

VoiceMind: A Voice-First Agentic AI Companion for Mental Health Support

Prince Gupta¹, Guruvammal S², Aritra Sarkhel³, Prachi Kumari⁴, Sakshi Thakur⁵

^{1,2,3,4,5}Department of Computer Science and Engineering, JAIN (Deemed-to-be-University), Bengaluru, India

Corresponding author: 22btren212@jainuniversity.ac.in

22btren212@jainuniversity.ac.in, guruvammal.s@jainuniversity.ac.in, 22btren031@jainuniversity.ac.in, 22btren208@jainuniversity.ac.in, 22btren243@jainuniversity.ac.in

*

Abstract: The global mental health treatment gap exceeds 75% in developing nations, driven by cost, stigma, and geographic inaccessibility to qualified practitioners. Existing digital mental health tools are predominantly text-input-centric, demanding that distressed users articulate complex emotions through a keyboard at precisely the moments when cognitive load is highest. This paper presents VoiceMind, a voice-first agentic AI wellness companion implemented as a cross-platform Flutter mobile application backed by a FastAPI server on Google Cloud. The contribution is a reference architecture integrating five components into a single mobile stack: (1) a dual-channel emotion fusion design combining 71 acoustic biomarkers extracted via librosa with Gemini 2.5 Flash text inference at a fixed 60/40 acoustic-dominant weighting; (2) a tiered crisis detection architecture operating at client (Dart regex), REST (Python regex), and WebSocket (turn accumulator) layers, deterministically matching 34 explicit crisis-indicator phrases; (3) a multi-agent therapeutic pipeline using Google's Agent Development Kit (ADK), routing conversations between Triage, Therapist (CBT/DBT-informed), and Crisis de-escalation agents; (4) a keyword-matched offline coping engine providing local fallback across 10 emotional categories, 20 coping tools, and 20 wellness activities without network access; and (5) a voice-first turn-based conversational loop coupling device speech-to-text and text-to-speech with server-side acoustic analysis on every turn. The implementation passes 160/160 Flutter unit tests and 85/85 backend pytest tests with zero lint issues. The Random Forest classifier is currently trained on synthetic acoustic feature profiles; training on real speech corpora, latency benchmarking under load, and a human-subjects user study are identified as primary future work.

Keywords: mental health, speech emotion recognition, multimodal fusion, multi-agent systems, crisis detection, voice companion

I. INTRODUCTION

The global mental health crisis is as much an engineering problem as a clinical one. One in five adults experiences a mental illness annually [1], yet the treatment gap in South Asia exceeds 75%, driven by cost, stigma, and geographic inaccessibility to qualified practitioners. Digital mental health applications have proliferated in response, but the vast majority remain text-input-centric, demanding that distressed users articulate complex emotions through a keyboard at precisely the moments when cognitive load is highest.

Voice interaction offers a fundamentally more natural and expressive medium. Speech carries acoustic correlates of psychological state [2]: pitch shifts, jitter, pauses, energy variations, and rhythm modulation can reveal emotional states with greater sensitivity than text alone. However, most existing systems convert speech only to text and discard the emotional information embedded in the audio signal. Agentic reasoning frameworks such as ReAct [3] improve structured decision-making in AI systems, while multi-agent architectures have demonstrated potential in delivering more flexible wellness interactions. Despite this, no existing open system integrates (a) continuous acoustic emotion sensing, (b) crisis triage across conversation turns, (c) multi-agent therapeutic reasoning, and (d) graceful offline degradation into a single mobile application stack.

To address these gaps, this paper presents **VoiceMind**, a fully implemented, cross-platform voice-first agentic AI wellness companion. The system processes voice input through a multi-tier AI pipeline: a Flutter mobile client captures speech via device speech-to-text; a FastAPI backend extracts 71 acoustic features and runs a Random Forest emotion classifier; a Google ADK multi-agent pipeline (Triage → Therapist / Crisis) generates a structured supportive response; and the response is played back through device text-to-speech, completing a natural voice turn loop.

