

Towards a clinical adaptation of the Whisper model for French

Haeul Hwang¹, Deok-Hee Kim-Dufor², Jalal Al-Tamimi³, Motasem Alrahabi¹,
Julien Desclés^{4,5}, Valeria Lucarini^{4,5}, Louis Pichereau², Christophe Lemey^{2,6,7}

¹ObTIC, Sorbonne Université, Paris, France

²URCP, CHU de Brest, Brest, France

³LLF, Université Paris Cité & CNRS, Paris, France

⁴IPNP, Université Paris Cité & INSERM U1266, Paris, France

⁵GHU Paris Psychiatrie et Neurosciences, Paris, France

⁶EA 7479 SPURBO, Université de Bretagne Occidentale, Brest, France

⁷Lab-STICC, IMT Atlantique & CNRS, Brest, France

haeul.hwang@sorbonne-universite.fr, dh.kimdufor@gmail.com,
jalal.al-tamimi@u-pariscite.fr, motasem.alrahabi@sorbonne-universite.fr,
julien.descles@inserm.fr, valeria.lucarini@inserm.fr,
louis.pichereau@etudiant.univ-brest.fr, christophe.lemey@chu-brest.fr

Abstract

Automatic speech recognition (ASR) applied to psychiatric interviews can support clinical assessment and facilitate the analysis of therapist–patient interactions. However, general-purpose ASR systems often struggle with the linguistic and acoustic specificities of psychiatric consultations. In this study, we fine-tune OpenAI’s Whisper model on approximately 48 hours of real-world French psychiatric consultations. The adapted model achieves a substantial improvement in transcription accuracy, reducing the word error rate from 26.9% to 13.7% after text normalization (a 13-point absolute reduction, or 49% relative) in comparison to the baseline model. We further analyze the remaining transcription errors after fine-tuning, highlighting persistent issues inherent to the Whisper model. Our findings demonstrate the effectiveness of domain adaptation for clinical ASR and provide insights into the potential applications and current limitations of deploying such models in mental health services.

1 Introduction

Automatic speech recognition (ASR) has recently achieved substantial progress through the development of large-scale pre-trained speech models. Among these, Whisper has demonstrated strong multilingual transcription capabilities and robustness across diverse acoustic conditions (Radford et al., 2023). These advances have opened new opportunities for applying ASR technologies in

healthcare settings, particularly in mental health services where large amounts of spoken clinical data are routinely generated. However, psychiatric consultations constitute a challenging domain for ASR systems. Unlike structured or read speech, psychiatric interviews involve spontaneous conversational speech characterized by hesitations, pauses, disfluencies, incomplete sentences, and emotionally variable speaking styles. In addition, consultations often contain specialized clinical vocabulary and long-form interactions between patients and clinicians. These characteristics can significantly degrade the performance of general-purpose ASR systems trained on broad-domain audio corpora (Koenecke et al., 2024). Reliable transcription of psychiatric consultations could nevertheless provide important benefits for clinical practice and research. Accurate ASR systems may support clinical documentation, longitudinal monitoring, conversational analysis, and large-scale computational studies of mental health disorders. However, achieving sufficient transcription accuracy for real-world deployment remains a major challenge, particularly in non-English clinical contexts such as French psychiatric interviews. To address this issue, this study investigates the domain adaptation of recent Whisper models using a corpus of real-world French psychiatric consultations. We evaluate two recent model variants, Whisper large-v3 and Whisper large-v3-turbo, and compare two fine-tuning strategies: Full Fine-Tuning (Full

FT), in which all model parameters are updated, and Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning approach designed to reduce computational cost. Beyond transcription accuracy, this work aims to identify the most suitable model and adaptation strategy for practical deployment in mental health services. In addition, we conduct a qualitative analysis of residual transcription errors after fine-tuning, focusing on persistent issues such as hallucinated word generation and the misrecognition of proper nouns. Through this analysis, we aim to better understand the current limitations of ASR systems in psychiatric contexts and evaluate their potential for future clinical applications.

