The Prosody of Clinical Speech: Annotation and Computational Analysis of Psychiatric Interviews

Claire Lawand[1], Haeeul Hwang[1], Jalal Al-Tamimi[2], Motasem Alrahabi[1,*], Mahshid Hozhabr[2], Deok-Hee Kim-Dufor[3], Christophe Lemey[3], Julien Desclés[4] and Valeria Lucarini[4]

[1]*ObTIC, Sorbonne Université, Paris 75006, France*
[2]*Université Paris Cité, CNRS, Laboratoire de linguistique formelle (LLF), Paris 75013, France*
[3]*Centre Hospitalier Universitaire (CHU) de Brest, Brest, France*
[4]*GHU Paris Psychiatrie et Neurosciences, Hôpital Sainte-Anne, PEPIT, F-75014 Paris, France*

Abstract

We present an exploratory analysis of silence and filled pause annotation in 39 semi-structured French-language clinical interviews between psychiatrists and patients assessed for psychosis risk, drawn from the APAISE+ corpus [1]. Participants were classified into three groups: at clinical high risk (AR), not at risk (NAR), and first-episode psychosis (FEP). We developed a manual annotation protocol using Praat TextGrid files, parsed 58,243 annotated intervals, and applied statistical methods to study silence distributions, stationarity, within-interview dynamics, and clinical group differences. Silence accounts for 15.5% of patient speaking time (median 1.01 s). Silence duration differs significantly across three clinical groups (AR, NAR, FEP; Kruskal-Wallis $H = 289.9$, $p < 0.001$), while filled pause rate does not ($H = 0.59$, $p = 0.75$). First-episode psychosis patients produce proportionally longer silences (>2 s) than at-risk patients, though the not-at-risk group also exhibits extended silence durations, complicating a straightforward clinical interpretation. These findings are invisible to automatic speech recognition, which suppresses over 90% of filled pauses and cannot represent silence, underscoring the irreplaceable role of manual annotation in clinical NLP.

Keywords

psychosis, speech analysis, silence, filled pauses, clinical NLP

1. Introduction

Much of clinical communication happens not in the semantic content of language but in the phenomena that surround it. In spoken psychiatric interviews, hesitations, self-corrections, and prolonged silences carry diagnostic potential as paralinguistic markers of speech planning difficulty and reduced fluency [3].

The APAISE+ project (*Apprentissage Profond pour l'Analyse Informatisée de la Subjectivité et des Émotions dans les troubles psychotiques émergents*) aims to develop computational tools for detecting early linguistic markers of schizophrenia and bipolar disorder through clinical speech analysis. Corpus annotation and statistical modelling are foundational elements of this pipeline. The annotation work operationalises theoretical constructs (what counts as a silence, a filled pause, a disfluency) and embeds linguistic and clinical knowledge into the corpus in a form accessible to computational methods.

Despite their clinical relevance, silence and filled pauses are systematically lost in automatic transcription via Whisper. A recent evaluation on this corpus showed that Whisper suppresses over 90% of filled pauses in at-risk patients and cannot represent silence at all [1]. A fine-tuned version substantially

NL4AI 2026: Ninth Workshop on Natural Language for Artificial Intelligence, October 6-9th, 2026, Perugia, Italy [2]
*Corresponding author.
clairelawand@gmail.com (C. Lawand); haeeul98@gmail.com (H. Hwang); jalal.al-tamimi@u-pariscite.fr (J. Al-Tamimi); motasem.alrahabi@sorbonne-universite.fr (M. Alrahabi); mahshidhozhabr98@gmail.com (M. Hozhabr); dh.kimdufor@gmail.com (D. Kim-Dufor); christophe.lemey@chu-brest.fr (C. Lemey); julien.descles@inserm.fr (J. Desclés); valeria.lucarini@inserm.fr (V. Lucarini)
http://www.llf.cnrs.fr/en/Gens/Al-Tamimi (J. Al-Tamimi); https://obtic.sorbonne-universite.fr/alrahabi/ (M. Alrahabi)
0000-0002-0926-6660 (J. Al-Tamimi); 0000-0001-5478-4283 (M. Alrahabi); 0009-0004-4233-7522 (M. Hozhabr); 0000-0002-0411-0093 (D. Kim-Dufor); 0000-0002-7308-7958 (C. Lemey); 0009-0007-3984-4760 (J. Desclés); 0000-0002-3553-8818 (V. Lucarini)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-05

CEUR-WS.org/Vol-4315/paper18.pdf

reduced the word error rate (WER) but did not fully recover filled pauses [4]. Manual annotation therefore remains the only reliable method for capturing these dimensions.

The contributions of this paper are threefold: (1) a reusable manual annotation protocol for silence, filled pauses, and disfluencies in French clinical speech; (2) a documented methodology for parsing Praat TextGrid files and computing prosodic statistics from clinical interview corpora; and (3) an exploratory statistical analysis of silence and filled pause distributions across three psychosis-risk groups, including group comparisons and within-interview temporal dynamics.

2. Related Work

Research on speech in psychotic and at-risk populations has consistently identified disruptions in pause behaviours as diagnostic signals. Andreasen [5] identified prolonged pauses, poverty of speech, and derailment as core diagnostic markers, providing the clinical framework within which our annotation categories operate.

Silence and filled pauses in clinical speech. Silence and disfluency are increasingly perceived as non-invasive markers of psychiatric and neurological disturbance. Tahir et al. [6] showed that non-verbal cues, such as pause frequency and duration, serve as objective measures of negative symptom severity in schizophrenia, motivating the use of temporal annotation as a complement to clinical assessment. Rapcan et al. [7] demonstrated that acoustic and temporal speech features constitute a potential biomarker for schizophrenia, while Parola et al. [8] conducted a systematic review and Bayesian meta-analysis confirming that voice and temporal patterns reliably distinguish schizophrenia from healthy controls across studies.

Beyond psychosis, pause and disfluency analysis has been applied to other neurological research. Dovetto et al. [9] analysed silences and disfluencies in patients with Alzheimer's disease, and Pistono et al. [10] examined disfluency patterns differentiating Alzheimer's disease from frontotemporal lobar degeneration, demonstrating the clinical significance of temporal disruptions in speech. In a crosslinguistic perspective, Betz et al. [11] examines the interplay between filled pauses and silences, motivating the joint annotation approach adopted here.

