Original Paper

Voice-Derived Speech Features of Anhedonia in Adults With Depression: Exploratory Secondary Analysis of a Semistructured Clinical Interview

Christopher G Beevers[1,2], PhD; Rochelle A Stewart[1], PhD; Olivia Brady[1], MA; Kean J Hsu[3], PhD

[1]Department of Psychology, Institute for Mental Health Research, The University of Texas at Austin, Austin, TX, United States

[2]Department of Psychology, University of Rhode Island, Kingston, RI, United States

[3]Department of Psychology, National University of Singapore, Queenstown, NA, Singapore

Corresponding Author:
Christopher G Beevers, PhD
Department of Psychology
University of Rhode Island
Chafee Hall, 142 Flagg Road
Kingston, RI, 02881
United States
Phone: 1 401 874 2193
Email: christopher.beevers@uri.edu

Abstract

Background: Anhedonia is a core, clinically consequential feature of depression, and its motivational and behavioral features may be difficult to observe through self-report alone. Speech is an effortful, dynamic, and socially contingent behavior that may index the motivational disruptions central to anhedonia, but few studies have examined anhedonia specifically using a broad, multimodal set of voice features.

Objective: We conducted an exploratory, hypothesis-generating analysis of whether voice-derived speech features from a semistructured clinical interview are associated with concurrent anhedonia severity and with anhedonia 4 weeks later, adjusted for baseline severity.

Methods: In a secondary analysis of a depression intervention trial, we analyzed recorded 17-item Hamilton Rating Scale for Depression interviews from adults with depression (anhedonia: n=44; depression severity: n=45) collected before participants initiated a behavioral intervention. Of these, 36 provided symptom data at week 4. Acoustic (pitch, spectral or mel-frequency cepstral coefficients, and jitter or shimmer), prosodic, linguistic (lexical richness and sentiment), topic-model, and automatic speech recognition quality features were extracted. After reducing the high-dimensional acoustic set to 9 theory-driven domain composites, 77 features were tested per outcome. Anhedonia (Mood and Anxiety Symptom Questionnaire, Anhedonic Depression Scale) and general depression severity (Quick Inventory of Depressive Symptomatology, Self-Report [QIDS-SR]) were examined with univariate Pearson correlations, with significance assessed by permutation tests and interpreted against a conservative P=.01 threshold, given the exploratory, multiple-testing context. For the week-4 analyses, baseline voice features were correlated, with week-4 symptom scores residualized for baseline symptoms and treatment condition. We report unadjusted P values and emphasize effect sizes, given the exploratory aims.

Results: Higher baseline anhedonia was associated with briefer, less redundant speech: higher hapax ratio and lexical diversity (r=0.46; P=.003) alongside fewer content and total words (rs≈–0.43; P≤.01) and shorter total speaking duration (r=–0.41; P=.007), all exceeding the P=.01 threshold. Weaker, subthreshold trends included more negative sentiment (compound r=–0.34; P=.02), less work or school topic content (r=–0.31; P=.04), and flatter pitch variation (fundamental frequency SD r=–0.29; P=.06). General depression severity (QIDS-SR) showed no associations that exceeded the threshold. In baseline-adjusted week-4 analyses, no feature was associated with anhedonia at P=.01, whereas lower baseline voice perturbation (jitter composite; r=–0.46; P=.005) was associated with higher-than-predicted week-4 depression severity.

Conclusions: Higher anhedonia aligned most robustly with a voice signature of reduced speech output and sparser but proportionally more diverse speech, with weaker, subthreshold trends toward flatter prosody and more negative sentiment, consistent with diminished behavioral engagement rather than uniform flattening of expressivity. Because the sample was small,

XSL•FO
RenderX

predominantly female, and drawn from a depressed sample, and because many features were tested, findings are preliminary and require replication with larger samples and more naturalistic speech.

(JMIR Form Res 2026;10:e94315) doi: 10.2196/94315

KEYWORDS

depression; anhedonia; speech analysis; acoustic features; natural language processing; digital phenotyping; automatic speech recognition; multimodal biomarkers

Introduction

Anhedonia, broadly defined as diminished interest or pleasure, is a core and clinically consequential feature of depression. Beyond its diagnostic relevance, anhedonia is associated with poorer psychosocial functioning, reduced treatment response, and elevated risk for relapse [1-3]. Contemporary models increasingly emphasize that anhedonia reflects not only deficits in hedonic experience but also disruptions in motivation, reward anticipation, and sustained engagement with the environment [4,5]. Despite its centrality, anhedonia remains challenging to characterize fully, as commonly used self-report measures rely on introspective judgments of interest, pleasure, and motivation rather than directly indexing how these disruptions are expressed in observable behavior.

Speech is a promising yet underused behavioral channel through which anhedonia may be expressed. Producing speech requires effortful engagement, unfolds dynamically over time, and is inherently social and reward-contingent, rendering it sensitive to motivational processes implicated in anhedonia [6,7]. Changes in how much individuals speak, how speech is structured, and how vocal features change during speech may reflect diminished drive to engage, reduced persistence, or altered allocation of effort. Accordingly, speech may offer a window into anhedonia that complements subjective symptom reports by indexing observable patterns of engagement and expression.

A growing literature has examined associations between vocal characteristics and depression, identifying links with features such as pitch variability, speech rate, pause structure, and spectral properties [6,8,9]. However, much of this work has focused on global depression severity rather than anhedonia specifically and has typically relied on a limited number of acoustic or prosodic features selected a priori. Such approaches may be insufficient for capturing the multifaceted ways in which anhedonia manifests in speech, particularly if anhedonia is expressed more strongly through temporal patterns, speech quantity, or coordination across multiple vocal domains rather than through uniform reductions in expressivity [10,11].

Multimodal speech analysis offers a means of addressing these limitations by modeling complementary aspects of vocal production, including acoustic dynamics, prosody, spectral structure, linguistic content, and speech quantity [7,12]. Examining features across these domains may be especially important for anhedonia, which is thought to involve altered motivation and engagement rather than simple flattening of affect [4,5]. Despite this potential, relatively few studies have adopted a multimodal approach to speech-based assessment of anhedonia, and fewer still have examined whether speech features prospectively relate to later anhedonic symptoms.

In the present study, we apply a comprehensive, multimodal voice-feature extraction pipeline to speech obtained during a semistructured clinical interview for depression. We examine both cross-sectional associations between voice-based features and anhedonia and longitudinal associations with residualized anhedonia 4 weeks later, controlling for baseline symptom severity. By focusing on anhedonia as a distinct symptom dimension and leveraging a broad feature set spanning multiple vocal domains, this work aims to clarify whether and how anhedonia is detectable in speech and to explore, in a small subsample, whether baseline vocal patterns are prospectively associated with later anhedonic symptoms.

Methods

Study Design

This was an exploratory secondary analysis of speech recordings from a completed randomized clinical trial, comprising a cross-sectional component (baseline voice features and concurrent symptoms) and a longitudinal component (baseline voice features and week-4 symptoms adjusted for baseline). We report the study in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guideline for observational research [13]. A completed checklist is provided as Multimedia Appendix 1.

