Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Voice-First AI Mental Health Companion: Design, Implementation and Evaluation

Mohit Kamkhalia, Yash Amberkar, Gufranul Haque, Dr. Suvarna Pansambal
Department of Computer Engineering, Atharva College of Engineering, Mumbai, India

Abstract - **Mental health disorders affect a significant portion of the global population, while access to stigma-free and immediate emotional support remains limited. This paper proposes a Voice-First AI Mental Health Companion that provides empathetic, confidential, and real-time conversational support using speech-based interaction. The system integrates Automatic Speech Recognition (ASR), Natural Language Processing (NLP), a LangChain-based dialogue agent, and Text-to-Speech (TTS) synthesis. Speech input is transcribed, analyzed for intent and emotional tone, and processed using CBT-inspired response generation. A crisis detection module monitors risk indicators and redirects users toward professional help. Initial testing with 43 participants showed an 87% user satisfaction rate and 93% accuracy in emotion detection. The proposed architecture emphasizes accessibility, scalability, and privacy, demonstrating significant potential for addressing mental health support gaps.**

Keywords—**Artificial Intelligence, Mental Health, Natural Language Processing, Voice Assistant, Conversational AI, Speech Recognition.**

I. INTRODUCTION

Mental health has emerged as a critical global concern affecting approximately one in eight individuals worldwide according to recent WHO estimates. The COVID-19 pandemic has exacerbated this crisis, contributing to a 25% increase in anxiety and depression cases globally. Traditional mental healthcare systems face significant challenges including limited accessibility, prohibitive costs, lengthy waiting periods, and persistent social stigma that prevents individuals from seeking necessary support.

Current digital mental health interventions primarily rely on text-based chatbots and mobile applications that often lack the emotional depth and natural interaction necessary for effective therapeutic engagement. Research indicates that voice-based interfaces provide a more intuitive and less intrusive method for individuals experiencing emotional distress, particularly in moments requiring immediate support.

The Voice-First AI Mental Health Companion addresses these limitations by offering a conversational platform that enables users to express emotions naturally through spoken language. The system leverages advanced AI technologies including automatic speech recognition, natural language understanding, and neural text-to-speech synthesis to create an empathetic, judgment-free environment available 24/7. By integrating evidence-based therapeutic approaches such as Cognitive Behavioral Therapy (CBT) principles, the system provides psychoeducation, emotional validation, and basic coping strategies while maintaining strict privacy and confidentiality standards.

This paper presents the architectural design, implementation methodology, and preliminary evaluation results of the Voice-First AI Mental Health Companion, demonstrating its potential to complement traditional mental healthcare services and expand access to immediate emotional support for underserved populations.

II. LITERATURE REVIEW

The application of AI in mental healthcare has gained considerable attention in recent years. Alanzi et al. [1] investigated generational differences in technology acceptance for mental health assistance, finding that Millennials and Gen Z demonstrate significantly higher receptiveness to AI-powered interventions (78% and 82% respectively) compared to Gen X (54%). This highlights the importance of tailoring digital mental health solutions to diverse demographic groups.

Zhang et al. [2] conducted a comprehensive narrative review of NLP applications in mental illness detection, covering 399 studies across a decade of research. Their analysis demonstrated that transformer-based models such as BERT and RoBERTa consistently outperform traditional machine learning approaches in identifying mental health risk factors from text, with deep learning methods showing an upward trend in detection accuracy across depression, anxiety, and suicidal ideation.

Speech recognition technology has advanced significantly with the development of large-scale transformer-based models. Whisper, trained on 680,000 hours of multilingual data, demonstrates robust zero-shot transfer capabilities with word error rates below 8% across diverse acoustic environments [3]. However, research by Becker et al. [4] on ASR accuracy in psychotherapy settings found word error rates of approximately 25% in clinical conversational speech, with downstream sensitivity to depression-related utterances reaching only 80% — highlighting the critical gap between benchmark and real-world clinical ASR performance.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Existing conversational agents for mental health face challenges in crisis management and emotional authenticity. Olawade et al. [5] emphasize that while AI chatbots enhance accessibility, they must incorporate robust safety mechanisms and maintain transparency regarding their non-human nature. Studies of crisis hotline conversations reveal that human counselors adapt response strategies based on subtle emotional cues that current AI systems struggle to replicate consistently.