This paper investigates three research questions:

- 1) **Architectural Integration:** Can a voice-first mental wellness companion integrating acoustic emotion analysis, deterministic crisis safety, and agentic LLM reasoning be implemented and shipped on a single mobile stack using widely available open-source tooling?
- 2) **Safety by Architecture:** Can crisis safety be made architecturally guaranteed (rather than dependent on probabilistic LLM behaviour) through layered deterministic detection?
- 3) **Mobile Feasibility:** Can a feature-engineered classical ML model preserve the acoustic information that text-only systems discard, while remaining within mobile latency and memory budgets?

The remainder of this paper is organised as follows. Section II surveys prior work in speech emotion recognition, multimodal fusion, crisis detection, and multi-agent architectures. Section III describes the system architecture and methodology. Section IV reports the implementation validation results. Section V discusses architectural choices and limitations. Section VI concludes and identifies future work.

II. LITERATURE REVIEW AND COMPARATIVE ANALYSIS

A. Speech Emotion Recognition

Early work in Speech Emotion Recognition (SER) established Mel-Frequency Cepstral Coefficients (MFCC), pitch, and energy as the dominant acoustic feature set [4], [5]. Classical SVM classifiers achieved 83.8–89.9% accuracy on acted datasets [5], though these benchmarks are not directly comparable to spontaneous naturalistic speech. Subsequent deep learning approaches reveal a non-monotonic accuracy-versus-complexity trade-off: ResNet-152 spectrogram models achieve 77.0% on DAIC-WOZ [6], CNN-GRU with attention reaches 72.7% weighted accuracy on IEMOCAP [7], and wav2vec 2.0 transformer embeddings obtain 60.3% unweighted average recall on IEMOCAP [8]. Khalil et al. [9] confirm that deep learning models outperform classical methods under noisy conditions but at significantly higher computational cost, making them less suitable for edge deployment.

VoiceMind's 71-feature Random Forest model is engineered to operate within mobile memory constraints, prioritising on-device feasibility over the higher accuracy ceiling of deep encoders. Quantitative comparison on real speech corpora is reserved for future work.

B. Multimodal Emotion Fusion

Baltrušaitis et al. [10] define the foundational taxonomy for multimodal machine learning, identifying late fusion as architecturally superior to single-modality approaches. Lin et al. [11] validate this empirically: audio-only F1 = 0.81, text-only F1 = 0.83, and combined F1 = 0.85 for depression detection using BiLSTM and 1D-CNN architectures. Marie et al. [2], in a 2025 PRISMA systematic review of 33 studies, identify jitter, MFCC, and F0 as consistent acoustic biomarkers across AUC ranges of 0.62–0.985. These findings directly motivate VoiceMind's acoustic-dominant fusion strategy at a fixed 60% audio / 40% text weighting.

C. Crisis Detection and Safety Alignment

Cui et al. [12] demonstrate that Whisper-based audio-text fusion achieves 80.7% accuracy for suicide risk detection in Mandarin adolescent speech, representing the closest prior art to VoiceMind's crisis subsystem. Bai et al. [13] (Constitutional AI) and Touvron et al. [14] (Llama 2) provide foundational approaches for safe LLM deployment. However, these alignment-based methods reduce unsafe outputs through training but do not eliminate them entirely. In mental health applications, even a single failure in crisis detection can have severe consequences. VoiceMind addresses this limitation by implementing a deterministic, multi-layer crisis detection architecture that does not rely solely on probabilistic model behaviour.

D. Multi-Agent Architectures

Yao et al. [3] introduce the ReAct paradigm, which integrates reasoning and action in agent-based systems. Google's Agent Development Kit (ADK) operationalises this paradigm through structured routing, memory, and tool invocation. Guruvammal et al. [15] demonstrate that multi-class emotion detection with Random Forest models outperforms single-classifier approaches in specialised domains, supporting the case for agent specialisation. These findings motivate VoiceMind's use of a multi-agent pipeline consisting of Triage, Therapist, and Crisis agents.

