Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Selective acoustic alterations and dissociable dysfluency patterns in people with ultra-high risk, early psychosis and their family members

Derya Çokal

deryacokal@gmail.com

University of Cologne https://orcid.org/0000-0002-5653-1412

Martin Villalba

University of Cologne

Cemal Demirlek

Harvard Medical School https://orcid.org/0000-0001-6195-6899

Burcu Verim

Dokuz Eylul University/Health Sciences Institute https://orcid.org/0000-0001-5798-570X

Berna Yalincetin

Dokuz Eylül University

Klaus von Heusinger

University of Cologne

Emre Bora

Dokuz Eylul University/Health Sciences Institute/Melbourne Neuropsychiatry Centre, University of Melbourne and Melbourne Healtht

Article

Keywords: Pauses, Dysfluency, Acoustic-prosodic features, Schizophrenia, First Episode of Psychosis, Ultra High-Risk, Natural Language Processing (NLP)

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9951289/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is no conflict of interest

Title: Selective acoustic alterations and dissociable dysfluency patterns in people with ultra-high risk, early psychosis and their family members

Short title: Dissociable speech markers across psychosis risk stages

Authors: Derya Çokal[1], Martin Villalba[1], Cemal Demirlek[2], Burcu Verim[3], Berna Yalınçetin[4], Klaus von Heusinger[1], Emre Bora[3-4]

[1]Department of German Language and Literature I, Linguistics, University of Cologne, Germany,

[2]Department of Psychiatry, McLean Hospital, Harvard Medical School, Belmont, MA, USA

[3]Department of Neurosciences, Health Sciences Institute, Dokuz Eylül University, İzmir, Turkey

[4] Department of Psychiatry, Dokuz Eylül University, İzmir, Turkey

Correspondence to: Derya Çokal, University of Cologne, Department of German Language and Literature I, Linguistics, Germany

dcokal@uni-koeln.de

Abstract

Speech provides a window into cognitive dysfunction and is a candidate marker of emerging psychosis, yet it remains unclear which components differ across ultra-high risk (UHR), first-episode psychosis (FEP), and family history of psychosis (FHP) groups. Most prior work has relied on aggregated dysfluency measures that do not distinguish pauses according to their position within speech. We automatically extracted acoustic-prosodic features, position-sensitive dysfluency features, and uniform information density (UID: the evenness of surprisal across words and over speaking time) from Turkish speakers with FEP (n = 51), UHR (n = 57), FHP (n = 37), and neurotypical controls NC (n = 34). Acoustic-prosodic alterations were selective rather than broadly distributed across domains. Voice-quality alterations indexed by shimmer were elevated in UHR and FHP but not FEP, whereas FEP showed a significantly shallower spectral slope relative to NC. Within-utterance unfilled pauses [e.g., a man (pause) ran] were increased in FEP and UHR with both automated and hand-annotated methods, and in FHP for longer pauses. Utterance-initial pauses [e.g., (pause) a man ran] were increased in all three groups with both methods when considered overall, but duration-specific effects were method-sensitive. Relative to NC, FEP participants produced significantly fewer fillers. Reduced shimmer variability was associated with greater negative symptom severity, whereas dysfluency measures were not associated with symptom severity or psychosocial functioning. UID did not distinguish groups and showed limited associations with dysfluency and acoustic-prosodic measures, suggesting that information-density regulation captures a dimension of speech production partly distinct from acoustic and temporal speech abnormalities. These findings support a multidimensional account of speech disturbances in psychosis risk groups.

Key words: Pauses, Dysfluency, Acoustic-prosodic features, Schizophrenia, First Episode of Psychosis, Ultra High-Risk, Natural Language Processing (NLP)

1. Introduction

Speech abnormalities in schizophrenia-spectrum disorders (SSD) extend across multiple levels of language production, including acoustic-prosodic properties, fluency, and higher-level linguistic organization. From a theoretical perspective, language has been proposed as a core transdiagnostic dimension of psychopathology, reflecting underlying disturbances in thought and communication processes[1]. Advances in computational and acoustic analysis have enabled increasingly precise and scalable quantification of these features, raising the possibility that speech may contribute to early detection, clinical stratification, and symptom monitoring [2–7]. However, despite this progress, a review of the literature reveals substantial heterogeneity, inconsistent replication, and limited interpretability, particularly regarding the mechanisms underlying observed speech abnormalities [8–10]. Large-scale and cross-linguistic evidence suggests that only a restricted subset of acoustic features are consistently replicated [9,10]. These findings suggest that speech abnormalities in SSD are selective, and multidimensional rather than reflective of a single generalized deficit**.**

A substantial body of work has examined acoustic-prosodic features in SSD, including pitch, shimmer, harmonics-to-noise ratio (HNR), and spectral characteristics. Several studies have reported these features distinguish patients from neurotypicals and relate to symptom dimensions [3,11]. Converging evidence from multimodal and speech-based machine learning studies suggest acoustic and prosodic features contribute to the prediction of symptom severity, although these effects are often modest when considered in isolation [12,13]. However, recent work found no significant associations between negative symptom severity and acoustic measures, including jitter, shimmer, F0, and HNR, despite relatively high test-retest reliability of these features [14]. Moreover, speech-based acoustic and prosodic features have recently been shown to be

satisfactorily stable across measurement points in neurotypical controls (NC) [15]. In contrast to acoustic-prosodic findings, fragmented speech, reduced speech rate, and increased pauses are reported in SSD [9], with temporal features among the most informative predictors of diagnosis [3, 16–18]. Speech rate, speech proportion, and unfilled pauses have also been identified as reliable measures associated with negative symptom severity across speech tasks and assessment points [14]. A 2017 computational study demonstrated that disruptions in temporal and organizational aspects of speech can predict diagnosis and symptom severity, often outperforming surface acoustic features [19].

Recent work indicates that speech abnormalities are best understood in dimensional terms. Speech and language features track symptom variation over time across multiple psychosis domains, with different speech measures relating to distinct clinical dimensions [6]. In addition, acoustic features have predicted negative schizotypy traits in non-clinical populations [20], although the features driving predictive accuracy were largely spectral rather than expected markers of blunted affect or alogia. Together, these findings suggest that different speech features may reflect partly dissociable cognitive and clinical processes.

A key unresolved question is whether these abnormalities show distinct profiles across psychosis-risk groups, including family history of psychosis (FHP), ultra-high risk (UHR), and first-episode psychosis (FEP). Individuals with FHP provide a critical comparison group. If speech abnormalities are associated with familial liability, similar patterns may be observable in this group. If they are more closely linked to current clinical status, effects may be attenuated or absent. Findings from early-risk populations are also mixed. While prosodic variability, pauses, and pitch

123 contours predict psychosis onset and are related to symptom severity [2], other work has reported no
124 significant acoustic-prosodic differences or robust clinical associations in clinical high-risk
125 samples [21]. These discrepancies suggest that identifying speech abnormalities in high-risk groups
126 may depend on symptom severity, transition risk, speech-elicitation task type, or linguistic context.
127
128 A major limitation of existing work is that dysfluency is often treated as a unitary construct, relying
129 on aggregated pause counts or speech rate. This approach may obscure potentially informative
130 differences in where dysfluency occurs during speech production. Unfilled pauses can arise from
131 processing demands at multiple levels of speech production, including speech planning, lexical
132 retrieval, syntactic and phonological encoding, monitoring, and memory [22]. These underlying
133 cognitive processes cannot be directly inferred from overt speech behaviour[22-29]. However,
134 dysfluency phenomena are commonly used as indirect indices of processing demands during
135 speech production. For example, pause occurrence is shaped by the structural and planning
136 demands of the unfolding utterance [22, 23]. Further experimental work shows that utterance-initiation
137 times increase with the length and syntactic complexity of the upcoming utterance, consistent with
138 planning before speech onset [23,30]. Within-utterance unfilled pauses (e.g., He is looking (PAUSE)
139 at a woman.) are associated with lexical retrieval difficulty and incremental speech formulation
140 demands (e.g., difficulties in phrase structure) [24, 25], while utterance-initial pauses [e.g., (PAUSE)
141 He is looking at a woman.] are linked with higher-level conceptual planning [26]. During ongoing
142 speech, hesitations are also more likely before less predictable words in context [27]. Fillers (e.g.,
143 Uhm, ahm.) can serve communicative functions such as floor-holding and self-monitoring [26, 28, 29].
144 Therefore, a particular processing difficulty is not necessarily restricted to a single utterance
145 position. Nevertheless, pause position remains theoretically informative since it indicates when

processing difficulty becomes manifest during the unfolding utterance, rather than uniquely identifying the processing level from which it originated. Despite the potential informativeness of pause position, relatively few studies [31] have differentiated unfilled pauses according to their location within the utterance (see Table 2 for examples). For example, pause patterns in SSD vary according to their position within the utterance, with utterance-initial pauses linked to increased planning demands, and increased filler use observed in individuals with FHP [30, 31]. Consistent with the need for theory-driven and low dimensional approaches [9, 32], these findings call for a more fine-grained analysis of dysfluencies, while avoiding a one-to-one mapping between pause location and specific cognitive mechanisms.

Present study

In this study, we addressed these limitations in two ways. First, we focused on a restricted set of theoretically motivated acoustic-prosodic features that have shown consistent effects in prior SSD work [10, 18], rather than employing large feature sets with limited interpretability (see Table 1 for our acoustic-prosodic features).

Table 1. Acoustic–prosodic features

Feature family	Feature	Definition	What it captures
Acoustic-prosodic features	Fundamental frequency mean (F0 mean)	The average pitch (fundamental frequency) of a speaker's voice over a speech segment.	Average pitch level
	F0 stddevNorm	The extent to which pitch varies over time.	Variability in pitch
Voice-quality features	Shimmer mean	The average cycle-to-cycle variation in vocal amplitude.	Average amplitude perturbation
	Shimmer stddevNorm	The variability of amplitude perturbation across speech.	Variability in amplitude perturbation
	HNR mean	The average ratio of harmonic (periodic) sound to noise,	Average voice clarity

Feature family	Feature	Definition	What it captures
		reflecting voice clarity or breathiness.	
	HNR stddevNorm	The extent to which voice clarity fluctuates over time.	Variability in voice clarity
	Spectral slope mean	The average slope of the spectral energy envelope between 500 and 1500 Hz, reflecting how energy decreases across higher frequencies.	Distribution of spectral energy
Temporal/dysfluency features	Within-utterance pauses	Unfilled pauses occurring within an utterance.	Processing difficulty that becomes observable during ongoing speech
	Utterance-initial pauses	Unfilled pauses occurring before utterance onset.	Processing difficulty that becomes observable before utterance onset
	Fillers	Filled hesitation markers (e.g., uhm, hmm).	Hesitation that may also serve communicative functions (e.g., monitoring, floor-holding)
Uniform Information Density (UID)	Text-only UID	Distribution of linguistic information based on text surprisal estimates from a language model.	Uniformity of information distribution in linguistic content
	Audio-enriched UID	Distribution of linguistic information using both language-model surprisal and word-level timing information.	Uniformity of information distribution incorporating temporal speech structure

Based on prior work, we predicted acoustic-prosodic effects to be selective. Specifically, we anticipated no robust group differences in mean pitch, consistent with meta-analytic evidence showing this feature does not reliably distinguish patients from NCs [9, 10]. We further predicted reduced pitch variability in clinical groups, particularly FEP, although previous findings suggest

that such effects are modest and dependent on task and language context [9, 10]. Prior evidence showing broadly similar acoustic profiles in FEP and chronic patients [10], suggests that some acoustic atypicalities may already be present at illness onset rather than being products of chronicity [10], although it remains unclear whether this holds for every individual acoustic feature.