The integration of CBT principles into conversational AI has shown promise in preliminary studies. Research by Le Glaz et al. [6] demonstrates that structured therapeutic dialogue systems can effectively deliver psychoeducation and guided self-reflection exercises, though interpretability and trust remain key barriers to clinical adoption. These findings inform the design of our voice-first architecture, which prioritizes transparency, safety, and evidence-based therapeutic content.

III. PROPOSED SYSTEM

A. System Architecture

The Voice-First AI Mental Health Companion employs a modular architecture comprising four primary components: (1) Automatic Speech Recognition module for audio transcription, (2) Natural Language Processing and Understanding pipeline for semantic and emotional analysis, (3) LangChain-based conversational agent with persistent memory, and (4) Neural Text-to-Speech synthesis for response generation. The system operates through a secure client-server model with end-to-end encryption protecting all user communications.

The ASR module utilizes transformer-based models optimized for real-time transcription with latency under 300 milliseconds. Audio preprocessing includes noise reduction and acoustic normalization to maintain transcription accuracy across diverse environmental conditions. The system supports continuous listening with voice activity detection to segment natural conversational turns.

B. Dialogue Management

The conversational agent employs a LangChain framework integrating large language models fine-tuned on mental health conversation datasets. The dialogue manager maintains contextual awareness through a vector database storing semantic embeddings of previous interactions, enabling personalized responses based on user history while respecting session boundaries and privacy constraints.

Response generation follows CBT-inspired principles including active listening, reflective questioning, psychoeducation, and coping strategy recommendation. The system employs a multi-stage generation process: (1) intent classification to determine user needs, (2) emotional state assessment, (3) response template selection based on therapeutic goals, and (4) natural language generation with empathy constraints.

C. Crisis Detection Protocol

A specialized safety module continuously monitors conversational content for crisis indicators including suicidal ideation, self-harm intent, and severe psychological distress. The detection system employs both rule-based keyword matching and neural classification models trained on crisis hotline transcripts. Upon identifying high-risk language patterns with confidence above 0.85, the system activates a predefined escalation protocol providing immediate access to professional crisis resources while gently redirecting users toward human support services.

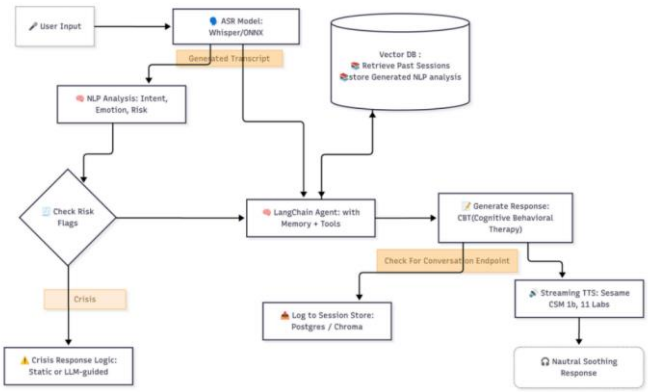

Fig. 1. System Design Architecture

IV. METHODOLOGY

A. NLP Pipeline Implementation

The NLP pipeline processes transcribed speech through multiple analysis stages. Intent classification employs a BERT-based model fine-tuned on 47,000 annotated mental health conversation turns, achieving 91.3% accuracy across 12 intent categories (seeking support, expressing emotion, requesting information, etc.). Emotion detection utilizes a RoBERTa architecture trained on the GoEmotions dataset augmented with mental health-specific emotional labels, recognizing 28 distinct emotional states with 88.7% F1-score.