2 Related Work

2.1 From rule-based systems to foundation models for ASR

Automatic speech recognition has moved from early constrained, rule-based systems to neural architectures capable of broad generalization. Early proofs of concept such as Audrey (Davis et al., 1952) could only discriminate isolated digits for a single speaker; from the 1970s–1980s, hidden Markov models combined with n-gram language models enabled large-vocabulary, multi-speaker systems. Deep learning brought some decisive gains, in particular the Transformer (Vaswani et al., 2017), whose multi-head attention captures long-range dependencies, and the Conformer (Gulati et al., 2020), which adds convolutional layers to model local acoustic dynamics such as coarticulation. Whisper (Radford et al., 2022) exemplifies the current foundation-model paradigm: trained on 680,000 hours of multilingual audio, it maps the spectral features of the signal to text through a Transformer encoder–decoder and generates transcriptions autoregressively. Reported Word Error Rates (WER) on read benchmarks are now very low: on LibriSpeech (Panayotov et al., 2015), strong self- and weakly supervised models reach roughly 1.8–2.7% on test-clean (Radford et al., 2022; Baeovski et al., 2020; Hsu et al., 2021), while weaker or LLM-free systems lag behind. These figures, however, transfer only partially to clinical settings (Jelassi et al., 2024), where commercial systems are often poorly documented and open-source models such as Whisper, including the Turbo variant, require domain adaptation (normalization, filtering, fine-tuning) to targeted corpora before they can be used reliably.

2.2 Fine-tuning Whisper for clinical and medical ASR

A growing body of work has fine-tuned Whisper for clinical and medical ASR, providing direct points of comparison for the present study. Le-Duc et al. (2025) released MultiMed, a corpus of real clinical conversations in five languages — including French — and fine-tuned Whisper under various encoder/decoder layer-freezing schemes, building on earlier monolingual medical efforts such as Vi-etMed (Le-Duc, 2024). Liao et al. (2023) proposed prompt-conditioning fine-tuning (Clairaudience), reporting WER reductions of up to 33% on out-of-domain test sets including medical conversation. On a Polish medical corpus, fine-tuning Whisper-medium lowered WER from 24.0% to 15.5%, and to 13.9% once anonymized and synthetic data were added (Czyżewski et al., 2025). Domain match is critical: on accented clinical English, Afonja et al. (2024) found that fine-tuning Whisper-medium on combined general- and clinical-domain speech reduced WER by 54% relative to fine-tuning on general-domain speech alone, confirming that exposure to in-domain data drives the gains. For French specifically, Jelassi et al. (2024) adapted Whisper large-v2 to radiology dictation, reaching a WER of roughly 17%, while Marie et al. (2026) showed that spontaneous French clinical dialogue still exceeds 30% WER, proposing iterative LLM-based post-processing for transcription and speaker attribution. Across these studies, in-domain fine-tuning yields large gains, yet accuracy frequently remains below the threshold required for unsupervised clinical use. This motivates parameter-efficient alternatives — notably Low-Rank Adaptation (LoRA; Hu et al., 2022), applied to Whisper for multilingual and domain adaptation (Song et al., 2024) — and frames the full fine-tuning versus LoRA comparison we pursue.

2.3 Linguistic and acoustic characteristics in psychiatry

In the speech of patients with psychiatric disorders, speech anomalies themselves constitute clinical markers (Lombardo et al., 2025). General-purpose ASR systems often fail in this context precisely because they aim to “clean” the signal, removing the very cues that clinicians seek to analyze (Hwang et al., 2025; Just et al., 2025).

