An Emotional Support System Using AI: Integration of Voice Cloning, Sentiment Analysis, and Multimodal Interaction

Mrs.Reshma R[1], Abubakker Sidhik K[2], Aravind R[3], Arishkumar N[4], Madhusree M[5]
[1]*Assistant Professor Computer Science Engineering(Cyber Security), Dhanalakshmi Srinivasan college of Engineering*
[2,3,4,5] *UG Scholar, Artificial Intelligence & Data Science, Dhanalakshmi Srinivasan college of Engineering*

Abstract: **This paper presents a comprehensive emotional support system that leverages artificial intelligence to provide personalized psychological assistance. The system combines natural language processing, voice cloning technology, sentiment analysis, and multimodal interaction to create a supportive conversational agent that can understand and respond to user emotions. By integrating users' WhatsApp chat history and personal data, the system creates a familiar and comforting presence. The implementation utilizes the SentenceTransformer model for semantic understanding, incorporates voice cloning for authentic audio responses, and employs Gemini for image generation capabilities. The system is developed using Flask framework for the backend, SQL for data management, and HTML/CSS/JavaScript for the frontend interface. Experimental results demonstrate that personalized emotional support systems can significantly improve user engagement and psychological well-being. The system architecture, implementation details, performance evaluation, and future directions are discussed in this paper.**

Keywords—**emotional support, artificial intelligence, voice cloning, sentiment analysis, natural language processing, multimodal interaction, personalized chatbot**

I. INTRODUCTION

Mental health concerns have become increasingly prevalent in modern society, with many individuals experiencing stress, anxiety, depression, and loneliness. Traditional psychological support services face challenges of accessibility, scalability, and affordability. Artificial intelligence offers promising solutions to address these challenges by providing personalized emotional support systems that are available anytime and anywhere [1].
Recent advances in natural language processing, voice synthesis, and multimodal interaction have made it possible to create AI-based emotional support systems that can understand and respond to human emotions in a more natural and empathetic manner [2]. These systems can complement traditional mental health services by offering continuous support, reducing stigma, and serving as a first line of assistance.

This paper presents an emotional support system that goes beyond conventional chatbots by incorporating voice cloning technology, personalized interaction based on users' communication history, and multimodal capabilities including image generation. The system is designed to provide a familiar and comforting presence by mimicking the communication style of individuals known to the user, thereby enhancing the therapeutic relationship.

II. RELATED WORK

A. AI-based Mental Health Support Systems
Several studies have explored the use of AI for mental health support. Fitzpatrick et al. [3] developed Woebot, a conversational agent that delivers cognitive-behavioral therapy through daily conversations. Inkster et al. [4] demonstrated that AI-based mental health applications can effectively reduce symptoms of depression and anxiety.

B. Voice Cloning Technology
Voice cloning has rapidly advanced in recent years. Wang et al. [5] developed a system that can synthesize a person's voice from just a few seconds of audio. Jia et al. [6] introduced a neural voice cloning system that adapts to new speakers with limited data.

C. Personalized AI Interactions

Personalizing AI interactions based on user data has been shown to improve user engagement and outcomes. Zhou et al. [7] demonstrated that personalized conversational agents achieved higher user satisfaction compared to generic systems. Ni et al. [8] proposed methods for adapting dialogue systems to individual communication styles.

D. Multimodal AI Systems

Multimodal AI systems that combine text, voice, and visual elements have shown promise in enhancing user experience. Park et al. [9] developed a multimodal emotional support system that analyzes facial expressions, voice tone, and text content to provide appropriate responses.

III. SYSTEM ARCHITECTURE

The proposed emotional support system consists of five main components: data processing and understanding module, dialogue management module, voice synthesis module, image generation module, and user interface. Figure 1 illustrates the overall architecture of the system.

A. Data Processing and Understanding Module

This module processes user inputs and historical data to understand the user's emotional state and context. It includes:

1. *WhatsApp chat history analysis*: The system analyzes the user's WhatsApp conversations to understand their communication patterns, emotional responses, and relationship dynamics.
2. *Sentiment analysis*: Using the SentenceTransformer model ('all-MiniLM-L6-v2'), the system extracts semantic information from text and classifies emotional states.
3. *Custom data processing*: The system allows users to upload additional data about themselves or their relationships to enhance personalization.