Reference norms and threshold analysis. For French spontaneous speech, Campione and Véronis [12] provided the foundational threshold analysis, establishing a low threshold of 200 ms and a high threshold of 2,000 ms, and showing how the mean pause duration changes as these thresholds are applied — a framework we replicate for our clinical corpus.

Acoustic feature extraction and explainability. The study of speech in medical research has been increasingly adopting feature-based approaches using toolkits such as openSMILE [13, 14]. Chenain et al. [15] applied an explainable AI approach to acoustic characterisation of emotional expression in Huntington's disease, demonstrating that feature-level extraction methods can serve as between-group differentiation.

Automatic speech recognition (ASR) in clinical settings. Hwang et al. [1] evaluate Whisper on the APAISE+ corpus and find that filled pauses account for 8 percentage points of total WER and are suppressed in over 90% of cases. Hwang et al. [4] show that fine-tuning on the 49-hour corpus reduces WER substantially, but filled pause suppression remains a challenge. The aim of this study is to present the manual transcription protocol developed for this study and used to quantify the silences and filled-pauses in this population. We present results specific to our population.

3. Corpus and Annotation Protocol

3.1. The APAISE+ Corpus

The full APAISE+ corpus [1] comprises 57 recordings of semi-structured psychiatric interviews conducted in French at the University Hospital of Brest (CHU de Brest). All participants were native speakers of French. The patients were assessed under the Comprehensive Assessment of At-Risk Mental

States (CAARMS) ([16]) and classified into three groups: AR (At Risk for psychosis, 40 patients), NAR (Not At Risk for psychosis, 10 patients), and FEP (First-Episode Psychosis, 7 patients). The total duration is approximately 49 hours; the mean interview duration is 49 minutes.

Of the 57 recordings in the full corpus, 39 (23 AR, 9 NAR, and 7 FEP) have been manually annotated to date; the remaining 18 are currently undergoing annotation following the same protocol. The annotated subset was not selected based on clinical criteria. All data are subject to an ethics agreement with the CHU de Brest; the files were anonymised before analysis.

3.2. Processing Pipeline

The pipeline proceeds through seven stages: audio conversion to WAV; noise reduction via Demucs [17]; automatic transcription via Whisper [18]; manual verbatim transcription produced by two trained research assistants and reviewed by an experienced linguist; forced alignment at word and phoneme level using the Montreal Forced Aligner (MFA) [19]; automatic speaker diarisation using pyannote [20]; and manual annotation in Praat [21], producing TextGrid files.

Two TextGrid format variants were encountered: the standard `ooTextFile` format and the chronological format, not supported by the `tgt` Python library; custom parsers were implemented for both. Tier-name inconsistencies and label errors across files were normalised programmatically. Files are encoded in UTF-16 with byte-order mark, requiring byte-level decoding.

3.3. Annotation Protocol

Each TextGrid contains six tiers: *words* and *phones* (forced alignment, excluded from analysis), *Patient* and *Psychiatre* (diarised speech intervals), *Silence* (SIL), and *Disfluences* (PR and DF).

Silent pauses (SIL). A SIL interval marks any gap of at least 80 ms during patient speech. A brief silence prior to a burst of plosive sound is not marked as a SIL, since it is a phonetic phenomenon, that is, the closing phase of the articulation of plosive consonants. We retained any discontinuities and long pauses within patient speech for subsequent analyses and excluded all response latencies.

Filled pauses (PR). PR (*pause remplie*) covers a fixed lexical inventory of French hesitation markers established at a calibration meeting: *euh*, *ben*, *bah*, *humm*, *enfin*, *fin*. Sighs are annotated as PR.

Disfluencies (DF). DF covers repetitions, self-corrections, false starts, and vowel lengthening. Disfluency intervals are annotated for completeness and future use and is not analysed inferentially in the present study. The SIL tier provides the primary material for the duration and temporal analyses in Sections 4 and 5; the PR tier provides the filled pause rate analysed in Section 5.4.

4. Corpus Analysis

4.1. Corpus Composition

Parsing 39 TextGrid files yielded 58,243 intervals. Table 1 summarises the distribution. Silence accounts for 15.5% of patient speaking time, or 8.85% of total conversation time (median 1.01 s, mean 1.17 s, maximum 17.9 s).

Table 1
Corpus composition (39 interviews, 58,243 intervals). Proportions are of patient speaking time (55,900 s). Fluent speech is the residual patient time not covered by any SIL, PR, or DF annotation.

Category	n	% of patient time
Silence (SIL)	7,390	15.5
Filled pause (PR)	11,426	8.9
Disfluency (DF)	5,126	5.2
Fluent speech	—	70.4

The 11,426 PR intervals represent a non-negligible amount: Whisper suppresses over 90% of filled pauses in this corpus [1], meaning the vast majority of these markers are invisible in raw ASR output.

4.2. Silence Duration and Threshold Analysis

The silence duration distribution is strongly right-skewed: most silences fall below 2 s but a long tail extends to nearly 18 s. Following Campione and Véronis [12], we apply a low threshold of 200 ms and a high threshold of 2,000 ms to classify silences into three categories (Figure 1).

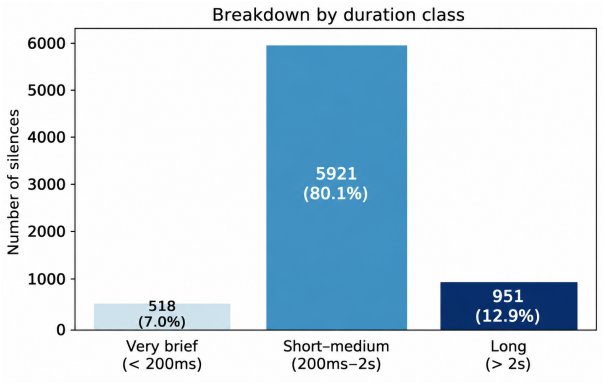


Figure 1: Silence duration classes.