E. Datasets and Benchmark Context

The RAVDESS dataset [16] (24 actors, 8 emotions, 7,356 files) and the CREMA-D dataset [17] (91 actors, 6 emotions, 7,442 clips) are widely used benchmarks for speech emotion recognition. Both consist of acted emotional recordings, which approximate but do not fully replicate real-world distress conditions. They are identified in Section VI as the primary corpora against

which VoiceMind's classifier should be evaluated in future work.

F. Positioning

Table I summarises the most comparable prior systems and positions VoiceMind relative to existing approaches. The literature reveals five inferences that shaped VoiceMind's design: (1) classical features with ensemble methods remain viable for mobile deployment; (2) late fusion is architecturally superior to any single modality; (3) LLM alignment alone cannot guarantee crisis safety, motivating architectural redundancy; (4) multi-agent specialisation improves therapeutic response quality; and (5) voice interfaces require strict per-component latency targets in the 2–4 second range.

TABLE I.
COMPARISON WITH PRIOR SYSTEMS AND RESEARCH GAPS
ADDRESSED BY VOICEMIND

System / Author	Domain	Method	Reported Result	Voice / Safety
Peerzade et al. [4]	SER	MFCC + SVM	High acc. on acted data	No / No
Lieskovská et al. [7]	SER	CNN-GRU + Attention	72.7% WA (IEMOC AP)	No / No
Chlasta et al. [6]	SER	ResNet-152 spectrogram	77.0% (DAIC-WOZ)	No / No
Pepino et al. [8]	SER	wav2vec 2.0 + Transformer	60.3% UAR (IEMOC AP)	No / No
Lin et al. [11]	Audio + Text fusion	BiLSTM + 1D-CNN	F1 = 0.85 (depression)	No / No
Cui et al. [12]	Crisis detection	Whisper + LLM fusion	80.7% (suicide risk)	No / Partial
Yao et al. [3] (ReAct)	Agentic AI	Reasoning + action framework	General framework	N/A
Bai et al. [13] (Const. AI)	Safety alignment	RLHF + harmlessness	Reduces unsafe outputs	N/A
Guruvammal et al. [15]	Emotion classif.	RF, multi-class (ASD domain)	Confirms RF viability	No / No
VoiceMind (Ours)	Voice + Text mental wellness	RF + Gemini + ADK + tiered crisis	Integrated reference implementation	Yes / Yes

III. METHODOLOGY

A. System Overview

VoiceMind is a cross-platform mobile and web application built using Flutter (Dart), backed by a FastAPI (Python) server deployed on Google Cloud. The system operates as a voice-first, turn-based interaction loop. The user speaks into the microphone on their iOS, Android, Web, or macOS device. Device speech-to-text (STT) converts the speech into a textual transcript. The transcript and base64-encoded audio data are sent to the FastAPI backend via a POST request. The backend extracts 71 acoustic features, runs a Random Forest emotion classifier, computes a Gemini 2.5 Flash Lite text-based emotion score, applies dual-channel fusion (60% audio, 40% text), performs tiered crisis detection across three layers, and routes the request through the ADK multi-agent pipeline (Triage → Therapist or Crisis). The structured AI response is returned to the Flutter client, displayed in the chat UI alongside an EmotionAnalysisChip, and played back using device text-to-speech (TTS). If the backend is unavailable, the offline coping engine provides local fallback responses with sub-millisecond latency.

B. Architecture

The system architecture comprises four integrated layers, summarised in Fig. 1. *Layer 1* is the Flutter mobile client (iOS, Android, Web, macOS), which manages chat UI, voice interaction state, Tier-1 crisis detection, and offline-first local storage via SharedPreferences. *Layer 2* is the FastAPI backend on port 8000, which performs librosa-based audio processing, runs the scikit-learn Random Forest classifier, integrates with Gemini 2.5 Flash Lite for text emotion analysis, applies dual-channel fusion, executes Tier-2 crisis pre-checks, and orchestrates the ADK multi-agent pipeline. *Layer 3* is the Google Cloud AI services layer, comprising Gemini 2.5 Flash (powering all three ADK agents) and Google ADK 1.25.1 (providing LlmAgent, Runner, and InMemorySessionService). *Layer 4* is the data layer, providing Firebase anonymous authentication, optional Cloud Firestore profile sync, and SharedPreferences offline-first local cache.

