Evaluating ASR in a Clinical Context : What Whisper Misses

Haeeul Hwang[1,2], Eric Jordan[1], Deok-Hee Kim-Dufor[3], Christophe Lemey[3], Motasem Alrahabi[1]

[1]Sorbonne Université, [2]Université de Paris Cité, [3]CHRU de Brest,

Correspondence: motasem.alrahabi@sorbonne-universite.fr

Abstract

Automatic Speech Recognition (ASR) powered by AI is rapidly advancing and finding applications across a wide range of domains. However, its application within domain specific contexts still represents a challenge, with the specific issues varying from one context to another.

In this paper, we examine the application of OpenAI's Whisper ASR system in the context of psychiatric interviews. First, through analysis of error rates in the automatic transcriptions then through analysis of the most common errors present in the transcriptions we found that Whisper achieved a Word Error Rate of 0.25 but failed to transcribe filler words most often associated with patient's hesitations during speech. We find that systems such as Whisper show great promise for applications in clinical contexts . However, due to the importance of filler words and other filled pauses from a clinical perspective, its application should be accompanied with fine-tuning and verification by specialists to ensure the best outcomes.

1 Introduction

Despite improvements in ASR models, the latest models fail to achieve the same impressive out of the box results when applied to specific domains. This can result in increased errors for minority groups or those with speech disfluencies (Koenecke et al., 2020, 2024) or poor performance on languages that are under-represented within the training data of the aforementioned models (San et al., 2024).

Among these domains healthcare is a highly sensitive area where the accuracy of speech recognition is critical. In psychiatric consultations, subtle nuances in language can carry important clinical implications, and even small transcription errors may influence diagnosis, treatment, or documentation quality (Liebenthal et al., 2023). In these cases evaluating the models' true performance remains a challenge, since widely used evaluation metrics, such as Word Error Rate (WER) and Character Error Rate (CER) quantify surface-level inconsistencies between the recognized text and the reference transcript. However, these metrics do not account for semantic integrity, context relevance, or the clinical impact of recognition errors (Miner et al., 2020).

To address these challenges, this study analyses ASR outputs from psychiatric consultations. Specifically, we measure error rates across 3 different clinical groups and investigate the most common errors of the ASR models.

Our contributions are as follows:

1. We present a dataset of real-world psychiatric consultation transcripts and their ASR outputs (described further below).

2. We conduct an analysis of the recognition errors, with a focus on words that were omitted or deleted in the ASR-generated transcripts.

1.1 Prior work

The application of ASR in a clinical context remains underdeveloped, notably due to recognition errors that undermine reliability. Within the setting of psychotherapy consultations, Miner et al. (2020) introduced a three-pronged framework for assessing ASR, emphasizing that conventional metrics such as WER alone do not capture clinically relevant nuances. Their study compared human and ASR-generated transcriptions of therapy sessions, evaluating performance from three perspectives: general linguistic accuracy (WER and semantic distance), recognition of depression-related vocabulary, and accuracy on passages with harm-related language (e.g., self-harm or violence).

While the general performance analysis showed an average WER of 25% performance on harm-related speech was significantly lower with a 34% WER, indicating poor reliability in safety-critical scenarios. This underscores how ASR systems can provide transcriptions that are mostly accurate, however they may miss finer details that are important for diagnosis. The authors conclude that these systems required further development before being ready for individual level safety surveillance.

Further work in the clinical domain has investigated the effectiveness of ASR systems for transcribing recordings of patients with Alzheimer's disease (Soroski et al., 2022; Akinrintoyo et al., 2025). Overall these studies showed promising results, with the transcripts produced being usable for distinguishing between patient and control groups (Soroski et al., 2022). However, the models' tendency to exclude filler words (e.g. *umm*, *uhh*) was noted, although this deficit could be made up for by fine-tuning Whisper on the patients' data (Akinrintoyo et al., 2025).

Additionally, some works have considered the use of ASR systems within the context of psychological experiments (Pfeifer et al., 2024; Ziman et al., 2018) finding error rates as low as 2.5 %, however these results only applied to studies with exclusively healthy participants (Pfeifer et al., 2024). Nonetheless, these works show the promise that these models can hold for application in psychological research.

