DETECCIÓN Y BIOMARCADORES

Vocal biomarkers for depression: a technical deep dive

SentirIA Research Papers · 2026

Infraestructura de detección temprana en salud mental

Este documento es parte de la base científica de SentirIA,
plataforma de detección temprana y monitoreo continuo de deterioro en salud mental.
No constituye diagnóstico clínico. La evaluación es responsabilidad del profesional.

Vocal biomarkers for depression: a technical deep dive

Voice analysis can detect depression with pooled accuracy of 81% and sensitivity of 84%, according to the largest meta-analysis to date (Maran et al., 2025; 105 studies). Specific acoustic features — reduced pitch variability, increased jitter and shimmer, slower speech rate, and longer pauses — correlate significantly with depression severity, with individual feature correlations reaching r = −0.54 for minimum pitch and effect sizes (η^2) up to 0.31 for pitch contour slope. For SentirIA's WhatsApp-based approach, the first published WhatsApp voice message depression study (Otani et al., 2026) achieved 91.67% accuracy for women and 80% for men using standard machine learning on Opus-compressed audio, proving the platform's viability. The field's dominant toolkit, openSMILE with the eGeMAPS 88-feature set, provides a standardized extraction pipeline validated across hundreds of clinical studies, while per-patient baseline models requiring as few as 10 longitudinal recordings can track individual deviations with clinically meaningful precision.

Acoustic features that reveal depression's vocal fingerprint

Depression alters voice production through psychomotor retardation, reduced respiratory support, and impaired vocal fold control. These physiological changes manifest as measurable shifts across multiple acoustic dimensions.

Pitch (F0) provides the strongest single-feature signal. A landmark multisite genetic study of 7,654 participants (Flint et al., 2024, Molecular Psychiatry) found that the interquartile range of F0 change speed showed β = −1.07 ($P_{FDR} = 6.8 \times 10^{-58}$), the largest effect size observed. Depressed individuals speak in a narrower pitch range with reduced variability — essentially more monotone. Menne et al. (2024, BMC Psychiatry) reported minimum pitch correlations with BDI-II of r = −0.53 to −0.54 ($p < 0.001$), while pitch contour slope showed the largest group-level effect size at $\eta^2 = 0.30–0.31$. Normal male F0 ranges from 80–175 Hz (mean ~113 Hz) and female F0 from 160–270 Hz (mean ~210 Hz); depressed speakers show reductions in both mean and variability. Treatment responders in a sertraline trial showed increased F0 coefficient of variation ($p = .01$), while non-responders showed no change.

Jitter and shimmer — cycle-to-cycle perturbations in pitch and amplitude — increase with depression severity, reflecting impaired neuromuscular control of the vocal folds. Healthy adults show local jitter of approximately 0.49–0.62% (threshold for pathology: >1.04%). Menne et al. (2024) found jitter correlated with BDI-II at r > 0.51 ($p < 0.001$) during positive storytelling, making it one of the strongest individual acoustic-depression correlations in the literature. Shimmer similarly correlated at r > 0.40, with APQ5 shimmer selected in every cross-validation iteration of their best-performing SVM model. Silva et al. (2021, Journal of Voice) confirmed significantly elevated jitter and shimmer in a depressed cohort (n = 54 vs. 90 controls).

Harmonics-to-Noise Ratio (HNR) typically ranges 15–20 dB in healthy speakers and decreases

with depression as more breathiness and aspiration enter the voice. Quatieri and Malyska (MIT/Columbia) found significant negative Spearman correlations between HNR and HAMD/QIDS scores, though some studies report null findings, suggesting HNR is less robust as a standalone marker than jitter or shimmer.

Formant frequencies reflect articulatory precision. Depressed speakers show reduced vowel space area — compressed F1/F2 distributions consistent with psychomotor retardation and reduced articulatory effort. Williamson et al. (2013) exploited formant coordination (the correlation structure of F1, F2, F3 trajectories) as a motor incoordination biomarker, achieving sensitivity/specificity of 0.86/0.64 and AUC of 0.70 with formant features alone.

