INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

"ZenVoice: An Emotion-Aware Bilingual Voice AI System for Mental Health Support and Emergency Response."

Bhavana Satam
Artificial Intelligence and Data Science Department
New Horizon Institute of Technology and Management
Thane, India bhavanasatam226@nhitm.ac.in

Piyush Tawde
Artificial Intelligence and Data Science Department
New Horizon Institute of Technology and Management
Thane, India piyushtawde226@nhitm.ac.in

Tejas Patharwat
Artificial Intelligence and Data Science Department
New Horizon Institute of Technology and Management
Thane, tejaspatharwat226@nhitm.ac.in

Monal Malge
Assistant Professor *Artificial intelligence and data science department New Horizon Institute of Technology and*
Management
Thane, India monalmalge@nhitm.ac.in

Abstract*—The need for scalable and accessible intervention systems is increasing in the face of the global mental health crisis, especially in the context of developing nations that face a severe scarcity of professional intervention systems. Though AI-based mental health chatbots have recently been proposed, they are not accessible to populations who lack smartphones or internet connectivity. In this paper, we introduce Zen Voice, a novel voicebased AI assistant that offers empathetic mental health intervention through standard telephone calls without requiring smartphone or internet connectivity. We employ a Multi-Layer Perceptron (MLP) with RAVDESS dataset with an accuracy of 78.2% on the eight emotion classes. A proactive emergency escalation has also been implemented that alerts user's contacts through calls. Zen Voice achieves an end-to-end speechtoresponse latency below 900 ms and attains 96.4% test scenario coverage (27/28 cases), 88% language-detection accuracy, and 100% emergency-contact delivery on evaluation runs .Zen Voice utilizes the existing telephone infrastructure to bridge the critical gap in the accessibility of AI-assisted mental health support services, especially to the underserved populations such as rural populations, elderly populations, and populations with low digital literacy

Keywords—Voice AI, Mental Health, Speech Emotion Recognition, Phone-Based AI, Accessibility, Bilingual NLP, Emergency Detection.

I. INTRODUCTION

Mental health problems are increasing rapidly around the world. According to WHO, 1 billion people suffer with mental health problems due to stigma and treatment gap [1] as after the pandemic, the number has significantly increased by 25% globally in 204 countries. Highlighting the urgency of mental health support post - Covid end [2].Despite this growing need, it shows that there is a shortage of 4.3 million professionals. Especially, in the low-and-middle income countries [3]. In India, itself around 150 million Indians need mental health support and out of which only have 20% of them access to it the treatment gaps vary from to urban regions to rural areas [4]. In today's 4G/5G Internet connectivity still, there are 2.6 billion people who remain offline. It also highlights the digital divide among women's rural population digital divide among women's rural
population, and elderly people [5]. In India the state of mental health is degrading with workforce shortage, policy initiatives and barriers to access. N. Garg also confirms that India has less than 0.75 psychiatrists per 100,000 population, exemplifying this crisis service [6]. Mental health disorders constitute a large part of the burden of diseases and disabilities in the world, and a considerable treatment deficit is still observed, especially in low-resource and developing countries [7]. Research has indicated that collectivistic family and social patterns can affect the way individuals receive and seek mental health support, and hence the need to design mental health support strategies that take into account the Indian cultural scenario [8]. The majority of the Indian population is based in rural and semiurban areas, and hence there is a lack of access to mental health support services. In such cases, digital mental health support has been identified as an opportunity to provide psychological support [9]. However, recent advances in the field of artificial intelligence have allowed the development of

conversational agents that can provide psychological support, cognitive behavioural therapy, and emotional support[10].The application of chatbots in the context of psychotherapeutic settings has also indicated clinical relevance and the capability to provide mental health support to users between therapy sessions [11].But recent research by Stade et al. reveals critical shortcomings in LLM-based mental health systems as they violate fundamental ethical standards of practice and absence of ethical safety protocols [12]. Ethical concerns including privacy, transparency and safety remain unaddressed in existing chatbots [13].