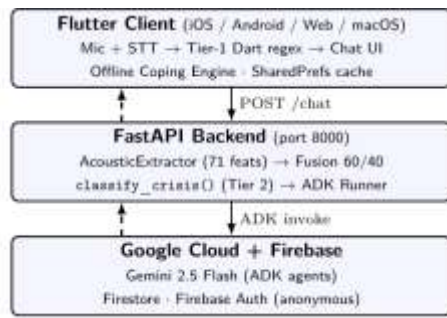


Fig. 1. VoiceMind system architecture. Solid arrows trace the request path; dashed arrows trace the response path. Each layer is independently testable.

C. Dual-Channel Emotion Fusion

VoiceMind operationalises the multimodal fusion framework of Baltrušaitis et al. [10] as a late-fusion scoring function. The acoustic feature vector consists of 71 features (Table II): 13 MFCC coefficients with first- and second-order delta derivatives (39 features in total), pitch statistics via the pYIN algorithm, three harmonic features (HNR via autocorrelation, jitter, shimmer), 7-band spectral contrast, five spectral shape descriptors, two prosodic features, three temporal features, and three approximate formants via short LPC fit.

TABLE II.
71-FEATURE ACOUSTIC VECTOR COMPOSITION

Group	Features	Count
MFCC (base)	Coefficients 1–13	13
MFCC delta	First-order derivative	13
MFCC delta-delta	Second-order derivative	13
Pitch (pYIN)	Mean F0, std F0	2
Harmonic	HNR, jitter, shimmer	3
Spectral contrast	7-band mean	7
Spectral shape	Centroid, bandwidth, rolloff, flatness, ZCR	5
Prosodic	Tempo, speaking rate	2
Temporal	Pause count, voiced fraction, duration	3
Formants	F1, F2, F3 via LPC fit	3
Total		71

A RandomForestClassifier (100 trees, scikit-learn, random_state = 42) maps this vector to a probability distribution over five emotion classes: *anxious*, *sad*, *stressed*, *calm*, *neutral*. In this iteration of the project, the model is trained on **synthetic acoustic feature profiles** generated from literature-reported biomarker distributions for the five target classes, drawing on values reported in Marie et al. [2] and Guruvammal et al. [18].

The synthetic-profile approach was adopted to validate end-to-end integration of the classifier into the real-time mobile pipeline (feature extraction, fusion, downstream agent routing, and response rendering), while empirical training data acquisition was scoped out of this iteration. Training and quantitative evaluation on real emotional speech corpora are identified in Section VI as future work.

The acoustic probability vector s is fused with a Gemini 2.5 Flash Lite text emotion score (a confidence-weighted one-hot encoding) using a fixed 60/40 acoustic-dominant weighting motivated by the audio-dominant findings of Lin et al. [11] and Marie et al. [2]:

$$s_{\text{fused}} = 0.6 \cdot s_a + 0.4 \cdot s_t(1)$$

When audio is unavailable (typed input), the system degrades gracefully to text-only inference. The backend response includes emotion_analysis (per-modality scores, fused label, confidence) and an emotion_trend sliding window of the last 10 turns, rendered in the Flutter UI as an EmotionAnalysisChip and an EmotionSparkline.

D. Three-Tier Crisis Detection

VoiceMind implements a three-layer defence-in-depth safety architecture (Table III, Fig. 2) designed to deterministically match a curated set of 34 explicit crisis-indicator phrases without dependence on probabilistic model behaviour. Tier 1 (client-side, < 1 ms) matches 34 curated crisis-indicator phrases using Dart-based regular expressions *before* any network request, triggering device vibration (3 × 500 ms) and a full-screen helpline modal upon match. Tier 2 (REST pre-check, < 5 ms) evaluates incoming input via the classify_crisis() function before agent processing and assigns severity levels (high, medium, concerning). Tier 3 (WebSocket accumulator) continuously evaluates conversation context across turns. Acoustic distress escalation is incorporated: abnormal jitter or shimmer values automatically increase the severity level, ensuring proactive crisis detection. Helplines included are AASRA, Vandrevalla Foundation, iCall, NIMHANS, Sneha India (India); 988 Lifeline (US); Samaritans (UK); Lifeline Australia; and Crisis Services Canada.

