

Vocal acoustic parameters as objective biomarkers of major depressive disorder and its severity: a case-control study using WhatsApp audio messages

Luiza Braconi Grilo

Faculdade de Ciências Médicas da Santa Casa de São Paulo

Victor H. O. Otani

Faculdade de Ciências Médicas da Santa Casa de São Paulo

Felipe O. Aguiar

Faculdade de Ciências Médicas da Santa Casa de São Paulo

Hudson S. Buck

Eólas Applied Neuroscience

Lucas M. Marques

`lucas.murrins@fcmsantacasasp.edu.br`

Faculdade de Ciências Médicas da Santa Casa de São Paulo

Ricardo R. Uchida

Faculdade de Ciências Médicas da Santa Casa de São Paulo

Research Article

Keywords: vocal biomarkers, major depressive disorder, formant, shimmer, acoustic analysis, WhatsApp

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10843800/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Depression lacks validated objective biomarkers in routine clinical practice. Voice is a non-invasive signal shaped by the neuromotor and psychomotor systems disrupted in depressive states, making it a promising candidate for objective assessment. We investigated whether acoustic parameters extracted from WhatsApp audio messages distinguish adults with major depressive disorder (MDD) from healthy controls and correlate with clinician-rated severity on the Montgomery-Åsberg Depression Rating Scale (MADRS). Seventy-two participants (31 MDD, 41 controls) sent WhatsApp audio in two tasks: structured counting (1-to-10) and spontaneous narrative (last week). Features were extracted with Praat (formants, shimmer, jitter, harmonic-to-noise ratio) and PyAudioAnalysis (68 features including MFCCs, spectral entropy, zero-crossing rate, and energy), and group comparisons were conducted using Generalised Estimating Equations with false discovery rate correction. Formant 1 was substantially reduced in MDD across both tasks (Cohen's $d = -0.95$ to -1.03 , $q < 0.003$). Shimmer was markedly elevated in the narrative task ($d = +1.08$ to $+1.14$, $q < 0.001$), and spectral entropy and zero-crossing rate were reduced (d up to -1.33 , $q < 0.001$). MADRS severity correlated most strongly with chroma-11 maximum (Pearson $r = +0.73$, $q < 0.001$) and energy entropy interquartile range ($r = +0.67$, $q < 0.001$). The spontaneous narrative task yielded 62 significant features versus 25 in the counting task. WhatsApp audio messages yield statistically robust vocal biomarkers of MDD detectable in naturalistic recordings on personal smartphones, with formant 1, shimmer, and spectral entropy emerging as the most promising candidates for scalable, non-invasive depression assessment.

1 Introduction

Major depressive disorder (MDD) is a leading cause of global disability, affecting approximately 280 million people worldwide and representing the highest prevalence burden in Latin America, where Brazil alone accounts for 5.8% of the affected population (WHO 2023; Hintz et al. 2023). Despite decades of neurobiological research, diagnosis continues to rely almost exclusively on clinical evaluation and psychometric scales, which are susceptible to recall bias, clinician heterogeneity, and the pervasive social stigma that suppresses symptom disclosure (Slavich et al. 2019). The absence of validated, objective biomarkers represents one of the most consequential gaps in contemporary psychiatry, delaying accurate diagnosis, complicating differentiation from bipolar depression, and limiting the objective monitoring of therapeutic response (Malhi and Mann 2018).

The human voice has long been recognised as a window onto mental state. Speech production engages more than 100 muscles coordinated by cortical motor, limbic, and autonomic circuits, all of which are compromised in MDD (Sobin and Sackeim 1997; Caligiuri and Ellwanger 2000). Psychomotor retardation, one of the most reproducible neurobiological signatures of depression, manifests as altered laryngeal control, reduced vocal fold tension, slower articulatory movement, and dampened prosodic variability (Quatieri and Malyska 2012; Cannizzaro et al. 2004). These alterations converge in measurable acoustic features: reduced fundamental frequency (F0) variability, lower first formant frequency (F1), elevated shimmer and jitter, decreased harmonic-to-noise ratio (HNR), reduced spectral entropy, and flattened

energy dynamics (Menne et al. 2024; Low et al. 2020). The mechanistic basis involves abnormal basal ganglia function, which modulates laryngeal motor coordination, and alterations in the hypothalamic-pituitary-adrenal axis and the autonomic nervous system that collectively affect subglottal pressure and vocal fold adduction (Cummins et al. 2015; Van Puyvelde et al. 2018).

A growing literature confirms that vocal acoustic markers can discriminate individuals with depression from healthy controls with clinically meaningful accuracy. A systematic review by Low et al. (Low et al. 2020) demonstrated that automated speech analysis achieved sensitivity and specificity values broadly comparable to conventional rating scales. More recent work has extended this evidence to naturalistic speech collected via smartphone applications, overcoming the ecological validity limitation of laboratory-based paradigms (Taguchi et al. 2018; Otani et al. 2026). In a parallel line of investigation, our research group demonstrated that machine learning models applied to WhatsApp audio messages could classify depressive profiles with an area under the receiver operating characteristic curve exceeding 0.90 for female voices (Otani et al. 2026), and that voice-affect coupling is selectively modulated by depressive symptom burden (Veiga et al. 2026). A systematic review and meta-analysis from the same group confirmed that fundamental frequency is significantly elevated under acute stress, particularly during spontaneous speech and in female participants, though publication bias warrants cautious interpretation (Veiga et al. 2025).

Despite this encouraging evidence base, critical methodological gaps remain. The overwhelming majority of published studies employ machine learning classification, which optimises discrimination but provides limited insight into which specific acoustic parameters drive the effect and which correlate with symptom severity rather than merely categorical diagnosis. Equally, most studies employ either scripted laboratory speech or uncontrolled naturalistic recordings without systematic comparison of structured versus spontaneous task conditions. Finally, few studies have applied rigorous inferential statistical frameworks that account for the hierarchical data structure inherent when multiple audio recordings are obtained from the same individual.

The present study addresses these gaps by applying Generalised Estimating Equations (GEE) with false discovery rate (FDR) correction to acoustic features extracted from WhatsApp audio messages in a diagnostically validated case-control sample. Two speech conditions were compared: a structured numerical counting task (1-to-10) and a semi-structured spontaneous narrative (last week). The study pursued two objectives: (i) to identify specific vocal acoustic parameters that distinguish MDD from healthy controls with statistically robust effects; and (ii) to determine which parameters correlate with MADRS severity. By deploying statistical inference rather than classification, this study positions vocal biomarker research within a framework accessible to clinicians and neuroscientists alike (Malhi and Mann 2018; Menne et al. 2024).