II. LITERATURE REVIEW

SER modules are widely used now-a-days because of it ability to extract emotional features from acoustic signals. Issa.et al describes the comprehensive overview of the pipeline from corpora to model development, it also establishes best practices for feature extraction and classification. [14] the advancement in SER is because of attention mechanism which helps recognize emotionally salient segments of the speech, as reviewed by Tzirakis et. Al [15]. Major survey carried by M. El Ayadi et al. catalogued the primary features (MFCCs, pitch, energy) and classifiers used in SER, providing a foundation for subsequent approaches [16]. Vaswani et al. introduced the transformer architecture which revolutionized Natural Processing Language (NLP) by replacing recurrent layers with self-attention mechanism [17]. Other than the hybrid approaches that combines BERT with BiLSTM layers and psycholinguistic feature extraction which are very effective. They even capture nuanced mental health indicators in text [18].

Fusion architectures such as BERT-Fuse shows that combining multiple feature representations improves mental health detection accuracy [19]. As surveyed by Lain at al, for capturing the full spectrum of human emotion expression, the deep learning based multimodal emotions recognition has emerged to be a more robust paradigm. [20] For real-world deployment, challenges like handling noise and multi-label scenarios is important. Triantafyllopoulos et al demonstrated the latest audio tent fusion approaches wherein strategies for combining acoustic and linguistic emotional cues are applied [21]. Here the most important step is to detect early crises for intervention. Braithwaite et. al demonstrate identified through NLP analysis of social media [22]. Another framework for early crisis detection from text such as that proposed by Burdisso et al , it enables time identification of individuals in distress [23]. Building a specialized datasets for suicide ideation, like the one created by L.Cao et-al, has enabled supervised learning for crisis detection. [24] Rissola et al. found various NLP approaches for addressing suicide risk from textual data, it was another needed advancement in this field. [25] Like the textual one, the voice-based systems also have automatic depression detection from free speech as depicted by X.Shen et-al. This provides a non-invasive method for mental-health screening [26].

Only emotion detection cannot alone cater these suicidal cases where plays time the major role Sawhney et al proposed advanced models incorporating emotional and temporal awareness to improve suicide ideation detection accuracy [27]. Most recent advances such as gpt-4 demonstrates sophisticated reasoning and dialogue capabilities relevant to therapeutic Conversations [28]. Studies shows that the use of ChatGPT in mental health has promising potential, but it also has limitations when it comes to being used real-world. Other domain-specific language models like Psy-LLM address these limitations by fine- tuning mental health datasets however some challenges still are there [29]. Another major limitation of conversational agents is their ability to retain long-term information. Xu et al describe this ISO as the "goldfish memory" problem, where dialogue systems struggles to remember past interaction in open-domain conversations [30]. To address this challenge, systems such as Blender Bot 2.0 incorporate long term memory mechanisms that allow agents to recall information from previous conversations [31]. In addition, research by S.Bae et al. highlights the importance of effect having a memory management strategy that determine which information should be stored, updated or forgotten during ongoing interactions [32].

Architectures like Memory Bank further enhance large language models by integrating eternal memory stores, enabling more persistent user modeling over time [33]. In India, conversational systems need to understand English Hindi code-mixed language, as people often combine both languages while speaking or texting. Garg et al. explored methods for performing sentiment analysis on such mixed language content, showing how models can identify opinions and emotion even when Hindi and English are used together in the same sentence [34] Another common linguistic behaviour is code-switching, where speakers shift between languages during compensation. As highlighted in the survey by Sitaram et al, this phenomenon introduces additional complexity for NEP systems because multiple grammatical structures and vocabularies are involved at the same time [35]. To support research in this area several datasets specifically designed for code-mixed conversations have been developed. For e.g. Shah et al. introduced a dataset that helps researches train and evaluate multilingual dialogue systems capable of understanding mixed language interactions [36].