Disfluencies and filler words. Conventional ASR systems frequently omit filler words (“uh”,

“well”, “hmm”), as well as repetitions, self-correction, and hesitations even though their frequency and position are clinical cues in psychiatry — for instance, an increase in filled pauses is associated with anxiety or heightened memory effort in patients with dementia. On psychiatric consultation transcripts, removing these markers artificially lowers WER by 6 to 8 points (Hwang et al., 2025), at the cost of clinically meaningful information. To address this, WhisperD (Akinrintoyo et al., 2025) adapts the tokenizer to include disfluencies explicitly in the vocabulary, improving their detection — notably in Alzheimer’s disease.

Acoustic parameters as clinical biomarkers. Beyond textual content, acoustic properties of the voice (pitch, intensity, speech rate, pausing, and voice quality) correlate with psychiatric symptoms (Lombardo et al., 2025; Jordan et al., 2025); speech emotion recognition (SER) systems exploit them, for example, to predict depression scores (Yang et al., 2016, 2013; Zhao et al., 2022) or to discriminate between a group of patients with schizophrenia and a group of healthy controls (Chakraborty et al., 2018). While such analyses operate directly on the acoustic signal, applying them at scale to clinical interviews would be more effective using a reliable transcription and segmentation layer to isolate patient speech, align measurements with conversational turns, and jointly exploit acoustic and lexical markers. The present study addresses this upstream component; the integration of acoustic features into the pipeline is left to future work.

2.4 Positioning of the present work

Building on this literature, our work adapts recent Whisper models (large-v3 and large-v3-turbo) to approximately 48 hours of real-world French psychiatric consultations, and makes three contributions. First, it provides one of the first fine-tuning studies of Whisper for French psychiatric speech — a domain that, unlike radiology dictation or read medical corpora, combines spontaneous, disfluent and emotionally variable speech with long clinician–patient interactions. Second, it systematically compares full fine-tuning and LoRA across two recent Whisper variants, in order to identify a configuration suitable for deployment in mental-health services. Third, it provides a qualitative analysis of residual errors after fine-tuning — in particular hallucinations and proper-noun misrecognition — which aggregate WER tends to obscure.

3 Methodology

3.1 Dataset

To evaluate the performance of an ASR system in the context of psychiatric consultations, we constructed a corpus composed of audio recordings of clinical interviews between help-seeking young subjects and a psychiatrist. In total, our sample included 58 help-seekers, consisting of 29 female participants, 28 males, and 1 gender-fluid. The 58 participants had one of the three clinical statuses according to the Comprehensive Assessment of At-Risk Mental States (CAARMS): AR (At Risk for Psychosis, $n = 41$), NAR (Not At Risk, $n = 10$), and FEP (First-Episode Psychosis, $n = 7$). All participants were native French speakers. The corpus cannot be shared: the research protocol does not permit redistribution of these identifiable clinical recordings, and pseudonymization does not eliminate re-identification risk in spontaneous speech.

The average recording duration was 49.31 minutes ($\text{std}=14.45$), representing a total of approximately 48 hours of audio data. This substantial duration reflects the complexity of the patient–clinician interactions analyzed. All interviews were fully transcribed by two trained research assistants following detailed annotation guidelines. The transcriptions explicitly include filled pauses (e.g., “uh”), pronunciation and grammatical errors, neologisms, and speech disfluencies in order to preserve the natural characteristics of clinical speech.

3.2 Pre-processing

Adapting an automatic speech recognition (ASR) model to a specialized clinical domain requires careful data preparation. In this context, preprocessing aimed to (1) ensure consistency between audio segments and their corresponding transcriptions, and (2) generate training units suitable for a sequence-to-sequence learning framework. The following preprocessing steps were implemented to ensure the reliability and relevance of the corpus used for fine-tuning. The transcriptions were automatically normalized and aligned with the audio recordings using the Montréal Forced Aligner (MFA; McAuliffe et al., 2017). MFA performed text normalization procedures including lower-case and punctuation removal, as well as word-level forced alignment, producing timestamped segments that enabled accurate ASR model training and evaluation. The audio recordings were then