For voice quality measures, we anticipated alterations in shimmer and HNR in clinical groups. On the Brief Negative Symptom Scale (BNSS), participants with SSD [11] have shown lower shimmer and elevated HNR relative to NC , with shimmer showing stable negative correlations with blunted affect and alogia subscales across emotion conditions. In addition, in a UHR sample a strong correlation between a composite of voice quality features (e.g., shimmer, jitter, and HNR) and negative symptom severity has been identified [2], though this was a multivariate result and the individual contributions of shimmer and HNR were not isolated. Whether voice-quality alterations would differ systematically across psychosis-risk groups or show partially dissociable profiles across FHP, UHR, and FEP remained an open question, given that acoustic effects in high-risk states have been observed in some samples [2], but not others [21]. Given evidence that this feature contributes to diagnostic classification in SSD, we predicted differences in spectral slope [3]. However, we did not make a strong directional prediction about illness-stage specificity for this feature, since existing work has not directly compared spectral slope across FEP, UHR, and FHP, and the stage-dependence of spectral effects remains untested.

Second, we extended prior work on dysfluency in chronic SSD by implementing an automated analysis of pause structure that captures pause position and duration across multiple psychosis risk stages (see Table 2). This enabled a theoretically motivated decomposition of dysfluency beyond aggregated pause counts [31]. We examined whether FEP would show dysfluency patterns, including

190 increased utterance-initial pauses and reduced filler use, similar to those reported in chronic
191 patients [31, 33, 34]. In addition, we explored whether FHP participants would show increased filler
192 use, potentially reflecting preserved interactive or compensatory mechanisms. However, because
193 filler use is highly sensitive to task demands and interactional context, and because previous
194 findings were obtained in different linguistic and task settings, we treated this prediction as
195 exploratory. Unlike previous studies relying primarily on manual annotation, our approach
196 provides scalable and reproducible detection of pause types defined by utterance position and
197 duration, as well as fillers, across all four groups.

198 Table 2. Dysfluency features with examples and explanation

	Definition	Example	Explanation
Unfilled pauses	Pauses without fillers	(Pause) O bakıyor.	Unfilled pause at the beginning of utterance.
Utterance position	Utterance-initial pauses (all)	(PAUSE) O bakıyor kadına. (PAUSE) He is looking at a woman.	Unfilled pause is at the beginning of the utterance**.**
	Within-utterance pauses (all)	O (PAUSE) bakıyor kadına. He (PAUSE) is looking at a woman.	Unfilled pause within the utterance.
Duration	Utterance-initial pauses < 1s & between 1-3s		The pause duration at the beginning of utterance (less than 1s and between 1-3s)
	Within-utterance pauses <1s & between 1-3s		The pause duration within the utterance (less than 1s and between 1-3s)
Fillers	Fillers were analysed independently of utterance position.	(ahm) O bakıyor kadına.	(ahm) is a filled pause.

(ahm) He is looking at
a woman.

Finally, using the Uniform Information Density (UID) framework, we examined whether acoustic-prosodic features and dysfluency relate to higher-level information-structural organization [36]. Unlike our text-only implementation of UID [37], the audio-enriched measure incorporates timing information. Text-only UID characterizes the distribution of surprisal across lexical units, whereas the audio-enriched measure incorporates surprisal per time unit. Without speech data, UID measures must assume language transmission "at a roughly constant rate, close to the channel capacity" [37]. Our audio-enriched measure moves beyond this simplifying assumption by dividing word surprisal for individual tokens by the timing information for the duration of said token. Thus, UID concerns the uniformity of information distribution rather than the absolute amount of information conveyed. Beyond acoustic-prosodic and dysfluency measures, disruptions in discourse organization including semantic similarity, semantic coherence and atypical referential structure have been identified as important dimensions of language disturbance in psychosis, both theoretically and using computational methods [3,33,38–44]. Although UID is not itself a measure of clinical disorganization, it can provide a complementary framework for examining the distribution of information in speech. This allowed us to test whether dysfluency patterns are associated with variation in information-structural organization or instead primarily reflect production-level constraints. Although surprisal-based measures were originally linked to processing costs for comprehenders, the UID hypothesis also makes a production-side claim: speakers adjust their productions to keep information density relatively uniform [45].

We therefore analysed picture descriptions across three domains (i.e., acoustic-prosodic features, dysfluency patterns, and UID) to examine whether speech abnormalities are selectively or broadly distributed, whether they differ across psychosis-risk groups, and whether dysfluency reflects an information-structural imbalance or production-level constraints.

2. Results

Acoustic-prosodic features

As a first step, we used generalized estimating equations (GEE) to estimate group differences in acoustic-prosodic features (see Table 1), controlling for gender, age, education, and task (for descriptive statistics comprising means and standard deviations, see Table S1). Average shimmer was significantly higher in the UHR (β = 0.096, q = .004) and FHP groups (β = 0.087, q = .014) relative to NC, with no difference observed in FEP (Table S3).

For HNR, the UHR group showed a lower mean than NC (β = −1.332, q = .044), alongside a significant gender effect (β = −2.259, q < .001). Spectral slope was significantly shallower (i.e., less negative) in FEP (β = 0.004, q = .002). A smaller shift toward a shallower spectral slope was also observed in UHR (β = 0.003, q = .042). FHP did not differ from NC. However, in gender-stratified sensitivity analyses, the UHR-related HNR and spectral slope effects were not robust after FDR correction in male-only or female-only analyses. In contrast, the UHR-NC shimmer effect remained significant across full sample and both gender-stratified analyses (males: β = 0.084, q = .045; females: β = 0.099, q = .041), providing additional support for the consistency of this finding across male and female participants. No group differences were observed for F0 variability (all q > .05).

Dysfluency features

As a second step, we used Poisson regression models to identify dysfluency across groups (e.g., fillers, and unfilled pause subtypes by duration and position). For descriptive statistics comprising mean rates of unfilled pauses see Table S2.

<Insert Figure 1. Dysfluency measures across groups using automated detection.>

In the automated analysis, all three groups produced more unfilled pauses than NC, although the pattern varied by pause position and duration (Figure 1; Table S4a). Within-utterance pauses showed the broadest group effect: overall within-utterance pausing was elevated in FEP (IRR = 1.57, $q < .001$), FHP (IRR = 1.48, $q < .001$), and UHR (IRR = 1.40, $q < .001$). This pattern was also observed for pauses <1 s (FEP: IRR = 1.46, $q < .001$; FHP: IRR = 1.40, $q = .001$; UHR: IRR = 1.35, $q = .001$) and for longer pauses of 1-3 s (FEP: IRR = 2.78, $q < .001$; FHP: IRR = 2.42, $q = .005$; UHR: IRR = 1.92, $q = .033$).

In contrast, utterance-initial pauses showed a more differentiated pattern. Overall, utterance-initial pausing was elevated in FEP (IRR = 1.53, $q < .001$), FHP (IRR = 1.34, $q = .010$), and UHR (IRR = 1.40, $q = .001$). The same three-group pattern was observed for pauses <1 s (FEP: IRR = 1.46, q < .001; FHP: IRR = 1.32, $q = .020$; UHR: IRR = 1.37, $q = .004$). However, longer utterance-initial pauses (1-3 s) were significantly elevated only in FEP (IRR = 2.62, $q = .012$). Reduced filler production was specific to FEP (IRR = 0.61, $q = .001$). Neither FHP ($q = .452$) nor UHR ($q = .172$) differed significantly from NC. Follow-up pairwise comparisons confirmed significantly lower filler production in FEP than in both FHP (IRR = 0.67, $q = .001$) and UHR (IRR = 0.73, $q = .007$) (see Table S4c).

Comparison of automated and hand-annotation pause analyses

Hand-annotated analyses showed both convergence with and divergence from the automated results (Table S4b; Figure S1). The two methods showed substantial convergence for within-utterance pauses. Longer within-utterance pauses (1-3 s) were significantly elevated in FEP, FHP, and UHR relative to NC with both methods. For overall within-utterance pauses and shorter pauses (<1 s), significant increases were observed in FEP and UHR with both methods, whereas FHP increases were significant only in the automated analysis. For utterance-initial pauses, both methods identified significant increases in all three groups (FEP, FHP, and UHR) when pauses were considered overall.

However, duration-specific effects showed greater method-dependent variation. For shorter utterance-initial pauses (<1 s), both methods identified a significant increase in FHP, whereas significant increases in FEP and UHR were observed only in the automated analysis. Conversely, for longer utterance-initial pauses (1-3 s), both methods identified a significant increase in FEP, whereas increases in FHP and UHR were significant only in the hand-annotated analysis. Overall, agreement between methods was higher for within-utterance pauses (7 of 9 group effects, i.e., three groups × three pause measures, significant with both methods) than for utterance-initial pauses (5 of 9).

Relationship between dysfluency patterns and UID

As a third step, we examined whether UID differed across groups and whether it was associated with dysfluency and acoustic-prosodic features. Neither audio-enriched nor text-only UID significantly distinguished FEP, UHR, or FHP from NC after FDR correction. For audio-enriched UID, GEE models (including age, education, gender, task, and number of utterances) revealed no

significant associations with dysfluency measures (all $q \geq .596$; Table S7) or acoustic-prosodic features (all $q \geq .095$; Table S8) after FDR correction. Text-only UID showed a largely similar pattern: none of the unfilled-pause measures were significantly associated with UID (all $q \geq .829$; Table S10), although filler use showed a small positive association with text-only UID irrespective of group ($\beta = .013$, $q = .043$; Table S10). Among the acoustic-prosodic features, only HNR stddevNorm was significantly associated with text-only UID after FDR correction ($q < .001$; Table S9); no other acoustic-prosodic features showed significant associations (all $q \geq .674$). The HNR stddevNorm association was not significant for audio-enriched UID ($q = .095$; Table S8). A higher number of utterances was consistently associated with lower UID values (audio-enriched: $\beta = -.013$; text-only: $\beta = -.020$ to $-.021$; all $q < .001$; Tables S7-S10).