2 Methods

2.1 Study design and participants

This cross-sectional case-control study was conducted at the Mental Health Department of the Santa Casa de São Paulo School of Medical Sciences between May 2023 and July 2024. Ethical approval was granted by the Research Ethics Committee of the University of Mogi das Cruzes (CAAE: 67711023.1.0000.5497), and all participants provided written informed consent in accordance with the Declaration of Helsinki.

Adults aged 18 years or older with proficiency in Brazilian Portuguese were eligible. The MDD group comprised outpatients meeting DSM-5 criteria for a current major depressive episode, confirmed by an experienced psychiatrist using the Mini International Neuropsychiatric Interview (MINI) (Sheehan et al. 1998). Depression severity was quantified with the MADRS (Montgomery and Åsberg 1979), with a supplementary hypersomnia item alongside the standard ten items. Exclusion criteria for both groups included: current manic or psychotic episode, substance use disorder (excluding tobacco), major neurological disorder, active upper respiratory infection, laryngeal or vocal fold pathology, head and neck neoplasm, and use of anabolic steroids. Healthy volunteers with no self-reported history of psychiatric treatment and no current depressive symptoms on the MINI were recruited through public advertisements. Smartphone type was not controlled, preserving ecological validity.

2.2 Audio acquisition and speech tasks

Each participant submitted WhatsApp audio messages for two tasks: (i) a spontaneous narrative describing their previous week (*narrative* task), with no imposed duration; and (ii) structured counting from 1 to 10 at an approximate one-second rhythm (*counting* task). All messages were recorded using the participant's personal smartphone in their natural environment; no acoustic booth or external microphone was imposed.

2.3 Audio preprocessing

Audio files received in OGG format were converted to WAV with a 44.1 kHz sampling rate and mono channel configuration. A silence-removal algorithm (pydub library, version 0.25.1) was applied with a threshold of -40 dB and a minimum silent passage duration of 100 ms, retaining only active speech frames. No additional noise-reduction or normalisation was applied, preserving ecological representativeness.

2.4 Acoustic feature extraction

Feature extraction followed two independent computational pipelines. The first used Praat (version 6.3) to extract clinically established phonation parameters: F0, F1, shimmer (APQ3, APQ5, local), jitter (PPQ5, RAP), and HNR. For each parameter, eight distributional statistics were computed (mean, median, SD, minimum, maximum, IQR, Q25, Q75). The second pipeline used the PyAudioAnalysis library (Giannakopoulos 2015) to extract 34 base features per 50-ms frame with a 25-ms step: temporal features (zero-crossing rate, ZCR; energy), spectral features (centroid, spread, entropy, flux, rolloff), Mel-frequency cepstral coefficients (MFCCs 1–13), and chroma coefficients (chroma 1–12 plus chroma SD). First-order temporal derivatives (Δ features) were calculated for all 34 base features, yielding 68 features

per frame. Eight distributional statistics were computed per feature across frames, yielding 536 candidate features per participant per task. All features were z-score standardised prior to analysis.

2.5 Statistical analysis

Group comparisons were conducted using Generalised Estimating Equations (GEE) with an exchangeable correlation structure (STATA 17.0), which appropriately accounts for within-participant correlation across multiple recordings. Separate models were estimated for each task and library. In each model, the standardised feature value was the outcome; group membership (MDD versus control) was the predictor, with age and sex as covariates. Effect sizes are reported as standardised GEE coefficients, corresponding to Cohen's d under this parameterisation. Statistical significance was assessed after Benjamini-Hochberg FDR correction (Benjamini and Hochberg 1995) within each library-task combination; $q < 0.05$ was the threshold for feature retention.

Dimensional associations between acoustic features and MADRS total scores were evaluated within the MDD group using Spearman r_s and Pearson r coefficients with FDR correction. Only features for which both estimators converged in direction are reported as primary findings. All analyses were conducted in STATA 17.0, with Python 3.12 used for preprocessing and feature extraction.

3 Results

3.1 Participant characteristics

Seventy-two adults participated (31 MDD, 41 controls). Groups did not differ in age (MDD: 34.2 ± 11.3 years; controls: 35.2 ± 14.0 years; $p = 0.74$) or sex distribution (MDD: 58.1% female; controls: 61.0% female; $p = 0.79$). Among MDD participants, mean MADRS total was 29.6 (SD = 8.7; range 17–43): 16.1% mild (7–19), 54.8% moderate (20–34), and 29.0% severe (35–60). Participant characteristics are summarised in Table 1.

Table 1
Participant characteristics stratified by diagnostic group

Variable	MDD group (n = 31)	Control group (n = 41)
Age, years – mean (SD)	34.2 (11.3)	35.2 (14.0)
Female sex – n (%)	18 (58.1)	25 (61.0)
MADRS total – mean (SD)	29.6 (8.7)	N/A
MADRS ≥ 20 – n (%)	26 (83.9)	N/A
MADRS classification		
Mild (7–19)	5 (16.1)	–
Moderate (20–34)	17 (54.8)	–
Severe (35–60)	9 (29.0)	–
Audio task – 1-to-10 (n)	31	41
Audio task – last week (n)	31	41

MDD major depressive disorder, MADRS Montgomery-Åsberg Depression Rating Scale, SD standard deviation, N/A not applicable

3.2 Group differences in vocal acoustics

The complete set of features surviving FDR correction in GEE group comparison models is shown in Table 2 for a representative selection. All significant features are listed in Online Resource 1.

Praat features. F1 was reduced in MDD across all distributional statistics and both tasks, with the largest effects in the *narrative* task (F1 mean: $d = -1.03$, $q < 0.001$; F1 Q75: $d = -1.00$, $q < 0.001$; F1 IQR: $d = -0.94$, $q < 0.001$). In the *counting* task, F1 reductions were also robust (F1 mean: $d = -0.95$, $q = 0.002$; F1 Q75: $d = -0.88$, $q = 0.003$; F1 median: $d = -0.73$, $q = 0.022$). Shimmer indices emerged exclusively in the *narrative* task, where all three measures were substantially elevated in MDD (APQ3: $d = +1.14$; local: $d = +1.09$; APQ5: $d = +1.08$; all $q < 0.001$). Jitter (PPQ5 and RAP) also showed significant elevation in the narrative condition (PPQ5: $d = +0.68$, $q = 0.024$; RAP: $d = +0.65$, $q = 0.035$). F0 showed a tendency towards lower values in MDD but did not survive FDR correction in either task, consistent with the heterogeneity in prior meta-analytic work (Veiga et al. 2025).