Participants

This report is a secondary analysis of speech recordings collected in a single-site randomized clinical trial that evaluated attention bias modification (ABM), a computerized cognitive training procedure, as an intervention for depression [14]. The parent trial was conducted at a large public university in the southern United States. Participants were recruited from the surrounding community and university through advertisements. Participants were eligible if they were 18 to 70 years of age, fluent in English, able to provide informed consent, and scored 13 or higher on the Quick Inventory of Depressive Symptomatology, Self-Report (QIDS-SR) at baseline. Those taking antidepressant medication were required to have had no change in medication or dose during the prior 12 weeks. Key exclusion criteria were active suicidal behavior or significant suicidal ideation within the prior 3 months (assessed with the Columbia-Suicide Severity Rating Scale); current or past bipolar or psychotic disorders and a current (past-12-month) alcohol or substance use disorder of at least moderate severity, both assessed with the Mini International Neuropsychiatric Interview; current opioid analgesic or systemic corticosteroid use; and ongoing psychotherapy [14]. The parent trial enrolled 145 adults.

XSL•FO
RenderX

The present study and its voice-feature analyses were not part of the original trial design. Recordings of the clinician-administered symptom interview had been retained for interview-fidelity monitoring and were analyzed retrospectively for this report.

Voice features were extracted from the baseline (week 0) 17-item Hamilton Rating Scale for Depression (HRSD-17) interviews, which had been recorded for interview-fidelity monitoring. Eligibility for the present analyses required a baseline recording together with at least 1 usable baseline self-report outcome (Mood and Anxiety Symptom Questionnaire, Anhedonic Depression Scale [MASQ-AD] or QIDS-SR). We excluded 11 participants who had a recording but neither outcome, as they could not contribute to any association analysis (these 11 were also excluded from the parent-trial outcome analyses, with treatment assignment recorded as missing and no week-4 data). The analytic sample therefore comprised 45 participants. Within this sample, cross-sectional correlations used the participants with a valid baseline score on the relevant outcome (anhedonia [MASQ-AD]: n=44; depression severity [QIDS-SR]: n=45), and 36 also provided self-reported symptom data at week 4 and contributed to the baseline-adjusted longitudinal analyses. Because this subsample was small, the longitudinal analyses should be considered very exploratory, with statistical power to detect only large effects. The baseline interviews analyzed here were recorded between October 9, 2017, and April 10, 2019.

Assessments

Anhedonia was assessed using the 10-item anhedonic depression scale from an adapted 30-item version of the Mood and Anxiety Symptom Questionnaire (MASQ-AD) [15]. The MASQ-AD emphasizes low positive affect and loss of interest or pleasure. Items are rated on a 1-5 Likert scale referencing the past week, and the 10 items are summed to yield a total score ranging from 10 to 50, with higher scores reflecting greater anhedonia. The MASQ-AD was the primary outcome for the present analyses.

Depression severity was assessed using the 16-item QIDS-SR [16]. The QIDS-SR is a brief self-report measure covering the 9 symptom criterion domains of the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (DSM-IV) over the past 7 days. Items are scored on 0-3 anchors and summed to yield a total score from 0 to 27, with higher scores indicating greater depression severity. The QIDS-SR was included as a secondary, comparison outcome representing global depression severity.

Clinician-rated depression severity was assessed with the HRSD-17, a widely used semistructured interview that served as the source recording for voice-feature extraction. The HRSD-17 comprises 17 items assessing symptoms such as depressed mood, guilt, suicidality, insomnia, work and activities, psychomotor change, anxiety, and somatic symptoms. Items are scored on either 5-point (0-4) or 3-point (0-2) anchors depending on the symptom and summed to a total score ranging from 0 to 52, with higher scores indicating greater severity. The MASQ-AD, QIDS-SR, and HRSD-17 were each administered at baseline, and the MASQ-AD and QIDS-SR were readministered at week 4.

Voice-Feature Source Interview

Voice features were extracted from the audio of the baseline HRSD-17 interview described earlier. The HRSD-17 is a brief semistructured interview that typically takes about 15-20 minutes to administer. Trained interviewers administered the HRSD-17, with weekly supervision by a licensed psychologist. Interrater reliability was high (intraclass correlation coefficient=0.98), and interviewers were blinded to participants' intervention condition. Only participant speech (not interviewer speech) was used for participant-level feature summaries.

Ethical Considerations

The parent trial was approved by the institutional review board of The University of Texas at Austin (protocol: 2016-06-0034), and all participants provided written informed consent for trial participation, including audio recording of the clinical interviews. The approved protocol and the informed consent obtained from participants explicitly provided for secondary analyses of these data. The present secondary analysis therefore fell within the scope of the original ethics approval and consent and did not require additional institutional review board review or additional participant consent. The present secondary analysis used these previously collected, approved recordings. No new data were collected. All audio recordings and derived feature tables were deidentified, with participants referred to only by study identifiers, and were stored on access-restricted, institutionally managed servers. Participant compensation for the parent trial is described in the original trial report [14].

Overview of the Processing Pipeline

Interview audio recordings were processed using a Python workflow designed (1) to standardize audio format and sampling characteristics; (2) to identify speakers and infer speaker roles (participant vs interviewer); (3) to generate time-aligned transcripts; and (4) to derive acoustic, prosodic, linguistic, sentiment, topic, and dyadic synchrony measures. The pipeline produces both segment-level datasets (eg, diarized turns, aligned words, and per-segment features) and participant-level summary tables.

Input Organization and Metadata Extraction

Raw interview recordings were provided in common formats (eg, .mp3, .wav, and .wma). For each recording, study identifier, participant identifier, and (when available) week or session number were parsed from the filename using pattern-based rules. When week or session information was absent, a default value was assigned. These metadata were carried through all derived tables.

Audio Standardization

All recordings were converted to a mono, 16-kHz waveform to ensure consistent downstream feature extraction. Conversion was performed with FFmpeg when available. If FFmpeg failed or produced an invalid output, decoding and resampling were performed using "librosa." Converted waveforms were cached in a derived folder and reused on subsequent runs when the cached file was nonempty and at least as recent as the source audio.

XSL•FO
RenderX

File-Level Signal Quality Metrics

To facilitate quality control, file-level signal metrics were computed from the 16-kHz waveform. Framewise root-mean-square (RMS) energy was estimated using 2048-sample frames and a 256-sample hop. From the RMS distribution, we extracted a noise-floor proxy (10th percentile), median RMS, and an upper-energy estimate (95th percentile), and expressed these in linear units and decibels (dB), computed as $20 \cdot \log_{10}(\mathrm{RMS}+\varepsilon)$. Clipping was quantified as the fraction of samples with absolute amplitude ≥0.999. Total recording duration (seconds) was also recorded.

Speaker Diarization and Role Assignment

Speaker diarization was conducted using the WhisperX diarization interface (a wrapper around a PyAnnote-based pipeline), yielding a sequence of speaker-labeled time segments (start and end times in seconds). Because diarization labels are anonymous, we inferred speaker roles (participant vs interviewer) using a heuristic applied within each file.

Role inference used per-speaker diarization statistics, including total speaking time, number of segments, the proportion of short utterances (≤1.5 seconds), and time of first speech. The interviewer was preferentially identified as the speaker who (1) tended to speak earlier, (2) produced relatively short prompts more frequently, and (3) exhibited higher question use. The participant was identified as the remaining speaker with the greatest speaking time and longer average segment duration. Speakers not meeting the participant or interviewer criteria were labeled as "other," and segments without an assigned role were retained with an "unknown" role label.