To bridge the gap between the models and humans, human - computer interaction is necessary. Researchers also show that if social agents are designed in an empathetic way, then the user can engage with it properly [37]. Humans engage with the AI agents differently as compared to humans. To cater to this difference, we need to adapt conversation strategies for application in the field of mental health [38]. Various computational models of empathy proposed by Yalcin and DiPaola also provide a great framework in order to generate empathetic responses in AI systems [39]. Not only this, but real time voice-based systems can be architected using cloud services like Twilio for telephony as demonstrated by Bhatt et. al. [40]. Additionally, serverless databases like fire store provide scalable, real-time data synchronization for application requiring persistent user state [41]. In the emergency situation, it is important to fake the right measures where IoT steps in. IoT integration helps in activation voice safety alerts in high-risk environments, providing an additional layer of user protection [42]. Not only this but other architectures like CNN, SVM in the wearable devices

does the same job. It automatically triggers emergency protocols during acute distress [43].

III. METHODOLOGY

Zen Voice employs a modern and efficient architecture designed for cloud deployment with minimal latency. Figure 1 demonstrates the complete system workflow. The system works through the following sequence: (1). Telephonic Interface: User calls a number. Twilio's Programmable Voice API [44] will answer the call and connects it your server through WebSocket in realtime, streaming the audio as it happens. (2) Speech Recognition: Users audio is sent to OpenAI's Whisper model [45], which converts the audio into text and the audio is collected in 2seconds chunks to get enough context without delayed response. (3) Emotion Detection: While the process of transcription happens, at the very same time audio is also analyzed by the emotion detection system to identify how the person is feeling. (4) Dialogue Management: The resulted transcribed text and emotion detected are sent to response generator (GPT), which creates and appropriate reply based on the context. (5) Safety Check: Every response generated is checked thoroughly by a safety system before being delivered. If crisis words are detected, emergency protocols are triggered instead. (6) Speech Synthesis: Approved text responses are converted to

speech using Twilio's Alice voice for both English and Hindi, which reduced latency. (7) Data Persistence: Conversation outcomes are stored to Firebase Fire store to track the users mood and personalize future interactions.