	NAR	AR	FEP	All groups
Mean age (std)	19.69 (2.02)	19.05 (2.82)	20.30 (4.01)	19.31 (2.82)
M	5	19	4	28
F	4	22	3	29
Gender-fluid	1	0	0	1
Subtotal	10	41	7	58

Table 1: Participant information

segmented into excerpts ranging from 10 to 20 seconds in duration. Segmentation was performed at the word level to avoid cutting words during pronunciation, ensuring that each segment contained complete speech units appropriate for model training. This segmentation strategy also helped maintain contextual coherence while keeping sequence lengths compatible with Whisper’s training framework. After segmentation, the corpus consisted of 6,827 training segments, 853 validation segments, and 854 test segments, corresponding approximately to an 80/10/10 split in terms of segment count. This distribution provided sufficient data for model learning while preserving validation and test sets representative of real-world clinical conditions.

3.3 Models

This study evaluates two recent large-scale ASR models based on the Whisper architecture developed by OpenAI: Whisper large-v3 (1,550M parameters) and Whisper large-v3-turbo (809M parameters). These models were selected because they represent the most recent and advanced versions of the Whisper family, offering strong multilingual speech recognition capabilities and improved robustness across diverse acoustic conditions. The choice of these two models is particularly motivated by the characteristics of the psychiatric consultation corpus used in this study. The recordings are relatively long and contain spontaneous conversational speech with frequent pauses, hesitations, disfluencies, and specialized clinical language. Such characteristics make transcription especially challenging for general-purpose ASR systems. Comparing Whisper large-v3 and large-v3-turbo therefore allows us to evaluate the trade-off between transcription accuracy and computational efficiency in a realistic clinical setting, while assessing the ability of recent Whisper models to handle complex psychiatric speech data.

3.4 Whisper fine-tuning

The ASR models were initialized from pre-trained Whisper checkpoints available on the Hugging Face Model Hub. Model training and evaluation were implemented using the PyTorch framework together with the Hugging Face Transformers library, following a standard sequence-to-sequence fine-tuning pipeline for speech recognition. The input audio data were resampled to 16 kHz and processed using the Whisper feature extractor, while transcripts were tokenized using the corresponding Whisper tokenizer. During preprocessing, each audio sample was converted into log-Mel spectrogram features and paired with its corresponding tokenized transcript. Language and task-specific decoding constraints were enforced by configuring the models for French transcription. Two adaptation strategies were investigated in this study: Full Fine-Tuning (Full FT) and Low-Rank Adaptation (LoRA). Full FT updates all model parameters, enabling complete adaptation of both acoustic and linguistic representations to the psychiatric domain. In contrast, LoRA performs parameter-efficient fine-tuning by introducing trainable low-rank matrices into selected layers while keeping most of the original model parameters frozen, thereby reducing computational and memory requirements. To ensure a fair comparison between the two adaptation strategies and across Whisper variants, the same training hyperparameters were applied to all experiments. Training was performed with a batch size of 8, for 3 epochs, using a learning rate of $5e-6$. For LoRA fine-tuning, the rank parameter was set to 16 and the scaling factor (α) to 32. Training was conducted using the Seq2SeqTrainer API with a custom data collator designed to dynamically pad both acoustic features and tokenized labels. Label padding tokens were masked with -100 to exclude them from the loss computation. To ensure reproducibility, the random seed was fixed at 42, checkpoint resumption was disabled, and the number of

	Whisper L-v3	Whisper L-v3 Turbo
Baseline	26.9	27.5
Full FT	13.7	15.5
LoRA FT	23.2	28.7

Table 2: Normalized WER (%) for Whisper L-v3 and L-v3 Turbo across fine-tuning strategies.

retained checkpoints was limited to three. The best-performing model checkpoint was selected based on validation performance and subsequently used for downstream evaluation.