Per the annotation protocol, all pauses $\geq$ 80ms were marked as SIL, yielding 7,390 annotated silences. For comparative analyses, we apply the low threshold of 200 ms and the 2,000 ms upper threshold from Campione and Véronis [12]. The latter corresponds to their analysis of very long pauses, which isolates a clinically salient minority of extended silences, rather than the 1,000 ms boundary of their three-category scheme. Of the 7,390 silences, 518 silences (7.0%) fall below 200ms and are set aside for threshold comparisons. The majority, 5,925 (80.1%), fall in the 200 ms–2 s range. A clinically salient 953 silences (12.9%) exceed 2 s, with a median duration of 2.59 s.

The distribution within the 200 ms–2 s range is near-uniform rather than peaked, contrasting with that of spontaneous speech by healthy French speakers which peaks around 600–800 ms [12]. This flatness suggests a broader planning-time profile consistent with the communicative demands of a psychiatric evaluation.

4.3. Internal Heterogeneity

The silence proportions range from 1% to 20% across the interviews. Two structural outliers are notable. Interview 040 contains 76% patient speech, the most talkative patient in the corpus. Interviews 020, 021, and 022 present anomalously high silence interval counts (778–1,479 vs. a corpus median of $\approx$80). Interviews 020 and 021 belong to the AR group; interview 022 belongs to the FEP group. To verify that these outliers do not drive the group comparison, we re-ran the Kruskal-Wallis test on silence duration excluding these three interviews ($n_{\text{AR}} = 2{,}514$, $n_{\text{NAR}} = 1{,}091$, $n_{\text{FEP}} = 432$ silence intervals). The result remains highly significant ($H = 76.1$, $p < 0.001$), with all three pairwise Dunn comparisons surviving Bonferroni correction, confirming that the group differences reported in Section 5 are not an artefact of these high-count interviews.

4.4. Temporal Dynamics

4.4.1. Stationarity of Silence Sequences

We applied the Augmented Dickey-Fuller (ADF) test [22] to the chronological sequences of silence durations per interview and divided each interview into 10 equal time window. Of 39 testable interviews,

36 are stationary ($p \leq 0.05$), indicating that silence patterns reflect stable individual speech habits rather than dynamic within-interview processes.

Three interviews are non-stationary: interview 011 ($p = 0.97$, AR), interview 013 ($p = 0.38$, AR), and interview 037 ($p = 0.31$, FEP). Figure 2 shows their detailed profiles. Interview 011 shows a monotonic upward drift in silence duration; interview 013 shows a bimodal pattern with peaks in windows 3 and 9; interview 037 (FEP) shows a front-loaded pattern with silence concentrated in windows 1–3 followed by rapid adaptation, representing an inverse of the corpus-wide trend.

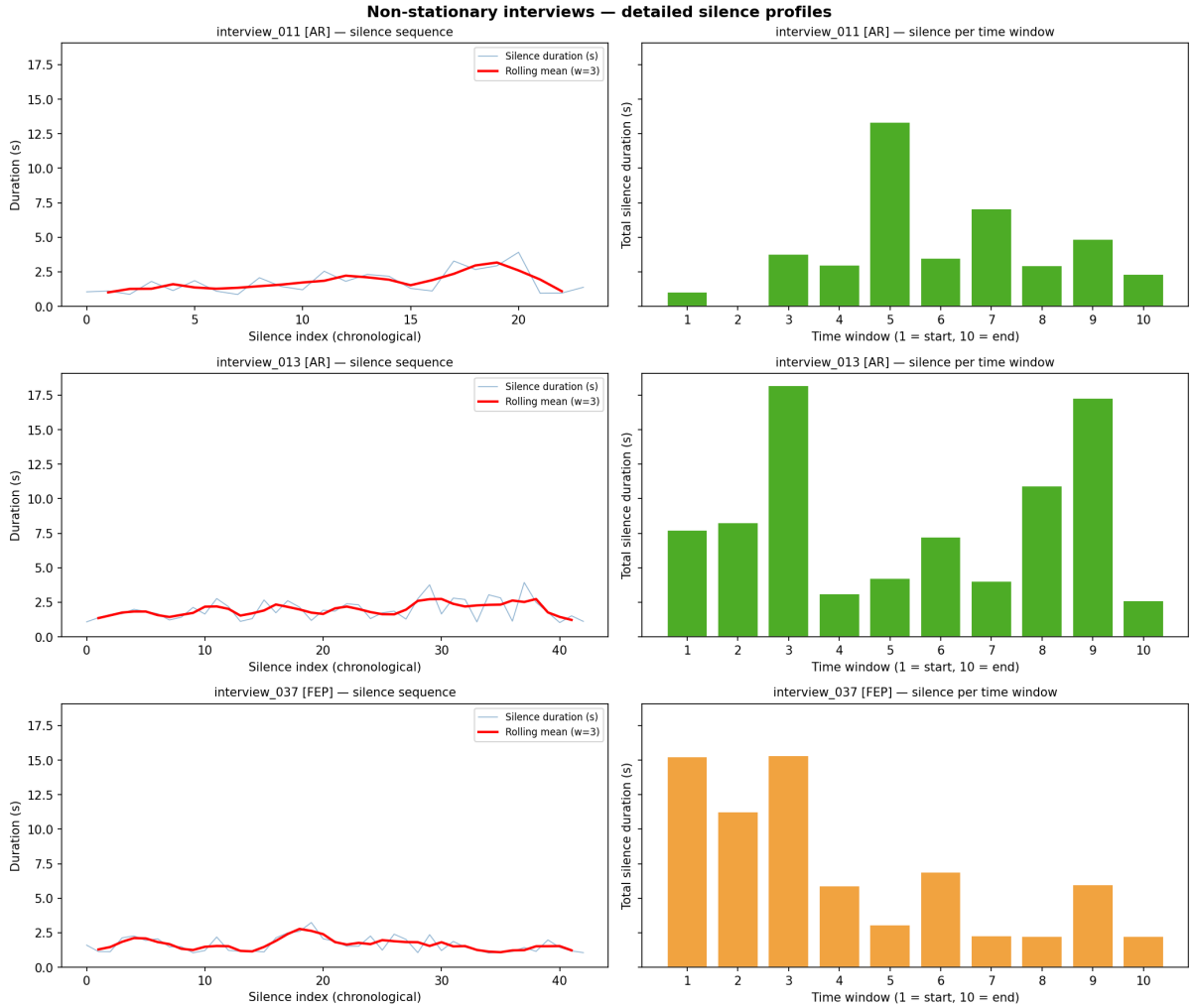


Figure 2: Non-stationary interviews. Left: chronological silence sequences with rolling mean ($w = 3$). Right: total silence per time window.