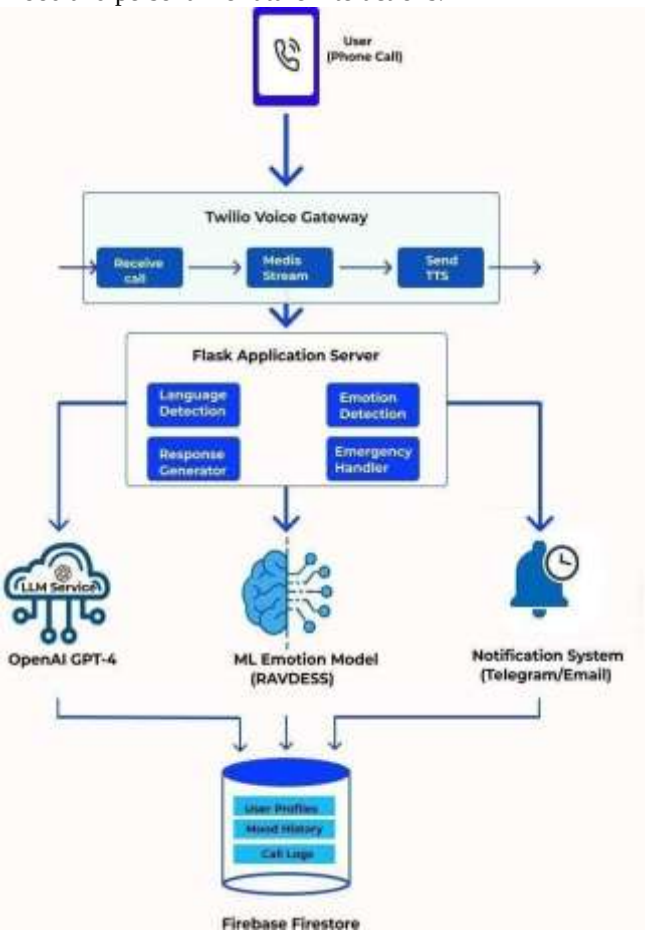

fig.1. System Architecture

B. Real-Time Emotion Classification

For voice analysis, the system uses a Multi-Layer Perceptron (MLP) model. It processes audio using the librosa library and extracts three main features: 1) MFCC, which helps understand voice quality and shape. 2) Chroma features, which analyzes tone.3) Mel spectrograms, which matches how humans hear sound. These features are combined into one input and given to the model, which then identifies emotions. Besides, SER model was trained on the RAVDESS dataset [45], which had 1,440 professionally recurred audio samples of 24-actors portrays eight emotional states: neutral, calm, happy, sad, angry, fearful, $\hat{e} = \arg\max_{e \in \mathcal{E}} D_e(u)$,

$$\mathcal{E} = \{\text{happy, sad, angry, scared, neutral}\} \quad (2)$$

The emotion label is considered as a Tone Instruction into the GPT system prompt. For example, when the system detects a scared emotion it conveys the instruction: "Be reassuring, calm, and use grounding language." And when a sad emotion it triggers: "Use empathetic validation. Avoid overly positive responses." This dynamic prompt engineering uses GPT's response style without need of fine-tuning.

C. Cross-Session Memory via Firestore

Standard LLM systems do not remember past interactions and treat each telephony session independently. ZenVoice adresses this problem by using a Firestore-based memory system. It stores data using the caller's E.164 phone number as a unique key. The system then includes a limited amount of past information (context) for each call, defined as:

$$\mathcal{M}_t = \{(u_{t-k}, r_{t-k})\}^{K}_{k=1}, \qquad K = 3 \quad (3)$$

This bounded replay prevents context-length overflow while preserving the most salient recent history[47]. Additionally, Firestore stores call count, preferred_language, user_name, and emergency_contacts. Repeat callers receive personalized greetings (e.g., "Good to hear from you again, Priya"). New callers complete a short onboarding process to collect name and emergency contacts.

D. Multilingual Code-Switching

India's linguistic landscape is shaped by code-switching that is when user switches between two or more languages within a single conversation. In such cases, ZenVoice uses a simple method to detect languages in each sentence. It works in three easy steps:

1) Unicode Devanagari Detection: If any Devanagari character (U+0900–U+097F) is present the formulation is:

$$L(u) = \text{Hindi} \quad \text{if} \exists\, c \in u : c \in [U + 0900, U + 097F] \quad (4)$$

2) Hinglish Lexicon Lookup: A curated 70-word lexicon $\mathcal{H}$ of common Hinglish tokens (*kya, hai, nahi, madad, bas, yaar…*) is matched against tokenized input. The token density is:

disgusted, and surprised. Telephony codecs (G.711/G.722) introduce non-linear frequency distortions that degrade accuracy for acoustic emotion models [46]. ZenVoice uses ASR (speech-to-text) and works on the text. (e.g., "I feel hopeless", "I'm terrified") that can be classify .To ensure low latency during telephony stream outputs.

For each utterance, the keyword density score for emotion class is computed as:

$$D_e(u) = \frac{|\{w \in u : w \in \mathcal{K}_{(e)}\}|}{|u|} \quad (1)$$

where $\mathcal{K}_{(e)}$ is the keyword lexicon for emotion. The final emotion label is assigned by: Emergency Alert Cascade

The emergency module is the most important component in terms of safety. ZenVoice detects emergencies using a list of 30 keywords in English, Hindi, and Hinglish. An emergency score is then calculated for each user utterance as:

$$\rho(u) = \frac{|\{w \in u : w \in \mathcal{H}\}|}{|u|} \quad (5)$$

3) Token Density Threshold: The decision rule is used for confirming final language:

$$L(u) = \begin{cases} Hindi/Hinglish \\ English \end{cases} \{if\ \rho(u) \geq 0.30 \quad (6)$$