2 Methodology

2.1 Dataset

Patient Group	Female	Male	Total
AR	22	19	41
NAR	5	5	11
FEP	3	4	7
Total	**30**	**28**	**59**

Table 1: Patient group counts by sex, including unknown and totals

Group	Mean	Std	Median	Min	Max
Agg	49	14	48	15	90
NAR	47	10	47	27	62
AR	50	14	48	16	90
FEP	48	18	53	15	74

Table 2: Overall and Group-Specific Recording Durations (in Minutes)

To assess ASR performance in psychiatric consultations, we compiled a dataset comprising audio recordings of 59 (30 female, 28 male, 1 undisclosed) patients at ultra-high risk for psychosis (UHR) with a psychiatrist. All participants were native speakers of French. The interview was a semi-structured conversation with predetermined questions on each patient's problems such as their background, family, social relationships, socio-professional insertion, emotional interactions, complaints about symptoms, and other issues brought up by the patient.

The average recording duration was 49 minutes (see Table 2). Two trained assistants transcribed the entire utterances verbatim, including filled pauses, mispronunciations, and neologisms, following clear guidelines. An experienced linguist then reviewed the transcripts to correct spelling errors – such as homophones and accents – without altering the verbatim content. The participants were grouped into three clinical categories according to the Comprehensive Assessment of At-Risk Mental States (CAARMS) (Yung et al., 2005) : AR (At Risk for psychosis, 41 patients), NAR (Not At Risk, 11 patients), and FEP (First-Episode Psychosis, 7 patients), see Table 1.

2.2 Data Preprocessing

2.2.1 Audio Processing

Prior to auto-transcription, all audio files – originally in formats such as WMA, M4A, MP4, and WAV – were converted to a uniform WAV format to ensure compatibility and consistency. The processed files were then transcribed using two ASR models: OpenAI's *Whisper medium* and *Whisper turbo* (Radford et al., 2023).

To improve evaluation accuracy, the start of each recording was trimmed to align the starting timepoints across files. Subsequently, a series of preprocessing steps were applied to both manual and ASR-generated transcripts before calculating error rates.

2.2.2 Text Processing

1. **Special character removal** : We first removed non-verbal annotations and special characters (e.g., #, [,]) that were used to distinguish between patient and clinician speech and to mark the proper names in the manual transcripts. This cleaned version of the reference transcripts was then used to compute baseline WER and CER scores.

2. **Normalization** : We applied the *Whisper nor-*

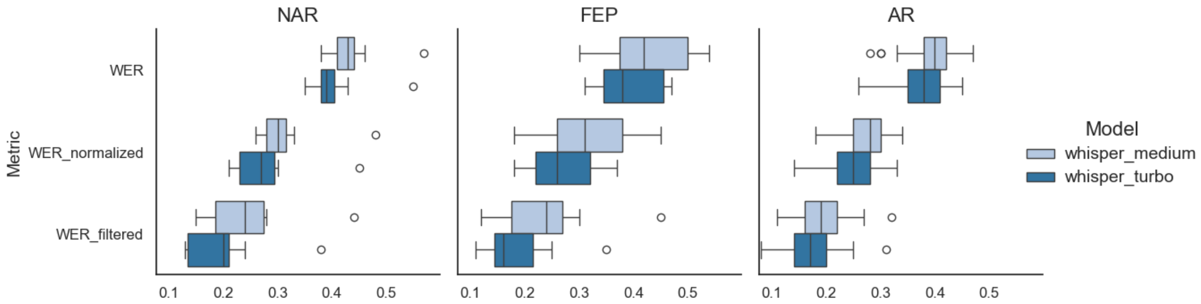

Figure 1: Comparison of WER Metrics Across Patient Groups for Whisper Medium and Whisper Turbo

Processing Stage	Aggregate	NAR	AR	FEP	Kruskal–Wallis p-value
WER	0.38	0.40	0.37	0.39	0.4889
WER_normalized	0.25	0.27	0.24	0.27	0.5981
WER_filtered	**0.17**	**0.19**	**0.17**	**0.19**	0.7969

Table 3: Mean WER Across Patient Groups With Kruskal–Wallis Test Results for Different Preprocessing Stages

malizer to both the reference and hypothesis transcripts in order to standardize formatting. This process involved removing punctuation, converting all text to lowercase, and normalizing common variants in transcription. The resulting metrics, calculated on these normalized texts, are reported as WER_normalized and CER_normalized.

3. **Stop word filtering** : Using the French language model from the SpaCy library, we removed stop words from both versions of the transcripts to emphasize semantically meaningful content.