PyAudioAnalysis features. In the *counting* task, 25 features survived FDR correction. Vocal energy was markedly increased in MDD (energy SD: $d = +1.00$, $q = 0.002$; energy maximum: $d = +0.94$, $q = 0.009$), along with MFCC-2 statistics (median: $d = +0.82$; mean: $d = +0.81$), spectral spread (Q75: $d = +0.85$; mean: $d = +0.76$), and reduced MFCC-1 IQR ($d = -0.73$) and MFCC-10 maximum ($d = -0.70$). In the *narrative* task, 62 features survived FDR correction. Spectral entropy was markedly reduced in MDD (mean: $d = -1.19$, $q < 0.001$), as was spectral rolloff (mean: $d = -1.05$, $q < 0.001$), ZCR (mean: $d = -0.96$,

$q < 0.001$), and MFCC-1 IQR ($d = -1.33$, $q < 0.001$), while MFCC-2 Q25 was elevated ($d = +1.18$, $q < 0.001$). The approximately 2.5-fold increase in significant features in the narrative relative to the counting task indicates that spontaneous speech elicits substantially richer acoustic differentiation.

Table 2

Selected acoustic features significantly associated with MDD in GEE models (FDR-corrected, $q < 0.05$)

Feature	Library	Task	Controls mean (SD)	MDD mean (SD)	Cohen's d	q (FDR)
Formant 1 – mean	<i>Praat</i>	1-to-10	664.8 (196.1)	558.7 (161.8)	–0.95	0.002
Formant 1 – Q75	<i>Praat</i>	1-to-10	877.1 (254.3)	679.2 (198.2)	–0.88	0.003
Formant 1 – IQR	<i>Praat</i>	1-to-10	505.9 (248.6)	319.4 (172.1)	–0.83	0.005
Formant 1 – std	<i>Praat</i>	1-to-10	390.9 (182.4)	298.2 (144.1)	–0.78	0.014
Formant 1 – median	<i>Praat</i>	1-to-10	544.9 (168.3)	458.4 (142.2)	–0.73	0.022
Formant 1 – mean	<i>Praat</i>	Last week	570.6 (175.4)	506.5 (154.2)	–1.03	< 0.001
Formant 1 – Q75	<i>Praat</i>	Last week	682.3 (210.8)	561.1 (173.4)	–1.00	< 0.001
Shimmer APQ3	<i>Praat</i>	Last week	0.041 (0.018)	0.058 (0.023)	+ 1.14	< 0.001
Shimmer local	<i>Praat</i>	Last week	0.099 (0.039)	0.130 (0.047)	+ 1.09	< 0.001
Shimmer APQ5	<i>Praat</i>	Last week	0.054 (0.021)	0.077 (0.028)	+ 1.08	< 0.001
Jitter PPQ5	<i>Praat</i>	Last week	0.011 (0.005)	0.013 (0.006)	+ 0.68	0.024
Jitter RAP	<i>Praat</i>	Last week	0.011 (0.005)	0.013 (0.006)	+ 0.65	0.035
Energy – SD	<i>PyAudio</i>	1-to-10	0.025 (0.009)	0.034 (0.011)	+ 1.00	0.002
Energy – max	<i>PyAudio</i>	1-to-10	0.133 (0.078)	0.222 (0.129)	+ 0.94	0.009
Spectral spread – Q75	<i>PyAudio</i>	1-to-10	0.177 (0.052)	0.187 (0.048)	+ 0.85	0.010
MFCC-2 – median	<i>PyAudio</i>	1-to-10	2.902 (0.741)	3.193 (0.754)	+ 0.82	0.012
MFCC-2 – mean	<i>PyAudio</i>	1-to-10	2.751 (0.712)	3.082 (0.741)	+ 0.81	0.016

Feature	Library	Task	Controls mean (SD)	MDD mean (SD)	Cohen's d	q (FDR)
Spectral spread – mean	PyAudio	1-to-10	0.141 (0.039)	0.158 (0.044)	+ 0.76	0.028
MFCC-1 – IQR	PyAudio	1-to-10	4.775 (1.282)	3.777 (1.149)	–0.73	0.022
MFCC-10 – max	PyAudio	1-to-10	0.700 (0.248)	0.575 (0.221)	–0.70	0.015
MFCC-2 – Q25	PyAudio	Last week	2.458 (0.612)	2.873 (0.638)	+ 1.18	< 0.001
MFCC-1 – IQR	PyAudio	Last week	3.773 (1.204)	2.820 (1.073)	–1.33	< 0.001
Spectral entropy – mean	PyAudio	Last week	0.222 (0.072)	0.140 (0.052)	–1.19	< 0.001
Spectral rolloff – mean	PyAudio	Last week	0.052 (0.018)	0.039 (0.014)	–1.05	< 0.001
Spectral spread – median	PyAudio	Last week	0.146 (0.041)	0.158 (0.044)	+ 1.05	< 0.001
ZCR – mean	PyAudio	Last week	0.036 (0.011)	0.028 (0.009)	–0.96	< 0.001

Values represent distributional statistics computed per participant across audio frames. Cohen's *d* derived from the standardised GEE regression coefficient; negative *d* indicates lower values in MDD relative to controls. MDD major depressive disorder, IQR interquartile range, MFCC Mel-frequency cepstral coefficient, ZCR zero-crossing rate, FDR false discovery rate, SD standard deviation. Complete feature list in Online Resource 1

3.3 Associations with depression severity

Correlations between acoustic features and MADRS total scores were examined within the MDD group (*n* = 31). Eleven Praat features and 28 PyAudioAnalysis features survived FDR correction (Table 3; complete list in Online Resource 2).

Praat severity correlates. Shimmer indices showed negative correlations with MADRS, the strongest being shimmer local (Spearman *r_S* = – 0.61, Pearson *r* = – 0.64, *q* = 0.002). Shimmer APQ3 and APQ5 yielded comparable magnitudes (*r* range – 0.56 to – 0.62, all *q* < 0.01). Minimum F0 correlated negatively with severity (Spearman *r_S* = – 0.59, Pearson *r* = – 0.53, *q* = 0.027). HNR showed a positive association (Spearman *r_S* = + 0.47, Pearson *r* = + 0.49, *q* = 0.037).

PyAudioAnalysis severity correlates. The strongest individual predictor was chroma-11 maximum (Spearman *r_S* = + 0.65, Pearson *r* = + 0.73, *q* = 1.3×10^{-7}). Notably, this is a distinct distributional statistic from chroma-11 mean reported as a group-comparison feature in Table 2 (*d* = + 0.82), indicating that

maximum tonal energy in the 11th chroma band specifically indexes severity, while its mean value differentiates groups. Energy entropy IQR was also strongly positively correlated (Spearman $r_s = +0.63$, Pearson $r = +0.67$, $q < 0.001$). Multiple MFCC-7 statistics showed consistent negative correlations (median: $r = -0.68$; minimum: -0.65 ; mean: -0.66 ; all $q < 0.005$), as did chroma-7 minimum ($r_s = -0.61$, $q = 0.029$) and energy entropy Q25 ($r = -0.65$, $q = 0.002$).