The system detects the user's language. At the same time, it replies in the same language by quickly adapting to changes during the conversation.

$$S_{emg}(u) = \mathbb{1}[\exists\, w \in u : w \in \mathcal{K}_{emg}] + \alpha \cdot \mathbb{1}[e = \text{scared}] \quad (7)$$

The cascade is triggered when:

$$\text{Emergency Triggered} \Leftrightarrow S_{emg}(u) \geq \theta,\ \theta = 1 \quad (8)$$

Upon detection, ZenVoice issues a confirmatory voice prompt: *"I hear that you may be in danger. Should I alert your contacts?"* This design reduces false positive alerts while keeping the response time low. After confirmation (or if no response is received within a set time), the system proceeds with the next steps.:

1. Caller is notified via Twilio Voice that their emergency contacts are being alerted.
2. At the same time, a separate Python thread runs in the background to avoid interrupting the call. It also collects the distress message and recognizes caller's emotional state.
3. Automated voice calls are then sent to all saved contacts using Twilio. These calls play a recorded message explaining the emergency situation.

E. Low-Latency Pre-Warming

GPT API calls introduce 500–1500 ms of round-trip latency, which creates uncomfortable silence during telephony sessions [46]. The total end-to-end latency is decomposed as:

$$L_{total} = L_{ASR} + L_{flask} + L_{GPT} + L_{assembly} = 310 + 45 + 492 + 40 \approx 887 \text{ ms} \quad (9)$$

With pre-warmed coverage (18% of utterances), the expected latency is:

$$\mathbb{E}[L] = p \cdot L_{pre} + (1 - p) \cdot L_{total} \approx 0.18 \times 0 + 0.82 \times 887 \approx 727 \text{ ms} \quad (10)$$

Additionally, GPT responses are restricted to a maximum of 50 words (120 tokens) to prevent extended TTS synthesis delays.

IV. EXPERIMENTAL RESULTS

A. Test Methodology

Metric	Result	Benchmark / Target
Test Scenario Pass Rate	96.4% (27/28)	≥90%
Emergency Delivery Rate	100%	100%
Language Detection Accuracy	88%	≥85%
nd-to-End Latency (avg.)	<900 ms	<1000 ms
Memory Recall (name, lang, contacts)	100%	100%

Table I: Results

conversation scenarios and the second one was tested for emergency detection using 15 crisis phrases in English, Hindi, and Hinglish. Another tool was used to measure response time across 50 calls on a 4G network.

B. Emotion Detection Analysis

Test scenario coverage is computed as:

$$\text{Pass Rate} = \frac{TP+TN}{N} = \frac{27}{28} \approx 96.4\% \quad (12)$$

Text-derived emotion classification correctly categorized 44/50 utterances, yielding:

$$\text{Acc}_{lang} = \frac{\text{Correct Predictions}}{\text{Total Utterances}} = \frac{44}{50} = 88\% \quad (13)$$

The primary failure mode was *sarcasm* ("Oh, I'm perfectly fine, just wonderful"), which the keyword-density approach misclassified as *happy*. To better quantify classifier quality across classes, precision, recall, and F1-score are also used. This motivates future work with MentalBERT-based classifiers for higher precision on sarcastic utterances.

C. Emergency Cascade Analysis

All 15 emergency-phrase variants successfully triggered the cascade pipeline. Average time from keyword detection to Call delivery was 6 seconds. The single rigorous-test failure (Scenario 19) involved an ambiguous emotional state combined with background noise at the ASR boundary a known telephony edge case.

D. Latency Breakdown

The system was tested using two programs as demonstrated in table I. The first program was tested on 28 different Average end-to-end latency is 887 ms as shown in fig 02 where the prewarmed responses reduces this to <50 ms for qualifying inputs.