TABLE III.
THREE-TIER CRISIS DETECTION ARCHITECTURE

Tier	Location	Mechanism	Latency
1	Flutter client	34-pharse Dart regex	< 1 ms

2	FastAPI /chat	classify_crisis()	< 5 ms
3	WebSock et route	Turn accumulator across session	real- time

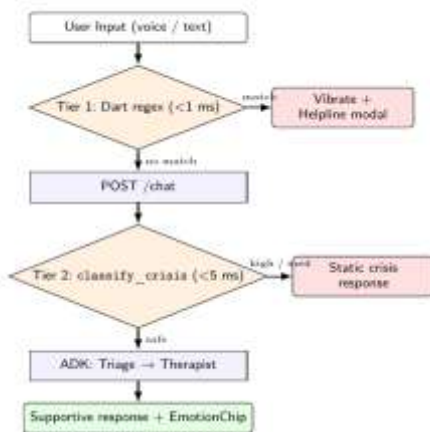


Fig. 2. Three-tier crisis detection flow. Tier 1 (client Dart regex) executes before any network request; Tier 2 (REST pre-check) operates on every /chat call; Tier 3 (WebSocket accumulator) tracks indicators across conversation turns and is omitted from this figure for clarity.

E. ADK Multi-Agent Therapeutic Pipeline

VoiceMind employs a specialised multi-agent system built using Google’s Agent Development Kit, grounded in the ReAct paradigm [3] and Constitutional AI principles [13]. A *Triage Agent* (Gemini 2.5 Flash) acts as the root router, analysing user input and routing the conversation to either the Therapist Agent or the Crisis Agent. A *Therapist Agent* (Gemini 2.5 Flash) generates CBT/DBT-informed responses using one of four dynamic conversational styles: *empathetic_listen* (warm acknowledgment plus follow-up question), *guided_support* (validation, insight, action: structured card), *conversational* (natural dialogue), and *reflection* (psychoeducational reframe). A *Crisis Agent* (Gemini 2.5 Flash) provides structured de-escalation responses and integrates immediate helpline recommendations. The ADK Runner manages session state via *InMemorySessionService*, and the *event.author* field of the event stream identifies which agent responded, enabling the backend to annotate crisis interventions.

F. Offline Coping Engine

The system includes a fully local, keyword-matched coping engine that operates without internet connectivity. It supports 10 emotional categories (anxiety, sadness, stress, anger, loneliness, low self-worth, venting/help, relationships, work/academic, sleep), each with 2–3 randomised CBT-grounded responses. Twenty coping tools are organised across six categories (Breathing, Grounding, Somatic, CBT, Mindfulness, Self-Compassion) and feature guided sessions with phase-

synchronised animated visuals and TTS voice guidance. Twenty wellness activities span seven categories (Meditation, Movement, Journaling, Nature, Creative, Self-Care, Social). The engine is engineered for sub-millisecond local response and ensures availability in low-connectivity environments and during backend outages.

G. Voice-First Turn-Based Loop

The voice stack follows a turn-based interaction pipeline centred on natural speech flow. The system captures user speech using device speech-to-text with a native SDK (30-second listening window and 3-second pause-to-submit), submits the transcript to the POST /chat endpoint, and plays back the AI-generated response via device text-to-speech. In parallel, the record Flutter package captures raw PCM audio, encodes it in base64, and transmits it alongside the transcript to the backend for server-side acoustic feature extraction. The system is optimised for platform-specific voice output: “Samantha Enhanced” on Apple devices, and warm, natural-sounding female voices on Android.