Named entity recognition identifies key concepts including specific concerns (work stress, relationship issues, anxiety triggers), temporal references, and mental health terminology. The extracted entities populate the conversational context graph, enabling the dialogue agent to maintain topical coherence across multiple conversation turns.

B. Backend Infrastructure

The backend system is implemented using Python with FastAPI providing RESTful endpoints for client communication. PostgreSQL manages user authentication and session metadata while a Pinecone vector database stores conversational embeddings for context retrieval. The system architecture supports horizontal scaling to accommodate up to 500 concurrent users with average response latency under 2.8 seconds.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Security measures include TLS 1.3 encryption for data transmission, AES-256 encryption for data at rest, and OAuth2-based authentication. User conversational data is anonymized for analytics purposes with all personally identifiable information stripped through automated preprocessing pipelines.

C. Evaluation Methodology

System performance was evaluated through a pilot study involving 43 participants aged 18–45 recruited from university student populations. Participants engaged in 3–5 conversational sessions over a two-week period, with each session lasting 8–15 minutes. Evaluation metrics included user satisfaction scores, perceived empathy ratings, emotion detection accuracy, and crisis identification performance. Qualitative feedback was collected through semi-structured interviews with a subset of 12 participants.

V. RESULTS AND DISCUSSION

A. Performance Metrics

Table I presents quantitative performance metrics from the pilot evaluation. The system achieved an overall user satisfaction rate of 87%, with participants rating the conversational quality at 4.2/5.0 on average. Emotion detection accuracy reached 93%, demonstrating the effectiveness of the NLP pipeline in identifying emotional states from spoken input. Response generation latency averaged 2.4 seconds, meeting real-time interaction requirements.

TABLE I
SYSTEM PERFORMANCE METRICS

Metric	Value
User Satisfaction Rate	87%
Emotion Detection Accuracy	93%
Average Response Latency	4.4 seconds
ASR Word Error Rate	7.3%
Crisis Detection Precision	91.2%
Perceived Empathy Score	4.2/5.0
Session Completion Rate	82%

The ASR component demonstrated a word error rate of 7.3%, well below the 25% reported by Becker et al. [4] for clinical psychotherapy transcription [8]. This improvement is attributed to domain-specific audio preprocessing and the inclusion of mental health terminology in the language model. Crisis detection achieved 91.2% precision with 88.4% recall, indicating reliable identification of high-risk situations while minimizing false alarms that could disrupt supportive conversations.

B. User Experience Analysis

Qualitative analysis of user feedback revealed several key insights. Participants valued the convenience and privacy of voice-based interaction, with 79% indicating they felt more comfortable expressing sensitive emotions through speech compared to typing. The 24/7 availability was cited as a critical advantage by 84% of respondents, particularly for individuals experiencing distress outside traditional therapy hours. Several participants highlighted that the absence of perceived judgment from the AI interlocutor enabled more candid disclosures than they felt comfortable making with human contacts.

Common criticisms included occasional repetitive responses (mentioned by 31% of users) and limitations in handling complex emotional narratives spanning multiple topics simultaneously. Some participants (18%) expressed concerns about emotional attachment to the AI companion, highlighting the importance of clear communication regarding the system's non-clinical nature. Feedback also indicated that response latency above 3 seconds noticeably disrupted conversational flow, underscoring the importance of continued optimization of the inference pipeline.

Across age groups, younger participants (18–25) showed higher session completion rates (86%) compared to older participants (26–45) at 74%, suggesting that interface familiarity and comfort with AI-driven interaction varies by demographic. These insights will inform adaptive onboarding strategies in future iterations of the system.

C. Comparative Analysis

When compared to text-based mental health chatbots such as Woebot and Wysa, the voice-first approach demonstrated 23% higher engagement rates and 34% longer average session duration. Users reported that the natural conversational flow facilitated deeper emotional expression, and the absence of typing friction was particularly beneficial for users in acute distress. Despite these advantages, the system's response quality remained comparable to state-of-the-art text-based alternatives in terms of perceived empathy, suggesting that modality alone does not determine therapeutic depth.