4.4.2. Corpus-Level Evolution

The mean total duration of silence and count were averaged across all 39 interviews, which were divided into time windows (Figure 3). Both measures decrease consistently from window 3 onward: total silence drops from ≈25 s (windows 2–3) to ≈16 s (window 9); silence count from ≈21 to 15. A slight uptick at window 10 may reflect a conversational closing phase and/or a moment of providing information on upcoming appointments and further assessments.

The decreasing trend is confirmed by Spearman rank correlations between window index and mean total silence duration ($\rho = -0.806$, $p = 0.005$) and mean silence count ($\rho = -0.891$, $p < 0.001$). We interpret this as a progressive adaptation effect: patients become more conversationally fluent as a rapport is being established. This is consistent with the dominant stationarity finding: the decrease is a corpus-wide tendency, while most individual silence profiles are stable within their own interview.

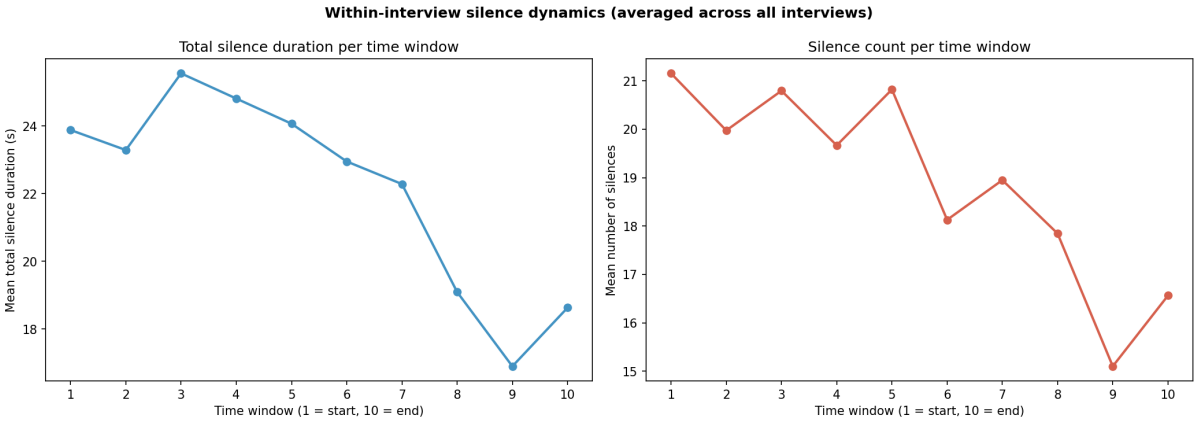


Figure 3: Within-interview silence dynamics averaged over 39 interviews. Left: total silence duration. Right: total silence count.

5. Clinical Group Comparison

5.1. Group Composition

The corpus composition shows between-group similarities: patient speech accounts for 59% (AR), 55% (NAR), and 56% (FEP) of annotated time. Silence as a proportion of annotated time is higher in AR (9.7%) than in FEP (8.7%) and NAR (7.2%), counter-intuitive if silence is assumed to increase with clinical severity.

5.2. Silence Duration

Table 2 summarises key metrics by group.

Table 2
Key metrics by clinical group. n: number of interviews. n_{SIL}: total silence intervals across the group. Med.: median silence duration per interval (seconds). Mean: mean silence duration per interval (seconds). PR/min: median filled pause rate per minute of patient speech time, computed per interview then aggregated.

	n	n_{SIL}	Med.	Mean	PR/min
AR	23	5,089	0.91 s	1.10 s	11.3
NAR	9	1,091	1.24 s	1.45 s	12.4
FEP	7	1,210	1.07 s	1.25 s	12.9

Silence duration differs significantly across groups: Kruskal-Wallis $H = 289.9$, $p < 0.001$. Post-hoc Dunn tests with Bonferroni correction ($\alpha = 0.0167$) confirm significance for all three pairs: AR vs. NAR $p < 0.001$; AR vs. FEP $p = 0.001$; NAR vs. FEP $p < 0.001$.

The most striking finding is that NAR patients present the longest median silence duration (1.24 s), exceeding both FEP (1.07 s) and AR (0.91 s). This is counter-intuitive: the group classified as not-at-risk produces the most extended silences, hereafter referred as NAR paradox. It should be noted that NAR designates patients assessed as not at risk for psychotic transition specifically: this group may nonetheless present other clinical difficulties, such as depressive episodes or suicidal ideation, that independently affect speech planning and silence duration. We discuss this in Section 6.

5.3. Threshold Analysis by Group

Figures 4 and 5 show the threshold analysis by clinical group.

FEP and NAR produce similar proportions of long silences (>2 s): 16.9% and 16.1% respectively, vs. 11.2% in AR. The median duration of long silences is nearly identical across groups (≈2.6 s): it