Speech rate drops approximately 10–30% in depressed speech versus healthy controls. Normal conversational English runs 4–5 syllables/second; one ambulatory study measured depressed speakers at 1.77 words/second (SD 0.57). Cummins et al. (2023, Journal of Affective Disorders) found that across a large multilingual clinical dataset, slower speech rate and articulation rate showed consistently stronger effect sizes than pause-related features.

Pause patterns elongate during depressive episodes. Mundt et al. (2012, Biological Psychiatry; n = 105) found treatment responders had significantly fewer pauses, less total pause time, and shorter recordings ($p < .01$ for all). Switching pause duration correlated with Hamilton Depression Rating Scale scores at $r = 0.47$ ($p = 0.05$) in women. Mean speech pause length in ambulatory samples averaged 0.26 seconds (SD 0.12), though pause features showed high within-person variability (ICC = 0.36).

MFCCs (Mel-Frequency Cepstral Coefficients) capture the spectral envelope of speech and are among the most discriminative features in multivariate models. Menne et al. (2024) reported MFCC correlations with BDI-II of $r = 0.32–0.40$, with SHAP analysis revealing MFCCs as the dominant contributors to their SVM classifier. Specific coefficients — MFCC2, MFCC4, MFCC7, and MFCC9 — showed significant differences between depressed and control groups across emotional tasks (Zhao et al., 2022).

Additional discriminative features include the alpha ratio (spectral energy balance above/below 1000 Hz; $\eta^2 > 0.20$), Cepstral Peak Prominence (CPP values halved in depressed women), spectral flux (reflecting spectral instability), and energy/loudness reduction ($\eta^2 = 0.19–0.21$ for mean loudness). Voice Onset Time (VOT) remains underexplored in depression specifically, though related articulatory timing measures show significant positive correlations with severity.

Quantitative thresholds and classification performance

No single threshold value reliably separates depressed from non-depressed speech — the field relies on multivariate models that combine features for clinical utility. Classification performance has been systematically evaluated in two major 2025 meta-analyses:

Meta-analysis	Studies	Pooled accuracy	Sensitivity	Specificity
Maran et al. (2025, JMIR)	105	0.81 (CI: 0.79–0.83)	0.84 (CI: 0.81–0.86)	0.83 (CI: 0.79–0.86)
Liu et al. (2024, JAMIA)	25	0.87	0.82	0.85
Briganti & Lechien (2025, J Voice)	12	78–96.5%	—	AUC: 0.71–0.93

The best individual study results come from Menne et al. (2024), whose 10-feature SVM model achieved AUC = 0.93 on a clinical sample of 96 participants, using pitch, shimmer, MFCC, loudness, and rate features. On the DAIC-WOZ benchmark, wav2vec 2.0 fine-tuned models reached 96.5% accuracy for binary classification, while multimodal fusion (wav2vec + BERT on audio + text) achieved F1 = 96.66% and PHQ-8 regression MAE = 2.88. However, a critical 2025 systematic review found that of 66 DAIC-WOZ papers assessed, only 5 met minimal reproducibility standards — many suffer from subject leakage that inflates performance.

The table below summarizes correlation coefficients between individual acoustic features and depression severity scales:

Feature	Correlation	Scale	Study
Minimum pitch (positive story)	r = −0.53	BDI-II	Menne et al. 2024
Minimum pitch (negative story)	r = −0.54	BDI-II	Menne et al. 2024
Jitter (positive story)	r > 0.51	BDI-II	Menne et al. 2024
Shimmer (positive/negative story)	r > 0.40	BDI-II	Menne et al. 2024
MFCC features	r = 0.32–0.40	BDI-II	Menne et al. 2024
Loudness (positive story)	r = 0.40	BDI-II	Menne et al. 2024
Switching pauses	r = 0.47	HRS-D	PMC 2023 (women)
F0 interquartile range change speed	β = −1.07	MDD diagnosis	Flint et al. 2024

For PHQ-9 regression, multivariate models predict total scores with MAE of approximately 4.37–4.65 (on a 0–27 scale). Automated voice biomarkers predicted PHQ-9 total scores with AUC = 0.821 for item 9 (suicidality screening). Most studies operate with binary classification (PHQ-9 ≥ 10 as cutoff for moderate-to-severe depression) rather than mapping to the five granular severity bands. The multi-cohort longitudinal study (2026) found distinct, non-overlapping feature sets for different symptom dimensions, with 38 features showing heterogeneous recovery trajectories — suggesting the relationship between acoustic features and severity is non-linear.