Associations between speech features and clinical symptoms

We computed participant-level partial Spearman correlations between speech features and clinical measures in the combined FEP and UHR sample, controlling for diagnostic group, age, gender, and education (Figures 2-3; Tables S5-S6). Acoustic-prosodic analyses included BNSS and PSP. Dysfluency analyses included BNSS, PSP, formal thought disorder, hallucinations, and delusions (see Methods).

For acoustic-prosodic features, negative symptom severity (BNSS) was negatively associated with shimmer variability (shimmer stddevNorm; $\rho = -.27$, $q = .030$), indicating that greater negative symptom severity was associated with reduced variability in vocal amplitude perturbation (see Figure 2). Dysfluency measures were not significantly associated with negative symptom severity,

formal thought disorder, hallucinations, delusions, or psychosocial functioning (see Figure 3; Table S6).

<Insert Figure 2. Heatmap of partial Spearman correlation matrices depicting associations between acoustic–prosodic features and clinical variables.>

< Insert Figure 3. Heatmap of partial Spearman correlations between dysfluency measures and clinical measures (BNSS, formal thought disorder, hallucinations, delusions, and PSP) in the combined FEP and UHR sample.>

Associations with medication dose

CPZ-equivalent dose was not significantly associated with any acoustic-prosodic feature (all $q \geq .164$; Table S11) or dysfluency outcome (all $q \geq .843$; Table S12). FEP participants produced fewer fillers than UHR participants in the primary analysis without medication adjustment (FEP vs UHR: $\beta = -0.315$, $q = .007$; Table S4c). The estimated difference was similar in magnitude after adjustment for medication dose (FEP vs UHR: $\beta = -0.300$), although the medication-adjusted comparison did not survive FDR correction across the seven dysfluency outcomes ($q = .117$; Table S12).

3. Discussion

The present study examined spontaneous speech in FEP, UHR, FHP, and NC groups, focusing on acoustic-prosodic features, dysfluency patterns, and UID. Three main findings emerged. First, acoustic alterations were selective rather than broadly distributed across acoustic domains, with UHR participants showing altered voice quality (particularly increased shimmer), whereas alterations in spectral slope were most pronounced in FEP. Second, dysfluency measures showed

both shared and group-specific alterations across psychosis-risk groups. Third, UID did not differ significantly across groups and showed little association with dysfluency features and acoustic-prosodic measures, suggesting that temporal and acoustic speech alterations may be partly dissociable from the organization of information distribution.

Regarding acoustic-prosodic features, we did not observe reduced pitch variability or reliable group differences in F0 mean, F0 variability, or HNR variability, indicating that the acoustic profile is not characterized by a broad acoustic-prosodic deficit [10]. Gender-stratified sensitivity analyses showed that the UHR-NC difference in HNR mean did not remain significant after FDR correction in gender-specific analysis as well, whereas the UHR-NC difference in shimmer mean remained significant in both male and female participants. This pattern provides additional support for the consistency of the shimmer finding across gender strata. The UHR voice-quality effects contrast with prior findings[21], which had shown no significant acoustic-prosodic differences in UHR vs NC.

Contrary to a simple severity-gradient account, shimmer mean elevations were most pronounced in UHR and FHP rather than in FEP. However, because shimmer measures are sensitive to recording conditions and phonatory variability, any physiological/mechanistic interpretation is tentative. In contrast, spectral slope showed the clearest acoustic alteration in FEP, suggesting differences in spectral or voice-quality-related properties. Notably, consistent with findings in a Dutch sample of patients with SSD[3], who showed a flatter mean spectral slope (500–1500 Hz) than NC, spectral slope in FEP was significantly shallower (less negative) than in NC. Voice quality and spectral features have also been observed in transdiagnostic contexts [46], suggesting

some acoustic alterations, particularly spectral and voice-quality-related features, may not be specific to psychosis.

In contrast to acoustic features, dysfluency measures provided the clearest group differences, showing shared but feature-specific patterns across groups. Unfilled pauses were increased across FEP, UHR, and FHP, although the distribution of effects differed according to utterance position, duration, and detection method. Within-utterance pauses showed the most consistent group differences, particularly for pauses lasting 1-3 s, which were significantly elevated in FEP, UHR, and FHP across both automated and hand-annotated analyses. For overall within-utterance pauses and pauses <1 s, effects in FEP and UHR were significant with both methods, whereas FHP effects were detected only by automated analysis. Within-utterance pauses may be associated with difficulties related to lexical retrieval [24, 25, 46] and phrase structure (i.e., connecting syntactic units). However, pause position alone does not identify the processing level responsible for the disruption.

For utterance-initial pauses, both methods identified significant increases across all three groups when pauses were considered overall, whereas duration-specific effects showed greater variation between detection methods. In automated analysis, utterance-initial pauses <1 s were elevated across all three groups, whereas longer pauses (1-3 s) were significantly elevated only in FEP, potentially reflecting greater disruption in speech planning before utterance onset. However, this pattern was not specific to FEP in the hand-annotated analysis, where all three groups showed significant increases, warranting caution in interpreting its clinical specificity. FEP also showed the largest estimated rate increases relative to NC across the automated pause categories, particularly for longer pauses (1-3 s), although this pattern was not consistently reproduced in hand-annotated analysis. In both methods, UHR participants showed altered speech fluency.

Specifically, the frequency and duration of within-utterance pauses (< 1 s and 1–3 s) substantially overlapped with those observed in FEP, suggesting that temporal speech disruption is already detectable in individuals at ultra-high risk for psychosis.

While some studies have reported minimal or no differences in dysfluency patterns in clinical high-risk samples [21], others support the diagnostic relevance of acoustic and temporal speech markers in psychosis [2, 3, 19]. These contradictory findings suggest speech abnormalities in high-risk groups may depend on symptom severity, risk enrichment, and task characteristics. Importantly, pausing alterations were also evident in the FHP group. Although some effects differed between annotation methods, across both methods, increases in longer within-utterance pauses (1-3 s) and overall utterance-initial pauses were consistently observed in FHP. These findings suggest that some temporal speech characteristics may be associated with familial vulnerability to psychosis rather than emerging only with overt illness. Because the groups were assessed cross-sectionally, these findings should not be interpreted as evidence of a developmental progression from FHP to UHR to FEP.

Another key finding is the dissociation between dysfluency types. FEP participants produced fewer fillers despite producing more unfilled pauses. Interestingly, greater filler use was associated with higher text-only UID values, although this association was not observed for audio-enriched UID. This suggests a possible relationship between filler use and information-density regulation. However, because UID does not directly measure discourse coherence, the present association does not establish a role for fillers in coherent discourse.

Reduced filler use in FEP may reflect diminished communicative adaptation [29], reduced metacognitive monitoring or self-regulation in speech[28], which is consistent with prior work

reporting reduced filler use in schizophrenia [34, 35]. Our exploratory hypothesis that FHP would show increased filler use was not supported, suggesting that such effects may depend on task demands or interactional context and may not be generalisable. In contrast to prior findings [31], the absence of increased filler use in FHP may reflect several differences between the studies. Although both used picture description tasks, the original finding of elevated filler use in FHP participants was observed within a chronic English-speaking SSD sample. These findings suggest that filler use may be context-sensitive rather than a stable marker of familial liability.

We found no significant associations between unfilled pauses and UID across either audio-enriched or text-only measures. Instead, UID was primarily associated with utterance segmentation. These findings suggest that temporal speech alterations and information-density regulation may capture partly distinct dimensions of speech production. Increased pausing in FEP and UHR may reflect production-related demands that are not directly captured by the UID measures examined here. However, the present findings do not establish the specific cognitive mechanisms underlying these alterations. A complementary, more fine-grained level of analysis would examine whether these temporal speech patterns are related to specific aspects of linguistic organization, such as semantic similarity and discourse coherence (e.g., referential expressions). Such analyses could help determine whether pausing is associated with lexical-semantic retrieval or discourse organization and how these processes relate to information-density regulation.

Within FEP and UHR, speech features showed selective associations with symptom severity and functioning. Shimmer (i.e., stddevNorm) showed a negative association with BNSS, suggesting reduced variability in vocal amplitude is linked to more severe negative symptoms. This is consistent with prior work linking voice quality to blunted affect and alogia [11], although the present

association involved shimmer variability rather than shimmer mean, suggesting distinct dimensions of amplitude disruption. Whereas elevated shimmer mean characterized UHR and FHP at the group level, lower shimmer variability was associated with greater negative symptom severity within the clinical sample. This dissociation suggests that group-level alterations in average voice quality and individual differences in the variability of voice production may capture partly distinct dimensions of acoustic disturbance. In addition, other acoustic-prosodic features were not associated with psychosocial functioning, suggesting that individual differences in voice quality, pitch, and spectral properties did not track concurrent social and occupational functioning in this early-stage sample.

This pattern is partly consistent with recent findings [14], which found no significant associations between negative symptom severity and acoustic measures, including F0, HNR, jitter, and shimmer in the English SSD participants (i.e., schizophrenia, schizophreniform disorder, schizoaffective disorder, unspecified psychotic disorder, bipolar disorder with psychotic features). However, the acoustic features examined in [14] were averaged over the entire speech segment, and the authors noted that such measures may be less able to capture the dynamics of vocal expression relevant to negative symptoms. The present association involved shimmer variability (stddevNorm), which indexes how amplitude perturbation fluctuates across speech rather than its average level, and may therefore capture aspects of reduced vocal expressivity not reflected in mean-level measures. The absence of significant associations for the remaining acoustic-prosodic features indicates that clinical associations were selective rather than reflecting a general relationship between acoustic variation and concurrent clinical severity.

Dysfluency measures showed no evidence of association with concurrent clinical severity (see Figure 3). In the combined FEP and UHR sample, none of the associations between dysfluency features and BNSS, formal thought disorder, hallucinations, delusions, or PSP was statistically significant after controlling for age, education, gender, and diagnostic group. This contrasts with a recent study [14], where unfilled pauses were associated with greater negative symptom severity across assessment points. Differences in the operationalization of pauses, speech-elicitation tasks, clinical characteristics, and analytic approach may contribute to this discrepancy. This study [14] examined an English-speaking sample with established SSD, whereas the present analyses combined FEP and UHR participants which represent earlier and less severe psychotic disorder. Therefore, the range of negative symptom severity in this cohort may be more restricted than in chronic samples. Our dysfluency findings suggest that the observed group differences in temporal speech organization are not readily explained by individual variation in concurrent symptom severity or psychosocial functioning. The presence of pause differences in FHP suggests the possibility that some temporal speech characteristics capture a vulnerability-related component that is partly dissociable from current symptom expression. However, the absence of symptom correlations does not establish trait status, and longitudinal studies are required to determine the stability of these features across changes in clinical state. These findings suggest that group-level speech alterations and individual differences in concurrent clinical severity are only partly overlapping, consistent with a multidimensional rather than unitary account of speech disturbance in psychosis risk.