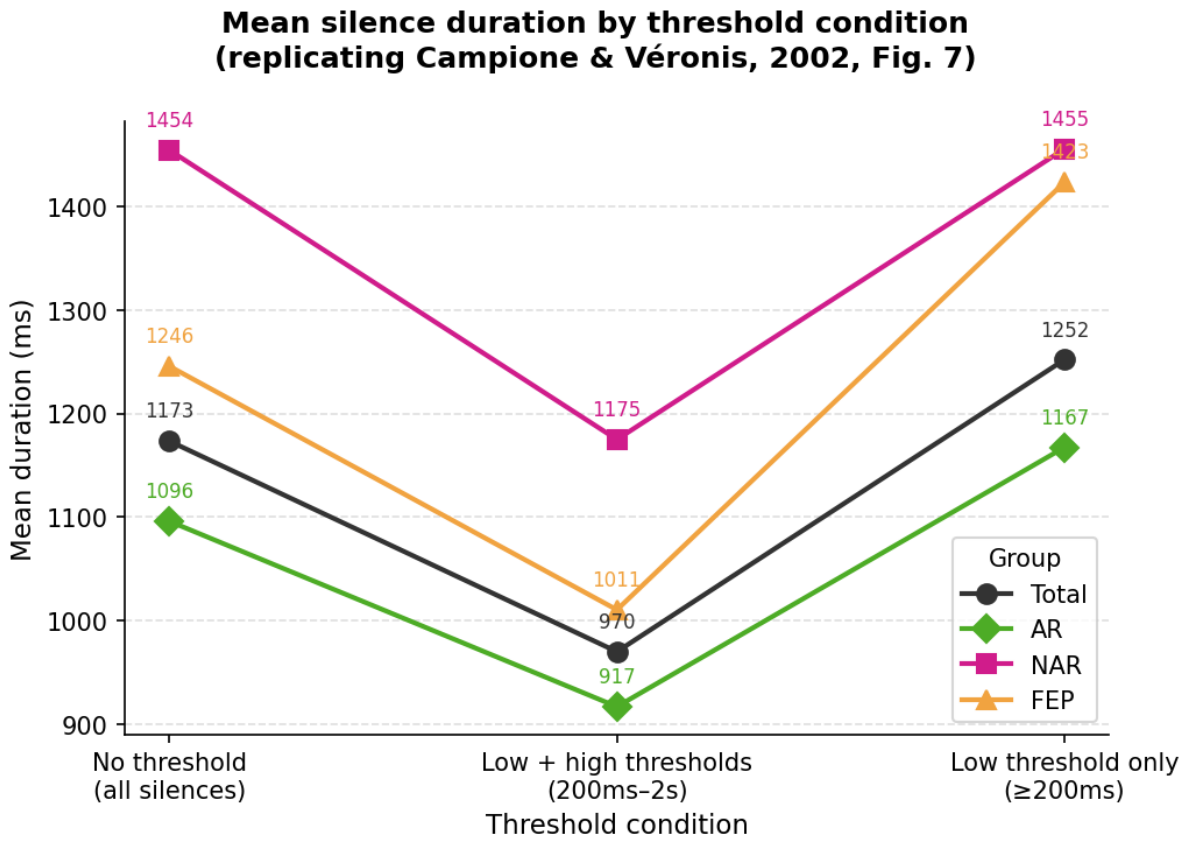


Figure 4: Mean silence duration per threshold condition, replicating the format of Campione and Véronis [12] Fig. 7. NAR consistently shows the highest mean. The V-shape reflects the threshold structure: the middle condition retains only short silences; the right condition isolates long silences only.

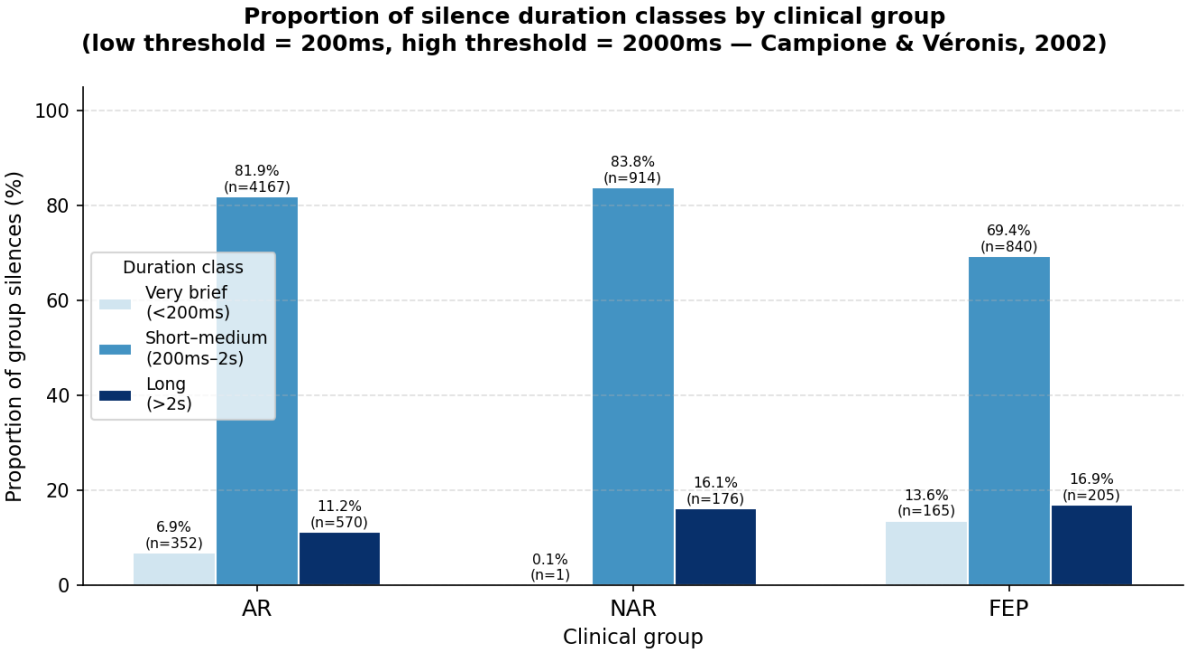


Figure 5: Proportion of three silence duration categories by clinical group.

is the *frequency* of long silences, not their duration, that distinguishes groups. The distribution of normal-range silences (200 ms–2 s) also differs: AR peaks early (≈0.5 s), NAR peaks later (≈1.2 s), and FEP shows a flatter profile (see Appendix Figure 7).

5.4. Filled Pause Rate

The filled pause rate does not differ significantly between groups: $H = 0.59$, $p = 0.75$. The group medians are AR 11.3/min, NAR 12.4/min, and FEP 12.9/min. The contrast between the significant difference in the silence duration and the non-significant difference in the PR rate suggests that the two phenomena likely reflect different underlying processes: filled pauses and silence may reflect partially distinct processes: while filled pause rate does not differ significantly across groups, silence duration does, though the specific mechanisms underlying this dissociation remain to be established.