Automatic Transcription and Word-Level Alignment

Automatic speech recognition (ASR) was performed using WhisperX. Each 16-kHz file was transcribed into text segments, and word-level timestamps were obtained via WhisperX alignment when enabled. Alignment used a wav2vec2-based alignment model (default: "facebook/wav2vec2-base-960h") on a central processing unit and was executed in fixed-size batches of transcript segments to reduce memory demands and stabilize long-running jobs. If alignment failed for a file, the pipeline fell back to unaligned ASR segments.

Aligned word outputs were organized into a word-level table containing word text and onset and offset times. These word timestamps served as the basis for per-segment transcript reconstruction and timing-derived measures (eg, speech rate and pause counts).

Refinement of Speaker Roles Using Question Use

To improve interviewer identification, question-mark counts were estimated from ASR text segments and attributed to diarized speakers by overlap: each ASR segment containing "?" was assigned to the diarized speaker whose segment window had the greatest temporal overlap with the ASR segment window, and question marks were accumulated by speaker. These per-speaker question counts were then incorporated into the role-inference heuristic to refine participant or interviewer assignments.

Segment-Level Text, Timing, and Linguistic Features

Diarized segments served as the primary unit of analysis ("turns"). For each diarized segment, aligned words were assigned to the segment when their start and end times fell within the diarized boundaries. Assigned words were concatenated to create a segment transcript and to compute timing measures.

Speech rate was computed as words per minute (WPM=60×number of words÷segment duration in seconds). Pauses were indexed as the number of interword gaps exceeding 250 ms (based on successive aligned word onset times). Linguistic features were computed from tokenized text using a consistent tokenization rule (alphabetic tokens with apostrophes) and stopword filtering. That is, all transcripts were processed with the same tokenization procedure and settings, so every linguistic feature (eg, word counts, lexical diversity, and sentiment counts) is computed on a comparable definition of "token." The same rules for splitting text into tokens (eg, handling punctuation, contractions, casing, and numbers) were applied uniformly across all samples, avoiding measurement differences caused by inconsistent preprocessing. Derived metrics included total token count, number of unique tokens, lexical diversity (number of unique tokens divided by total tokens), hapax count and ratio (tokens occurring once), average word length, and the proportion or count of content words (nonstopwords). Normalized text (lowercased, tokenized, and stopwords removed) was retained for downstream lexical similarity computations.

Sentiment Features

To estimate the sentiment features, valence was computed from each segment transcript using a rule-based sentiment model, Valence Aware Dictionary and Sentiment Reasoner. Segment-level outputs included positive, negative, neutral, and compound sentiment scores. For speaker-level summaries, segment scores were aggregated using word counts as weights to reduce the influence of very short segments.

Acoustic Feature Extraction

Acoustic features were extracted using openSMILE with the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPSv02 version 2.0) feature set at the "functionals" level, producing file-level summary descriptors. In addition, speaker-specific acoustic summaries were computed by concatenating all diarized segments for each speaker role (participant and interviewer) within a file and re-extracting the same eGeMAPS functionals from the resulting role-specific waveform. Acoustic summary statistics included pitch, loudness, spectral, and voice-quality measures summarized using standard statistics such as means, variability, and percentiles. Speaker-level acoustic tables also included total speaking time (sum of diarized segment durations) and the number of segments contributing to the summary.

Prosodic Features

Prosodic features were computed for each diarized segment from the 16-kHz waveform. Fundamental frequency (F0) was estimated using "librosa.pyin" (65-400 Hz) with 2048-sample frames and a 256-sample hop. For each segment, we summarized

XSL•FO
RenderX

F0 mean, median, SD, minimum, and maximum across voiced frames and computed voiced fraction (proportion of voiced frames). Intensity-related features were summarized from the RMS envelope (mean, SD, minimum, and maximum), and spectral-temporal activity was indexed using the mean zero-crossing rate. Segment-level prosody was aggregated to speaker-level summaries within file and role.

Integrated Quality Segments and Summary Indices

To support downstream filtering and sensitivity analyses, segment-level text or timing features were merged with segment-level prosody and file-level quality metrics. Segment-level signal-to-noise ratio (SNR; dB) was computed as follows:

$$20 \cdot \log_{10}\left(\frac{\mathrm{RMS}_{\mathrm{seg}}+\varepsilon}{\mathrm{RMS}_{\mathrm{noise}}+\varepsilon}\right)$$

Additional derived indicators included segment RMS in decibels and a binary marker indicating whether the segment contained any aligned words. Summary quality indices were computed by speaker role, including mean or median SNR, proportion of segments with low SNR (<5 dB), proportion with no aligned text, mean WPM, and mean clipping fraction.

Topic Modeling of Participant Speech

Topic modeling was conducted on participant segments meeting a minimum token threshold (default ≥5 tokens). Sentence embeddings were computed using a transformer encoder (default: “all-MiniLM-L6-v2”) and clustered with k-means (default k=12; reduced automatically when fewer segments were available). For interpretability, topic descriptors were computed using term frequency-inverse document frequency to identify top terms per topic, and each segment was assigned a topic label along with its similarity to the assigned centroid. Participant-level topic features were represented as topic-use proportions (topic counts normalized within participant and file).

Dyadic Synchrony Metrics

Two complementary synchrony indices were computed to characterize coordination between participant and interviewer. First, a turn-pair synchrony dataset was created by pairing each participant segment with the temporally nearest interviewer segment (nearest center time). For each paired turn, we computed absolute differences in timing (center-time distance), prosody (difference in mean F0 and mean RMS), speech rate (WPM difference), affect (difference in compound sentiment), and lexical alignment (Jaccard similarity over normalized token sets). Second, prosodic synchrony was summarized in fixed temporal bins (default 1-second bins) by averaging participant and interviewer features within each bin and computing within-file correlations and mean absolute differences between roles for F0, RMS, and sentiment. These synchrony summaries were exported at the participant-file level.

Data Outputs

The pipeline exported intermediate and derived datasets as both Parquet and CSV files, including diarization segments, aligned words, segment-level text metrics, acoustic and prosodic features, sentiment summaries, topic outputs, quality-control variables, synchrony metrics, and a consolidated participant-level feature table. Participant-specific copies of each dataset were additionally written to participant folders to support participant-level analyses and auditing.

Total Features Used in Analyses

For statistical analyses, we constructed a consolidated, analysis-ready feature matrix by merging participant-level summaries across domains using participant ID and recording filename as keys. This matrix comprised 161 participant-level derived features: acoustic descriptors (n=90), prosody (n=13), linguistic content and fluency measures (n=10), sentiment (n=6), quality-control metrics (n=17; 9 segment-level summary indices and 8 file-level signal-quality metrics), topic composition (n=12 topic-proportion features), and dyadic synchrony (n=13; 6 prosodic synchrony indices and 7 turn-level synchrony summary indices). As described in the Statistical Analysis section, the acoustic descriptors were subsequently summarized by domain composites, and the feature set was screened and reduced before analysis. The full list of features entered into the analyses is provided in Table S7 in Multimedia Appendix 2.