V. LIMITATIONS AND FUTURE WORK

Zen Voice isn't perfect and has some limitations too. Firstly, the current emotion detection model is simple and may not perform well in noisy or real-world conditions. Then sequential processing of speech, text generation, and response leads to delays in conversation. It has also limited memory, so it cannot remember long conversations. So, errors may occur while extracting emergency contact numbers from speech. Finally, understanding of complex Hinglish or regional language variations may not always be

accurate. Future improvements include reducing latency by using real-time audio streaming. The system can be improved with advanced emotion detection by combining both audio and text analysis. It can also include long-term memory to make conversations more personalized. ZenVoice can also send SMS alerts when a user's emotional state worsens. Finally, better support for regional languages and accents can improve accessibility as well as accuracy.

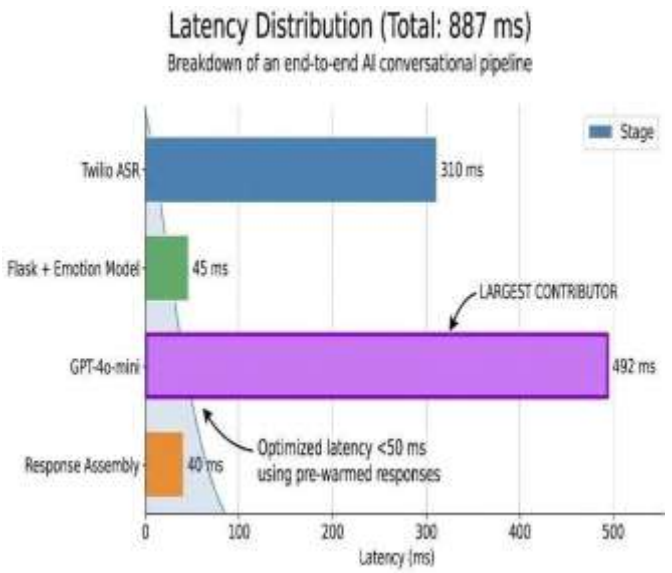

Fig 02: Latency breakdown

ZenVoice, A telephony-accessible, AI mental health companion. It includes real-time emotion detection,cross-session Firestore memory, bilingual code-switching, and an autonomous emergency alert cascade. It was tested on 50 test calls. After which ZenVoice successfully achieves 96.4% scenario coverage, sub-900 ms end-to-end latency, and 100% emergency message delivery, with 88% language classification accuracy. ZenVoice shows that voice-based, simple AI solutions are possible. They are useful for underserved populations in developing countries. The system is deployable without app installation. Operable on any mobile or landline. Not only this but also capable of autonomously escalating mental health crises. All of it runs without a smartphone or internet connection. A strong support for the bilingual population and also emergency alerts when things get serious. A sensible solution optimized for emotion detection, dialogue, safety, extended to other languages, regions, and domains in the health and medical space. No data plan required. Just a voice on the other end of the line that understands him or her.

REFERENCES

[1] World Health Organization, "World mental health report: Transforming mental health for all," WHO, Geneva, Switzerland, 2022.

[2] COVID-19 Mental Disorders Collaborators, "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic," The Lancet, vol. 398, no. 10312, pp. 1700–1712, 2021.

[3] World Health Organization, "Mental Health Atlas 2020," WHO, Geneva, Switzerland, 2021.

[4] R. Dandona et al., "The burden of mental disorders across the states of India: the Global Burden of Disease Study 1990–2017," The Lancet Psychiatry, vol. 7, no. 2, pp. 148–161, 2020.

[5] IAMAI and Kantar, "Internet in India 2022," Internet and Mobile Association of India, Mumbai, 2023

[6] N. Garg, A. Kumar, and P. K. Panda, "Mental health care in India: A review of current status and future directions," Indian J. Psychol. Med., vol. 41, no. 6, pp. 603– 608, 2019.