5.5. Controlling for Interactional Confounds

To test whether the NAR paradox (Section 5) reflects interactional structure rather than clinical status, we computed, per interview, patient speech rate (words/min, MFA-aligned), turn count, mean turn duration, and psychiatrist-to-patient speech-time ratio, using the same 39 annotated interviews analysed throughout this paper. Spearman correlations between these variables and median silence duration per interview were computed ($n = 37$; 2 interviews lacked a usable silence estimate).

Table 3
Spearman correlations with median silence duration per interview ($n = 37$).

Predictor	ρ	p
Speech rate (wpm)	-0.206	0.221
Turn count	-0.032	0.851
Mean patient turn duration	-0.149	0.380
Mean psychiatrist turn duration	-0.024	0.890
Psychiatrist/patient time ratio	$+0.206$	0.222

None of these correlations reached significance. An ANCOVA comparing clinical group alone ($R^2 = 0.024$) to group plus these covariates ($R^2 = 0.318$, $n = 37$) showed no significant improvement in fit (nested F-test, $F = 2.50$, $p = 0.053$). Within the full model, speech rate showed a marginal negative association with median silence duration ($\beta = -0.005$, $p = 0.046$), suggesting that faster speakers tend to produce shorter silences, which is consistent with the expected inverse relationship between speech rate and pause duration, though the overall model did not reach significance. These results neither support nor rule out interactional structure as a confound of the NAR paradox; the reduced, partly incomplete sample limits statistical power, and this question should be revisited with the full corpus. The absence of a significant confound may itself partly reflect the small, imbalanced sample rather than a true absence of effect.

6. Discussion

Scale of ASR blindness The 11,426 manually annotated PR intervals, against Whisper's >90% suppression rate, represent a body of clinically meaningful information invisible to raw ASR pipelines. This finding directly motivates the hybrid annotation approach of APAISE+ and the continued development of clinically adapted ASR systems.

Stable individual habits The dominant finding from an ADF test is that individual patients' silence patterns are consistent in their 50-minute interviews. This supports a trait-level view of silence behaviour. The three non-stationary interviews would be clinically more meaningful precisely because they deviate from this norm and warrant a closer qualitative examination alongside the interview transcripts to verify whether silence peaks co-occur with specific topics or affective moments.

Adaptation, not fatigue The corpus-wide decrease in silence from window 3 onward is consistent with a progressive adaptation interpretation. This is clinically relevant: the first third of a psychiatric interview may be the richest period to capture silence-based markers before a rapport and conversational routinisation attenuate them.

The NAR paradox The NAR patients produce the longest median silences despite being classified as not-at-risk. Several interpretations are possible. First, the NAR group is small ($n = 9$) and may not be representative. Second, the CAARMS classification is based on symptom thresholds that do not include certain speech characteristics: subclinical speech planning differences visible in silence behaviour may not cross the clinical thresholds. Third, NAR patients in a psychiatric interview may exhibit conversational hesitancy for contextual reasons unrelated to psychotic risk. Notably, prior work on voicing/silence duration in monologue speech found that speech-rate effects were strongest among highly motivated speakers actively working to be understood in a demanding communicative situation [23] — a condition that may not hold uniformly across psychiatric interviews, which could further explain the weak confound effects observed here. This finding warrants a further investigation with larger samples and integration with the Positive And Negative Syndrome Scale (PANSS) scores (medical scale used for measuring positive and negative symptoms and their severity in patients with schizophrenia), which were not available for this analysis.

FEP and long silences The FEP patients produce disproportionately longer silences (>2 s) compared with the AR patients, consistent with the severe disruption of speech production associated with active psychosis [24]. The fact that the duration of these long silences is similar between the 3 groups — it is their frequency that sorts out the FEP — suggests a threshold effect: once planning failure reaches a certain level, the silence is prolonged to approximately 2.6 s, regardless of the group, but FEP patients cross this threshold more often.

Limitations The annotation subset (39 of 57 interviews) is not fully balanced across the clinical groups, particularly for FEP ($n = 7$). Clinical metadata (PANSS, CAARMS scores) are not yet integrated. The Inter-Annotator Agreement (IAA) has not yet reached the target of $\kappa = 0.70$ for SIL; post-calibration analyses may revise the interval counts. The analysis is exploratory; no predictive modelling has been attempted. The interactional-confound check (Section 5.5) was likewise limited by the sample size ($n = 37$) and missing data.

7. Conclusion

We have presented a manual annotation protocol, computational pipeline, and exploratory results for silence and filled pause behaviour in 39 clinical interviews with patients assessed for psychosis risk. Our main contributions are: (1) a documented annotation protocol for silences, filled pauses, and disfluencies in French psychiatric interviews; (2) a Python pipeline for parsing, normalising, and analysing Praat TextGrid files at scale; and (3) empirical findings showing that silence duration differs significantly between the clinical groups ($H = 289.9$, $p < 0.001$), while the filled pause rate does not — both dimensions invisible to automatic transcriptions.

Future work will integrate the PANSS and CAARMS scores to test whether silence duration correlates with negative symptom severity, extend annotation to the full 57-interview corpus, and explore dynamic modelling using Generalised Additive Mixed Models [25]. Future analyses will also examine psychiatrist speech prosody — including filled pauses and disfluencies — to test for possible interactive alignments or mimicry effects that could influence patient silence patterns independently of clinical status. The DF annotation tier, not analysed in the present study, will be examined in relation to both silence patterns and clinical severity scores. A key limitation of the current analysis is that silence is treated as a homogeneous category, characterised only by duration. Yet, silences of similar length may correspond to distinct underlying phenomena — a hesitation reflecting lexical search, a pause marking emotional processing, or a break signalling disrupted speech planning — and these are unlikely to share the same acoustic signature. Future work will extract prosodic and voice-quality features (e.g. eGeMAPS, 26) around each SIL interval — pre-/post-silence pitch and intensity contours, since F0 and other

voicing-dependent measures are undefined during silence — as well as breathing artefacts and creak where measurable within the silence, to test whether acoustic sub-types of silence can be distinguished computationally. This direction is already being explored in parallel work using openSMILE feature sets (eGeMAPS, emobase, ComParE16). The protocol and pipeline developed here provide a reusable foundation for these analyses.