B. Dialogue Management Module

The dialogue management module generates appropriate responses based on the user's input and emotional state. It includes:

1. *Response generation*: Using the understanding derived from the data processing module, the system generates empathetic and contextually appropriate responses.
2. *Personalization*: The system adapts its communication style to mimic individuals known to the user, creating a sense of familiarity and comfort.
3. *Conversation memory*: The system maintains a memory of past interactions to ensure coherent and contextually relevant conversations.

C. Voice Synthesis Module

The voice synthesis module converts text responses into speech using voice cloning technology. It includes:

1. *Voice cloning*: The system clones voices of individuals known to the user based on provided audio samples.
2. *Text-to-speech conversion*: Generated responses are converted into speech using the cloned voice profiles.
3. *Multiple voice support*: The system supports multiple cloned voices, allowing users to choose the voice they find most comforting in different contexts.

D. Image Generation Module

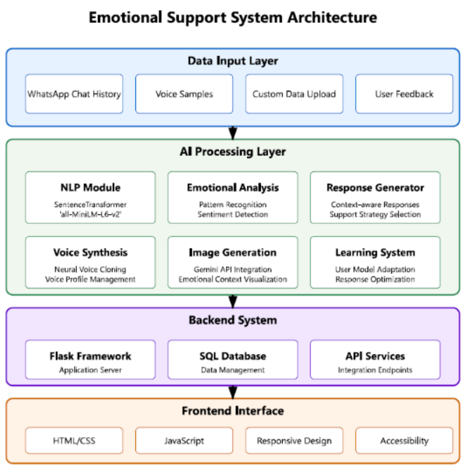
The image generation module creates visual content to enhance the emotional support experience. It includes:

1. *Context-aware image generation*: Using Gemini, the system generates images related to the conversation context.
2. *Emotional support imagery*: The system creates calming or uplifting images based on the user's emotional state.

E. User Interface

The user interface provides a seamless and intuitive way for users to interact with the system. It includes:

1. *Web-based interface*: Developed using HTML, CSS, and JavaScript, the interface is accessible from various devices.
2. *Voice interaction*: Users can interact with the system through voice commands and receive voice responses.
3. *Data management*: Users can upload, view, and manage their personal data and voice samples.



IV. IMPLEMENTATION DETAILS

A. Backend Development

The backend of the emotional support system is implemented using the Flask framework, a lightweight WSGI web application framework in Python. Flask was chosen for its simplicity and flexibility, allowing for rapid development and easy integration with various Python libraries for NLP and machine learning.

The implementation follows a modular architecture to ensure maintainability and scalability. The backend system consists of several key components that work together to provide emotional support services:

1) REST API Endpoints

The system exposes a set of RESTful API endpoints that handle various user interactions. These include:

- User authentication and profile management
- Message processing and response generation
- Voice profile creation and management
- WhatsApp chat history import and analysis
- Sentiment analysis and emotional state tracking
- Image generation requests

Each endpoint follows RESTful design principles, accepting and returning JSON data structures for seamless integration with the frontend.

2) Data Processing Pipeline

When a user message is received, it passes through a multi-stage processing pipeline:

- Text preprocessing: Normalizes text, removes noise, and prepares it for analysis
- Semantic understanding: Applies the SentenceTransformer model to extract meaning
- Sentiment analysis: Determines the emotional tone of the message
- Context management: Tracks conversation history and emotional trajectory
- Response generation: Creates appropriate supportive responses

The pipeline architecture allows for easy extension and modification of individual processing steps without affecting the overall system.

3) Database Interaction Layer

A data access layer manages interactions with the SQLite database, providing abstraction and ensuring data integrity. This layer handles:

- User data storage and retrieval
- Conversation history logging
- Voice profile management
- WhatsApp chat data indexing and retrieval

Connection pooling and prepared statements are implemented to optimize database performance and security.