The dominant classification approaches include SVMs (historically strongest with hand-crafted features; accuracy 75–93%), Random Forests (used as AVEC baselines), and increasingly deep learning: CNNs on spectrograms, LSTMs on temporal features, and transformer-based models (wav2vec 2.0, HuBERT, WavLM) that learn representations directly from raw audio. The current

frontier involves multimodal LLMs — Zhao et al. (2025) built on Qwen2-Audio-7B with three-stage training to jointly process audio and visual features, achieving state-of-the-art results on DAIC-WOZ.

The Annals of Family Medicine study and Kintsugi's trajectory

The most clinically relevant validation study is Mazur et al. (2025), published in the *Annals of Family Medicine* (23(1):60-65, DOI: 10.1370/afm.240091). This cross-sectional study evaluated Kintsugi Voice v1 on 14,898 unique adults (training: 10,442; validation: 4,456) recruited via social media between February 2021 and July 2022. Participants responded to "How was your day?" with ≥25 seconds of free-form speech recorded on personal devices, converted to 16-kHz linear PCM for standardization. The model analyzed purely acoustic biomarkers — pitch, intonation, cadence, hesitations, pauses — without any linguistic content processing. Against PHQ-9 ≥ 10 as ground truth, the system achieved:

- Overall: Sensitivity 71.3% (CI: 69.0–73.5), Specificity 73.5% (CI: 71.5–75.5), PPV 69.3%, NPV 75.3%
- Women: Sensitivity 74.0%, Specificity 68.9%
- Men: Sensitivity 59.3%, Specificity 83.9%
- Age < 60: Sensitivity 71.9%, Specificity 71.8%
- Age ≥ 60: Sensitivity 63.4%, Specificity 86.8%
- Hispanic/Latine: Sensitivity 80.3% (highest by ethnicity)

The system generated a three-tier output — "Signs Detected" (output > 0.5631), "Not Detected" (< 0.4449), and "Further Evaluation Recommended" (0.4449–0.5631, returned for ~20% of samples). Key limitations include three of five authors being Kintsugi employees with equity, PHQ-9 self-report as ground truth rather than clinician-administered SCID, convenience sampling skewing toward younger female participants (average PHQ-9 of 9.7 versus population norms), and no comorbidity data collected.

Kintsugi's broader trajectory illustrates both the promise and challenge of commercializing vocal biomarkers. The company raised ~$30M, trained its Depression and Anxiety Model (DAM) on ~35,000 individuals and 863 hours of speech, and built a GCP-based microservices architecture (Python/Go, FastAPI, gRPC, Kubernetes). Their API-first platform integrated with telehealth workflows, requiring as little as 20 seconds of speech. A health payer deployment found standard questionnaires detected depression in 3% of members while Kintsugi flagged 33% — a finding that could reflect either superior sensitivity or elevated false positives. After spending approximately $16M over 4 years on FDA De Novo classification without achieving clearance, Kintsugi ceased commercial operations in early 2026 and open-sourced its technology on Hugging Face (`KintsugiHealth/dam` and `KintsugiHealth/dam-dataset`), framing it as "a $30M gift to the global mental health community." The pivotal clinical trial (NCT06809907) enrolled up to 1,000 adults with SCID-5 gold-standard interviews but was not completed before shutdown.