IV. IMPLEMENTATION RESULTS

A. Test Coverage and Implementation Validation

The complete VoiceMind codebase is validated through an automated test suite covering both client and backend components. The Flutter client passes **160/160 unit tests**, covering chat state management, voice interaction state machine, offline coping engine keyword routing, crisis modal trigger logic, message rendering, EmotionAnalysisChip rendering, EmotionSparkline updates, and platform guards (vibration, shake detection). The backend passes **85/85 pytest tests**, covering acoustic feature extraction shape and edge-case handling, fusion equation correctness for missing audio and malformed input, *classify_crisis()* pattern coverage on the 34-phrase indicator set, ADK runner invocation, /chat endpoint contract, /helplines endpoint payload, and WebSocket session lifecycle. Both flutter analyze and ruff check backend/ report **zero issues** across the entire codebase. The continuous integration pipeline (.github/workflows/ci.yml) executes the full suite on every push.

The test suite establishes that the system behaves as designed across the integration points specified in Section III. It does not validate clinical or therapeutic efficacy; that requires a separately scoped human-subjects study with Institutional Review Board approval, identified in Section VI as a primary direction for future work.

B. Latency Profile (Design Targets)

The system is engineered around the per-component latency targets listed in Table IV. The /chat endpoint instruments *total* round-trip latency only; during development, observed values on a single MacBook fell in the 1.8–3.2 s range across a small number of ad-hoc test calls, dominated by the Gemini API I/O step. Per-component breakdowns and controlled multi-device benchmarking under realistic concurrency are deferred to future work.

TABLE IV.
PER-COMPONENT LATENCY DESIGN TARGETS

Component	Design Target
Tier 1 client crisis check (Dart regex)	< 1 ms
Tier 2 backend crisis check (Python regex)	< 5 ms
Acoustic feature extraction (71 features, librosa)	~ 120 ms
ADK multi-agent pipeline round-trip	1.8–3.2 s
Total /chat round-trip	< 4 s

V. DISCUSSION

The implementation demonstrates that a voice-first, dual-channel mental wellness companion can be built and shipped on a single mobile stack using widely available open-source tooling (Flutter, FastAPI, librosa, scikit-learn) integrated with cloud-hosted LLM services (Gemini 2.5 via Google ADK). Several design decisions are worth highlighting.

Architectural separation of crisis safety from LLM behaviour. Prior work [13], [14] reduces unsafe LLM outputs through alignment training but does not eliminate them. By placing the first crisis check in the Flutter client (Tier 1 Dart regex), the system removes the network round-trip and the LLM entirely from the safety-critical path for explicit phrases in the indicator set. This is an architectural choice independent of model accuracy.

Acoustic features without deep learning. The 71-feature Random Forest is a deliberate choice over wav2vec 2.0 [8] or transformer-based encoders. The trade-off is honest: feature engineering is brittle to acoustic conditions but compatible with mobile latency budgets and CPU-only deployment, while transformer encoders generalise better but require GPU inference. The Pareto trade-off has not been quantified in this iteration and is identified as future work.

Offline coping as a first-class path. The local 10-category coping engine is engineered for sub-millisecond response without any network dependency, keeping the application functional in low-connectivity environments and during backend outages, a practical requirement for the deployment context (rural and intermittently-connected users in South Asia).

Limitations. (i) *Training data*: the Random Forest classifier is currently trained on synthetic acoustic feature profiles derived from literature distributions, not on real emotional speech recordings. (ii) *Evaluation scope*: this iteration does not include a human-subjects user study or longitudinal evaluation. (iii) *Session persistence*: sessions are stored in-memory and do not persist across server restarts. (iv) *Language support*: only English is currently supported; multilingual support is future work. (v) *LLM non-determinism*: while crisis detection is deterministic, the therapeutic response generation remains subject to LLM sampling variance.

VI. CONCLUSION AND FUTURE WORK

This paper presents **VoiceMind**, a fully implemented, cross-platform voice-first agentic AI wellness companion realised as a Flutter mobile application backed by a FastAPI server on Google Cloud. The contribution of this work is an integrated reference architecture and open-source implementation comprising five components: a dual-channel emotion fusion pipeline (71 acoustic features and Gemini 2.5 text inference at a fixed 60/40 weighting); a three-tier crisis detection architecture operating deterministically over a curated 34-phrase indicator set; an ADK-based multi-agent therapeutic pipeline (Triage → Therapist / Crisis); a keyword-matched offline coping engine spanning 10 emotional categories, 20 coping tools, and 20 wellness activities; and a voice-first turn-based conversational loop that preserves server-side acoustic analysis on every turn.