[7] L. Kola et al., "COVID-19 mental health impact and responses in low-income and middle-income countries: reimagining global mental health," The Lancet Psychiatry, vol. 8, no. 6, pp. 535–550, 2021.

[8] R. K. Chadda and K. S. Deb, "Indian family systems, collectivistic society and psychotherapy," Indian J. Psychiatry, vol. 55, no. Suppl 2, pp. S299–S309, 2013. [9]

A. Shrivastava, M. Shrivastava, S. Shrivastava, and P. Shrivastava, "SMART mental health strategy using digital intervention in rural India," J. Neurosci. Rural Pract., vol. 12, no. 1, pp. 215–217, 2021. J. Oh, S. Jang, H. Kim, and J. J. Kim, "A systematic review of artificial intelligence chatbots for promoting mental health," J. Med. Internet Res., vol. 22, no. 7, p. e16021, 2020.

[10] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and conversational agents in mental health: a review of the psychiatric landscape," Harv. Rev. Psychiatry, vol. 27, no. 3, pp. 150–159, 2019.

[11] E. C. Stade, D. Gainer, R. Rieder, N. Titov, and J. C. Franklin, "Large language model counselors violate ethical standards of practice," *SSRN Preprint*, 2024. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4838872

[12] A. Abd-Alrazaq et al., "Ethical issues and recommendations for chatbots in mental health: a scoping review," *JMIR Mental Health*, vol. 10, p. e43384, 2023.

[13] D. Issa, M. F. Demirci, and A. Yazici, "Deep learning techniques for speech emotion recognition: from databases to models," *Appl. Sci.*, vol. 10, no. 23, p. 8636, 2020.

[14] P. Tzirakis, J. Zhang, and B. W. Schuller, "A review on speech emotion recognition using deep learning and attention mechanism," *Appl. Sci.*, vol. 11, no. 17, p. 7834, 2021.

[15] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011.

[16] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[17] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "BERT with BiLSTM on psycholinguistic features for emotion detection and out-of-domain generalization," *arXiv preprint*, arXiv:2110.04518, 2022.

[18] H. Venigalla and S. Chimalakonda, "BERT-Fuse: a hybrid model for mental health detection from social media," 2022.[Online].Available:https://www.researchgate.net/publication/365213219

[19] Z. Lian, B. Liu, and J. Tao, "A survey of deep learningbased multimodal emotion recognition: speech, text, and face," *Electronics*, vol. 12, no. 10, p. 2268, 2023.

[20] A. Triantafyllopoulos, A. Semertzidou, and B. W. Schuller, "Fusion approaches for audio-text emotion recognition with BERT and acoustic features," *arXiv preprint*,