Table 3
Acoustic features significantly correlated with MADRS severity (FDR-corrected, $q < 0.05$), MDD group ($n = 31$)

Feature	Library	Spearman r	Pearson r	p (raw)	q (FDR)
Shimmer local	<i>Praat</i>	-0.61	-0.64	0.004	0.002
Shimmer APQ5	<i>Praat</i>	-0.57	-0.61	0.009	0.004
Shimmer APQ3	<i>Praat</i>	-0.56	-0.62	0.011	0.006
F0 minimum	<i>Praat</i>	-0.59	-0.53	0.006	0.027
HNR mean	<i>Praat</i>	+ 0.47	+ 0.49	0.036	0.037
Chroma-11 – max	<i>PyAudio</i>	+ 0.65	+ 0.73	< 0.001	1.3×10^{-7}
Energy entropy – IQR	<i>PyAudio</i>	+ 0.63	+ 0.67	< 0.001	< 0.001
MFCC-7 – median	<i>PyAudio</i>	-0.67	-0.68	< 0.001	0.003
MFCC-7 – min	<i>PyAudio</i>	-0.66	-0.65	< 0.001	0.010
Chroma-7 – min	<i>PyAudio</i>	-0.61	-0.57	< 0.001	0.029
Energy entropy – Q25	<i>PyAudio</i>	-0.56	-0.65	0.002	0.002
MFCC-7 – mean	<i>PyAudio</i>	-0.57	-0.66	< 0.001	0.004
Energy entropy – SD	<i>PyAudio</i>	+ 0.52	+ 0.62	0.003	0.002
Spectral spread – min	<i>PyAudio</i>	-0.40	-0.51	0.025	0.025

Only features with convergent direction of Spearman and Pearson r are presented. MADRS Montgomery-Åsberg Depression Rating Scale, MFCC Mel-frequency cepstral coefficient, HNR harmonic-to-noise ratio, F0 fundamental frequency, IQR interquartile range, FDR false discovery rate. Complete list in Online Resource 2

[Fig. 1 – INSERT HERE]

Figure 1 Forest plot of Cohen's d effect sizes (95% confidence intervals) for acoustic features surviving FDR correction ($q < 0.05$) in GEE group comparisons, stratified by speech task (counting vs. narrative) and acoustic library (*Praat* vs. *PyAudioAnalysis*). Features are ordered from largest to smallest absolute effect size within each panel. Negative d indicates lower acoustic values in MDD; positive d indicates higher

values. Error bars represent 95% CIs derived from GEE model standard errors. MDD major depressive disorder, FDR false discovery rate, CI confidence interval. Figure created by the authors

[Fig. 2 – INSERT HERE]

Figure 2 Radar plots comparing standardised mean acoustic profiles (z-scores) between the MDD group (red shading) and the control group (blue shading) across FDR-significant features from each library-task combination. a Praat features, counting task. b PyAudioAnalysis features, counting task. c Praat features, narrative task. d PyAudioAnalysis features, narrative task. Each radial axis represents one acoustic feature. The control group polygon lies near the centre ($z \approx 0$ on all axes). MDD major depressive disorder, FDR false discovery rate. Figure created by the authors

[Fig. 3 – INSERT HERE]

Figure 3 Scatter plots of the two strongest MADRS severity correlates from each acoustic library, restricted to the MDD group ($n = 31$). Panels a and b show PyAudioAnalysis features (blue circles); panels c and d show Praat features (green circles). a Chroma-11 maximum vs. MADRS total (Pearson $r = + 0.73$, $q < 0.001$); b energy entropy IQR vs. MADRS total ($r = + 0.67$, $q < 0.001$); c shimmer local vs. MADRS total ($r = - 0.64$, $q = 0.002$); d MFCC-7 median vs. MADRS total ($r = - 0.68$, $q = 0.003$). Lines show linear regression with 95% confidence bands. MADRS Montgomery-Åsberg Depression Rating Scale, IQR interquartile range, MFCC Mel-frequency cepstral coefficient. Figure created by the authors

4 Discussion

This study provides statistically rigorous, ecologically grounded evidence that vocal acoustic parameters extracted from WhatsApp audio messages reliably distinguish adults with MDD from healthy controls and correlate dimensionally with clinician-rated depression severity. The use of GEE with FDR correction enables feature-level inference that complements machine learning classification, clarifying which specific acoustic dimensions are responsible for group separation and how they track along the severity continuum.

4.1 Formant 1 reduction as a signature of depressive phonation

The most consistent group-level finding was a substantial reduction in F1 in MDD across both tasks and all distributional statistics. In spoken language, F1 is primarily determined by jaw opening and tongue body elevation: lower F1 reflects a constricted vocal tract configuration consistent with reduced articulatory effort and jaw mobility (Saggio and Costantini 2022). These characteristics align with the psychomotor retardation model of depression, wherein fine motor control of the speech apparatus is attenuated through disrupted cortico-striatal-thalamo-cortical circuitry (Quatieri and Malyska 2012). The effect magnitude ($d = - 1.03$ in the narrative task) is large by psychiatric biomarker standards, and its consistency across all F1 statistics and both task types indicates a stable, generalisable alteration rather

than a task-dependent artefact. Prior machine learning studies identified F1 alterations in their feature importance rankings (Otani et al. 2026), but without quantifying the effect independently. The present finding establishes F1 reduction as a stand-alone, statistically robust biomarker in Brazilian Portuguese speakers, complementing analogous observations in other languages (Menne et al. 2024).

4.2 Shimmer and jitter as markers of vocal fold instability

Shimmer and jitter, capturing cycle-to-cycle variability in amplitude and frequency respectively, were elevated in MDD exclusively during spontaneous narrative speech. The selective emergence of these features in the narrative condition is theoretically meaningful: during simple numerical counting, the rhythmic constraints of the task may partially override the irregular fold dynamics that depression produces in unscripted speech. Elevated shimmer in MDD is consistent with a hyperfunction model of depressive dysphonia, wherein increased laryngeal muscle tension without adequate dampening from dopaminergic motor circuits produces irregularly amplified but aperiodic vocal fold vibrations (Van Puyvelde et al. 2018). These findings extend prior meta-analytic evidence that vocal perturbation measures are elevated in depression (Menne et al. 2024) by demonstrating that the effect is task-dependent and detectable in everyday ecological voice messages.

A counterintuitive aspect of the severity analysis was the negative correlation between shimmer and MADRS scores. A plausible explanation is that the most severely depressed individuals, exhibiting pronounced psychomotor retardation and affective blunting, produced more stereotyped and uniform vocalisations with less moment-to-moment perturbation, consistent with hypoprosodic speech in melancholic depression. Moderately depressed participants may exhibit greater vocal fold agitation without the complete flattening of output seen at higher severity levels, producing an inverse severity gradient that warrants investigation in larger, severity-stratified cohorts.