Statistical Analysis

Analyses were conducted at the participant level using baseline (week 0) and follow-up (week 4) symptom outcomes, the MASQ-AD and QIDS-SR. Although the focus of this report is anhedonia, we included the QIDS-SR for comparison. The feature matrix comprised acoustic, prosodic, linguistic, sentiment, topic, synchrony, and ASR quality measures. Data quality was summarized for each participant (low SNR in >30% of files, empty transcripts or ASR failures in >50% of files, or mean transcript length <3 words). All participants were retained to preserve power, and analyses are interpreted with this data-quality caveat in mind. In sensitivity analyses of the cross-sectional anhedonia associations, we (1) re-estimated the associations after excluding participants flagged for potentially problematic data quality and (2) computed partial correlations adjusting for total speech quantity (total tokens); these sensitivity analyses used parametric correlation tests (Tables S1 and S2 in Multimedia Appendix 2).

Feature preprocessing included removal of predictors with >20% missingness or zero or near-zero variance, followed by median imputation of remaining missing values. Because the acoustic feature set was high-dimensional relative to the sample size, the acoustic domain was represented by theory-driven domain composites constructed for each acoustic subfamily (pitch, jitter, shimmer, mel-frequency cepstral coefficients [MFCCs], formant, loudness, spectral, voice quality, and prosody) by averaging *z*-scored features within each domain. These composites replaced the individual acoustic variables in the predictor set. Because several segment-level prosodic and temporal features were also retained as individual predictors, the prosody composite overlaps with those features. The domain composites are therefore intended as coarse summaries of feature families rather than mutually independent predictors. After this reduction, 77 features were entered into the analyses for each outcome (68 nonacoustic features and 9 acoustic domain composites).

XSL•FO
RenderX

Primary inference relied on univariate Pearson correlations between each feature and each outcome, with statistical significance assessed using permutation tests (1000 label permutations). For week-4 outcomes, the week-4 symptom score was residualized on baseline symptom level and ABM assignment, and each voice feature was correlated with these residuals. These week-4 associations therefore reflect where a participant stood at week 4 relative to the level predicted by their baseline severity and treatment assignment, rather than an individual change in symptoms over time. Significance for the adjusted associations was evaluated with the same 1000-permutation procedure. We summarized the strongest associations for each outcome and used volcano plots for visualization. Analyses were performed in R (R Foundation for Statistical Computing), with fixed random seeds to ensure reproducibility.

Because these analyses are explicitly exploratory and hypothesis-generating, we report unadjusted *P* values and emphasize effect sizes and consistency across outcomes. To limit the multiple-testing burden, we took 2 complementary steps: we reduced the number of tests by representing the acoustic domain with 9 composites, and we adopted a conservative interpretive threshold of *P*=.01 for identifying associations of interest. We note the exact number of tests: each outcome was tested against 77 features, and the 4 primary analyses reported here (MASQ-AD and QIDS-SR, each examined cross-sectionally at baseline and in baseline-adjusted week-4 models) comprise 308 univariate correlation tests. We did not apply a formal familywise or false-discovery correction, which would be overly conservative for an exploratory screen of this kind. Permutation *P* values provide an empirical null tailored to the data and help calibrate significance under minimal distributional assumptions, but they do not themselves correct for multiple testing. We did not fit multivariate models because the high feature-to-sample ratio risks unstable estimates and overfitting, so we prioritized transparent univariate screening as a first-pass exploratory step. All findings are interpreted as preliminary signals that require independent replication.

Results

Preliminary Analyses

Participant Characteristics

In the analytic sample (N=45), mean participant age was 23.1 (SD 4.9; range 18.1-35.2) years. Of the 45 participants, 36 (80%) identified their sex at birth as female and 9 (20%) as male. Mean baseline symptoms were MASQ-AD 31.1 (SD 6.2), QIDS-SR 15.5 (SD 2.2), and HRSD-17 16.3 (SD 5.2). In the parent trial, ABM assignment among these participants was active: 18 (40%); placebo: 14 (31%); and waitlist: 13 (29%). The distributions of MASQ-AD and QIDS-SR at week 0 are presented in Figure S1 in Multimedia Appendix 2.

Participant-Level Data Quality

Among the 45 participants in the analytic sample, 7 had low SNR in more than 30% of files, indicating background noise that could distort acoustic measures; 6 had empty transcripts or failed ASR in more than 50% of files, suggesting frequent speech-to-text failures or silence; and 9 had an average transcript length of less than 3 words. Accounting for overlap across these indicators, 14 participants were identified as having potentially problematic data (conversely, 31/45, 69% had good-quality data). All participants were retained to maximize statistical power, but results should be interpreted with this data-quality caveat in mind (see the Limitations section).

Feature-Level Data Quality

Of the 161 derived features extracted from the interview data, 2 were missing in 20% or more participants, and 1 had zero or very low variance. These 3 features were removed from the set of predictors: the correlation and the mean absolute difference between participant and interviewer sentiment (sentiment synchrony and high missingness) and the mean time difference, in seconds, between paired participant and interviewer turns (turn synchrony and zero or near-zero variance).

Acoustic Feature Representation

Because the acoustic feature set was high-dimensional relative to the sample size, the 90 acoustic features were summarized by 9 theory-driven domain composites (pitch, jitter, shimmer, MFCC, formant, loudness, spectral, voice quality, and prosody), each computed by averaging *z*-scored features within its acoustic domain. These composites replaced the individual acoustic variables in all subsequent analyses. After this reduction, 77 features were entered into the analyses for each outcome (68 nonacoustic features and 9 acoustic domain composites).

Topic Modeling Interpretation

Topic modeling produced a small set of interpretable themes (Table S4 in Multimedia Appendix 2) that loosely reflect the types of depression-related symptoms that are discussed during the Hamilton Rating Scale for Depression. These included check-ins or recency; conversational fillers; energy, pain, and health; activities or enjoyment; eating or appetite; mood or sadness; symptom-rating language; time and scheduling; plans and intentions; general feelings or doing; work, school, and tasks; and sleep and bedtime routines. These labels are derived from the top terms or segments per topic (Table S4 in Multimedia Appendix 2), so associations can be interpreted as spending more or less interview time on those specific themes.

Associations Between Week 0 Symptoms and Voice Features

Anhedonia

The top 20 strongest correlations between anhedonia symptoms measured by the MASQ and voice features are presented in Table 1. Correlations between all features and anhedonia symptoms are presented in the top panel of the volcano plot in Figure 1.

Adopting a more conservative interpretive threshold of *P*=.01, the associations that exceeded this threshold all centered on lexical richness and speech quantity. Higher anhedonia was associated with higher hapax ratio and higher lexical diversity (both *r*=0.46; *P*=.003), alongside fewer content words (*r*=–0.44; *P*=.006), fewer tokens per segment (*r*=–0.43; *P*=.004), fewer total tokens and words (*r*=–0.43; *P*=.008), fewer characters

XSL•FO
RenderX

(*r*=–0.43; *P*=.008), and shorter total speaking duration (*r*=–0.41; *P*=.007). This pattern is consistent with shorter, less redundant speech.

Several additional associations fell in the small-to-moderate range but did not reach the *P*=.01 threshold and are reported as effect-size-level trends. Higher anhedonia was nonsignificantly associated with more negative sentiment (compound *r*=–0.34; *P*=.02), a lower share of work or school topic content (topic 10; *r*=–0.31; *P*=.04), and flatter pitch variation (F0 SD; *r*=–0.29; *P*=.06), together with a slightly higher pitch floor (higher F0 minimum, mean, and median). No acoustic domain composite was significantly associated with anhedonia.