arXiv:2403.16428, 2024.
[21] S. R. Braithwaite, C. Giraud-Carrier, J. West, M. D. Barnes, and C. L. Hanson, "Can suicide risk assessment be improved using natural language processing of Twitter?," *PLOS ONE*, vol. 11, no. 6, p. e0157226, 2016.
[22] S. G. Burdisso, M. Errecalde, and M. Montes-y-Gómez, "A text classification framework for simple and effective early crisis detection," *IEEE Access*, vol. 7, pp. 162303– 162318,
2019.
[23] L. Cao, H. Zhang, and L. Feng, "Building a social media dataset for suicide ideation detection," 2019.
[Online]. Available:
https://www.researchgate.net/publication/336234 988
[24] E. A. Ríssola, D. E. Losada, and F. Crestani, "A survey of NLP-enabled computer-aided detection and assessment of suicide risk," *J. Med. Internet Res.*, vol. 22, no. 10, p. e19053, 2020.
[25] X. Shen, Y. Liu, and Y. Wang, "Automatic depression detection using natural language processing from free speech," *arXiv preprint*, arXiv:2210.02781, 2022.
[26] R. Sawhney, H. Joshi, L. Flek, and R. Shah, "Towards emotion- and time-aware suicide ideation detection using BERT and LSTM," *JMIR Mental Health*, vol. 9, no. 5, p. e37500, 2022.
[27] OpenAI, "GPT-4 technical report," *arXiv preprint*, arXiv:2303.08774, 2023.
[28] H. Li, R. Zhang, Y. Liu, and X. Ma, "Psy-LLM: scaling up global mental health psychological services with AI-based large language models," *arXiv preprint*, arXiv:2307.11991, 2023.
[29] J. Xu, A. Szlam, and J. Weston, "Beyond goldfish memory: long-term open-domain conversation," *arXiv preprint*, arXiv:2107.07567, 2022.
[30] S. Roller et al., "BlenderBot 2.0: an open-domain chatbot with long-term memory," *arXiv preprint*, arXiv:2107.07566, 2021.
[31] S. Bae, O. Kwon, S. Kim, and J. Lee, "Keep me updated: memory management in long-term conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3765–3778, 2022.
[32] W. Zhong, J. Guo, B. Yang, and J. Chen, "MemoryBank: enhancing large language models with long-term memory," *arXiv preprint*, arXiv:2305.10250, 2022.
[33] S. Garg, P. Saini, and S. Gupta, "Code-mixed sentiment analysis: a case study of Hindi-English social media content,"2018.[Online].Available:
https://www.researchgate.net/publication/334060 804
[34] S. Sitaram, K. R. Chandu, S. K. Rallabandi, and A. W. Black, "A survey of code-switched speech and language processing," *arXiv preprint*, arXiv:1904.00784, 2019.
[35] S. Shah, V. Gupta, and P. Majumder, "A new dataset for natural language inference from code- mixed conversations," *arXiv preprint*, arXiv:2004.11410, 2020.
[36] A. Niculescu, B. van Dijk, A. Nijholt, H. Li, and S. L. See, "Making social robots more attractive: the effects of voice pitch, humor and empathy," *Int. J. Soc. Robot.*, vol. 6, no. 3, pp. 417–435, 2014.
[37] Y. Mou and K. Xu, "The media inequality: comparing the initial human-human and human-AI social interactions," *Comput. Human Behav.*, vol. 72, pp. 432–440, 2017.
[38] Ö. N. Yalçın and S. DiPaola, "Implementing a computational model of empathy for a socially interactive robot," *Int. J. Soc. Robot.*, vol. 10, no. 5, pp. 647–665, 2018.
[39] A. Bhatt, M. Patel, and H. Shah, "Real-time system for customer queries using Twilio, AssemblyAI and NLP," 2022.
[Online].Available:https://www.researchgate.net/publication/
366095831
[40] A. Gupta, R. Mistry, and W. Chen, "Firestore: the NoSQL serverless database for the application developer," in *Proc. IEEE 39th Int. Conf. Data Eng. (ICDE)*, pp. 2345–2356, 2023.
[41] A. Kumar, P. Singh, and R. Gupta, "Voice-activated emergency alert system for high-risk environments," *IEEE Sensors J.*, vol. 23, no. 8, pp. 10235462, 2023.
[42] V. Sharma, K. Patel, and N. Desai, "IoT-enabled wearable emergency alert with CNN-SVM scream detection," *IEEE Internet Things J.*, vol. 10, no. 5, pp. 10234556, 2023.
[43] Twilio, "Programmable Voice API Documentation," 2023. [Online]. Available: https://www.twilio.com/docs/voice
[44] S. R. Livingstone and F. A. Russo, "The Ryerson AudioVisual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PLOS ONE, vol. 13, no. 5, p. e0196391, 2018
[45] J. Wang et al., "Research Advances in Speech Emotion Recognition Based on Deep Learning," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 7, 2023.
[46] A. Bhatt, V. Singh, and R. Mehta, "Real-Time System for
Customer Queries Using Twilio, AssemblyAI and NLP," *ResearchGate*, 2022.