Limitations

Several limitations should be considered. First, the cross-sectional design precludes conclusions about longitudinal progression and transition to psychosis. Second, the use of a structured picture-description task may limit generalizability, particularly for interactional features such as fillers. The communicative functions of fillers, such as maintaining the conversational floor and managing turn-taking, may be less prominent in structured picture-description tasks that elicit relatively constrained monologic speech with limited turn-taking demands. Therefore, future studies should examine whether the present filler patterns generalize to less constrained speech paradigms, including interactive conversational tasks (e.g., dialogues), in which speakers have greater responsibility for discourse generation and turn management.

Third, although automated methods enhance scalability and reproducibility, they may introduce measurement noise, especially in the detection of dysfluencies. Across pause categories, automated and hand-annotated analyses agreed most consistently for within-utterance pauses, whereas duration-specific utterance-initial effects differed between detection methods. Overall utterance-initial pauses were replicated in all three groups with both methods. Thus, the within-utterance pattern and overall utterance-initial pauses appear relatively robust across annotation approaches, whereas finer-grained duration-specific classifications remain sensitive to detection and segmentation procedures. Importantly, the automated and hand-annotated analyses used independently derived utterance counts as offsets. Consequently, differences between methods may reflect variation in both pause detection and utterance segmentation, rather than just pause detection alone. Because automated and hand-annotated pause events were not temporally aligned on an event-by-event basis, agreement should be interpreted at the level of group-pattern convergence rather than exact event-level correspondence. Future work should directly benchmark automated pause detection against temporally aligned hand annotations and establish the cross-

linguistic robustness of automated dysfluency pipelines across corpora, tasks, and recording conditions.

Fourth, this framework should be extended to examine whether acoustic-prosodic and dysfluency features are associated with more fine-grained measures of linguistic organization (including semantic similarity and coherence), and whether these relationships complement those observed for information-density regulation.

Fifth, chlorpromazine-equivalent dosage provides an index of overall antipsychotic exposure but does not capture differences in receptor-binding profiles across medications, including dopamine D2 receptor affinity. Because a sufficiently standardized D2-affinity metric was not available for all the antipsychotic combinations used in the present cohort, medication effects related to specific pharmacological properties cannot be excluded. Future studies should examine receptor-specific medication effects where sufficiently detailed pharmacological data are available.

Finally, although the omnibus group difference in gender distribution was not statistically significant, the NC group contained a numerically higher proportion of female participants than the clinical groups. Although gender was included as a covariate in all models, covariate adjustment does not fully address potential effects of this numerical imbalance or possible gender-by-group interactions. Future studies should prioritise more balanced gender distributions across groups, given evidence that several acoustic speech measures, including fundamental frequency, jitter, and shimmer, can vary by gender [47-50].

Conclusion

Our results support a multidimensional account of speech disturbance in psychosis risk, involving partly separable alterations in voice quality and temporal fluency. Voice-quality alterations, particularly shimmer mean, were most evident in UHR and FHP, whereas temporal dysfluency showed both shared and feature-specific alterations across FEP, UHR, and FHP. Within-utterance pauses showed the most consistent agreement between automated and hand-annotated analyses. They were elevated in FEP and UHR with both methods, whereas in FHP only longer within-utterance pauses (1-3 s) were elevated with both methods. Duration-specific utterance-initial effects were more sensitive to detection method. Reduced filler use was also specific to FEP. Dysfluency measures were not significantly associated with concurrent symptom severity or psychosocial functioning in the combined FEP/UHR sample. The presence of pause alterations in FHP, together with limited associations with current clinical severity, is consistent with a vulnerability-related component of temporal speech disruption, although longitudinal data are required to determine whether these characteristics are stable traits or vary with clinical state. UID did not reliably distinguish the groups and showed little association with acoustic-prosodic or dysfluency measures, suggesting that information-density regulation may capture a partly distinct dimension of speech production. The present findings further demonstrate the feasibility of applying an automated dysfluency-detection pipeline to characterize theoretically motivated pause types across psychosis-risk populations. Overall, these findings indicate that speech abnormalities across psychosis-risk groups are multidimensional, with partially dissociable acoustic, temporal, and information-density patterns rather than a single generalized speech disturbance.

Methods

Figure 4 illustrates the workflow of the study.

<Insert Figure 4. Schematic of the workflow of the study>

Participants

The sample was comprised of 51 individuals with first-episode psychosis (FEP), 57 ultra-high-risk (UHR) participants, 37 individuals with a family history of psychosis (FHP), and 34 neurotypical control participants (NC). Within the FHP group, 81% had at least one first-degree relative diagnosed with schizophrenia, while the remaining participants had multiple second-degree relatives with psychotic disorders. Participants with FEP were recruited from the First-Episode Program of the Early Intervention in Psychiatric Disorders Unit (ETAP) at the Department of Psychiatry, Dokuz Eylül University Hospital in İzmir. These participants met criteria for a schizophrenia spectrum disorder (SSD) and had experienced their first psychotic episode within the preceding 12 months. At the time of assessment, all FEP participants were clinically stable, which is defined as the absence of acute psychotic or affective episodes, no changes in antipsychotic medication due to symptom worsening in the past four weeks, a reduction of more than 30% in total positive symptom scores, and a Hamilton Depression Rating Scale (HDRS) score below 12.

Help-seeking individuals in the UHR group were also recruited through ETAP and met criteria for one of three prodromal syndromes as assessed with the Structured Interview for Psychosis-Risk Syndromes and the Scale of Psychosis-Risk Symptoms (SIPS/SOPS): (1) brief intermittent psychotic symptom syndrome, (2) attenuated positive symptom syndrome, or (3) genetic risk and deterioration syndrome. Forty-three UHR participants were classified as having attenuated positive symptoms, eight as having brief intermittent psychotic symptoms, and six as meeting criteria for genetic risk and deterioration syndrome. NCs were recruited from the local community

via advertisements and had no current or past psychiatric diagnoses, nor a family history of psychotic disorders.

All participants completed clinical interviews using the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) [52]. Current positive and negative symptoms were assessed using the Scale for the Assessment of Positive Symptoms (SAPS) [53] and the Brief Negative Symptom Scale (BNSS)[53-55]. The BNSS is a next-generation instrument with strong psychometric properties, including good internal consistency as well as convergent and discriminant validity, and offers advantages over traditional measures such as the SANS by differentiating between anticipatory and consummatory anhedonia. Antipsychotic medication dosages were converted to chlorpromazine equivalents [57], and the duration of untreated psychosis (DUP) was calculated. Functional outcomes were assessed using the Personal and Social Performance (PSP) scale, which yields a total score based on four domains: (1) socially useful activities, (2) personal and social relationships, (3) self-care, and (4) disturbing and aggressive behaviors [57, 58]. Higher scores in the PSP indicate better social functioning.

The study protocol was approved by the Dokuz Eylül University Hospital Ethics Committee for Non-Interventional Studies, and all participants provided written informed consent. All participants were native Turkish speakers. Exclusion criteria for all groups included significant medical conditions, current alcohol or substance abuse, or a history of neurological disorders.

The groups did not differ significantly in age or gender distribution (see Table 3). Years of education differed significantly across groups, with NC participants having more years of education than the UHR, FHP, and FEP groups, which did not differ from one another. Age, gender, and education were included as covariates in all analyses.

The clinical groups differed significantly in negative symptom severity (BNSS), depressive symptoms (HDRS), and psychosocial functioning (PSP). BNSS scores were higher in FEP than UHR ($p = .006$), and both groups had higher scores than FHP (both ps < .001). HDRS scores were higher in UHR than FHP ($p = .003$) and FEP ($p = .005$), whereas FEP and FHP did not differ ($p = .299$). PSP scores were significantly different between all three groups, with the highest scores in FHP, followed by UHR, and FEP (all ps < .001). FEP also had higher hallucination ($p = .007$) and delusion scores ($p < .001$) than UHR, whereas formal thought disorder did not differ between FEP and UHR ($p = .635$).

Table 3. Demographic and clinical data of neurotypical controls (NC), ultra-high risk (UHR), family history of psychosis (FHP), and first-episode psychosis (FEP) participants

	NC (n = 34)	UHR (n = 57)	FHP (n = 37)	FEP (n = 51)	Group comparison
	Mean ($s.d$)	Mean (s.d)	Mean ($s.d$)	Mean ($s.d$)	
Age	22.33 (4.25)	20.59 (3.99)	22.08 (3.65)	21.02 (4.47)	H (3) = 5.50, p = .139
Gender (m/f)	10/24	28/29	16/21	25/26	χ^2 (3) = 4.07, p = .254
Education	14.44 (2.51)	11.82 (2.12)	12.41 (2.25)	12.41 (3.50)	H (3) = 21.48, p < .001
BNSS	—	21.73 (15.16)	3.89 (8.00)	31.24 (17.35)	H (2) = 58.10, p < .001
HDRS	—	6.85 (4.83)	3.72 (4.42)	4.14 (3.95)	H (2) = 14.42, p < .001
PSP	—	51.60 (12.98)	76.95 (15.93)	39.90 (15.15)	H (2) = 66.10, p < .001
SAPS	—	—	—	8.78 (7.57)	—
SAPS FTD	—	0.30 (0.57)	—	0.45 (0.88)	W = 1369, p = .635
Hallucinations	—	0.45 (0.60)	0.03 (0.16)	1.00 (1.08)	H (2) = 30.15, p < .001

Delusions	—	0.98 (0.59)	0.05 (0.23)	1.90 (1.12)	H (2) = 75.11, p < .001

Note. Values are means (SD) unless otherwise indicated. NC = neurotypical controls; UHR = ultra-high risk; FHP = family history of psychosis; FEP = first-episode psychosis; BNSS = Brief Negative Symptom Scale; HDRS = Hamilton Depression Rating Scale; PSP = Personal and Social Performance Scale; SAPS = Scale for the Assessment of Positive Symptoms; FTD = formal thought disorder. Continuous variables were compared using Kruskal–Wallis tests, except SAPS FTD, for which only UHR and FEP data were available, and a Wilcoxon rank-sum test was used. Gender was compared using Pearson's χ^2 test. Significant omnibus tests were followed by pairwise Wilcoxon rank-sum tests with Holm correction. Analyses were based on available cases for each measure. Demographic and clinical data are shown for the full sample. Four FHP participants were excluded from the automated dysfluency and UID analyses because automated segmentation failed (see Data Analysis). Automated dysfluency analyses were therefore based on 175 participants (FHP n = 33).