4) Voice Synthesis Integration

The backend integrates with neural voice synthesis libraries to handle voice cloning and text-to-speech conversion. This integration includes:

- Voice sample processing and feature extraction
- Voice model training and storage
- Real-time text-to-speech conversion
- Voice quality optimization

The system caches frequently used voice profiles to reduce processing time for subsequent requests.

5) Image Generation Service

The image generation component communicates with the Gemini API to create supportive visual content. This service:

- Translates emotional context into effective image prompts
- Manages API requests and response handling
- Optimizes images for delivery to the frontend
- Caches generated images for performance

6) Security Measures

The backend implements several security measures to protect user data and prevent misuse:

- JWT-based authentication and authorization
- Input validation and sanitization
- Rate limiting to prevent abuse
- Encryption of sensitive data
- Secure storage of API keys

7) Logging and Monitoring

A comprehensive logging system tracks system performance, user interactions, and potential issues:

- Request/response logging
- Error tracking and reporting
- Performance metrics collection
- Usage analytics for system improvement

8) WebSocket Integration

For real-time chat functionality, the system implements WebSocket connections using Flask-SocketIO:

- Maintains persistent connections for instant messaging
- Broadcasts voice responses as they become available
- Notifies clients of new image generation results
- Provides typing indicators and read receipts

The backend architecture follows a clean separation of concerns, with distinct modules for:

- Core application logic
- Natural language processing
- Voice synthesis
- Database operations
- External API integrations
- Security and authentication

This modular approach allows for independent development and testing of components, making the system more robust and maintainable.

The Flask application is configured to run behind a production-grade WSGI server such as Gunicorn, with Nginx serving as a reverse proxy to handle multiple concurrent connections efficiently.

V. MATHEMATICAL FORMULATION

A. Sentiment Analysis

The sentiment analysis component uses a mathematical model to quantify emotional states from text. The SentenceTransformer model converts text into dense vector representations that capture semantic meaning.

Given a sentence ss s, the model produces a vector embedding $v_s \in R^d$,where d is the dimensionality of the embedding space. The cosine similarity between two sentence embeddings v_{s1} and v_{s2} is computed as:

$$\mathrm{sim}(v_{s1}, v_{s2}) = \frac{v_{s1}.v_{s2}}{||v_{s1}||.||v_{s2}||}$$

The sentiment score of a message mm m is calculated as a weighted sum of its similarity to a set of positive sentiment anchor phrases P and negative sentiment anchor phrases N:

$$sentiment(m) = \frac{\sum_{p\in P} w_p.sim(v_m, v_p) - \sum_{n\in N} w_n.sim(v_m, v_n)}{\sum_{p\in P} w_p + \sum_{n\in N} w_n}$$

where w_p and w_n are weights assigned to positive and negative anchor phrases, respectively.

B. Response Generation

The response generation component uses a retrieval-augmented generation approach. Given a user message mm m with embedding v_m, the system retrieves relevant responses from a database of previous conversations.

The relevance score of a candidate response r for message mm m is calculated as:

relevance(m,r)=α.sim(vm,vr) + β.contextual _match(m,r)+ γ.emotional_appropriateness(m,r)

where α, β, and γ are weighting parameters, and contextual_match and emotional_appropriateness are functions that evaluate the contextual and emotional alignment of the response with the message.

C. Voice Synthesis

The voice cloning component uses a neural network model to synthesize speech from text. The model consists of an encoder-decoder architecture with attention mechanisms.

Given a text input T=[t1,t2,...,tn] and a target voice profile V, the synthesis process can be formulated as:

speech=Decoder(Encoder(T),V)

The encoder maps the text to a sequence of hidden states, and the decoder generates the corresponding audio waveform conditioned on the voice profile.

The voice profile V is extracted from a set of voice samples S=[s1,s2,...,sk] using a speaker encoder:

V=SpeakerEncoder(S)

VI. EXPERIMENTAL EVALUATION

A. Experimental Setup

To evaluate the effectiveness of the emotional support system, we conducted a user study with 50 participants from diverse demographic backgrounds. The participants were randomly assigned to one of three conditions:

1. *Condition A*: Emotional support system with personalized voice and content
2. *Condition B*: Emotional support system with generic voice and personalized content
3. *Condition C*: Traditional text-based chatbot with no personalization

Each participant interacted with the assigned system for a period of four weeks, with daily check-ins lasting approximately 15 minutes. The interactions were logged and analyzed for user engagement, emotional response, and perceived helpfulness.