8. Acknowledgments

This work was supported by the Fondation de l'Avenir, the Fondation de France, and the PsyCARE project (ANR-18-RHUS-0014), funded by the "Investissements d'Avenir" program, and the IdEx program ANR-18-IDEX-0001, implemented by Université Paris Cité and funded by the "France 2030" program.

9. Declaration on Generative AI

During the preparation of this work, the authors used Claude Sonnet 4.6 in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] H. Hwang, E. Jordan, D.-H. Kim-Dufor, C. Lemey, M. Alrahabi, Evaluating ASR in a clinical context: What Whisper misses, in: Proceedings of ICNLSP 2025, Odense, Denmark, 2025.

[2] A. Bondielli, G. Bonetta, E. Leonardelli, I. Siragusa, Preface to the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026) co-located with 24th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2026), 2026.

[3] H. Ding, Y. Zhang, Speech prosody in mental disorders, Annual Review of Linguistics 9 (2023) 335–355.

[4] H. Hwang, D.-H. Kim-Dufor, J. Al-Tamimi, M. Alrahabi, J. Desclés, V. Lucarini, L. Pichereau, C. Lemey, Towards a clinical adaptation of the whisper model for french, in: Proceedings of the 9th International Conference on Natural Language and Speech Processing (ICNLSP-2026), Association for Computational Linguistics, University of Trento, Italy, 2026. To appear.

[5] N. C. Andreasen, Scale for the assessment of thought, language, and communication (TLC), Schizophrenia Bulletin 12 (1986) 473–482.

[6] Y. Tahir, Z. Yang, D. Chakraborty, N. Thalmann, D. Thalmann, Y. Maniam, N. A. binte Abdul Rashid, B.-L. Tan, J. Lee Chee Keong, J. Dauwels, Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia, PLOS ONE 14 (2019) 1–17. URL: https://doi.org/10.1371/journal.pone.0214314. doi:10.1371/journal.pone.0214314.

[7] V. Rapcan, S. D'Arcy, S. Yeap, N. Afzal, J. Thakore, R. B. Reilly, Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia, Medical Engineering & Physics 32 (2010) 1074–1079. URL: https://www.sciencedirect.com/science/article/pii/S1350453310001657. doi:https://doi.org/10.1016/j.medengphy.2010.07.013.

[8] A. Parola, A. Simonsen, V. Bliksted, R. Fusaroli, Voice patterns in schizophrenia: A systematic review and bayesian meta-analysis, Schizophrenia Research 216 (2020) 24–40. URL: https://www.sciencedirect.com/science/article/pii/S0920996419305389. doi:https://doi.org/10.1016/j.schres.2019.11.031.

[9] F. M. Dovetto, A. Guida, A. Chiara Pagliaro, R. Guarasci, Silences and disfluencies in a corpus of patients with alzheimer's disease (CIPP-MA), in: Proceedings of DiSS 2021, 2021.

[10] A. Pistono, J. Pariente, M. Jucla, Disfluency patterns in alzheimer's disease and frontotemporal lobar degeneration, Clinical Linguistics & Phonetics 38 (2024) 345–358. URL: https://doi.org/10.1080/02699206.2022.2112085. doi:10.1080/02699206.2022.2112085.

[11] S. Betz, N. Bryhadyr, L. Kosmala, L. Schettino, A crosslinguistic study on the interplay of fillers and silences, in: Proceedings of DiSS 2021, Paris, France, 2021.
[12] E. Campione, J. Véronis, A large-scale multilingual study of silent pause duration, in: Proceedings of Speech Prosody 2002, Aix-en-Provence, France, 2002, pp. 199–202.
[13] N. Cummins, A. Baird, B. W. Schuller, Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning, Methods 151 (2018) 41–54. URL: https://www.sciencedirect.com/science/article/pii/S1046202317303717. doi:https://doi.org/10.1016/j.ymeth.2018.07.007, health Informatics and Translational Data Analytics.
[14] D. M. Low, K. H. Bentley, S. S. Ghosh, Automated assessment of psychiatric disorders using speech: A systematic review, Laryngoscope Investigative Otolaryngology 5 (2020) 96–116. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/lio2.354. doi:https://doi.org/10.1002/lio2.354.
[15] L. Chenain, A.-c. Bachoud-Lévi, C. Clavel, Acoustic characterization of huntington's disease emotional expression: An explainable ai approach, in: ACIIW 2024 - 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, Glasgow, United Kingdom, 2024. URL: https://inria.hal.science/hal-04934178.
[16] A. R. Yung, H. P. Yuen, P. D. McGorry, L. J. Phillips, D. Kelly, M. Dell'olio, S. M. Francey, E. M. Cosgrave, E. Killackey, C. Stanford, K. Godfrey, J. Buckby, Mapping the onset of psychosis: The comprehensive assessment of at-risk mental states., Australian & New Zealand Journal of Psychiatry 39 (2005) 964—-971.
[17] A. Défossez, N. Usunier, L. Bottou, F. Bach, Music source separation in the waveform domain, 2021. ArXiv preprint arXiv:1911.13254.
[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: Proceedings of ICML 2023, 2023.
[19] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, Montreal forced aligner: Trainable text-speech alignment using Kaldi, in: Proceedings of Interspeech 2017, 2017, pp. 498–502.
[20] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, M.-P. Gill, pyannote.audio: neural building blocks for speaker diarization, in: ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing, Barcelona, Spain, 2020.
[21] P. Boersma, D. Weenink, A. Shchupak, Praat: Doing phonetics by computer [Computer program], 2026.
[22] D. A. Dickey, W. A. Fuller, Distribution of the estimators for autoregressive time series with a unit root, Journal of the American Statistical Association 74 (1979) 427–431.
[23] A. Astolfi, A. Carullo, L. Pavese, G. E. Puglisi, Duration of voicing and silence periods of continuous speech in different acoustic environments, The Journal of the Acoustical Society of America 137 (2015) 565–579. URL: https://doi.org/10.1121/1.4906259. doi:10.1121/1.4906259.
[24] M. Rohanian, R. Hüppi, F. Nooralahzadeh, N. Dannecker, Y. Pauli, W. Surbeck, I. Sommer, W. Hinzen, N. Langer, M. Krauthammer, et al., Uncertainty modeling in multimodal speech analysis across the psychosis spectrum, NPJ digital medicine 9 (2026) 218.
[25] S. N. Wood, Generalized Additive Models: An Introduction with R, 2nd ed., CRC Press, 2017.
[26] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, K. P. Truong, The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing, IEEE Transactions on Affective Computing 7 (2016) 190–202. doi:10.1109/TAFFC.2015.2457417.