In summary, the associations exceeding the *P*=.01 threshold indicate that higher anhedonia is characterized by shorter, less redundant speech. The accompanying higher lexical-diversity ratios most plausibly reflect this reduced speech quantity rather than richer vocabulary (see the Discussion section). In a sensitivity analysis excluding the 14 quality-flagged participants (n=30), the hapax-ratio and lexical-diversity associations were essentially unchanged in magnitude (both *r*=0.46; *P*=.01), whereas the speech-quantity associations attenuated (*r*=–0.27 to –0.33) but retained the same direction (Table S1 in Multimedia Appendix 2).

Table 1. Top 20 voice-feature correlations with anhedonia symptoms (MASQ-AD[a]) and permutation-test *P* values.

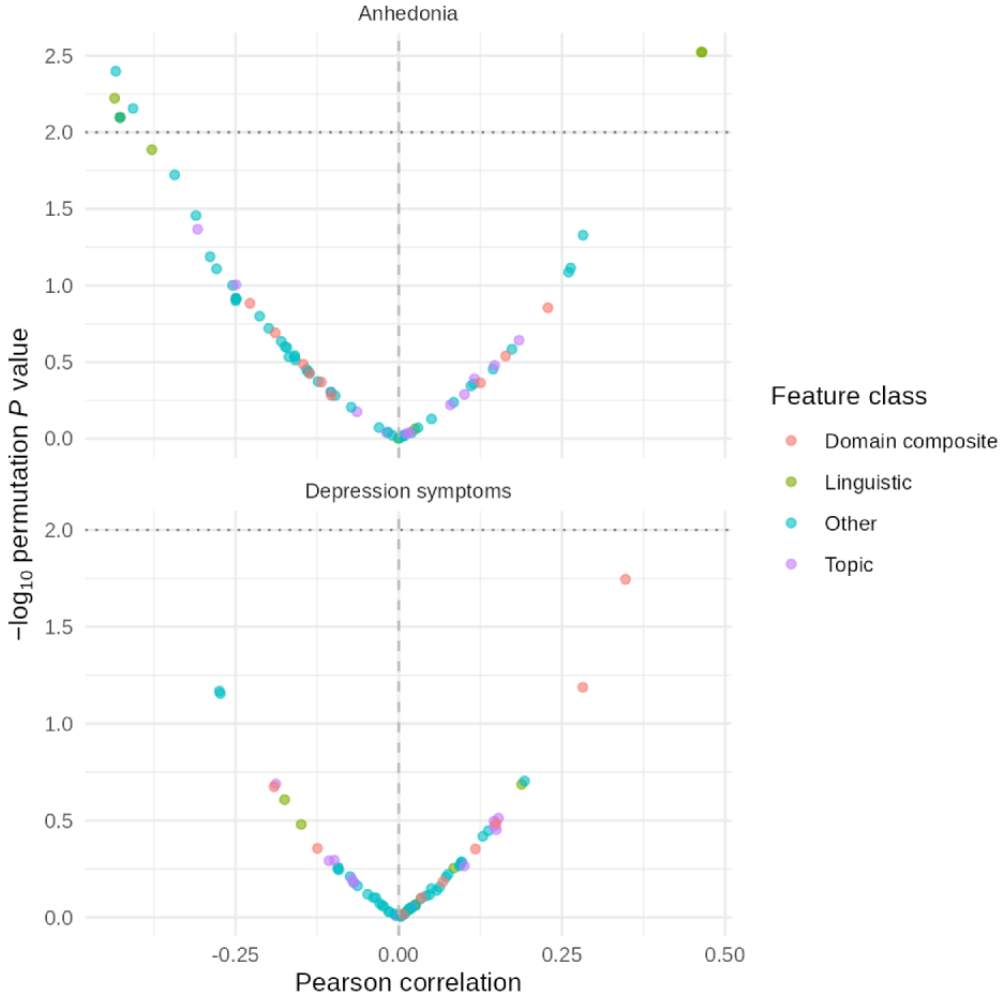
Feature	*r*	*P* value
Hapax ratio	0.46	.003
Lexical diversity	0.46	.003
Content word count	–0.44	.006
Mean token count	–0.43	.004
Total tokens	–0.43	.008
Total words	–0.43	.008
Sentiment word count	–0.43	.008
Character count	–0.43	.008
Total duration (seconds)	–0.41	.007
Hapax count	–0.38	.013
Sentiment compound (mean)	–0.34	.02
Mean sentiment (quality)	–0.31	.03
Topic 10: work or school	–0.31	.04
F0[b] SD (Hz)	–0.29	.06
F0 minimum (Hz)	0.28	.047
F0 synchrony absolute difference	–0.28	.08
F0 median (Hz)	0.26	.08
F0 mean (Hz)	0.26	.08
Recording duration (seconds)	–0.25	.10
Segment count	–0.25	.13

[a]MASQ-AD: Mood and Anxiety Symptom Questionnaire, Anhedonic Depression Scale.

[b]F0: fundamental frequency.

XSL•FO
RenderX

Figure 1. Volcano plot of the correlations between voice features and anhedonia and depression symptoms at week 0. The horizontal dotted line indicates the *P*=.01 interpretive threshold based on permutation tests. Each point is one of the 77 features tested per outcome. *P* values are unadjusted for multiple comparisons.



General Depression Severity

The 10 strongest correlations between general depression severity measured by the QIDS-SR and voice-based features are presented in Table 2. Correlations between all features and depression severity are presented in the bottom panel of the volcano plot in Figure 1.

No voice feature was associated with general depression severity at the *P*=.01 threshold. The strongest associations were small and mostly acoustic: higher QIDS-SR severity was nonsignificantly associated with a higher MFCC domain composite (*r*=0.35; *P*=.02), a higher shimmer domain composite (*r*=0.28; *P*=.06), and lower F0 mean and median (*r*s=–0.27; *P*s=.07). A higher MFCC composite indicates a greater MFCC-derived spectral or cepstral envelope, and slightly higher shimmer with slightly lower pitch is consistent with a somewhat flatter, less stable delivery. Remaining associations, including the pitch and spectral domain composites, lexical measures, topic proportions, and synchrony indices, were weak and nonsignificant.

In summary, for general depression severity, no associations exceeded the *P*=.01 threshold, with only small, acoustically oriented effects and minimal linguistic contributions. This contrasts with the more coherent and statistically significant linguistic signal observed for anhedonia.

XSL•FO
RenderX

Table 2. Top 10 voice-feature correlations with depression severity (QIDS-SR[a]) and permutation-test *P* values.

Feature	*r*	*P* value
MFCC[b] domain	0.35	.02
Shimmer domain	0.28	.06
F0[c] mean (Hz)	–0.27	.07
F0 median (Hz)	–0.27	.07
Lexical synchrony (Jaccard)	0.19	.20
Pitch domain	–0.19	.21
Average word length	0.19	.21
Topic 0: recent weeks	–0.19	.20
Hapax ratio	–0.17	.25
Topic 7: time spent or scheduling	0.15	.31

[a]QIDS-SR: Quick Inventory of Depressive Symptomatology, Self-Report.

[b]MFCC: mel-frequency cepstral coefficient.

[c]F0: fundamental frequency.