4.3 Spectral complexity and energy dynamics

The PyAudioAnalysis findings capture dimensions not accessible through Praat parameters, providing a complementary and convergent picture. Reduced spectral entropy in MDD during the narrative task reflects a more predictable, less complex spectral distribution, a computational operationalisation of what clinicians recognise as a monotone and hypoprosodic voice quality (Takano et al. 2023). ZCR reduction, reflecting slower oscillation of the signal, aligns with the reduced high-frequency content of a voice that has lost prosodic modulation.

The elevation of vocal energy in the counting task is interpretable in the context of task demands. When instructed to count at a fixed rhythm, MDD participants may have exerted greater muscular effort to maintain pacing, producing transiently elevated energy despite reduced voluntary prosodic elaboration. This effort-compensation phenomenon has been noted in motor retardation research, where increasing task demands paradoxically increase force output due to impaired internal rhythm generation.

4.4 Severity correlates: chroma, energy entropy, and the MADRS gradient

The strong positive correlation between chroma-11 maximum and MADRS severity (Pearson $r = +0.73$, $q = 1.3 \times 10^{-7}$) is the most striking dimensional finding of this study. Chroma features capture the distribution of spectral energy across the 12 pitch classes of Western tonal music; in the context of voice, elevated chroma-11 values at higher MADRS scores may reflect a paradoxical sharpening of tonal content under severe depression, possibly representing an acoustic signature of extreme laryngeal tension as psychomotor inhibition deepens. Although the clinical interpretation of chroma in voice analysis is still developing, its relevance as a machine learning feature has been noted (Otani et al. 2026), and the present finding provides the first inferential characterisation of its relationship to severity as a continuous gradient.

Energy entropy IQR as a severity correlate is more immediately interpretable: higher inter-frame variability in spectral energy distribution among more severely depressed individuals suggests greater moment-to-moment dysregulation of vocal output, consistent with a disorganised rather than uniformly suppressed vocal motor system. The negative severity correlation for MFCC-7 across multiple statistics may reflect reduced mid-frequency cepstral energy at higher severity levels, consistent with the spectral flattening documented in severe depression.

4.5 Spontaneous versus structured speech

The narrative task consistently yielded more significant features and larger effect sizes than the counting task, with 62 versus 25 PyAudioAnalysis features surviving FDR correction, and shimmer features appearing exclusively in the narrative condition. This finding is consistent with meta-analytic evidence that spontaneous speech is more sensitive to affective state than standardised speech (Veiga et al. 2025), and with the theoretical argument that unscripted speech allows full expression of depressive psychomotor alterations without the scaffolding of a fixed lexical template. The counting task retains value as a replicable, language-neutral, cross-cultural phonation probe, but researchers and clinicians seeking maximal sensitivity should prioritise spontaneous narrative tasks. This has direct implications for the design of voice-based clinical screening tools: a brief spontaneous message of the kind that patients already send to their physicians via WhatsApp in Brazilian clinical practice may be more diagnostically informative than a structured vocal exercise.

4.6 Positioning within the vocal biomarker landscape

This study complements rather than replaces machine learning-based approaches. Where classification algorithms optimise group separation, the present GEE-based approach identifies which specific acoustic parameters are responsible for that separation and characterises their relationship to severity along a continuous gradient. This inferential clarity is essential for biomarker translation: a diagnostic tool that cannot specify what it measures and why cannot be evaluated for mechanistic plausibility or withstand regulatory scrutiny. The convergence of the present inferential findings with the feature importance rankings reported by Otani et al. (Otani et al. 2026), including the prominence of MFCC-2 and spectral rolloff, provides mutual validation across analytic paradigms. Similarly, the superior sensitivity of spontaneous speech aligns with the meta-analytic conclusions of Veiga et al. (Veiga et al. 2025) that

naturalistic contexts yield larger vocal effects than controlled paradigms. The present study also contextualises findings from Abutara et al. (Abutara et al. 2026) and Galvão et al. (Galvão et al. 2025), demonstrating that acoustic features observed in non-clinical WhatsApp samples survive rigorous FDR correction in a diagnostically validated clinical cohort.

4.7 Limitations

Several limitations constrain the scope of inference. The sample size of 72 participants, while sufficient to detect large effects with GEE, may be inadequate to detect smaller effects or characterise subgroup heterogeneity by depression subtype, comorbid anxiety, or medication class. The cross-sectional design precludes causal inference and does not allow examination of whether acoustic changes normalise with treatment. Smartphone and microphone type were not controlled, introducing acoustic variability; however, the survival of large-effect features across this variability strengthens the argument for ecological robustness. The absence of a validated Brazilian Portuguese speech corpus limits contextualisation of absolute feature values, and participants were recruited from a single metropolitan centre, limiting generalisability. Finally, the strong correlation between chroma-11 maximum and MADRS severity requires independent replication and biological mechanistic investigation before clinical application can be responsibly considered.

5 Conclusions

WhatsApp audio messages contain robust, statistically valid acoustic signals that distinguish individuals with MDD from healthy controls with large effect sizes and correlate dimensionally with clinician-rated severity. Formant 1 reduction, elevated shimmer, reduced spectral entropy and zero-crossing rate, and altered MFCC dynamics constitute a multi-dimensional vocal signature of depression detectable in naturalistic audio captured on personal smartphones. Spontaneous narrative tasks elicit substantially richer and more sensitive acoustic information than structured numerical tasks, an important design consideration for future clinical tools. Chroma-11 maximum and energy entropy emerge as novel, inferentially characterised correlates of MADRS severity, complementing machine learning feature importance findings. Future longitudinal studies with larger, diagnostically stratified cohorts are needed to determine whether these biomarkers track treatment response and predict relapse, and whether their discriminative validity extends across linguistic communities and depression subtypes.

Declarations

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Research Ethics Committee of the University of Mogi das Cruzes (CAAE: 67711023.1.0000.5497).

Consent to participate

Informed consent was obtained from all individual participants included in the study.

Consent to publish

The participants have consented to the submission of anonymised data for publication.

Funding

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil. The funding source had no role in study design, data collection, analysis, interpretation, or the decision to publish.

Author Contribution

Luiza Braconi Grilo: Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – review and editing. Victor H. O. Otani: Conceptualization, Formal analysis, Methodology, Supervision, Writing – review and editing. Felipe O. Aguiar: Data curation, Formal analysis, Methodology, Software, Writing – review and editing. Hudson S. Buck: Formal analysis, Methodology, Software, Writing – review and editing. Lucas M. Marques: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Writing – review and editing. Ricardo R. Uchida: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – review and editing.

Acknowledgements

[Acknowledgements removed for blind review]

Data Availability

The dataset cannot be shared publicly owing to ethical restrictions. Requests for access to de-identified data should be directed to the corresponding author.