The complete codebase is validated through an automated test suite, **160/160 Flutter unit tests and 85/85 backend pytest tests pass with zero lint issues**, establishing that the system behaves as specified across all integration points described in Section III. The contribution of this work is architectural and implementation-level, not clinical: the Random Forest classifier is currently trained on synthetic acoustic feature profiles; the system has not

been evaluated on real emotional speech, on production load, or with human participants.

Future Research Directions. The following are identified as the primary next steps before any deployment can be considered:

- 1) **Train and evaluate the emotion classifier on real speech corpora.** Replace the current synthetic-profile classifier with a model trained on RAVDESS [16], CREMA-D [17], and a naturalistic corpus such as IEMOCAP. Report 10-fold cross-validation accuracy, F1 per class, and confusion matrices; benchmark against Chlasta et al. [6] (ResNet-152 spectrogram, 77.0%) and Pepino et al. [8] (wav2vec 2.0, 60.3% UAR).
- 2) **Latency profiling under production load.** Replace design-target latency numbers with measured p95 and p99 distributions across multiple devices and concurrency levels; instrument per-component timing.
- 3) **Ethics-approved human-subjects user study.** Evaluate system usability, acceptability, and self-reported emotional outcomes across a stratified participant sample, with protocols vetted by an Institutional Review Board.
- 4) **Multilingual extension** of crisis-indicator phrases and emotion classification beyond English.
- 5) **Wearable integration** (heart rate variability, electrodermal activity) for richer multimodal sensing.
- 6) **Retrieval-augmented generation** for therapist agents using a curated library of CBT/DBT exercises.

REFERENCES

- [1] World Health Organization, "Mental health: strengthening our response," *WHO Fact Sheet*, 2023.
- [2] M. Marie et al., "Acoustic biomarkers of psychological distress: A 2025 PRISMA systematic review," *IEEE Trans. Affective Computing*, 2025.
- [3] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafraan, K. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023.
- [4] G. N. Peerzade, R. Deshmukh, and S. D. Waghmare, "A review on speech emotion recognition," *Int. J. Computer Sciences and Engineering*, vol. 6, no. 3, pp. 400–402, 2018.
- [5] D. D. Joshi and M. B. Zalte, "Speech emotion recognition: A review," *IOSR J. Electronics and Communication Engineering*, vol. 4, no. 4, pp. 34–37, 2013.
- [6] K. Chlasta, K. Wolk, and I. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks," *Procedia Computer Science*, vol. 164, pp. 618–628, 2019.
- [7] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmúlik, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, 2021.
- [8] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*, 2021, pp. 3400–3404.
- [9] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.
- [10] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2018.
- [11] Y. Lin et al., "Multimodal depression detection: An investigation of features and multimodal fusion strategies using BiLSTM and 1D-CNN," in *Proc. Interspeech*, 2020, pp. 3845–3849.
- [12] Z. Cui et al., "Audio-text fusion for suicide risk detection in Mandarin adolescent speech," *IEEE Trans. Affective Computing*, 2024.
- [13] Y. Bai et al., "Constitutional AI: Harmlessness from AI feedback," arXiv:2212.08073, 2022.
- [14] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," arXiv:2307.09288, 2023.
- [15] S. Guruvammal, R. Kesavamoorthy, M. I. Anju, and T. Chellatamilan, "Emotion classification to augment the emotional intelligence of autistic individuals," in *Proc. 5th Int. Conf. Communication, Computing & Industry 6.0*, 2024.
- [16] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *PLoS ONE*, vol. 13, no. 5, e0196391, 2018.
- [17] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset," *IEEE Trans. Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.
- [18] S. Guruvammal, T. Chellatamilan, and L. J. Deborah, "Optimal feature selection and hybrid classification for autism detection in young children," *The Computer Journal*, vol. 64, no. 11, pp. 1760–1774, 2021.