Non-Archival Abstract

Benchmarking Automatic Speech Recognition for Psychotherapy Session Transcription Under Real-World Recording Conditions

Viktoria Lavrynovska[1], Thang Vu[1], Seemab Hassan[2], Christoph Nikendei[2], Hans-Christoph Friederich[2,3], Joe J. Simon[2,3]

[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Department of General Internal Medicine, Psychosomatics and Psychotherapy, Centre for Psychosocial Medicine, University Hospital Heidelberg, Germany
[3]DZPG (German Centre for Mental Health – Partner Site Heidelberg/Mannheim/Ulm), Germany
{first.last}@ims.uni-stuttgart.de[1]; {first.last}@med.uni-heidelberg.de[2]

Introduction The preparation of psychotherapy reports required by the German health insurance system is time-consuming and a substantial administrative burden for psychotherapists. Recent advances in automatic speech recognition (ASR) and large language models (LLMs) enable new approaches to partially automate this documentation process. However, psychotherapy recordings pose significant challenges for current ASR systems due to spontaneous and overlapping speech, domain-specific terminology, long recording durations, and variable acoustic conditions. Recent ASR models such as Whisper (Radford et al., 2023) achieve strong performance across languages and domains, yet their robustness in clinical settings remains insufficiently studied. Previous research has demonstrated the potential of ASR for clinical conversations, while also revealing substantial challenges in real-world conversational speech (Miner et al., 2020; Schubert et al., 2024; Taubitz et al., 2025). Furthermore, privacy requirements limit the applicability of cloud-based ASR services in sensitive medical contexts, motivating the investigation of locally deployable approaches. While prior studies have examined individual aspects of clinical speech processing, a systematic evaluation of integrated transcription pipelines on large-scale datasets of real-world psychotherapy recordings is still lacking. To address this gap, we systematically evaluate a privacy-preserving transcription pipeline on a large-scale dataset of German psychotherapy recordings, examining the effects of ASR, speaker diarization, and processing strategies on transcript quality.

Method The dataset provided by the Heidelberg Institute for Psychotherapy (HIP) contains approximately 250 hours of manually transcribed video-recorded psychotherapy sessions. The extracted audio is processed using the pipeline shown in Figure 1 to generate transcripts with speaker labels. We benchmark multiple state-of-the-art ASR and diarization models and investigate the impact of audio preprocessing, transcript postprocessing, and parameter-efficient fine-tuning. Audio preprocessing strategies address variable recording quality through normalization and enhancement methods, while transcript postprocessing explores strategies for improving ASR outputs through disfluency handling, text normalization, and LLM-assisted refinement of transcript content and speaker attribution. Performance is evaluated using word-level, speaker-aware, and semantic metrics.

Preliminary Results Preliminary experiments demonstrate the potential of automated psychotherapy session transcription while highlighting the challenges posed by real-world clinical recordings. The WhisperX (Bain et al., 2023) baseline, combining Whisper large-v3 (Radford et al., 2023) with pyannote speaker diarization (Bredin, 2023; Plaquet and Bredin, 2023), achieves a word error rate (WER) of 24.4%. Parakeet-TDT-0.6B-v3 (Sekoyan et al., 2025) achieves improved transcription quality with a WER of 22.6%, while other ASR approaches show varying robustness on long-form psychotherapy conversations. For speaker diarization, preliminary results indicate that recent approaches such as DiariZen (Han et al., 2025a,b, 2026) show improvements in speaker-aware evaluation metrics.

Conclusion This work provides a systematic evaluation of a privacy-preserving psychotherapy transcription pipeline and investigates how model selection, processing and adaptation strategies influence transcription quality in realistic clinical conditions. Preliminary results indicate that ASR errors remain the key challenge, motivating further investigation of parameter-efficient fine-tuning and preprocessing techniques for psychotherapy speech.

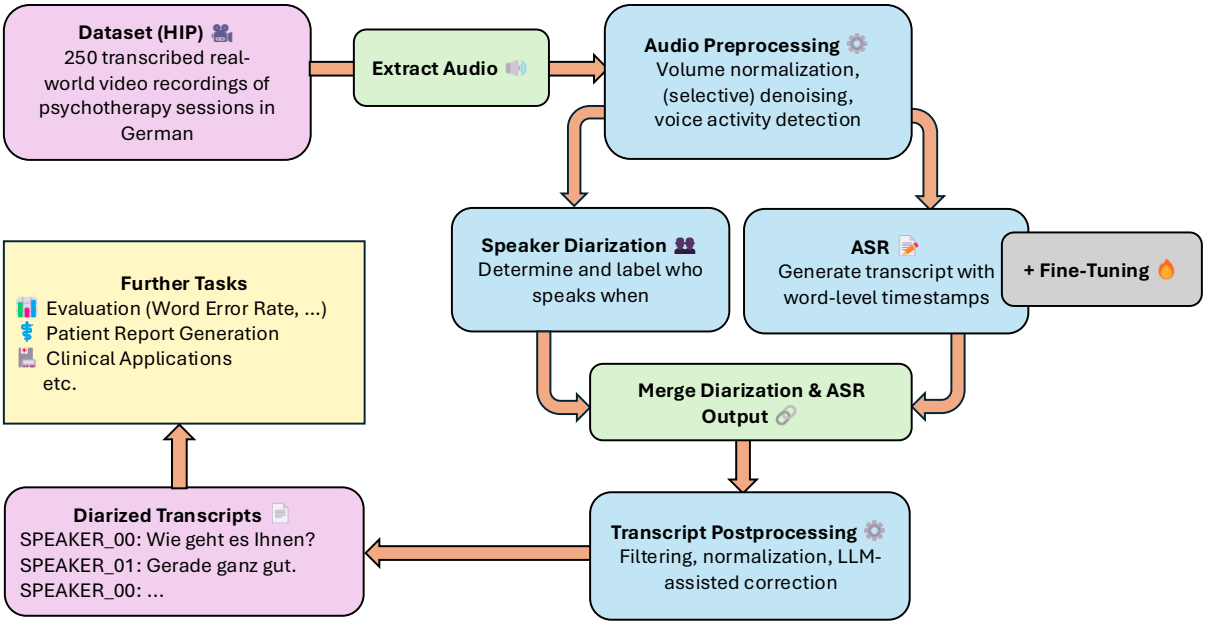

Figure 1: Overview of the proposed privacy-preserving psychotherapy transcription pipeline. This work systematically evaluates the components highlighted in blue (audio preprocessing, speaker diarization, ASR, and transcript postprocessing) to assess their impact on transcription quality.

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *Preprint*, arXiv:2303.00747.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, and Lukáš Burget. 2025a. Leveraging self-supervised learning for speaker diarization. In *Proc. ICASSP*.

Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, Jan Cernocky, and Lukas Burget. 2025b. Fine-tune Before Structured Pruning: Towards Compact and Accurate Self-Supervised Models for Speaker Diarization. *arXiv preprint arXiv:2505.24111*.

Jiangyu Han, Petr Pálka, Marc Delcroix, Federico Landini, Johan Rohdin, Jan Černocký, and Lukáš Burget. 2026. Efficient and Robust Speaker Diarization via Structured Pruning of Self-Supervised Models. *IEEE Transactions on Audio, Speech and Language Processing*, 34:1903–1914.

Adam S Miner, Albert Haque, Jason A Fries, Scott L Fleming, Denise E Wilfley, G Terence Wilson, Arnold Milstein, Dan Jurafsky, Bruce A Arnow, W Stewart Agras, and 1 others. 2020. Assessing the accuracy of automatic speech recognition for psychotherapy. *NPJ digital medicine*, 3(1):82.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Martha Schubert, Yamini Sinha, Julia Krüger, and Ingo Siegert. 2024. Speech Recognition Errors in ASR Engines and Their Impact on Linguistic Analysis in Psychotherapies. In *Elektronische Sprachsignalverarbeitung 2024, Tagungsband der 35. Konferenz, Regensburg, 6.-8. März 2024*, pages 203–210. TUDpress.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1B-v2 Parakeet-TDT-0.6B-v3: Efficient and High-Performance Models for Multilingual ASR and AST. *Preprint*, arXiv:2509.14128.

Friedrich-Samuel Taubitz, Linus HA Sehn, and Georg W Alpers. 2025. Zum Diktat, Language Model: Transkription, Sprecher*Innenzuordnung und Verarbeitung von Therapiegesprächen mit KI. *Verhaltenstherapie*.