References

1. World Health Organization (2023) Depressive disorder (depression). <https://www.who.int/news-room/fact-sheets/detail/depression>
2. Hintz AM, Gomes-Filho IS, Loomer PM et al (2023) Depression and associated factors among Brazilian adults: the 2019 national healthcare population-based study. *BMC Psychiatry* 23:704. <https://doi.org/10.1186/s12888-023-05133-9>
3. Slavich GM, Taylor S, Picard RW (2019) Stress measurement using speech: recent advancements, validation issues, and ethical and privacy considerations. *Stress* 22:408–413. <https://doi.org/10.1080/10253890.2019.1584180>
4. Malhi GS, Mann JJ (2018) Depression. *Lancet* 392:2299–2312. [https://doi.org/10.1016/S0140-6736\(18\)31948-2](https://doi.org/10.1016/S0140-6736(18)31948-2)
5. Sobin C, Sackeim HA (1997) Psychomotor symptoms of depression. *Am J Psychiatry* 154:4–17. <https://doi.org/10.1176/ajp.154.1.4>
6. Caligiuri MP, Ellwanger J (2000) Motor and cognitive aspects of motor retardation in depression. *J Affect Disord* 57:83–93. [https://doi.org/10.1016/S0165-0327\(99\)00074-6](https://doi.org/10.1016/S0165-0327(99)00074-6)
7. Quatieri TF, Malyska N (2012) Vocal-source biomarkers for depression: a link to psychomotor activity. *Proc Interspeech 2012*. ISCA, Portland
8. Cannizzaro M, Harel B, Reilly N et al (2004) Voice acoustical measurement of the severity of major depression. *Brain Cogn* 56:30–35. <https://doi.org/10.1016/j.bandc.2004.05.003>
9. Menne F, Dörr F, Schröder J et al (2024) The voice of depression: speech features as biomarkers for major depressive disorder. *BMC Psychiatry* 24:794. <https://doi.org/10.1186/s12888-024-06253-6>
10. Low DM, Bentley KH, Ghosh SS (2020) Automated assessment of psychiatric disorders using speech: a systematic review. *Laryngoscope Investig Otolaryngol* 5:96–116. <https://doi.org/10.1002/lio2.354>
11. Cummins N, Sethu V, Epps J et al (2015) Analysis of acoustic space variability in speech affected by depression. *Speech Commun* 75:27–49. <https://doi.org/10.1016/j.specom.2015.09.003>
12. Van Puyvelde M, Neyt X, McGlone F, Pattyn N (2018) Voice stress analysis: a new framework for voice and effort in human performance. *Front Psychol* 9:1994. <https://doi.org/10.3389/fpsyg.2018.01994>
13. Taguchi T, Tachikawa H, Nemoto K et al (2018) Major depressive disorder discrimination using vocal acoustic features. *J Affect Disord* 225:214–220. <https://doi.org/10.1016/j.jad.2017.08.038>
14. Otani VHO, Aguiar FO, Justino TP et al (2026) ML-based detection of depressive profile through voice analysis in WhatsApp™ audio messages of Brazilian Portuguese speakers. *PLOS Ment Health* 3:e0000357. <https://doi.org/10.1371/journal.pmen.0000357>
15. Veiga DL, Almeida TM, Uchida RR, Cordeiro Q (2026) Dynamic coupling between voice and affective change: experimental evidence for modulation by depressive symptoms. *Manuscript under review*
16. Veiga DL, Almeida TM, Uchida RR, Cordeiro Q (2025) The fundamental frequency of voice as a potential stress biomarker: a systematic review and meta-analysis. *Stress Health* 41:e70112.

<https://doi.org/10.1002/smi.70112>

17. Sheehan DV, Lecrubier Y, Sheehan KH et al (1998) The Mini-International Neuropsychiatric Interview (M.I.N.I.): development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 59(Suppl 20):22–33
18. Montgomery SA, Åsberg M (1979) A new depression scale designed to be sensitive to change. *Br J Psychiatry* 134:382–389. <https://doi.org/10.1192/bjp.134.4.382>
19. Giannakopoulos T (2015) pyAudioAnalysis: an open-source Python library for audio signal analysis. *PLoS ONE* 10:e0144610. <https://doi.org/10.1371/journal.pone.0144610>
20. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57:289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
21. Saggio G, Costantini G (2022) Worldwide healthy adult voice baseline parameters: a comprehensive review. *J Voice* 36:637–649. <https://doi.org/10.1016/j.jvoice.2020.08.028>
22. Albuquerque L, Valente ARS, Teixeira A et al (2021) Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan. *PLoS ONE* 16:e0248842. <https://doi.org/10.1371/journal.pone.0248842>
23. Takano T, Mizuguchi D, Omiya Y et al (2023) Estimating depressive symptom class from voice. *Int J Environ Res Public Health* 20:3965. <https://doi.org/10.3390/ijerph20053965>
24. Kappen M, Vanderhasselt MA, Slavich GM (2023) Speech as a promising biosignal in precision psychiatry. *Neurosci Biobehav Rev* 148:105121. <https://doi.org/10.1016/j.neubiorev.2023.105121>
25. Shin D, Cho WI, Park CHK et al (2021) Detection of minor and major depression through voice as a biomarker using machine learning. *J Clin Med* 10:3046. <https://doi.org/10.3390/jcm10143046>
26. Abutara A, Kissimoto A, Aguiar FO et al (2026) Acoustic signatures of anxiety and depression in academic stress contexts: evidence from voice analysis in pre-university students. *Smart Health* 40:100659. <https://doi.org/10.1016/j.smhl.2026.100659>
27. Galvão APFC, Soares SM, Aguiar FO et al (2025) Detection of anxiety in university students through voice acoustics using WhatsApp chatbot audio: a machine learning study. *Scire Salutis* 15(2). <https://doi.org/10.6008/CBPC2236-9600.2025.002.0010>
28. Briganti G, Lechien JR (2025) Voice quality as digital biomarker in bipolar disorder: a systematic review. *J Voice*. <https://doi.org/10.1016/j.jvoice.2025.01.002>
29. Parola A, Simonsen A, Lin JM et al (2023) Voice patterns as markers of schizophrenia: a cross-linguistic and meta-analysis based investigation. *Schizophr Bull* 49(Suppl 2):S125–S141. <https://doi.org/10.1093/schbul/sbac128>
30. Lim GY, Tam WW, Lu Y et al (2018) Prevalence of depression in the community from 30 countries between 1994 and 2014. *Sci Rep* 8:2861. <https://doi.org/10.1038/s41598-018-21243-x>