4. **Single-letter token removal** : We excluded tokens consisting of a single character (e.g., “c” from *c’est*, “d” from *d’accord*), as these are often misrecognized and lack standalone semantic value.

5. **Filled pauses and filler word removal** : After analyzing omissions in the ASR-generated transcripts, we found that filled pauses and filler words were the most frequently omitted across all clinical groups (AR, NAR, and FEP). Based on this, we identified and removed 11 common words – such as *euh*, *bah*, *humm*, *oui*, *ok* and *non* – which consistently appeared missing in the ASR outputs. This enabled the calculation of refined metrics (WER_filtered, CER_filtered) that better capture recognition quality in a clinical context.

For quantitative analysis between ASR-generated and manual transcripts, we used the `jiwer` library to compute standard evaluation metrics, including WER and CER. Beyond providing overall error rates, `jiwer` also facilitated the automatic extraction and categorization of specific error types such as substitutions, deletions, and insertions. This enabled a more fine-grained analysis of both the frequency and the nature of recognition errors across different models.

3 Results and Discussions

3.1 Model-Level WER Comparison

Figure 1 shows the distribution of WER scores for both Whisper models (medium and turbo), aggregated across all patients and preprocessing stages. Overall, the Whisper turbo model consistently outperformed the Whisper medium model across all metrics.

The most notable performance improvement occurred after removing filled pauses (e.g., *euh, bah, humm*), with WER decreasing by approximately 6–8 percentage points for both models. This highlights the strong influence of spontaneous speech markers on surface-level error rates in psychiatric dialogues.

3.2 Patient Group Comparison

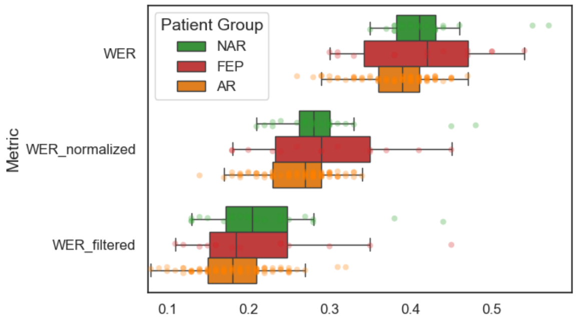

Figure 2: Aggregate Distributions of WER by Patient Type

Figure 2 presents the WER distributions for each patient group: FEP, AR, and NAR. The differences between patient groups were more pronounced than those between models. Specifically, the FEP group exhibited the highest WERs throughout all preprocessing stages, indicating greater transcription difficulty.

3.3 Effect of Preprocessing

Initially, raw transcripts contained non-verbal annotations, special characters, and clinically irrelevant elements such as filled pauses and filler words, which can artificially inflate error rates if not accounted for. By systematically removing these components, we obtained a cleaner reference set that better reflected the meaningful content of the speech.

3.4 Statistical Analysis of Group Differences

To assess whether the differences in WER across clinical groups were statistically significant, we conducted a Kruskal–Wallis H test for each version of the metric: raw WER, WER_normalized, and WER_filtered. The results are summarized in Table ??.

None of the comparisons reached statistical significance, with p-values of 0.4889 (WER), 0.5981 (WER_normalized), and 0.7969 (WER_filtered), respectively. These results indicate that while median WER values differed slightly between the groups, the intra-group variability remained high, preventing clear group-level differentiation.

In particular, the wide spread of WER scores within the FEP group (as seen in Figure 2) suggests that individual differences – such as speech disfluency, cognitive state, and acoustic environment – may play a stronger role than diagnostic group alone in shaping ASR performance.

As outlined above, our initial WER scores fell short of those presented in (Miner et al., 2020). This difference was largely explained by the presence of filler words within the reference transcripts that were ignored by Whisper. While these words can seem superfluous from a linguistic perspective, they can be deemed important from a clinical perspective. The frequency, manner and timing of these filled pauses can give an important insight into the mental state of a given patient. With this in mind, we plan to investigate fine-tuning transcription models to produce outputs that more closely resemble verbatim transcriptions, as discussed in (Akinrintoyo et al., 2025). Additionally, we aim to examine alternative evaluation metrics that may better capture clinically relevant transcription fidelity.

4 Conclusion

This study shows that the *medium* and *turbo* Whisper models perform well when applied in the context of clinical consultations in French, achieving an aggregate WER of 0.25 after applying Whisper normalisation (no statistical significance between patient groups was observed).