Associations Between Baseline-Adjusted Week-4 Symptoms and Baseline Voice Features

Anhedonia

The top 20 strongest correlations between baseline voice features and week-4 anhedonia symptoms (MASQ-AD residualized for baseline MASQ-AD and ABM) are presented in Table S5 in Multimedia Appendix 2. Correlations between all features and baseline-adjusted week-4 anhedonia are presented in the top panel of the volcano plot in Figure S2 in Multimedia Appendix 2. These associations index where a participant stood at week 4 relative to the level predicted by their baseline severity and treatment assignment.

In these baseline-adjusted week-4 MASQ-AD analyses, no voice feature was associated with week-4 anhedonia at the *P*=.01 threshold. The strongest association was with a lower share of the "recent weeks or check-ins" topic (topic 0; *r*=–0.31; *P*=.06), followed by shorter interview duration (*r*=–0.29; *P*=.08) and larger participant-interviewer pitch mismatch (lower F0 synchrony; *r*=0.27; *P*=.09), all of which did not reach statistical significance. Given that none of these associations met the *P*=.01 threshold in a small (n=36) and underpowered longitudinal subsample, they should be regarded as preliminary signals requiring replication rather than reliable predictors.

General Depression Severity

The top 20 strongest correlations between baseline voice features and baseline-adjusted week-4 general depression symptoms (QIDS-SR residualized for baseline QIDS-SR and ABM) are presented in Table S6 in Multimedia Appendix 2. Correlations between all features and baseline-adjusted week-4 depression severity are presented in the bottom panel of the volcano plot in Figure S2 in Multimedia Appendix 2.

In these baseline-adjusted week-4 QIDS-SR analyses, 1 feature was associated with week-4 depression severity at the *P*=.01 threshold: the jitter domain composite (*r*=–0.46; *P*=.005), such that higher-than-predicted week-4 severity was associated with lower baseline voice perturbation, suggesting reduced cycle-to-cycle instability. Subthreshold trends included the shimmer domain composite (*r*=–0.27; *P*=.10), turn-level pitch synchrony (smaller participant-interviewer mean F0 difference; *r*=–0.31; *P*=.07), and a higher share of the time or scheduling topic (topic 7; *r*=0.29; *P*=.09).

Overall, higher baseline-adjusted week-4 depression severity showed a single significant association reflecting reduced baseline voice perturbation (lower jitter). The pattern suggests small, acoustically oriented associations rather than broad linguistic ones and, like the anhedonia analyses, should be treated as preliminary, given the small longitudinal subsample.

Discussion

Principal Findings

In this study, we conducted an exploratory analysis of whether speech-derived features from a semistructured clinical interview are associated with individual differences in anhedonia severity in a sample of adults with elevated depression, and whether baseline features are associated with anhedonia 4 weeks later after adjusting for baseline severity. Using a multimodal framework spanning acoustic, prosodic, linguistic, and topic-based features, and applying a conservative *P*=.01 interpretive threshold, we observed a coherent cross-sectional profile associated with higher baseline anhedonia: reduced speech output (fewer content and total words and shorter speaking duration) alongside higher lexical-diversity ratios. Weaker, nonsignificant associations pointed toward reduced activity-related (work or school) content, more negative sentiment, and attenuated pitch variability.

In baseline-adjusted week-4 analyses, no baseline feature was associated with week-4 anhedonia at the *P*=.01 threshold, although the strongest nonsignificant effects again involved reduced speech output and engagement. The only week-4 association to exceed the significance threshold was for general

XSL•FO
RenderX

depression severity, where lower baseline voice perturbation (jitter) predicted higher-than-expected week-4 symptoms.

Together, these preliminary findings are consistent with the possibility that anhedonia is expressed in speech more through reduced behavioral engagement and altered linguistic efficiency than through uniform flattening of vocal expressivity, and they are compatible with a dimensional approach to speech-based phenotyping of depressive symptoms. Because these associations were estimated in a small sample across many features, they should be regarded as hypothesis-generating rather than confirmatory even under the conservative threshold applied here.

Comparison With Prior Work

One notable finding was that higher anhedonia was associated with lexical brevity and diversity (shorter speech, higher hapax ratio and lexical diversity, and fewer content words). We interpret the diversity finding cautiously, because hapax ratio and lexical diversity (type-token ratio) are mechanically sensitive to text length. With fewer total words, the proportion of unique words tends to be higher simply because there are fewer opportunities to repeat, so apparently "greater diversity" may be largely an artifact of reduced speech quantity rather than richer content. In a semistructured interview, short responses to multiple prompts can introduce new vocabulary without much repetition, further inflating diversity ratios. The concurrent finding of fewer content words is consistent with this account. We therefore treat the lexical-diversity associations as potentially artifactual and mechanically driven by speech quantity rather than as evidence of more elaborate or information-dense speech [17]. Consistent with this, a sensitivity analysis adjusting for total speech quantity substantially attenuated the lexical-diversity associations (Table S2 in Multimedia Appendix 2).

Reduced speech output was particularly evident when participants responded to the HRSD-17 interview item assessing recent activity, a domain closely tied to behavioral engagement in depression. Topic-model features pointed in the same direction, though this association was nonsignificant at our P=.01 criterion: reduced work- or school-related content at baseline (r=–0.31; P=.04) and, in the baseline-adjusted week-4 analyses, less check-in-related content among participants with higher-than-predicted week-4 anhedonia (r=–0.31; P=.06). These patterns are consistent with contemporary models of anhedonia in depression that emphasize impairments in motivation, effort allocation, and goal-directed behavior, often described as motivational or anticipatory anhedonia, rather than a simple reduction in hedonic capacity [4]. Reduced verbal output and narrower topical focus during activity-oriented clinical interviews may therefore index clinically meaningful aspects of behavioral activation and reward pursuit [18], although our cross-sectional design cannot establish that these speech features are caused by these processes.

In addition to less speech, anhedonia was associated with less expressive speech. Affective flattening is often described as reduced variability in vocal affect, with less modulation of pitch, intensity, and prosodic contours [17,19,20]. The observed flatter F0 variability, although nonsignificant at the P=.01 criterion, is directionally consistent with this account, suggesting a narrower expressive range, and the slightly higher pitch floor could indicate a constrained pitch band anchored higher but with less dynamic movement. Stratified analyses in Table S3 in Multimedia Appendix 2 indicated that the pitch associations were broadly consistent across sex. However, given that only 9 participants were male, these within-sex analyses are themselves underpowered and should not be taken as strong evidence that the findings are sex-invariant. Together, these prosodic features offer convergent but tentative evidence that anhedonia may be reflected not only in what participants say but in how vocally expressive their speech sounds.

The baseline-adjusted week-4 associations were weaker than the cross-sectional findings and based on a much smaller subsample. For anhedonia, no baseline feature was associated with week-4 symptoms at the P=.01 threshold. For general depression severity, one association exceeded the threshold: lower baseline voice perturbation (the jitter domain composite) was associated with higher-than-predicted week-4 severity. Notably, across analyses, anhedonia was more consistently associated with temporal and quantity-based measures than with pitch-based indices, which are known to vary across tasks and samples [21,22].