Speech recording

Speech was recorded in a quiet room using head-worn Sennheiser HSP 4 microphones. Focusrite Scarlett 2i2 (3rd Gen) was used as audio interface. Microphones were worn by both the participants and a trained interviewer and were positioned at 2 cm from the lips at an angle of 45°- 90° away from the mouth. The recordings were acquired with a sampling frequency of 44.1 kHz with 16-bit quantization.

Speech pre-processing

Crosstalk (i.e., speech from the interviewer captured on the participant's audio channel) was manually identified and removed in PRAAT [60]. Only the interviewer's speech was removed. The silent interval between the end of the interviewer's turn and the participant's speech onset was retained, so that utterance-initial pauses were preserved. Prior to acoustic feature extraction, the recordings were also manually reviewed by a trained research assistant. During this quality-control procedure, identifiable non-speech environmental disturbances, such as sounds caused by doors opening or another person entering the interview room, were manually removed when they

occurred during periods in which the participant was not speaking. Participant speech was not removed or modified during this procedure. Following removal of interviewer speech and non-speech disturbances, the remaining participant-speech segments were concatenated into a new audio file containing only participant speech.

Linguistic Testing

All subject interviews took place in the Early Intervention in Psychiatric Disorders Unit (ETAP) at the Department of Psychiatry, Dokuz Eylül University Hospital. Participants were shown eight pictures from the Thematic Apperception Test (TAT) [61] in sequence and asked to talk about each picture for one minute. If participants stopped speaking before the minute was up, they were given a non-directive prompt such as, 'Can you say more?' or 'What else do you see?'. Such non-directive prompts produced by the interviewer were manually removed from the audio recordings. The entire process lasted approximately nine minutes. All interviews were recorded on a laptop and subsequently pseudonymized. Each participant was represented by a unique study number, and directly identifiable information was removed from analysed material. Audio files were transcribed using Whisper (https://openai.com/whisper/) [62]. Every participant's recording was processed using both Whisper [62] and CrisperWhisper [63] to auto-transcribe and obtain word-level timing information and speech disfluencies. Each recording was segmented into utterances automatically from the CrisperWhisper transcription. The primary dysfluency analyses used automatically generated utterance boundaries, while hand-annotated segmentation was used for complementary validation analyses. Below we report the results of automatically segmented utterances. Following the same methodology [41,64], statistical analysis is based on speech generated in response to four selected pictures from the TAT: (1) a farm scene, (2) a man with a pipe, (3) a

637 woman opening a living room door, and (4) a human figure/man next to a streetlamp post in the
638 dark (see Figure S2).

639 *Speech and language variables*

640 Our speech features are divided into three groups: acoustic-prosodic parameters, dysfluency
641 measures, and uniform information density (UID) (See Tables 1 and 2). Acoustic-prosodic features
642 capture continuous properties of the speech signal, whereas dysfluency features reflect temporal
643 structuring and speech planning.

644

645 The first set of features were exclusively acoustic-prosodic parameters (i.e., the sound signal
646 instead of its content). These features were selected based on prior meta-analyses identifying them
647 as having modest effect sizes in psychopathology studies [10, 18]. They were drawn from the extended
648 Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [65] and extracted using the openSMILE
649 library (Eyben et al., 2010). F0, shimmer, and HNR variability were indexed using the eGeMAPS
650 stddevNorm functionals extracted with openSMILE. We therefore refer to these measures as F0
651 stddevNorm, shimmer stddevNorm, and HNR stddevNorm, respectively, to distinguish them from
652 conventional raw standard deviation.

653 The second set of features consisted of dysfluency measures (i.e., unfilled pauses and fillers; see
654 Table 2). We classified pauses following the method in [31]. Pauses shorter than 250 ms or longer
655 than 3 s were excluded. Pauses below 250 ms were considered phonatory gaps linked to respiration
656 rather than linguistically meaningful pauses, and are difficult to segment reliably algorithmically.
657 Pauses of 3 s or longer were excluded as likely reflecting task disengagement rather than speech-
658 internal pausing.

659

We identified pauses using the Pydub library [66] and a two-step process. The first step identified a threshold volume to classify individual audio segments into either audio or silence. We extracted the average volume of every 1ms of audio (in dBFS), identified the mean and standard deviation of these values, and then defined the threshold as every value whose volume fell 0.95 standard deviations below the mean. Specifically, we processed the audio file one ms at a time, keeping both a rolling window with 137ms of rolling window length and a classification of audio/silence for the current segment. When the 99% classification criteria was met (i.e., 99% of the segments in the window disagreed with the current classification), said classification was flipped from audio to silence (and vice versa). This smoothing procedure was implemented to reduce sensitivity to transient acoustic fluctuations, breathing noise, and micro-segmentation artifacts that may otherwise inflate pause counts in automated pipelines. These hyperparameters (threshold, window length, and classification criteria) were selected on a validation subset of five participants and then fixed; the testing was performed on the test set, not the validation set, by comparing automated with hand-annotated pause counts for each participant and picture (Supplementary Table S14). Utterance boundaries were derived from the automatically segmented Whisper transcription output. Manual inspection was conducted to assess segmentation quality but did not alter the automatically generated utterance boundaries or utterance counts used in the automated analyses. Utterance-initial pauses were defined as silent intervals occurring immediately before utterance onset, whereas within-utterance pauses occurred between lexical constituents within the same utterance.

Using CrisperWhisper, we identified and extracted the number of fillers (e.g., hmm, err) per participant. We defined fillers as the word segments identified by CrisperWhisper as such

(represented with the tokens [UH] and [UM]) plus any segment containing the words “yani”, “ee”, “şey”, and “hmm”. These filler categories were selected based on common Turkish hesitation markers and prior linguistic work on filled pauses [67–69]. Given that Whisper is known to hallucinate in non-speech segments [70], we only counted a given segment as a valid dysfluency if the timestamp between its start and the end of its predecessor token was 250ms or higher.

To assess whether automated and hand-annotated pause analyses converged at the group level, all recordings were additionally coded by a trained human annotator blind to group membership. The annotator identified unfilled pauses and classified them according to utterance position (utterance-initial vs. within-utterance) and duration using the same classification criteria as the automated pipeline (see Table 2). Automated and hand-annotated pause measures were compared at the level of group-pattern effects rather than through exact one-to-one temporal alignment. Filler detection was not included in this comparison because fillers were identified using the automated token-based procedure described above. Hand-annotated pause data were available for 178 participants (51 FEP, 56 UHR, 37 FHP, and 34 NC), compared with 179 participants in the primary sample. Automated segmentation failed for four FHP participants (zero utterances despite speech confirmed by hand annotation). As the log-utterance offset is undefined for zero counts, they were excluded from the automated dysfluency and UID analyses (n = 175; 33 FHP). Acoustic-prosodic analyses included all 179 participants.

To assess the accuracy of Whisper-based utterance segmentation, we compared automatically generated segmentations with corresponding human transcriptions serving as the reference standard. Human transcriptions contained an average of 11.35 utterances per recording. Defining segmentation error as the number of automatically generated segments minus the number

of human-annotated segments, the 95% confidence interval for the mean error was -0.94 to -0.41 segments. Thus, relative to human annotation, the automated procedure showed a modest tendency to – on average – under-segment by less than one utterance. Accordingly, the two approaches should be interpreted as complementary rather than numerically equivalent measures of dysfluency.

The third set of features are based on the Uniform Information Density (UID) Hypothesis [45]. We followed the UID operationalization in [37], defining the UID of a given text with the formula

$$UID^{-1}(u) = \frac{1}{N}\sum_{n=1}^{N} s\,(u_n)^k (k > 1)$$

where $s(u_n)$ represents the surprisal of word u_n as given by a language model and *k* controls the strength of the super-linearity effect of surprisal on processing effort. Following results[37] we defined k=1.25 to be the value that best explains psychometric data. The UID proposes that language users prefer information content to be distributed relatively uniformly across a linguistic signal. Importantly, UID concerns the distribution of information rather than the absolute amount of information conveyed: utterances with different overall information densities may nevertheless show similar degrees of uniformity. We operationalized UID using two different methods:

1. Text-only UID: This measure characterized the distribution of information across the participant's speech transcription. Conditional probabilities of linguistic units given their context (i.e., surprisal values) were estimated using a Turkish-only GPT-2 model [71] and applying the super-linear operationalization (k=1.25) in [37]

2. Audio-enriched UID: This measure extended the text-only operationalization by incorporating word-level timing information. Specifically, rather than considering surprisal only per lexical unit, the audio-enriched measure divides the surprisal associated with each word by the time over which

that word is produced, thereby characterizing information distribution in relation to the temporal structure of speech. Information Density was calculated by combining transcriptions of participant's speech, GPT-2 word probabilities, and word-level timing information obtained using Whisper [62]. Since the Whisper transcriptions were not fully accurate, additional steps were required to align its output with the manual transcriptions. Manual and auto-generated transcriptions were aligned using the Python library minineedle to apply the Needleman-Wunsch algorithm [72].

The analysis in this section was performed using Python 3.11.2 and the libraries are as follows: lhotse (1.30.3), numpy (1.26.4), opensmile (2.5.1), pydub (0.25.1), statsmodels (0.14.5), torch (2.7.1), openai-whisper (20250625), and crisperwhisper (2.0).

Data Analysis

Statistical analysis was performed in R (R Core Team, 2023)[73] using RStudio (version 2021.09.1; Posit Team, 2023) [74]. Data preparation was performed using the *dplyr* package. To account for repeated observations within participants across multiple tasks, acoustic-prosodic features were analysed using generalized estimating equations (GEE) implemented in the *geepack* package. Model results were formatted and exported using *broom*, *purrr*, *gt*, *flextable*, and *officer*.

Unless otherwise specified, all models included group, age, gender, education, and task as covariates, and p-values were adjusted using the Benjamini-Hochberg false discovery rate (FDR) procedure.

Seven acoustic features were examined separately: F0 mean, F0 stddevNorm, shimmer mean, shimmer stddevNorm, HNR mean, HNR stddevNorm, and spectral slope (see Table 1). To assess distributional properties, we computed skewness using the moments package. Audio_enriched_uid showed moderate positive skew (skewness = 1.40). Accordingly, UID measures were analysed using GEE with a Gamma distribution and log link, without log-transforming the dependent variables (see UID analysis below). For each of the seven acoustic-prosodic outcomes, a separate GEE model was fitted using a Gaussian distribution with an identity link, accounting for repeated observations within participants through an exchangeable correlation structure:

Geeglm (Y ~ group + gender + age + education + task, id = patient, data = dat_gee,
family = gaussian (link = "identity"), corstr = "exchangeable")

Acoustic features were analysed on their original scale. No offset term (i.e., log number of utterances) was included, because acoustic-prosodic features are averaged signal properties rather than counts.