These results show that these models can be integrated into the interview process. Although the error rate would suggest that some human supervision and correction is still required, these automated transcriptions can provide a starting point that could significantly reduce the work load when transcribing interviews.

However, particular attention should still be paid to certain types of errors made by the ASR models. For example, we found 8 percentage points of the total error came from omitted filler words used during verbal pauses. These words can be essential to clinicians when diagnosing patients, and so these omissions are of the utmost importance. With this in mind, further work should follow the example of (Akinrintoyo et al., 2025) and investigate the possible fine tuning of ASR models to improve their accuracy vis-à-vis those linguistic details which are relevant to clinicians' diagnoses. This work provides valuable insight into ASR performance in a linguistic context such as French, as, to our knowledge, few studies have evaluated systems like Whisper in non-English clinical settings.

Limitations

As outlined in Section 2.1 above, the dataset used for this experiment has a limited sample size and is not balanced in terms of the CAARMS groups. Table 1 shows that the number of AR patients is twice that of the other two groups combined. However, this distribution is typical of the clinical population, with a majority of patients falling into the at-risk category. The data presented here constitute a subset of our dataset for which both the recordings and transcriptions have been finalized. We are continuing to expand the dataset, with the aim of this experiment being to investigate the use of ASR to accelerate the transcription process.

Additionally, our analysis of transcription errors was conducted at the level of the entire transcription. This meant that no distinction was made between the patients' speech and the therapists' speech. Further work will aim to speaker turns to get a more accurate representation of performance. Finally, silent pauses were not analyzed in this study. In clinical interviews, the duration and timing of silences—alongside filled pauses—can provide meaningful cues about cognitive or emotional states. Incorporating silence duration into future ASR evaluations could offer a more comprehensive understanding of patient behavior and support finer-grained clinical insights.

References

Emmanuel Akinrintoyo, Nadine Abdelhalim, and Nicole Salomons. 2025. WhisperD: Dementia Speech Recognition and Filler Word Detection with Whisper. *arXiv preprint*. ArXiv:2505.21551 [eess].

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1672–1681, New York, NY, USA. Association for Computing Machinery.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Einat Liebenthal, Michaela Ennis, Habiballah Rahimi-Eichi, Eric Lin, Yoonho Chung, and Justin T. Baker. 2023. Linguistic and non-linguistic markers of disorganization in psychotic illness. *Schizophrenia Research*, 259:111–120. Language and Speech Analysis in Schizophrenia and Related Psychoses.

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, Li Fei-Fei, and Nigam H. Shah. 2020. Assessing the accuracy of automatic speech recognition for psychotherapy. *npj Digital Medicine*, 3(1):82.

Valeria A. Pfeifer, Trish D. Chilton, Matthew D. Grilli, and Matthias R. Mehl. 2024. How ready is speech-to-text for psychological language research? Evaluating the validity of AI-generated English transcripts for analyzing free-spoken responses in younger and older adults. *Behavior Research Methods*, 56(7):7621–7631.

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.

Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams, and Dan Jurafsky. 2024. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens.

Thomas Soroski, Thiago da Cunha Vasco, Sally Newton-Mason, Saffrin Granby, Caitlin Lewis, Anuj Harisinghani, Matteo Rizzo, Cristina Conati, Gabriel Murray, Giuseppe Carenini, Thalia S Field, and Hyeju Jang. 2022. Evaluating web-based automatic transcription for alzheimer speech data: Transcript comparison and machine learning analysis. *JMIR Aging*, 5(3):e33460.

Alison R. Yung, Alison R. Yung, Hok Pan Yuen, Patrick D. Mcgorry, Lisa J. Phillips, Daniel Kelly, Margaret Dell'olio, Shona M. Francey, Elizabeth M. Cosgrave, Eoin Killackey, Carrie Stanford, Katherine Godfrey, and Joe Buckby. 2005. Mapping the onset of psychosis: The comprehensive assessment of at-risk mental states. *Australian & New Zealand Journal of Psychiatry*, 39(11-12):964–971. PMID: 16343296.

Kirsten Ziman, Andrew C. Heusser, Paxton C. Fitzpatrick, Campbell E. Field, and Jeremy R. Manning. 2018. Is automatic speech-to-text transcription ready for use in psychological experiments? *Behavior Research Methods*, 50(6):2597–2605.