Figures

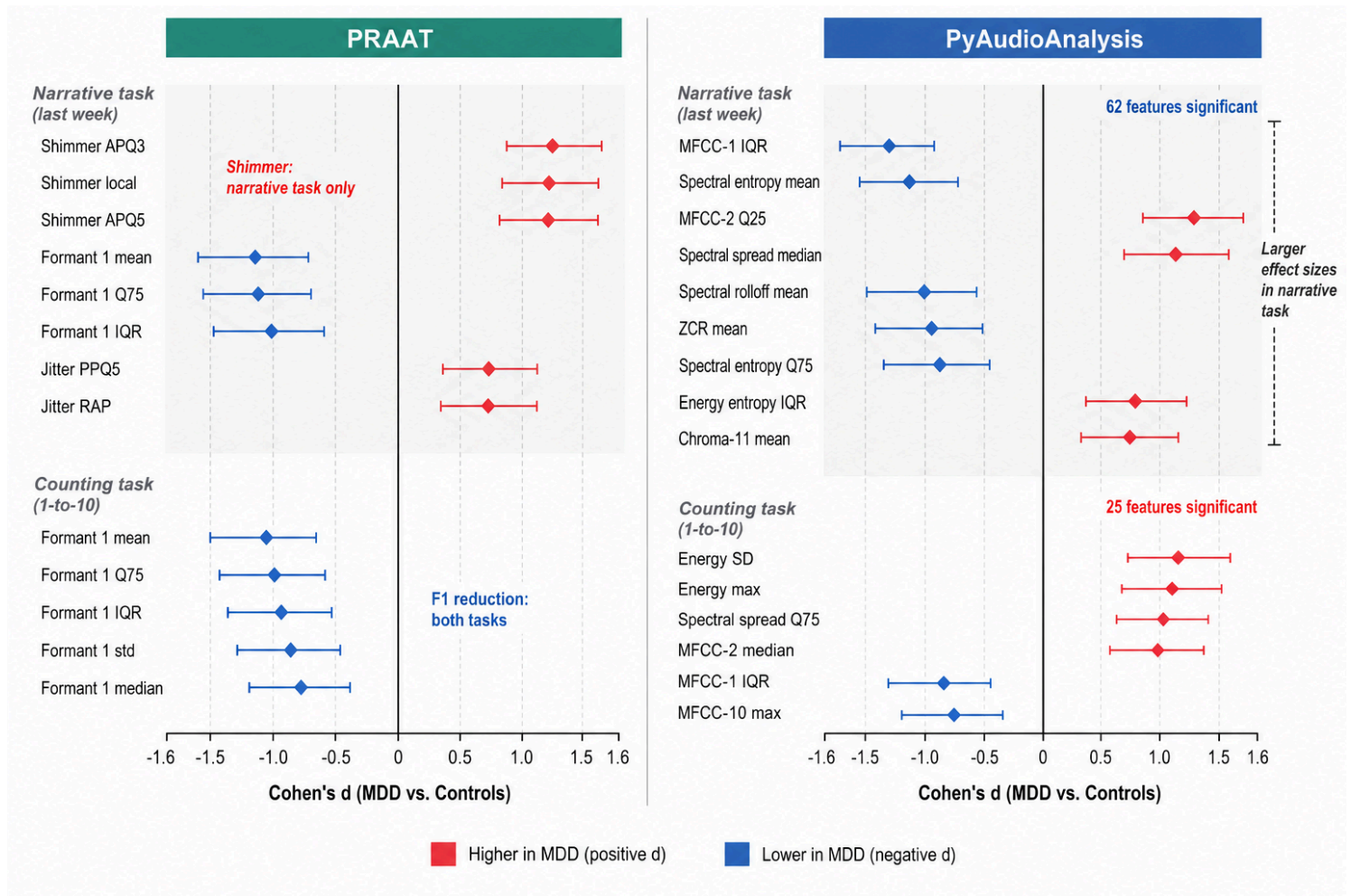


Figure 1

Forest plot of Cohen's d effect sizes (95% confidence intervals) for acoustic features surviving FDR correction ($q < 0.05$) in GEE group comparisons, stratified by speech task (counting vs. narrative) and acoustic library (Praat vs. PyAudioAnalysis). Features are ordered from largest to smallest absolute effect size within each panel. Negative d indicates lower acoustic values in MDD; positive d indicates higher values. Error bars represent 95% CIs derived from GEE model standard errors. MDD major depressive disorder, FDR false discovery rate, CI confidence interval. Figure created by the authors