Implications

These findings are consistent with the broader argument for targeting specific symptom dimensions rather than global depression severity when developing speech-based models. Depression is a heterogeneous syndrome comprising partially dissociable symptom clusters that may differ in their behavioral and neurobiological expression [23-25]. In the current study, speech features showed somewhat stronger and more interpretable associations with anhedonia than with overall depression severity, which would be consistent with speech more directly indexing motivational and behavioral engagement processes central to anhedonia. We note, however, that this comparison is not definitive: it could partly reflect restricted range in total depression scores, given that study entry required elevated depression (Figure S1 in Multimedia Appendix 2), as well as measurement differences between the self-reported MASQ-AD and QIDS-SR. If replicated, speech-derived markers of verbal engagement and topic content could eventually complement symptom scales as low-burden, ecologically valid indices of motivational engagement. This clinical potential remains aspirational at the current stage of evidence. Future studies should assess additional depressive dimensions (eg, psychomotor retardation, cognitive control, rumination, and anxiety-related arousal) to determine whether distinct speech signatures map onto different facets of depression [26].

Limitations

Several limitations temper these conclusions. First, and most importantly, the sample was small (analytic: N=45; cross-sectional: n=44 for anhedonia and n=45 for depression severity; longitudinal: n=36). With this many derived features, small samples make individual associations vulnerable to instability and to both false positives and false negatives, and they preclude reliable multivariate or machine learning models. The longitudinal analyses in particular were powered to detect

XSL•FO
RenderX

only large effects and rest on at most a single significant association, so they are best read as hypothesis-generating.

Second, we tested many features, and although we reduced the feature set and adopted a conservative P=.01 threshold, we did not apply a formal correction for multiple comparisons. Each outcome was correlated with 77 features, and the 4 reported analyses comprise 308 univariate tests. Although permutation testing provides a data-driven null, it does not by itself control the familywise or false-discovery rate. The associations should therefore be interpreted as preliminary and exploratory rather than confirmed effects.

Third, the sample was predominantly female (36/45, 80%) and was drawn entirely from a depression intervention trial. The sex imbalance is a substantive limitation because several vocal features, particularly pitch, are sex-sensitive. Our within-sex analyses were underpowered and cannot rule out that some associations differ by sex. Because all participants had elevated depression and the speech came from a depression interview, we also cannot determine whether the observed features are specific to anhedonia versus depression more broadly, and the restricted range in depression severity limits comparisons across symptom dimensions. These features of the sample limit generalizability to community, nonclinical, or transdiagnostic populations, and the claims in this report should be read as applying to adults with depression assessed in a clinical interview rather than to speakers in general.

Fourth, data quality was variable. Of the 45 analytic participants, 14 were flagged as having potentially problematic data (low SNR, frequent ASR failures, or very short transcripts); yet, all were retained to preserve power. Retaining lower-quality recordings adds measurement noise and further increases uncertainty about the stability of the reported associations. In a sensitivity analysis excluding the flagged participants, the lexical-diversity associations were essentially unchanged, whereas the speech-quantity associations attenuated but retained the same direction (Table S1 in Multimedia Appendix 2). Results should nonetheless be interpreted with this in mind, and future work should preregister quality thresholds.

Finally, the speech data were derived from a semistructured clinical interview, which may constrain the range and expression of vocal and linguistic features relative to more naturalistic contexts. Interview prompts can shape speech quantity, topic distribution, and prosody in ways that may not generalize to everyday communication, so the observed associations may reflect how depressive symptoms manifest under evaluative, goal-directed conversation rather than spontaneous speech. Future studies incorporating more ecologically valid speech samples, such as passive audio capture or ambulatory voice diaries, will be important for assessing generalizability.

Conclusions

Within the constraints of a small, exploratory, predominantly female clinical sample, this study adds to an emerging literature, suggesting that anhedonia may be detectable in the speech of adults with depression. Higher anhedonia was associated most robustly with briefer speech (fewer total and content words and shorter duration), with weaker, nonsignificant associations with more negative sentiment, flatter pitch variation, and less work or school topic content, a pattern more consistent with reduced behavioral and verbal engagement than with uniform flattening of expressivity. The accompanying increase in lexical diversity is most plausibly an artifact of reduced speech quantity. More broadly, the results illustrate both the promise and the current limits of speech-based phenotyping of specific symptom dimensions: if these signals replicate in larger, more diverse, and more naturalistic samples with appropriate multiplicity control, voice-derived indices of verbal engagement could become useful, low-burden complements to existing symptom measures and may also provide an important insight into how anhedonia manifests in natural behavior.

Acknowledgments

The authors thank Kayla Caffey, Jocelyn Labrada, and the many volunteer research assistants who contributed to this study. The authors are especially grateful to the individuals who participated in the study and allowed them to learn from their data.

Funding

Preparation of this article was supported by the National Institute of Mental Health, National Institutes of Health (grants R01MH131750 and R33MH109600). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The speech recordings analyzed in this study contain identifiable voice data from clinical interviews and therefore cannot be shared publicly to protect participant confidentiality. Deidentified, derived participant-level feature tables and the analysis code that supports the findings are available in the Texas Data Repository [27]. Requests for access to additional deidentified data may be directed to the corresponding author and will be considered subject to a data use agreement and the original ethics approval.

Generative AI Use

The authors used a generative AI assistant, Anthropic's Claude Opus 4.8, to support language editing and to help organize and document analysis code. Generative AI was not used to generate study data, conduct the statistical analyses, or produce results.

XSL•FO
RenderX

All AI-assisted text and code were reviewed, verified, and edited by the authors, who take full responsibility for the content of the manuscript.

Authors' Contributions

Conceptualization: CGB
Data curation: CGB, KJH
Formal analysis: CGB
Funding acquisition: CGB
Methodology: CGB, KJH
Software: CGB
Supervision: CGB
Writing—original draft: CGB
Writing—review and editing: CGB, RAS, OB, KJH

Conflicts of Interest

None declared.

Multimedia Appendix 1

STROBE checklist.
[PDF File (Adobe PDF File), 96 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Baseline symptom score distributions, sensitivity analyses (excluding quality-flagged participants and adjusting for speech quantity), sex-stratified pitch associations, topic model interpretation, baseline-adjusted week-4 associations with anhedonia and depression severity, and the full list of voice features analyzed.
[PDF File (Adobe PDF File), 77 KB-Multimedia Appendix 2]