The DAIC-WOZ dataset remains the field's primary benchmark: 189 semi-structured interviews conducted by the animated virtual interviewer "Ellie" at USC's Institute for Creative Technologies, with PHQ-8 labels (≥10 = depressed). The dataset splits into 107 training, 35 development, and 47 test participants (mean age 31.5, range 18–63). The AVEC Depression Sub-Challenges (2013, 2014, 2016, 2017, 2019) drove methodological progress — from hand-crafted LBP-TOP features and PLS regression toward BERT-based text analysis and wav2vec 2.0 audio encoders. AVEC 2019 best results reached CCC = 0.430 (test), up from a 0.120 baseline, using multi-scale temporal dilated CNNs with deep contextual BERT features. A separate landmark: Flint et al. (2024, Molecular Psychiatry) studied 7,654 Chinese women in the CONVERGE study, achieving AUC = 0.90 for depression classification using 16 replicated voice pitch features, some of which proved heritable — suggesting a genetic basis for vocal depression biomarkers.

The eGeMAPS feature set and extraction toolkits

The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), defined by Eyben et al. (2016, IEEE TAFFC), has become the de facto standard for clinical speech analysis. Designed by 11 international experts, it compresses the feature space to 88 parameters selected for their ability to index affective physiological changes, proven discriminative value, and reliable automatic extractability. The architecture extracts 25 low-level descriptors organized into three categories:

Frequency parameters (6 LLDs): logarithmic F0 on semitone scale, jitter, and formant frequencies F1–F3 with F1 bandwidth. Energy/amplitude parameters (5 LLDs): shimmer, loudness (perceived signal intensity from auditory spectrum), HNR, and F2–F3 bandwidths (the latter two added in eGeMAPS over the base GeMAPS). Spectral parameters (14 LLDs): alpha ratio (energy ratio above/below 1000 Hz), Hammarberg index (peak energy ratio 0–2 kHz vs. 2–5 kHz), spectral slopes (0–500 Hz, 500–1500 Hz), formant relative energies (F1/F2/F3 energy relative to F0), harmonic differences (H1–H2, H1–A3), MFCCs 1–4, and spectral flux — with MFCCs and spectral flux being the eGeMAPS extensions that specifically improve valence detection.

These 25 LLDs are processed through statistical functionals to produce exactly 88 features:

- Arithmetic mean + coefficient of variation applied to all LLDs over voiced segments (~50 features)
- Eight additional functionals (20th/50th/80th percentiles, 20-80 range, rising/falling slope mean and SD) applied to F0 and loudness only (16 features)
- Means of spectral parameters over unvoiced segments: alpha ratio, Hammarberg index, spectral slopes 0–500 and 500–1500 Hz (4 features)
- Six temporal features: rate of loudness peaks, mean/SD of continuously voiced regions, mean/SD of unvoiced regions, number of continuous voiced regions per second (pseudo-syllable rate)
- Equivalent sound level and additional adjustments for the remaining features

The GeMAPS base set contains 62 features from 18 LLDs; eGeMAPS adds MFCC 1–4, spectral flux, and F2/F3 bandwidth for 88 total. The most discriminative eGeMAPS features for depression are voiced segments per second, spectral flux, F0 percentiles and coefficient of variation, loudness mean and variability, unvoiced region duration, and jitter/shimmer.

openSMILE (Speech & Music Interpretation by Large-space Extraction) is the dominant extraction toolkit, with 2,650+ citations and standard use in all AVEC challenges. It offers configuration sets ranging from GeMAPS (62 features) through eGeMAPS (88) to ComParE_2016 (6,373 features). Its C++ core with Python wrapper (`pip install opensmile`) supports real-time processing on mobile platforms including Android and iOS — critical for SentirIA's deployment. The toolkit comparison landscape:

Aspect	openSMILE	Praat	pyAudioAnalysis	librosa
Feature count	62–6,373	~10–40 (script-dependent)	34 short-term	~20+ configurable
Speech-specific features	✓ F0, jitter, shimmer, HNR, formants	✓ Gold-standard	✗ No jitter/shimmer/formants	✗ No speech features
Clinical feature sets	✓ eGeMAPS, ComParE, IS09–13	✓	✗	✗
Real-time capable	✓	✓	Limited	✓
ML pipeline integration	Good (pandas)	Poor	Good (built-in classifiers)	Excellent (numpy native)
Depression studies	Dominant (majority of papers)	Many clinical studies	Some studies	Academic prototypes
License	Free research; commercial license	GPL (free)	Apache 2.0	MIT

The optimal pipeline for a production system combines openSMILE (eGeMAPS) for standardized clinical features with librosa for deep learning spectrogram input and Praat (via the parselmouth Python wrapper) for supplementary voice quality analysis when needed.

Building per-patient baselines for longitudinal monitoring

Individual baseline models represent the most clinically promising approach for depression monitoring, addressing the fundamental challenge that between-subject vocal variation (male F0 range: 85–180 Hz; female: 165–255 Hz) dwarfs depression-related changes. Wadle et al. (2024, JMIR Mental Health) conducted the first intensive longitudinal within-person study, finding that

pitch variability, speech pauses, and speech rate significantly predicted momentary depression severity fluctuations within 30 patients tracked over 3 weeks (~32 assessments per patient). This within-person approach increases ecological validity and sidesteps the heterogeneity that limits population-level models.

Practical implementation requires 10–14 baseline recordings over 2–4 weeks to establish a reliable personal norm. The ReMAP study (2025–2026) used ≥10 entries to train person-specific idiographic models on weekly voice diaries from 284 adults, achieving MAE = 4.65 (R^2 = 0.34) for BDI prediction using LLM sentence embeddings combined with acoustic features. Deviation detection relies on z-score normalization ($z = (x - \mu_{personal}) / \sigma_{personal}$), with deviations >1.5–2 standard deviations from baseline flagged for clinical review. Change-point detection algorithms can identify abrupt shifts in vocal feature trajectories, while multilevel linear regression models capture gradual within-person trends.

Recording standardization is critical for WhatsApp deployment. Audio should be decoded from Opus/OGG to 16-bit PCM at ≥16 kHz before feature extraction. The logMMSE algorithm (logarithmic minimum mean square error) provides effective noise reduction, and conservative Voice Activity Detection filtering isolates speech-dense segments. Stasak and Epps (2017) showed that per-manufacturer normalization reduces accuracy variance to approximately 5% across different smartphones, while cepstral mean variance normalization (CMVN) provides device-invariant features. Spectral features (MFCCs) are more vulnerable to device variability than prosodic features (F0, speech rate) — an important consideration for feature prioritization in SentirIA's pipeline.

The WhatsApp platform introduces Opus codec lossy compression (8–16 kHz sampling, variable bitrate 6–510 kbit/s), but Otani et al. (2026, PLOS Mental Health) demonstrated this does not preclude effective classification. Their study of 160 Brazilian Portuguese speakers using WhatsApp voice messages achieved 91.67% accuracy (AUC 91.9%) for women using SVM on acoustic features, with spontaneous speech ("describe your past week") outperforming structured counting tasks (82% for women, 78% for men). Minimum recording duration should target 20–30 seconds of speech content after VAD filtering; Kintsugi validated on 25-second samples, while shorter 15-second clips achieved 70–77% accuracy. A combination of structured (counting) and semi-structured (describe your week) prompts is optimal.

Several systems have operationalized longitudinal baseline tracking. Grünerbl et al. (2015) monitored 10 bipolar patients via smartphone over 12 weeks, achieving 97% precision and recall for state change detection using voice features including HNR and F0 variance. Sonde Health's mental fitness app collects daily 30-second voice journals. The MONARCA system achieved 72–81% bipolar state recognition from smartphone data. Faurholt-Jepsen et al. (2021) collected voice data from 180 participants averaging 157 days each, demonstrating long-term feasibility. The multi-cohort analysis (2026) identified 23 non-redundant representative features across cohorts, with spectral-shape and modulation markers showing higher temporal sensitivity than energy and voice-quality features — suggesting these should be prioritized for longitudinal tracking.