Because significant gender effects were observed for HNR and shimmer, we conducted gender-stratified sensitivity analyses restricted to UHR and NC participants. Separate GEE models were fitted for male-only and female-only subsamples using the same covariate structure as the primary acoustic models, excluding gender as a covariate. FDR correction was applied across three model specifications (the full UHR-NC sample, males only, and females only) separately for each acoustic outcome.

To analyze dysfluency patterns across group, we fitted GLMs with a Poisson distribution and log link. For each pause measure we fitted two models: one using hand-annotated data and the other using automatically detected pauses. Each model included group, age, gender, education, and task as predictors, with the logarithm of the number of utterances included as an offset. For the hand-annotated analyses, utterance counts were derived from manual segmentation, whereas the automated analyses used utterance counts obtained from Whisper-generated boundaries**,** preserving the independence of the automated analysis from manual utterance counts. We compared the direction, magnitude, and FDR-adjusted statistical significance of group coefficients across the two annotation methods (Table S4b; Figure S1). These comparisons assessed convergence in group-level patterns rather than formally testing method-by-group interactions. We then contrasted the results for both models using the emmeans R package and calculated their 95% confidence interval using the confint function of the stats R package.

Using GEE, we examined whether audio-enriched and text-only UID were associated with dysfluency and acoustic-prosodic features. Both UID measures were analysed using a Gamma distribution with a log link to accommodate positively skewed continuous outcomes. All models included group, age, education, gender, task, and number of utterances as covariates. The number of utterances was included to account for variation in utterance counts across participants and recordings.

UID-alone model:

Geeglm (audio_enriched_uid_num ~ group + age + education + gender + task + number of utterances, id = patient, data = dat_model_gamma, family = Gamma (link = “log”), corstr = “exchangeable”)

Relation between UID and dysfluency features:

Geeglm (audio_enriched_uid_num ~ group + unfilled_pause + age + education + gender + task + number of utterances, id = patient, data = dat_model, family = Gamma (link = "log"), corstr = "exchangeable")

Tables S7–S8 report audio-enriched UID analyses, while S9–S10 report text-only UID analyses.

Clinical measures included in the correlation analyses were selected according to the theoretical and empirical literature relevant to each speech-feature domain. For acoustic-prosodic features, analyses focused on negative symptom severity (BNSS) and psychosocial functioning (PSP), given prior evidence linking vocal acoustic characteristics, including pitch and voice-quality measures, to negative symptoms and functional outcomes. For dysfluency measures, we examined a broader set of clinical dimensions (i.e., negative symptoms, formal thought disorder, hallucinations, delusions, and psychosocial functioning) because pause measures differentiated by utterance position, as well as fillers, have been linked to language processes involved in conceptual planning, lexical retrieval, speech monitoring, and discourse organization.

Associations between speech features (acoustic-prosodic and dysfluency measures) and clinical measures were examined using participant-level partial Spearman correlations. Analyses were restricted to FEP and UHR participants. Acoustic-prosodic features were examined in relation to BNSS and PSP, whereas dysfluency features were additionally examined in relation to formal thought disorder, hallucinations, and delusions. Speech features were averaged across picture descriptions for each participant. Duration-based measures of unfilled pauses were excluded when they were collinear with the corresponding pause-count measures and therefore did not provide

independent interpretive value. For each speech-clinical association, the speech and clinical variables were rank-transformed and separately residualized for diagnostic group, age, education, and gender using linear regression. Pearson correlations between the resulting residuals were then computed to obtain partial Spearman correlation coefficients (ρ), with statistical inference accounting for the number of covariates. Missing observations were handled on a pairwise basis, such that participants with missing data for one clinical measure were retained in analyses of the remaining measures. P-values were adjusted using the Benjamini-Hochberg FDR procedure.

To examine medication effects, we conducted additional analyses restricted to FEP and UHR, as medication dose information was available only for these groups. For acoustic outcomes, GEE models with a Gaussian distribution and identity link were fitted, including group (UHR as reference), gender, age, education, task, and medication dose (i.e., chlorpromazine [CPZ] equivalent) as predictors. An exchangeable correlation structure accounted for repeated measures within participants, and Wald tests were used for inference.

For medication-adjusted dysfluency analyses, Poisson regression models with a log link were fitted separately for six unfilled-pause measures and filler production, restricted to FEP and UHR participants with available CPZ-equivalent dose information. Each model included group (UHR as reference), age, gender, education, task, and CPZ-equivalent dose as predictors, with the logarithm of the number of utterances included as an offset. Benjamini-Hochberg FDR correction was applied jointly across the seven FEP–UHR group comparisons, comprising six unfilled-pause outcomes and filler production.

Data availability

Raw audio recordings cannot be shared publicly because they contain potentially identifiable information. They are available from the corresponding author upon reasonable request and subject to approval by the Dokuz Eylül University ethics committee.

The data and analysis code is available at

https://osf.io/b246u/overview?view_only=e6b454affdd64a88be69eab77c8218aa

Conflicts of Interest

The authors declare no competing interests

Author contribution: Conceptualization: D.Ç, M.V, Methodology: D.Ç, Formal analysis: D.Ç, M.V; Software & Algorithm development: M.V; Investigation: D.Ç, M.V., KvH, Data curation: C.D., B.V., B.Y. & E.B., Writing-original draft: D.Ç & M.V. Writing-reviewing and editing: D.Ç, E.B., M.V, & KvH. Funding acquisition: D.Ç, KvH, & E.B. .

Funding

This work was supported by the University of Cologne, Excellence Research Program (PI: D.Ç.; KvH). The Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) -Project-ID 281511265 -SFB 1252 Prominence in Language -at the University of Cologne, Department of German Language and Literature I, Linguistics (PI: KvH)

References

1. Hinzen, W. & Palaniyappan, L. The 'L-factor': Language as a transdiagnostic dimension in psychopathology. *Progress in Neuro-Psychopharmacology & Biological Psychiatry* **131**, 1–8 (2024).
2. Agurto, C. *et al.* Analyzing acoustic and prosodic fluctuations in free speech to predict psychosis onset in high-risk youths. *Annual International Conference of the IEEE*

Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference **2020**, 5575–5579 (2020).

3. de Boer, J. N. *et al.* Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool. *Psychol Med* **53**, 1302–1312 (2023).
4. Zaher, F. *et al.* Speech markers to predict and prevent recurrent episodes of psychosis: A narrative overview and emerging opportunities. *Schizophrenia Research* **266**, 205–215 (2024).
5. Melshin, G. *et al.* Taking a look at your speech: identifying diagnostic status and negative symptoms of psychosis using convolutional neural networks. *NPP—Digital Psychiatry and Neuroscience* **3**, 19 (2025).
6. Tang, S. X. *et al.* Automated speech and language markers of longitudinal changes in psychosis symptoms. *NPP—Digit Psychiatry Neurosci* **3**, 13 (2025).
7. Worthington, M. *et al.* Measurement of schizophrenia symptoms through speech analysis from PANSS interview recordings. *Front Psychiatry* **16**, 1571647 (2025).
8. Hitczenko, K., Mittal, V. A. & Goldrick, M. Understanding Language Abnormalities and Associated Clinical Markers in Psychosis: The Promise of Computational Methods. *Schizophr Bull* **47**, 344–362 (2021).
9. Parola, A., Simonsen, A., Bliksted, V. & Fusaroli, R. Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis. *Schizophrenia Research* **216**, 24–40 (2020).
10. Parola, A. *et al.* Voice Patterns as Markers of Schizophrenia: Building a Cumulative Generalizable Approach Via a Cross-Linguistic and Meta-analysis Based Investigation. *Schizophr Bull* **49**, S125–S141 (2023).

11. Zhao, Q. *et al.* Vocal acoustic features may be objective biomarkers of negative symptoms in schizophrenia: A cross-sectional study. *Schizophrenia Research* **250**, 180–185 (2022).
12. Chuang, C.-Y. *et al.* Multimodal Assessment of Schizophrenia Symptom Severity From Linguistic, Acoustic and Visual Cues. *IEEE Trans. Neural Syst. Rehabil. Eng.* **31**, 3469–3479 (2023).
13. Voleti, R. *et al.* Language Analytics for Assessment of Mental Health Status and Functional Competency. *Schizophrenia Bulletin* **49**, S183–S195 (2023).
14. Spilka, M. J. *et al.* Identification of robust speech-based markers of negative symptom severity in schizophrenia spectrum disorders. *Schizophr* https://doi.org/10.1038/s41537-026-00788-1 (2026) doi:10.1038/s41537-026-00788-1.
15. Çokal, D. *et al.* What is the retest reliability of computationally extractable speech and language markers? *Computer Speech & Language* **100**, 101981 (2026).
16. Alpert, M., Kotsaftis, A. & Pouget, E. R. At issue: speech fluency and schizophrenic negative signs. *Schizophr Bull* **23**, 171–177 (1997).
17. Rapcan, V. *et al.* Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia. *Med Eng Phys* **32**, 1074–1079 (2010).
18. Voppel, A. E., de Boer, J. N., Brederoo, S. G., Schnack, H. G. & Sommer, I. E. C. Semantic and Acoustic Markers in Schizophrenia-Spectrum Disorders: A Combinatory Machine Learning Approach. *Schizophrenia Bulletin* **49**, S163–S171 (2023).
19. Mota, N. B., Copelli, M. & Ribeiro, S. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *NPJ Schizophr* **3**, 18 (2017).

20. Cohen, A. S. *et al.* High Predictive Accuracy of Negative Schizotypy With Acoustic Measures. *Clinical psychological science : a journal of the Association for Psychological Science* **10**, 310–323 (2022).
21. Bianciardi, B. *et al.* Investigating temporal and prosodic markers in clinical high-risk for psychosis participants using automated acoustic analysis. *Early Intervention in Psychiatry* **17**, 327–330 (2023).
22. Lickley, R. J. Fluency and Disfluency. 445–474 (2015) doi:10.1002/9781118584156.ch20.
23. Ferreira, F. Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language* **30**, 210–233 (1991).
24. Goldman-Eisler, F. Psycholinguistics; experiments in spontaneous speech: Frieda Goldman Eisler Academic Press, London & New York 1968. viii, 169 pp. https://www.sciencedirect.com/science/article/pii/0024384170900288 (1968).
25. Kircher, T. T. J., Brammer, M. J., Levelt, W., Bartels, M. & McGuire, P. K. Pausing for thought: engagement of left temporal cortex during pauses in speech. *NeuroImage* **21**, 84–90 (2004).
26. Levelt, W. J. M. *Speaking: From Intention to Articulation*. (The MIT Press, Cambridge, MA, US, 1989).
27. Beattie, G. W. & Butterworth, B. L. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech* **22**, 201–211 (1979).
28. Clark, H. Using uh and um in spontaneous speaking. *Cognition* **84**, 73–111 (2002).
29. Corley, M. & Stewart, O. W. Hesitation Disfluencies in Spontaneous Speech: The Meaning of um. *Language and Linguistics Compass* **2**, 589–602 (2008).