References

1. McMakin DL, Olino TM, Porta G, Dietz LJ, Emslie G, Clarke G, et al. Anhedonia predicts poorer recovery among youth with selective serotonin reuptake inhibitor treatment-resistant depression. J Am Acad Child Adolesc Psychiatry. 2012;51(4):404-411. [FREE Full text] [doi: 10.1016/j.jaac.2012.01.011] [Medline: 22449646]
2. Watson D, Weber K, Assenheimer JS, Clark LA, Strauss ME, McCormick RA. Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom scales. J Abnorm Psychol. 1995;104(1):3-14. [doi: 10.1037//0021-843x.104.1.3] [Medline: 7897050]
3. Rizvi SJ, Pizzagalli DA, Sproule BA, Kennedy SH. Assessing anhedonia in depression: potentials and pitfalls. Neurosci Biobehav Rev. 2016;65:21-35. [FREE Full text] [doi: 10.1016/j.neubiorev.2016.03.004] [Medline: 26959336]
4. Treadway MT, Zald DH. Reconsidering anhedonia in depression: lessons from translational neuroscience. Neurosci Biobehav Rev. 2011;35(3):537-555. [FREE Full text] [doi: 10.1016/j.neubiorev.2010.06.006] [Medline: 20603146]
5. Pizzagalli DA. Depression, stress, and anhedonia: toward a synthesis and integrated model. Annu Rev Clin Psychol. 2014;10:393-423. [FREE Full text] [doi: 10.1146/annurev-clinpsy-050212-185606] [Medline: 24471371]
6. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 2015;71:10-49. [doi: 10.1016/j.specom.2015.03.004]
7. Schuller BW, Batliner A. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. New Jersey. John Wiley & Sons; 2013.
8. Low LS, Maddage NC, Lech M, Sheeber LB, Allen NB. Detection of clinical depression in adolescents' speech during family interactions. IEEE Trans Biomed Eng. 2011;58(3):574-586. [FREE Full text] [doi: 10.1109/TBME.2010.2091640] [Medline: 21075715]
9. Alghowinem S, Goecke R, Wagner M, Epps J, Breakspear M, Parker G. Detecting depression: a comparison between spontaneous and read speech. 2013. Presented at: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 26-31, 2013:7547-7551; Vancouver. [doi: 10.1109/icassp.2013.6639130]
10. Mao K, Wu Y, Chen J. A systematic review on automated clinical depression diagnosis. Npj Ment Health Res. 2023;2(1):20. [FREE Full text] [doi: 10.1038/s44184-023-00040-z] [Medline: 38609509]
11. Cummins N, Dineley J, Conde P, Matcham F, Siddi S, Lamers F, et al. Multilingual markers of depression in remotely collected speech samples: a preliminary analysis. J Affect Disord. 2023;341:128-136. [FREE Full text] [doi: 10.1016/j.jad.2023.08.097] [Medline: 37598722]

XSL•FO
RenderX

12. Quatieri TF, Malyska N. Vocal-source biomarkers for depression: a link to psychomotor activity. 2012. Presented at: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH); September 9-13, 2012:1059-1062; Portland. [doi: 10.21437/Interspeech.2012-311]
13. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453-1457. [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
14. Hsu KJ, Shumake J, Caffey K, Risom S, Labrada J, Smits JAJ, et al. Efficacy of attention bias modification training for depressed adults: a randomized clinical trial. Psychol Med. 2022;52(16):3865-3873. [FREE Full text] [doi: 10.1017/S0033291721000702] [Medline: 33766151]
15. Wardenaar KJ, van Veen T, Giltay EJ, de Beurs E, Penninx BW, Zitman FG. Development and validation of a 30-item short adaptation of the Mood and Anxiety Symptoms Questionnaire (MASQ). Psychiatry Res. 2010;179(1):101-106. [doi: 10.1016/j.psychres.2009.03.005] [Medline: 20472297]
16. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry. 2003;54(5):573-583. [doi: 10.1016/s0006-3223(02)01866-8] [Medline: 12946886]
17. Lin Q, Wu X, Lei Y, Cheng W, Huang S, Wang W, et al. Acoustic signatures of depression elicited by emotion-based and theme-based speech tasks. BMJ Ment Health. 2025;28(1):e301858. [FREE Full text] [doi: 10.1136/bmjment-2025-301858] [Medline: 41027679]
18. Barch DM, Pagliaccio D, Luking K. Mechanisms underlying motivational deficits in psychopathology: similarities and differences in depression and schizophrenia. Curr Top Behav Neurosci. 2016;27:411-449. [doi: 10.1007/7854_2015_376] [Medline: 26026289]
19. Wadle LM, Ebner-Priemer UW, Foo JC, Yamamoto Y, Streit F, Witt SH, et al. Speech features as predictors of momentary depression severity in patients with depressive disorder undergoing sleep deprivation therapy: ambulatory assessment pilot study. JMIR Ment Health. 2024;11:e49222. [FREE Full text] [doi: 10.2196/49222] [Medline: 38236637]
20. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J Neurolinguistics. 2007;20(1):50-64. [FREE Full text] [doi: 10.1016/j.jneuroling.2006.04.001] [Medline: 21253440]
21. Zhao Q, Fan HZ, Li YL, Liu L, Wu YX, Zhao YL, et al. Vocal acoustic features as potential biomarkers for identifying/diagnosing depression: a cross-sectional study. Front Psychiatry. 2022;13:815678. [FREE Full text] [doi: 10.3389/fpsyt.2022.815678] [Medline: 35573349]
22. Tan EJ, Neill E, Kleiner JL, Rossell SL. Depressive symptoms are specifically related to speech pauses in schizophrenia spectrum disorders. Psychiatry Res. 2023;321:115079. [FREE Full text] [doi: 10.1016/j.psychres.2023.115079] [Medline: 36716551]
23. Veal C, Tomlinson A, Cipriani A, Bulteau S, Henry C, Müh C, et al. Heterogeneity of outcome measures in depression trials and the relevance of the content of outcome measures to patients: a systematic review. Lancet Psychiatry. 2024;11(4):285-294. [doi: 10.1016/S2215-0366(23)00438-8] [Medline: 38490761]
24. Fried EI, Nesse RM, Zivin K, Guille C, Sen S. Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. Psychol Med. 2014;44(10):2067-2076. [FREE Full text] [doi: 10.1017/S0033291713002900] [Medline: 24289852]
25. Fried EI, Nesse RM. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. BMC Med. 2015;13:72. [FREE Full text] [doi: 10.1186/s12916-015-0325-4] [Medline: 25879936]
26. Weisenburger RL, Mullarkey MC, Labrada J, Labrousse D, Yang MY, MacPherson AH, et al. Conversational assessment using artificial intelligence is as clinically useful as depression scales and preferred by users. J Affect Disord. 2024;351:489-498. [doi: 10.1016/j.jad.2024.01.212] [Medline: 38290584]
27. Beevers CG. Files to reproduce manuscript results [dataset]. Version 1. Texas Data Repository. 2026. URL: https://dataverse.tdl.org/dataset.xhtml?persistentId=doi:10.18738/T8/1GTUDA [accessed 2026-11-04]

Abbreviations

ABM: attention bias modification
ASR: automatic speech recognition
DSM-IV: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
eGeMAPS: extended Geneva Minimalistic Acoustic Parameter Set
F0: fundamental frequency
HRSD-17: 17-item Hamilton Rating Scale for Depression
MASQ-AD: Mood and Anxiety Symptom Questionnaire, Anhedonic Depression Scale
MFCC: mel-frequency cepstral coefficient
QIDS-SR: Quick Inventory of Depressive Symptomatology, Self-Report

XSL•FO
RenderX

RMS: root mean square
SNR: signal-to-noise ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
WPM: words per minute

Edited by L MacNeill; submitted 27.Feb.2026; peer-reviewed by K Cho; comments to author 27.May.2026; revised version received 04.Sep.2026; accepted 04.Sep.2026; published 10.Nov.2026

Please cite as:
Beevers CG, Stewart RA, Brady O, Hsu KJ
Voice-Derived Speech Features of Anhedonia in Adults With Depression: Exploratory Secondary Analysis of a Semistructured Clinical Interview
JMIR Form Res 2026;10:e94315
URL: https://formative.jmir.org/2026/1/e94315
doi: 10.2196/94315
PMID:

©Christopher G Beevers, Rochelle A Stewart, Olivia Brady, Kean J Hsu. Originally published in JMIR Formative Research (https://formative.jmir.org), 10.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX