supervision without the user.** Synthetic speech can target the user's difficult phonemes without any recording effort [11, 4], and unlabelled everyday speech can be exploited through uncertainty-filtered self-training [7, 6].
- **Cheaper and partly automated annotation.** An LLM judge can pick the most plausible transcript among candidates sampled from the Bayesian model, either to pre-rank the top-k suggestions shown to the user or to create preference pairs for DPO without user involvement [21, 18, 12, 1]. Multi-fidelity active search offers a principled way to split a feedback budget between cheap ratings and costly corrections [10].
- **Closing the loop.** Combine these components into one continuous pipeline in Adapt4Me [17], decide when to retrain and which source to trust, and evaluate the result longitudinally in terms of error rate per unit of user effort, error severity and calibration.

Context & Prior Work

This project builds on a line of work in our group: Semantic Re-chaining for personalization from little data [14], uncertainty-guided sampling with the Phoneme Difficulty Score [16], Bayesian fine-tuning with VI-LoRA [15], the Adapt4Me application [17], and multi-fidelity active search for discovering severe errors [10]. In the wider literature, Bayesian active learning selects data by its expected information about the model parameters [5, 3], and proactive learning extends this to several imperfect oracles with different costs [2]. For ASR, unlabelled speech has been exploited through uncertainty-driven self-training [7, 6], LLM-based error correction [1] and reinforcement learning from LLM feedback [12], while TTS-based augmentation has improved dysarthric ASR [11, 4]. These approaches have mostly been studied in isolation and on normative or multi-speaker data; how they interact in a continuous, single-speaker personalization loop is still open.

References

[1] Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng Siong Chng. HyPoradise: An open baseline for generative speech recognition with large language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023.

[2] Pinar Donmez and Jaime G. Carbonell. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pages 619–628. ACM, 2008. doi:10.1145/1458082.1458165.

[3] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR, 2017.

[4] Enno Hermann and Mathew Magimai-Doss. Few-shot dysarthric speech recognition with text-to-speech data augmentation. In *Proc. Interspeech 2023*, pages 156–160, 2023. doi:10.21437/Interspeech.2023-2481.

[5] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[6] Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Chengwei Qin, Pin-Yu Chen, Eng Siong Chng, and Chao Zhang. Self-taught recognizer: Toward unsupervised adaptation for speech foundation models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi:10.52202/079017-0930.

[7] Sameer Khurana, Niko Moritz, Takaaki Hori, and Jonathan Le Roux. Unsupervised domain adaptation for speech recognition via uncertainty driven self-training. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6553–6557, 2021. doi:10.1109/ICASSP39728.2021.9414299.

[8] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 7026–7037, 2019.

[9] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 2024. doi:10.1145/3630106.3658996.

[10] Orlando Leone, Niclas Pokel, Pehuén Moure, Yingqiang Gao, and Roman Boehringer. Micro: Multi-fidelity active search for severe error discovery, 2026. URL https://arxiv.org/abs/2609.26025.

[11] Wing-Zin Leung, Mattias Cross, Anton Ragni, and Stefan Goetze. Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis. In *Proc. Interspeech 2024*, pages 2494–2498, 2024. doi:10.21437/Interspeech.2024-1645.

[12] Shaoshi Ling and Guoli Ye. Customizing speech recognition model with large language model feedback. *arXiv preprint arXiv:2506.11091*, 2025.

[13] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pages 13991–14002, 2019.

[14] Niclas Pokel, Pehuén Moure, Roman Boehringer, and Yingqiang Gao. Adapting foundation speech recognition models to impaired speech: A semantic re-chaining approach for personalization of german speech. In *12th edition of the Disfluency in Spontaneous Speech Workshop (DiSS 2025)*, pages 82–86, 2025. doi:10.21437/DiSS.2025-17.

[15] Niclas Pokel, Pehuén Moure, Roman Boehringer, Shih-Chii Liu, and Yingqiang Gao. Variational low-rank adaptation for personalized impaired speech recognition. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18457–18461, 2026. doi:10.1109/ICASSP55912.2026.11462593.

[16] Niclas Pokel, Pehuén Moure, Roman Böehringer, and Yingqiang Gao. Data-Efficient ASR Personalization for Non-Normative Speech Using an Uncertainty-Based Phoneme Difficulty Score for Guided Sampling. In *Interspeech 2026*, pages 855–860, 2026. doi:10.21437/Interspeech.2026-776.

[17] Niclas Pokel, Yiming Zhao, Pehuen Moure, Yingqiang Gao, and Roman Boehringer. Demonstration of adapt4me: An uncertainty-aware authoring environment for personalizing automatic speech recognition to non-normative speech. In *Proceedings of the 8th ACM Conference on Conversational User Interfaces*, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400727412. URL https://doi.org/10.1145/3816046.3816317.

[18] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi:10.52202/075280-2338.

[19] Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison, Department of Computer Sciences, 2009.

[20] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024. doi:10.1038/s41586-024-07566-y.

[21] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023.