Clinical deployment: from PHQ-9 mapping to ethical safeguards

Mapping acoustic features to PHQ-9 severity bands (minimal 0–4, mild 5–9, moderate 10–14, moderately severe 15–19, severe 20–27) remains an active research frontier. Most published work uses binary classification at the PHQ-9 ≥ 10 threshold. MFCC4 and MFCC7 showed positive correlations with PHQ-9 total scores across emotion tasks, with MFCC7 predicting PHQ-9 scores at $\beta = 0.90$ ($p = 0.01$). The multi-cohort analysis (2026) found that distinct, non-overlapping feature sets correspond to different symptom dimensions (somatic vs. depressed mood vs. cognitive), with somatic and depressed-mood dimensions yielding the most stable acoustic markers. This suggests SentirIA should consider mapping specific feature clusters to individual PHQ-9 items rather than predicting total scores alone — an approach that would differentiate the product from existing binary-classification tools.

Real-world deployment faces several well-documented challenges. Background noise at 5 dB SNR degrades standard features significantly, though robust features (DOCCs — Damped Oscillator Cepstral Coefficients) maintain performance (Mitra et al., 2016, SRI International). Language and accent affect model generalizability — SVMs trained on one English-speaking country showed substantial degradation on another despite the same language. Gender remains a persistent confound: women consistently show higher classification accuracy, attributed partly to training data imbalance (Kintsugi's dataset was 69% female) and partly to traditional masculine norms suppressing vocal expressivity. The WhatsApp study showed 91.67% vs. 80% accuracy for women vs. men respectively.

Ethical considerations are paramount for SentirIA's design:

- Informed consent: Voice data enables re-identification even after de-identification measures; "passive screening" implies persistent monitoring that users may not fully comprehend; consent for incidentally recorded third parties must be addressed
- Bias mitigation: Gender, age, cultural, and socioeconomic biases propagate from training data to model outputs — Kintsugi's lower sensitivity for men (59.3% vs. 74.0%) and older adults (63.4% vs. 71.9%) is instructive
- Regulatory positioning: Kintsugi's $16M FDA pursuit without clearance argues strongly for positioning as clinical decision support rather than standalone diagnostic; voice biomarker technology currently lacks a specific regulatory category
- False positive/negative risk: Current pooled sensitivity (0.84) and specificity (0.83) mean roughly 1 in 6 depressed individuals will be missed and 1 in 6 non-depressed individuals will be falsely flagged — appropriate clinical workflows must account for this
- Data governance: HIPAA and GDPR compliance for voice recordings, encrypted storage, data-minimization practices, and clear retention policies are essential

Despite these challenges, the opportunity is significant: only 4% of primary care patients are currently screened for depression. Even an imperfect automated tool operating through a

ubiquitous platform like WhatsApp (2+ billion users) could represent a transformative increase in screening coverage, particularly in underserved populations where access to mental health professionals is limited.

Conclusion

Three architectural decisions emerge as critical for SentirIA's vocal biomarker pipeline. First, the eGeMAPS 88-feature set via openSMILE provides the optimal balance of clinical validation, computational efficiency, and discriminative power — with pitch variability, jitter, speech rate, MFCC clusters, and spectral flux as the highest-value features. Second, within-person longitudinal tracking using z-score deviation detection from personal baselines (established over 10+ recordings across 2–4 weeks) outperforms population-level classification for ongoing monitoring and aligns naturally with WhatsApp's asynchronous communication pattern. Third, the Opus codec compression inherent to WhatsApp does not preclude clinically meaningful classification — the Otani et al. study demonstrates >90% accuracy is achievable, provided audio is decoded to 16-kHz PCM and preprocessed with logMMSE noise reduction and VAD filtering. The critical differentiation opportunity lies in moving beyond binary classification to granular PHQ-9 item-level mapping using symptom-dimension-specific feature clusters, combining continuous vocal monitoring with conversational PHQ-9 assessment to create a validation feedback loop that no existing system provides.