3.5 Evaluation Metric

The evaluation metric was the Word Error Rate (WER), computed on generated predictions after text normalization using Whisper’s multilingual basic normalizer. WER is a standard evaluation metric in automatic speech recognition that measures the proportion of insertion, deletion, and substitution errors relative to the total number of reference words. Lower WER values indicate higher transcription accuracy. Before evaluation, both reference and predicted transcriptions were normalized using Whisper’s built-in multilingual text normalizer through the `basic_normalize()` function provided by the tokenizer. This normalization step standardizes transcriptions by applying operations such as lowercase, punctuation removal, whitespace normalization, and character normalization across multilingual text, including French. The purpose of this preprocessing is to reduce superficial orthographic variations that do not affect semantic content, thereby enabling a more reliable computation of the Word Error Rate (WER). Special tokens and leading or trailing whitespace were also removed prior to metric computation. To complement this quantitative evaluation, we additionally conducted a qualitative analysis of transcription errors using the three error categories of all speech recognition metrics, *i.e.* insertion, deletion, and substitution.

4 Evaluation and results

4.1 Word Error Rate

Table 2 presents the ASR performance of the different Whisper configurations evaluated on the psychiatric consultation corpus.

Among all evaluated systems, Whisper large-v3 with Full Fine-Tuning (Full FT) achieved the best

overall performance, reaching a normalized WER of 13.72%. Whisper large-v3-turbo with Full FT obtained comparable performance, with a normalized WER of 15.50%. These results indicate that full fine-tuning enables the models to better adapt to the linguistic and acoustic characteristics of psychiatric speech. In contrast, LoRA Fine-Tuning (LoRA FT) produced lower performance for both Whisper variants. Whisper large-v3 with LoRA FT achieved a normalized WER of 23.20%, while Whisper large-v3-turbo with LoRA FT reached 28.75%. Although parameter-efficient fine-tuning reduces computational requirements, it appeared less effective in capturing the complexity of spontaneous clinical conversations. The base models without domain adaptation showed substantially higher error rates, with normalized WERs of 26.91% for Whisper large-v3 and 27.50% for Whisper large-v3-turbo. Compared with the corresponding base models, full fine-tuning achieved a relative reduction in normalized WER of approximately 49% for Whisper large-v3 and 44% for Whisper large-v3-turbo.

4.2 Error analysis (Qualitative analysis)

In the transcriptions generated by the baseline Whisper Turbo model, deletion errors turned out to be the most frequent in all three categories of patients, *i.e.* not-at-risk (NAR), at-risk (AR), and First-Episode Psychosis (FEP). The highest deletion rate was observed in the NAR group, and the most deleted element was filled pauses in the NAR and AR groups with 42.29% and 44.58% respectively, and the FEP group showed a relatively lower rate of 20.39%. The proportion of deletion dropped in the fine-tuned model as shown in Table 3 whereas that of substitution only slightly decreased in the AR and FEP groups. Fine-tuning likely “added” elements as shown in insertion, regardless of the clinical statuses. Hallucinations, *i.e.* texts generated by a model that are nonsensical or unfaithful to the provided source (Ji et al., 2023), or repeated groups of words were found in the generated output. The hallucination rate was higher in the Full fine-tuned model than the Turbo model. Disfluency is a speech element crucial to clinical understanding but is often cleared away by ASR models for the sake of properly understandable texts. We therefore paid particular attention to errors in filled pauses (Table 4). The transcriptions generated by the Turbo Model removed most filled pauses, in particular in the AR group with 93.6%

and slightly fewer in the other groups. No insertion of filled pauses was observed. As for the Full FT model, the deletion problem has improved, but some insertions were found in the three groups.

4.3 Discussion

The results demonstrate that domain-specific adaptation substantially improves the performance of ASR systems on psychiatric consultations. Across both Whisper variants, fine-tuning consistently reduced the Word Error Rate and the rate of filled pause deletion compared with the corresponding base models. These results confirm that general-purpose ASR systems benefit significantly from exposure to specialized clinical speech data. More specifically, our full fine-tuning (Full FT) model has largely corrected the systematic cleanup of disfluency by the baseline models. Some substitutions observed are mainly linked to phonetic similarity or approximation, especially when it comes to medicine names, toponyms, and speech overlaps. Insertions are often function words omitted in speaking but put by the models as if to “guarantee” the grammaticality of the generated text. Both models in some cases replaced silences by stopwords, adverbial phrases, or filled pauses. The hallucinations found in our models could be explained by the contextual particularity of the interviews and the functioning of many ASR models. As for the context, the speech data used in this study are recorded psychiatric consultations with background noise, and patients sometimes have unusually longer silences and emotional overload. Whisper models that entirely rely on maximum likelihood estimation with no confidence threshold generate output text for each audio segment despite very low confidence scores. Without integrated voice activity detection (VAD) they cannot ignore silences and ambient noise. The models then interpret these non-speech data as speech and generate irrelevant texts as shown in a study on aphasic speech with long silences (Koenecke et al., 2024). This could lead to hallucinations and looping (incorrect repetitions of previously recognized text, Barański et al., 2025). These flaws could be expected to be corrected, namely with more samples and a VAD system integration. Overall, these findings highlight the importance of domain-specific adaptation for ASR systems applied to psychiatric consultations and suggest that Full FT remains the most effective strategy for modeling specialized clinical speech. The comparison between Full FT

and LoRA Fine-Tuning (LoRA FT) further reveals important trade-offs between transcription accuracy and computational efficiency. Although LoRA FT offers a parameter-efficient alternative with reduced memory and training costs, its performance remained substantially below that of Full FT for both Whisper variants. This suggests that psychiatric speech may require broader adaptation of the model’s internal acoustic and linguistic representations than that can be achieved through low-rank parameter updates alone. Interestingly, the performance gap between Whisper large-v3 and Whisper large-v3-turbo under Full FT remained relatively small. While Whisper large-v3 achieved the best normalized WER overall, Whisper large-v3-turbo produced comparable results despite being a lighter and more computationally efficient model. This observation suggests that Whisper large-v3-turbo may represent a practical alternative for real-world deployment scenarios where GPU memory, inference speed, or computational cost are important constraints. In clinical environments requiring scalable or near real-time transcription, the turbo variant could therefore offer a favorable balance between efficiency and transcription quality.

4.4 Limitations

Several limitations should nevertheless be acknowledged. First, the corpus size remains relatively modest — approximately 48 hours of recordings from a homogeneous population of French-speaking patients assessed for psychosis risk — compared with the large-scale datasets used during Whisper pre-training; this limited volume may constrain the model’s ability to capture the full variability of clinical speech. Second, the generalizability of the findings to other clinical contexts remains to be evaluated, including different psychiatric services, diverse patient profiles or varying levels of French-language proficiency, and different interaction styles of the therapist. Third, the performance of the fine-tuned model on patients with severe speech disorders — which are prevalent in psychiatric settings — has not been fully characterized and warrants specific attention in future work. Finally, the absence of an integrated VAD component contributes to hallucination errors on silent or noisy segments.

4.5 Conclusion and Future work

This study demonstrates that domain-specific fine-tuning of the Whisper model on real-world French

	NAR		AR		FEP	
	Turbo	FT	Turbo	FT	Turbo	FT
Deletion	73.92	43.06	66.16	45.83	49.87	22.92
Substitution	19.00	34.00	21.84	20.29	39.17	34.13
Insertion	6.90	22.74	9.94	17.84	11.63	29.81
Hallucinations	1.70	5.23	2.42	8.84	3.71	6.09

Table 3: Comparison of error rates between Whisper Turbo and Fine-Tuned Whisper Turbo models

	Manual transcription		Whisper Turbo			Whisper Full FT		
	W	FP (%)	D (%)	S (%)	I	D (%)	S (%)	I
NAR	5209	580 (11.13)	471 (81.21)	1 (0.17)	0	137 (23.62)	12 (2.07)	40
AR	5737	516 (8.99)	483 (93.60)	14 (2.17)	0	208 (40.31)	1 (0.19)	17
FEP	8220	324 (3.94)	241 (74.69)	1 (0.31)	0	25 (7.72)	3 (0.93)	42

Table 4: Errors in filled pauses per patient group and per model. W: total number of words; FP: filled pauses (% of W); D: deletion; S: substitution; I: insertion. Percentages for D, S, and I are relative to the number of filled pauses.

psychiatric consultations substantially improves transcription accuracy, reducing the normalized WER from approximately 27% to 13.7% after full fine-tuning of the large-v3 variant — a relative reduction of nearly 49%. These results, based on 48 hours of clinical recordings, highlight the effectiveness of domain adaptation for psychiatric ASR even with limited annotated data, a common constraint in clinical research settings. This fine-tuned ASR model, characterized by a substantially reduced word error rate, represents a meaningful advance for computational linguistics research applied to psychiatry. Beyond the significant reduction in manual transcription resources, it opens the possibility of collecting and analyzing far larger volumes of clinical speech data in a shorter timeframe. This increased capacity could directly support clinical assessment by providing clinicians with richer and more actionable information to guide diagnostic and therapeutic decision-making. Future work could investigate larger and more diverse multilingual corpora, the integration of VAD to reduce hallucinations on silent segments, speaker diarization methods for improved attribution in multi-speaker conversations, and hybrid adaptation strategies combining parameter-efficient fine-tuning with selective full-model optimization.

Acknowledgments

This work benefited from the Fondation de l’Avenir, the Fondation de France, the PsyCARE project

ANR-18-RHUS-0014, which is funded by the "Investissements d’Avenir" program, and the "Initiative d’excellence" program IdEx ANR-18-IDEX-0001, implemented by Université Paris Cité and funded by the "France 2030" program.

References

- Tejumade Afonja, Tobi Olatunji, Sewade Ogun, Naome A. Etori, Abraham Owodunni, and Moshood Yekini. 2024. [Performant asr models for medical entities in accented speech](#). In *Proc. Interspeech 2024*, pages 2315–2319.
- Emmanuel Akinrintoyo, Nadine Abdelhalim, and Nicole Salomons. 2025. WhisperD: Dementia speech recognition and filler word detection with whisper. In *Proc. Interspeech 2025*. ArXiv:2505.21551.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12449–12460.
- Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. 2025. [Investigation of Whisper ASR Hallucinations Induced by Non-Speech Audio](#). In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Debsubhra Chakraborty, Zixu Yang, Yasir Tahir, Tomasz Maszczyk, Justin Dauwels, Nadia Thalmann, Jianmin Zheng, Yogeswary Maniam, Nur Amirah, Bhing Leet Tan, and Jimmy Lee. 2018. [Prediction of](#)

- Negative Symptoms of Schizophrenia from Emotion Related Low-Level Speech Signals. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 1215. IEEE.
- Andrzej Czyżewski, Sebastian Cygert, Karolina Marciniuk, Maciej Szczodrak, Arkadiusz Harasimiuk, Piotr Ody, Marina Galanina, Piotr Szczuko, Bożena Kostek, Beata Graff, Dariusz Szplit, Mariusz Budzisz, and Krzysztof Narkiewicz. 2025. [A comprehensive polish medical speech dataset for enhancing automatic medical dictation](#). *Scientific Data*, 12:1436.
- K. H. Davis, R. Biddulph, and S. Balashek. 1952. [Automatic recognition of spoken digits](#). *The Journal of the Acoustical Society of America*, 24(6):637–642.
- Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. [Conformer: Convolution-augmented transformer for speech recognition](#). In *Proc. Interspeech 2020*, pages 5036–5040.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. [Hubert: Self-supervised speech representation learning by masked prediction of hidden units](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations (ICLR)*. ArXiv:2106.09685.
- Haeul Hwang, Eric Jordan, Deok-Hee Kim-Dufor, Christophe Lemey, and Motasem Alrahabi. 2025. [Evaluating ASR in a clinical context: What whisper misses](#). In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 374–378, Odense, Denmark. Association for Computational Linguistics.
- Mariem Jelassi, Oumaima Jemai, and Jacques Demongeot. 2024. [Revolutionizing radiological analysis: The future of french language automatic speech recognition in healthcare](#). *Diagnostics*, 14(9):895.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of Hallucination in Natural Language Generation](#). *ACM Computing Surveys*, 55(12):248:1–248:38.
- Eric Jordan, Raphaël Terrisse, Valeria Lucarini, Motasem Alrahabi, Marie-Odile Krebs, Julien Desclés, and Christophe Lemey. 2025. [Speech Emotion Recognition in Mental Health: Systematic Review of Voice-Based Applications](#). *JMIR Mental Health*, 12(1):e74260.
- Sandra Anna Just, Brita Elvevåg, Shrankhla Pandey, Ivan Nenchev, Anna-Lena Bröcker, Christiane Montag, and Sarah E. Morgan. 2025. [Moving beyond word error rate to evaluate automatic speech recognition in clinical samples: Lessons from research into schizophrenia-spectrum disorders](#). *Psychiatry Research*, 352:116690.
- 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)*, pages 1672–1681.
- Khai Le-Duc. 2024. [VietMed: A dataset and benchmark for automatic speech recognition of vietnamese in the medical domain](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17365–17370.
- Khai Le-Duc, Phuc Phan, Tan-Hanh Pham, Bach Phan Tat, Minh-Huong Ngo, Thanh Nguyen-Tang, and Truong-Son Hy. 2025. [MultiMed: Multilingual medical speech recognition via attention encoder decoder](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1113–1150, Vienna, Austria. Association for Computational Linguistics.
- Feng-Ting Liao, Yung-Chieh Chan, Yi-Chang Chen, Chan-Jan Hsu, and Da-shan Shiu. 2023. [Zero-shot domain-sensitive speech recognition with prompt-conditioning fine-tuning](#). In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.
- Clara Lombardo, Giulia Esposito, Silvia Carbone, Salvatore Serrano, and Carmela Mento. 2025. [Speech analysis and speech emotion recognition in mental disease: A scoping review](#). *Frontiers in Psychology*, 16.
- Ambre Marie, Thomas Bertin, Guillaume Dardenne, and Gwenolé Quéllec. 2026. [Iterative llm-based improvement for french clinical interview transcription and speaker diarization](#).
- Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. [Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi](#). In *Proc. Interspeech 2017*, pages 498–502.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. [Librispeech: An asr corpus based on public domain audio books](#). In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *arXiv preprint arXiv:2212.04356*.

- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pages 28492–28518. PMLR.
- Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. [LoRA-Whisper: Parameter-efficient and extensible multi-lingual asr](#). In *Proc. Interspeech 2024*, pages 3934–3938.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008.
- Tsung-Hsien Yang, Chung-Hsien Wu, Kun-Yi Huang, and Ming-Hsiang Su. 2016. [Detection of mood disorder using speech emotion profiles and LSTM](#). In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*.
- Ying Yang, Catherine Fairbairn, and Jeffrey F. Cohn. 2013. [Detecting Depression Severity from Vocal Prosody](#). *IEEE Transactions on affective computing*, 4(2):142–150.
- Qing Zhao, Hong-Zhen Fan, Yan-Li Li, Lei Liu, Ya-Xue Wu, Yan-Li Zhao, Zhan-Xiao Tian, Zhi-Ren Wang, Yun-Long Tan, and Shu-Ping Tan. 2022. [Vocal Acoustic Features as Potential Biomarkers for Identifying/Diagnosing Depression: A Cross-Sectional Study](#). *Frontiers in Psychiatry*, 13:815678.