30. Ferreira, F. Prosody and performance in language production. *Language and Cognitive Processes* **22**, 1151–1177 (2007).
31. Çokal, D. *et al.* Disturbing the rhythm of thought: Speech pausing patterns in schizophrenia, with and without formal thought disorder. *PLOS ONE* **14**, e0217404 (2019).
32. Zimmerer, V. *et al.* The relationship between language disorder and thought disorder: Comparing micro- and macrostructure of spoken narratives of people with aphasia and people with schizophrenia. *Cortex* **195**, 81–95 (2026).
33. Parola, A. *et al.* Speech disturbances in schizophrenia: Assessing cross-linguistic generalizability of NLP automated measures of coherence. *Schizophrenia Research* **259**, 59–70 (2023).
34. Howes, C., Lavelle, M., Healey, P. G. T., Hough, J. & McCabe, R. Disfluencies in dialogues with patients with schizophrenia. in *CogSci* (2017).
35. Matsumoto, K. *et al.* Frequency and Neural Correlates of Pauses in Patients with Formal Thought Disorder. *Frontiers in Psychiatry* **4**, (2013).
36. Jaeger, T. F. & Levy, R. P. *Speakers Optimize Information Density Through Syntactic Reduction*. (MIT Press, 2006).
37. Meister, C. *et al.* Revisiting the Uniform Information Density Hypothesis. in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (eds Moens, M.-F., Huang, X., Specia, L. & Yih, S. W.) 963–980 (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021). doi:10.18653/v1/2021.emnlp-main.74.
38. Corcoran, C. M. *et al.* Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry* **17**, 67–75 (2018).

39. Elvevåg, B., Foltz, P. W., Weinberger, D. R. & Goldberg, T. E. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research* **93**, 304–316 (2007).
40. Bedi, G. *et al.* Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophr* **1**, 15030 (2015).
41. Çokal, D. *et al.* Three dimensions of speech coherence in people with early psychosis and their family members. *Schizophr* **12**, 2 (2025).
42. Çokal, D. *et al.* Automated detection of referential features in schizophrenic speech using large language models. *Neuropsychologia* **230**, 109505 (2026).
43. Palaniyappan, L. Dissecting the neurobiology of linguistic disorganisation and impoverishment in schizophrenia. *Seminars in Cell & Developmental Biology* **129**, 47–60 (2022).
44. Ditman, T. & Kuperberg, G. R. Building coherence: A framework for exploring the breakdown of links across clause boundaries in schizophrenia. *Journal of Neurolinguistics* **23**, 254–269 (2010).
45. Levy, R. & Jaeger, T. F. Speakers optimize information density through syntactic reduction. https://doi.org/10.7551/mitpress/7503.003.0111 (2007) doi:10.7551/mitpress/7503.003.0111.
46. Olah, J. *et al.* Online speech assessment of the psychotic spectrum: Exploring the relationship between overlapping acoustic markers of schizotypy, depression and anxiety. *Schizophr Res* **259**, 11–19 (2023).
47. Butterworth, B. Hesitation and the production of verbal paraphasias and neologisms in jargon aphasia. *Brain and Language* **8**, 133–161 (1979).

48. Brockmann, M., Drinnan, M. J., Storck, C. & Carding, P. N. Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *J Voice* **25**, 44–53 (2011).
49. Brockmann, M., Storck, C., Carding, P. N. & Drinnan, M. J. Voice Loudness and Gender Effects on Jitter and Shimmer in Healthy Adults. *Journal of Speech, Language, and Hearing Research* **51**, 1152–1160 (2008).
50. Hillenbrand, J. M. & Clark, M. J. The role of f 0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics* **71**, 1150–1166 (2009).
51. Martinho, D. H. D. C. & Constantini, A. C. Acoustic Measures According to Speaker Gender Identity: Differences and Correlation With Vocal Satisfaction. *Journal of Voice* S0892199724004442 (2025) doi:10.1016/j.jvoice.2024.12.016.
52. First, M., Spitzer, R. L., Gibbon, M. & Williams, J. B. W. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition (SCID-I/P)*. (New York: Biometrics Research, New York State Psychiatric Institute., 2002).
53. Andreasen, N. C. *The Scale for the Assessment of Positive Symptoms (SAPS).* (The University of Iowa Press, Iowa City, IA., 1984).
54. Kirkpatrick, B. *et al.* The Brief Negative Symptom Scale: Psychometric Properties. *Schizophrenia Bulletin* **37**, 300–305 (2011).
55. Polat Nazlı, I. *et al.* Validation of Turkish version of brief negative symptom scale. *International Journal of Psychiatry in Clinical Practice* **20**, 265–271 (2016).
56. Weigel, L. *et al.* The Brief negative Symptom Scale (BNSS): a systematic review of measurement properties. *Schizophrenia* **9**, 1–9 (2023).

57. Leucht, S. *et al.* Dose Equivalents for Second-Generation Antipsychotics: The Minimum Effective Dose Method. *Schizophrenia Bulletin* **40**, 314–326 (2014).
58. Aydemir, Ö. *et al.* The validation of Turkish version of personal and social performance scale (PSP) Bireysel ve sosyal performans ölçeği'nin Türkçe sürümünün geçerlilik ve güvenilirlik çalişmasi. *Klinik Psikofarmakoloji Bulteni* **19**, (2009).
59. Morosini, P.-L., Magliano, L., Brambilla, L., Ugolini, S. & Pioli, R. Personal and Social Performance Scale. https://doi.org/10.1037/t38751-000 (2000).
60. Boersma, P., & Weenink, D. (1999-2013). Praat Doing phonetics by computer. Software. - References - Scientific Research Publishing.
61. Murray, H. A. *Thematic Apperception Test.* (Harvard University Press, Cambridge, MA, US, 1943).
62. Radford, A. *et al.* Robust speech recognition via large-scale weak supervision. in *Proceedings of the 40th International Conference on Machine Learning* vol. 202 28492–28518 (JMLR.org, Honolulu, Hawaii, USA, 2023).
63. Zusag, M., Wagner, L. & Thallinger, B. CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions. in 1265–1269 (2024). doi:10.21437/Interspeech.2024-731.
64. Çokal, D. *et al.* Referential noun phrases distribute differently in Turkish speakers with schizophrenia. *Schizophrenia Research* **259**, 104–110 (2023).
65. Eyben, F. *et al.* The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* **7**, 190–202 (2016).
66. Robert, J., Webbie, M. & others. Pydub. (2018).

67. Ruhi, Ş. (Editorial) Interactional Markers in Turkish: A corpus-based perspective. *Mersin Üniversitesi Dil ve Edebiyat Dergisi* **10**, 1–7 (2013).
68. Özbek, N. Yani, işte, ya: interactional markers of Turkish. *Studies on Turkish and Turkic Languages* 393–401 (2000).
69. Erkan, Y. A Pragmatic Analysis of Turkish particles: yani, işte, and şey. (2004).
70. Wang, Y., Alhmoud, A., Alsahly, S., Alqurishi, M. & Ravanelli, M. Calm-Whisper: Reduce Whisper Hallucination On Non-Speech By Calming Crazy Heads Down. Preprint at https://doi.org/10.48550/arXiv.2505.12969 (2025).
71. Kesgin, H. T. *et al.* Introducing cosmosGPT: Monolingual Training for Turkish Language Models. Preprint at https://doi.org/10.48550/arXiv.2404.17336 (2024).
72. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443–453 (1970).
73. R: The R Project for Statistical Computing. https://www.r-project.org/.
74. Posit | Data Science Platform for Enterprise Teams. https://posit.co/.

Figure 1. Dysfluency measures across groups using automated detection.

Note: Models include group as the predictor and control for age, gender, education, and task. Offset of log (number of utterances) was added to the model to control differences in utterance numbers across participants. Colours denote participant groups (NC = neurotypical controls [blue], FEP = first-episode psychosis [red], FHP = family history of psychosis [green], UHR = ultra-high risk [black]).

Figure 2. Heatmap of partial Spearman correlation matrices depicting associations between acoustic–prosodic features and clinical variables. Values represent partial Spearman correlation coefficients (ρ). P-values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) procedure. F0 SD, Shimmer SD, and HNR SD refer to the corresponding stddevNorm features extracted using openSMILE, rather than conventional raw standard deviations.

Figure 3. Heatmap of partial Spearman correlations between dysfluency measures and clinical measures (BNSS, formal thought disorder, hallucinations, delusions, and PSP) in the combined FEP and UHR sample. Correlations control for diagnostic group, age, education, and gender. Values represent partial Spearman correlation coefficients (ρ). *P*-values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) procedure.