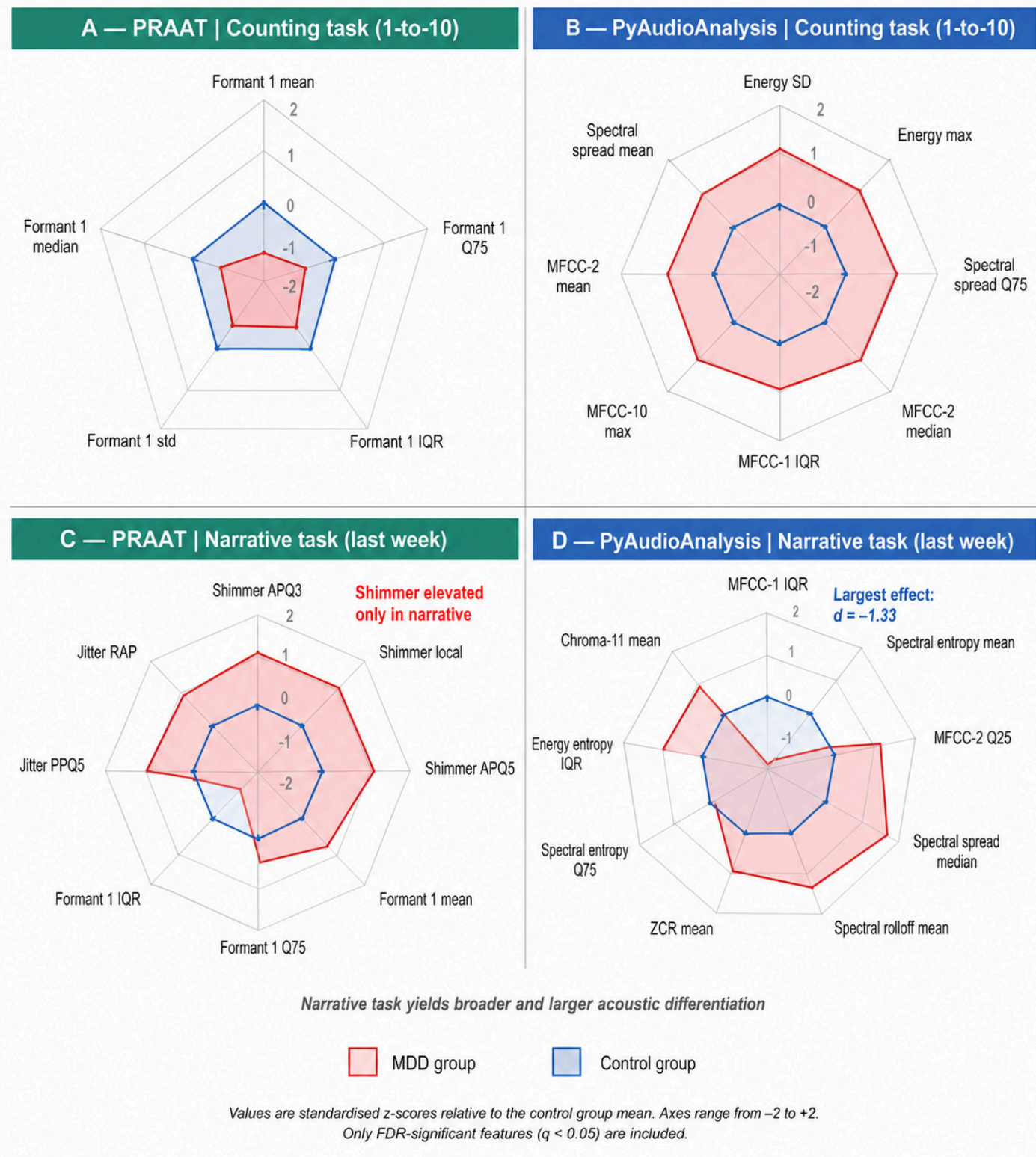


Figure 2

Radar plots comparing standardised mean acoustic profiles (z-scores) between the MDD group (red shading) and the control group (blue shading) across FDR-significant features from each library-task combination. a Praat features, counting task. b PyAudioAnalysis features, counting task. c Praat features, narrative task. d PyAudioAnalysis features, narrative task. Each radial axis represents one acoustic feature.

The control group polygon lies near the centre ($z \approx 0$ on all axes). MDD major depressive disorder, FDR false discovery rate. Figure created by the authors

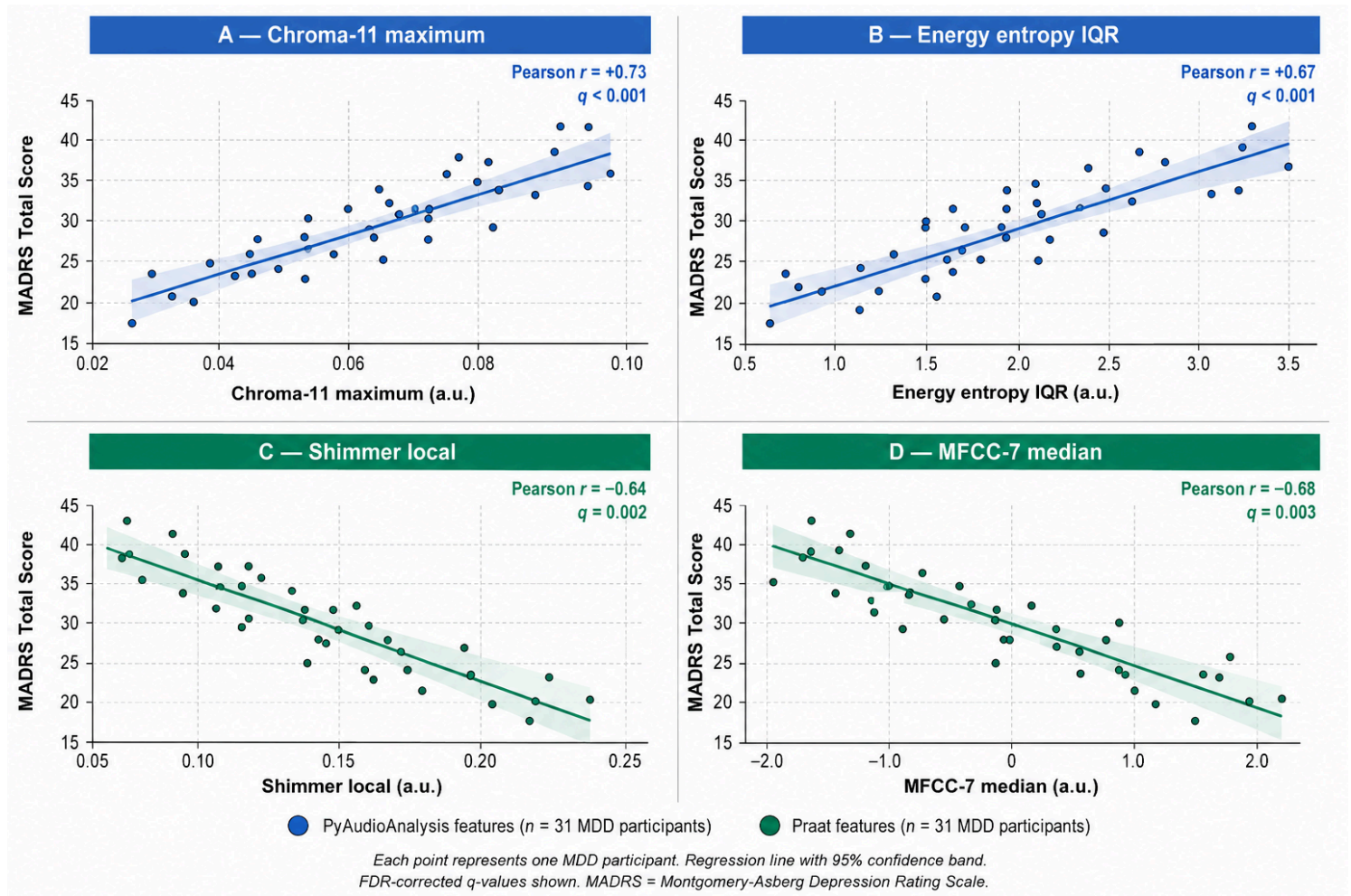


Figure 3

Scatter plots of the two strongest MADRS severity correlates from each acoustic library, restricted to the MDD group ($n = 31$). Panels a and b show PyAudioAnalysis features (blue circles); panels c and d show Praat features (green circles). a Chroma-11 maximum vs. MADRS total (Pearson $r = +0.73$, $q < 0.001$); b energy entropy IQR vs. MADRS total ($r = +0.67$, $q < 0.001$); c shimmer local vs. MADRS total ($r = -0.64$, $q = 0.002$); d MFCC-7 median vs. MADRS total ($r = -0.68$, $q = 0.003$). Lines show linear regression with 95% confidence bands. MADRS Montgomery-Åsberg Depression Rating Scale, IQR interquartile range, MFCC Mel-frequency cepstral coefficient. Figure created by the authors

Supplementary Files

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

- [supplementaryeurarch.docx](#)