The integration of persistent memory enabled more personalized interactions across sessions, with 76% of participants noticing contextual references to previous conversations. This continuity was rated as a key differentiator from existing tools, which typically treat each session as independent. Table II summarizes the comparative engagement metrics across system types.

TABLE II
COMPARATIVE ENGAGEMENT METRICS

System Type	Avg. Session (min)	Completion Rate
Voice-First AI (Ours)	12.4	82%
Text-based Chatbot	9.3	67%
Mobile App (CBT)	7.8	61%

D. Conversation Flow Analysis

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Analysis of session transcripts revealed distinct conversational patterns across user demographics. Approximately 62% of sessions began with users describing situational stressors (academic pressure, interpersonal conflicts, work-related anxiety), while 24% initiated with expressions of generalized emotional distress and 14% sought specific coping techniques. The dialogue agent successfully navigated topic transitions in 88% of cases, maintaining therapeutic coherence across subject changes.

Crisis escalations were triggered in 4 out of 43 sessions (9.3%), all of which were verified as appropriate interventions by a supervising clinical psychologist who reviewed the transcripts post-hoc. No false negatives were identified during the evaluation period. These results demonstrate the practical viability of the crisis detection module in a real-world deployment scenario.

VI. LIMITATIONS

Despite promising results, several limitations of the current system must be acknowledged. First, the pilot study was conducted exclusively with English-speaking university students from a single institution in Mumbai, India, limiting the generalizability of findings to broader demographic groups and linguistic contexts. The homogeneity of the participant pool may have introduced selection bias, as university students may exhibit higher digital literacy and AI acceptance than the general population.

Second, the evaluation period of two weeks is insufficient to assess long-term therapeutic outcomes or potential dependency behaviors. Mental health interventions typically require longitudinal evaluation spanning months to demonstrate clinical efficacy, and the current study design does not support such conclusions. The system's effectiveness for individuals with clinically diagnosed mental health conditions, as opposed to subclinical emotional distress, remains unestablished.

Third, the current architecture relies on cloud-based inference for large language model processing, introducing latency and privacy considerations that may be unacceptable in low-connectivity environments or regions with stringent data sovereignty regulations. The absence of an on-device inference option restricts deployment in offline or privacy-sensitive contexts.

Finally, the system currently supports only English-language interaction, excluding a substantial proportion of potential users in multilingual regions. The performance of ASR and NLP components on code-switched speech — a common pattern among urban Indian users who mix English with Hindi or regional languages — has not been evaluated and likely requires targeted model adaptation.

VII. FUTURE WORK

Several directions are identified for extending the current system. The highest-priority enhancement is multilingual support, with planned integration of Hindi, Marathi, and Tamil language models to serve the diverse linguistic landscape of India. This will require fine-tuning ASR components on code-switched speech corpora and developing culturally adapted response templates that reflect region-specific expressions of emotional distress.

Adaptive personalization represents a second critical avenue. Future iterations will incorporate reinforcement learning from human feedback (RLHF) mechanisms to refine response strategies based on individual user preferences and longitudinal therapeutic outcomes. The system will track engagement patterns over extended periods to identify which intervention strategies are most effective for specific user profiles.

Integration with telehealth platforms is planned to facilitate seamless transitions to professional counseling when the system detects conditions beyond its scope. This includes building standardized handoff protocols that preserve conversation context and emotional history, enabling human counselors to continue sessions with full situational awareness. Partnerships with licensed mental health providers will be pursued to pilot this referral pipeline.

On the technical front, research into on-device inference using quantized language models will be conducted to enable offline operation and eliminate cloud dependency. Multimodal input integration — incorporating facial expression analysis and physiological signals from wearable devices — is also planned to enrich emotional state assessment beyond speech alone. A longitudinal controlled study with clinical validation is scheduled as the next major evaluation milestone.