Figure 4. Schematic of the workflow of the study

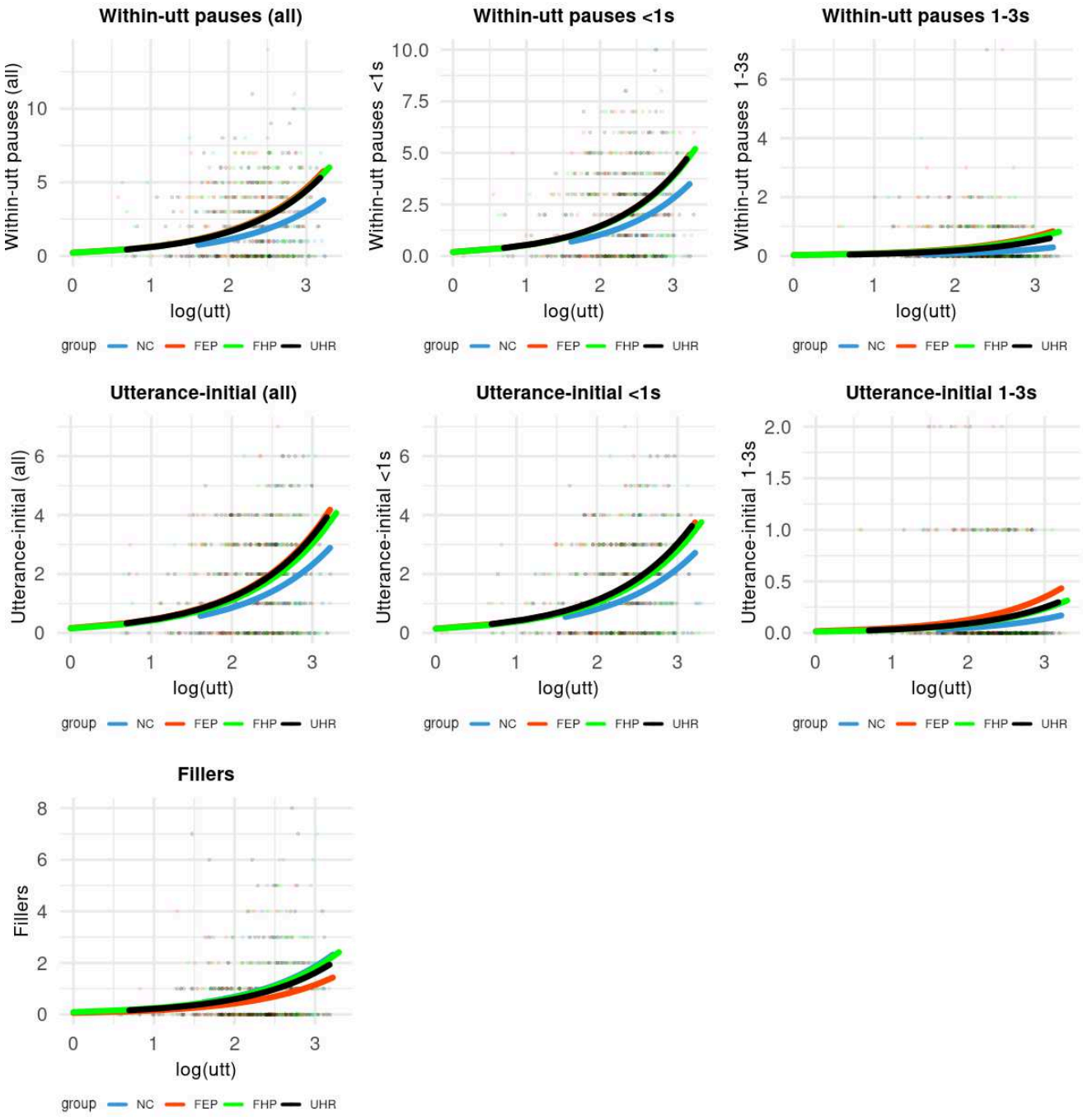

Figure 1. Dysfluency measures across groups using automated detection.

Note: Models include group as the predictor and control for age, gender, education, and task. Offset of log (number of utterances) was added to the model to control differences in utterance numbers across participants. Colours denote participant groups (NC = neurotypical controls [blue], FEP =

first-episode psychosis [red], FHP = family history of psychosis [green], UHR = ultra-high risk [black]).

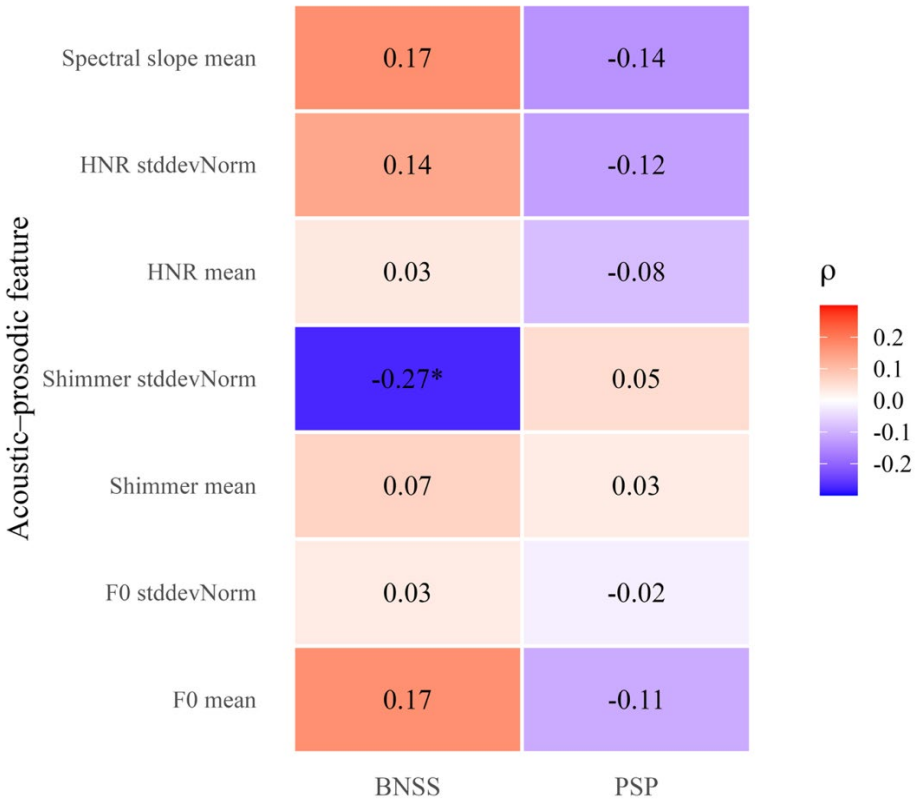

Figure 2. Heatmap of partial Spearman correlation matrices depicting associations between acoustic–prosodic features and clinical variables. Values represent partial Spearman correlation coefficients (ρ). P-values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) procedure. F0 SD, Shimmer SD, and HNR SD refer to the corresponding stddevNorm features extracted using openSMILE, rather than conventional raw standard deviations.

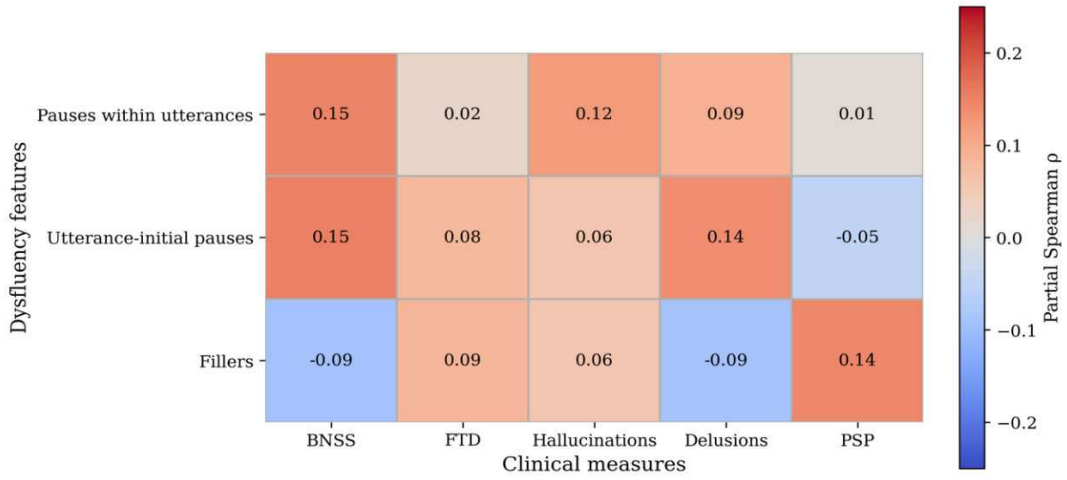

Figure 3. Heatmap of partial Spearman correlations between dysfluency measures and clinical measures (BNSS, formal thought disorder, hallucinations, delusions, and PSP) in the combined FEP and UHR sample. Correlations control for diagnostic group, age, education, and gender. Values represent partial Spearman correlation coefficients (ρ). *P*-values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) procedure.

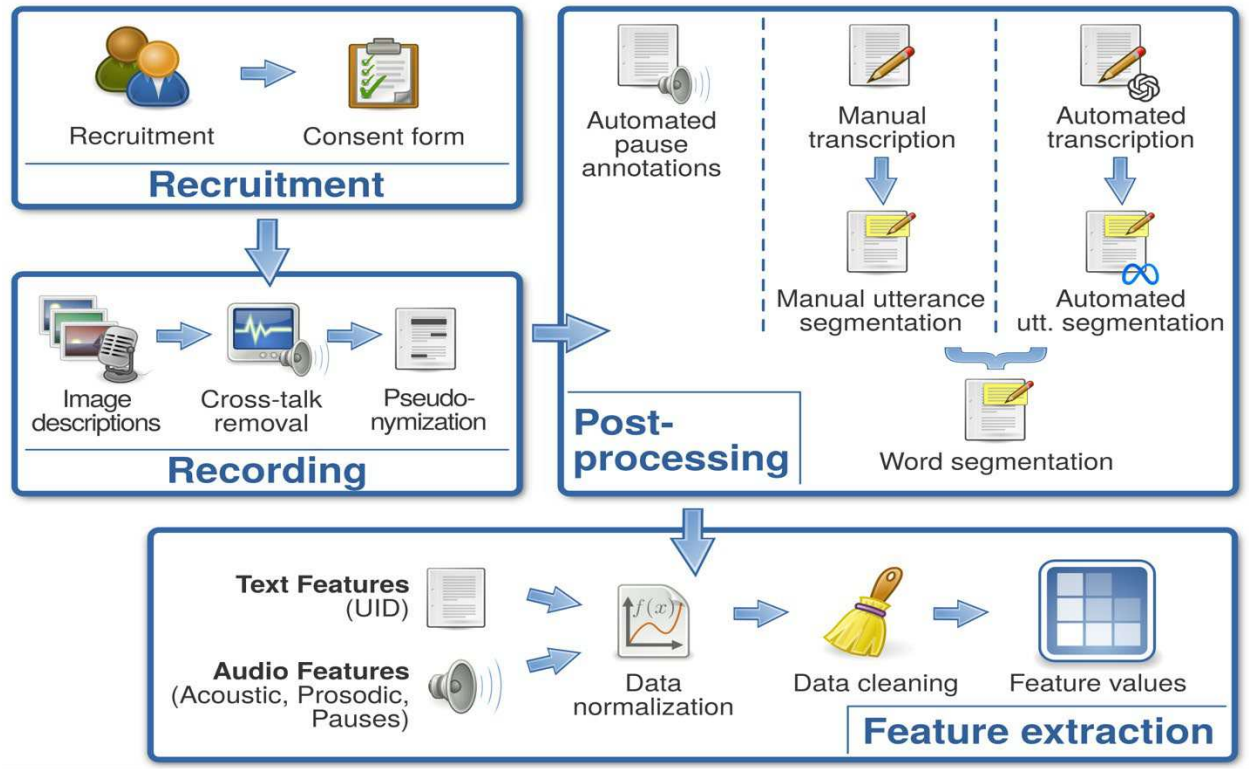

Figure 4. Schematic of the workflow of the study

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterials.docx
- Supplementarymaterials.doc
- Supplementarymaterials.doc