10. Appendix

A. Supplementary Figures

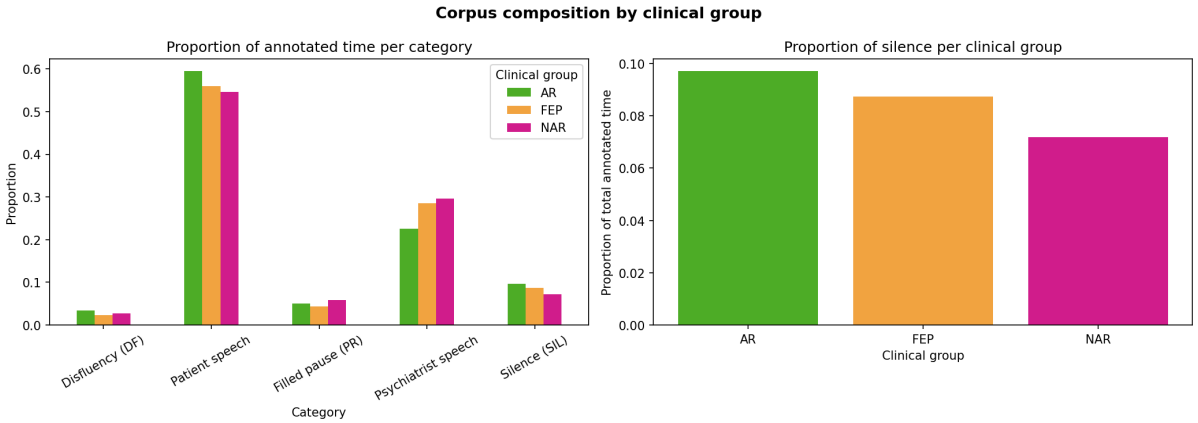


Figure 6: Corpus composition by clinical group. Left: proportion of annotated time per category (AR/NAR/FEP). Right: proportion of silence per group. AR shows the highest silence proportion (9.7%); NAR the lowest (7.2%).

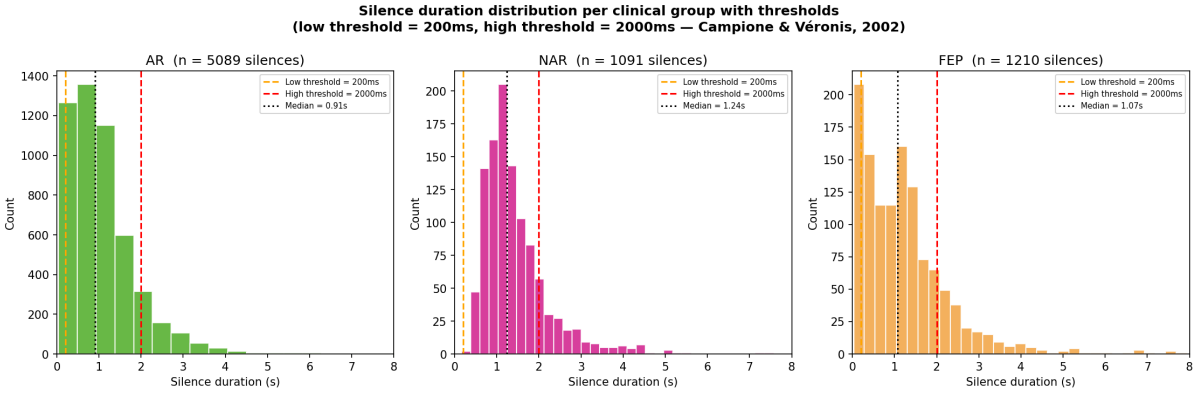


Figure 7: Silence duration histograms per clinical group with 200 ms and 2,000 ms threshold lines. AR peaks early (≈0.5 s); NAR later (≈1.2 s); FEP shows a flatter profile. Medians: AR 0.91 s, NAR 1.24 s, FEP 1.07 s.

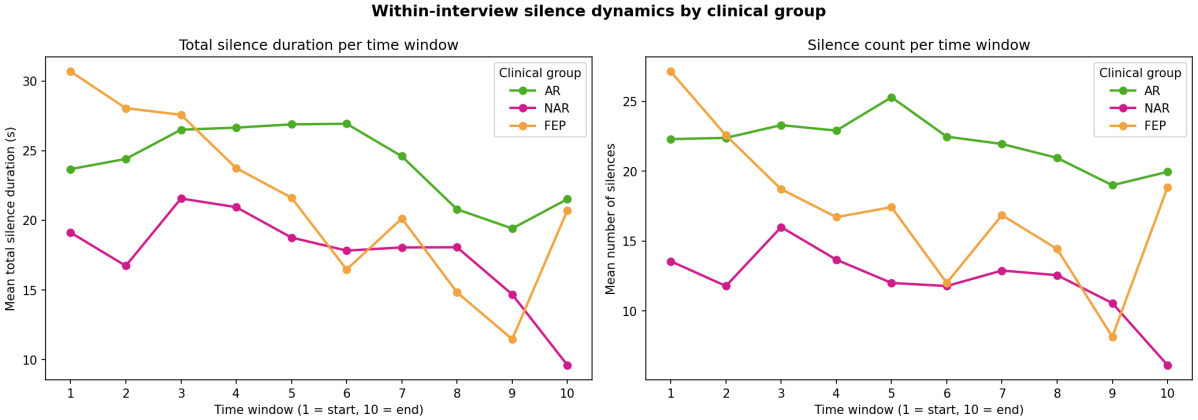


Figure 8: Within-interview silence dynamics by clinical group. FEP shows the highest total silence in early windows and the steepest initial drop; AR is the most stable across windows.