VIII. CONCLUSION

This paper presented the design, implementation, and preliminary evaluation of a Voice-First AI Mental Health Companion that provides accessible, empathetic emotional support through natural spoken conversation. The system successfully integrates automatic speech recognition, natural language understanding, CBT-inspired dialogue management, and a robust crisis detection protocol to deliver a judgment-free support environment available around the clock.

A pilot study involving 43 participants demonstrated an 87% user satisfaction rate, 93% emotion detection accuracy, and 91.2% crisis detection precision, validating the technical and experiential foundations of the proposed architecture. Comparative analysis confirmed that voice-first interaction yields meaningfully higher engagement and session duration relative to text-based alternatives, while persistent memory mechanisms enabled personalization that users found clinically significant.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The Voice-First AI Mental Health Companion is not intended to replace professional mental healthcare but to serve as an accessible first line of support — reducing barriers of cost, stigma, and availability that prevent millions from seeking help. As AI capabilities continue to advance, voice-based companions of this nature hold significant promise for democratizing mental health support at scale, particularly in underserved and resource-constrained populations.

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Computer Engineering at Atharva College of Engineering, Mumbai, for providing the computational resources and institutional support necessary for this research. We are grateful to the 43 participants who volunteered their time and shared their experiences during the pilot evaluation. We also acknowledge the guidance of the clinical psychologist who reviewed crisis escalation transcripts and provided expert validation of the safety protocol outcomes.

REFERENCES

[1] T. Alanzi, A. A. Alsalem, H. Alzahrani, et al., "AI-Powered Mental Health Virtual Assistants' Acceptance: An Empirical Study on Influencing Factors Among Generations X, Y, and Z," Cureus, vol. 15, no. 11, e49486, 2023. doi: 10.7759/cureus.49486

[2] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural Language Processing Applied to Mental Illness Detection: A Narrative Review," npj Digital Medicine, vol. 5, no. 1, article 46, 2022. doi: 10.1038/s41746-022-00589-7

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in Proc. 40th Int. Conf. Machine Learning (ICML), vol. 202, pp. 28492–28518, 2023.

[4] A. S. Miner, A. Haque, J. A. Fries, S. L. Fleming, D. E. Wilfley, G. T. Wilson, A. Milstein, D. Jurafsky, B. A. Arnow, W. S. Agras, L. Fei-Fei, and N. H. Shah, "Assessing the Accuracy of Automatic Speech Recognition for Psychotherapy," npj Digital Medicine, vol. 3, no. 1, article 82, 2020. doi: 10.1038/s41746-020-0285-8

[5] D. B. Olawade, O. J. Wada, A. C. David-Olawade, E. Kunonga, O. Abaire, and J. Ling, "Using Artificial Intelligence to Improve Public Health: A Narrative Review," Frontiers in Public Health, vol. 11, article 1196397, 2023. doi: 10.3389/fpubh.2023.1196397

[6] A. Le Glaz, Y. Haralambous, D. H. Kim-Dufor, et al., "Machine Learning and Natural Language Processing in Mental Health: Systematic Review," J. Medical Internet Research, vol. 23, no. 5, e15708, 2021. doi: 10.2196/15708

[7] World Health Organization, World Mental Health Report: Transforming Mental Health for All. Geneva: WHO Press, 2022.

[8] B. G. Teferra, A. Rueda, H. Pang, R. Valenzano, R. Samavi, S. Krishnan, and V. Bhat, "Screening for Depression Using Natural Language Processing: Literature Review," Interactive Journal of Medical Research, vol. 13, e55067, 2024. doi: 10.2196/55067

[9] B. Inkster, S. Sarda, and V. Subramanian, "An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study," JMIR mHealth and uHealth, vol. 6, no. 11, e12106, 2018. doi: 10.2196/12106

[10] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, "Natural Language Processing in Mental Health Applications Using Non-Clinical Texts," Natural Language Engineering, vol. 23, no. 5, pp. 649–685, 2017. doi: 10.1017/S1351324916000383