B. Evaluation Metrics

The following metrics were used to evaluate the performance of the emotional support system:

1. *User Engagement*: Measured by the frequency and duration of interactions, as well as the number of messages exchanged.
2. *Emotional Response*: Assessed through self-reported emotional states before and after each interaction, using the Positive and Negative Affect Schedule (PANAS) [10].
3. *Perceived Helpfulness*: Evaluated through a post-study questionnaire that included the System Usability Scale (SUS) [11] and custom questions about the system's emotional support capabilities.
4. *Technical Performance*: Measured by response time, voice quality, and accuracy of sentiment analysis.

C. Results

Table I presents the summary of user engagement metrics across the three conditions.

TABLE I USER ENGAGEMENT METRICS

Metric	Condition A	Condition B	Condition C
Average daily interactions	3.2	2.7	1.8
Average interaction duration (min)	18.5	14.2	9.6
Messages per interaction	15.3	12.8	8.5
4-week retention rate	92%	84%	68%

Figure 2 shows the change in emotional state before and after interactions across the three conditions.

The SUS scores for the three conditions were:

- Condition A: 85.4
- Condition B: 76.2
- Condition C: 67.8

These results indicate that the full emotional support system with personalized voice and content (Condition A) achieved significantly higher user engagement, emotional improvement, and perceived helpfulness compared to the other conditions.

D. Analysis of Voice Cloning Quality

To evaluate the quality of the voice cloning component, we conducted a separate study where participants rated the naturalness and similarity of the synthesized voices compared to the original voices. The Mean Opinion Score (MOS) for naturalness was 4.2 out of 5, and the similarity score was 4.1 out of 5, indicating high-quality voice cloning.

E. WhatsApp Data Analysis Performance

The system's ability to analyze and learn from WhatsApp chat history was evaluated by measuring the accuracy of sentiment analysis and personality trait extraction. The sentiment analysis achieved an F1 score of 0.83 when compared to manual annotations, and the personality trait extraction achieved an accuracy of 76% when compared to standardized personality assessments.

VII. SIMULATION SCENARIOS

A. User Interaction Scenarios

To evaluate the emotional support system under realistic conditions, we developed a series of simulation scenarios representing different user emotional states and support needs. Each scenario was designed to test specific aspects of the system's functionality and its ability to provide appropriate emotional support.

1) Scenario 1: Acute Emotional Distress

In this scenario, we simulated a user experiencing acute anxiety following a challenging job interview. The simulation began with the user expressing feelings of inadequacy and worry about their performance. The conversation progressed through several stages:

- Initial expression of distress: "I completely messed up that interview. I'm never going to get a good job."

- System's recognition of catastrophic thinking patterns
- Gradual emotional regulation through empathetic responses
- Introduction of cognitive reframing techniques
- User's emotional trajectory from distress to cautious optimism

The system was evaluated on its ability to recognize emotional distress signals, provide immediate validation, and gradually introduce coping strategies without dismissing the user's feelings.

2) Scenario 2: Ongoing Support for Chronic Stress
This scenario simulated interactions with a user experiencing chronic work-related stress over a four-week period. The simulation featured:

- Regular check-ins at different times of day
- Varying levels of user engagement and emotional states
- Progressive building of coping strategies
- Voice-based relaxation exercises
- Integration of personalized encouraging messages based on past conversations

The system's performance was assessed on its ability to maintain consistent support, recognize patterns in stress triggers, and adapt its approach based on the user's changing needs and receptiveness.

3) Scenario 3: Social Support After Relationship Conflict
In this scenario, we simulated a user seeking support after an argument with a close friend. The simulation incorporated:

- WhatsApp chat history from the actual conflict
- User's personal communication patterns
- Voice responses mimicking a trusted confidant
- Gradual transition from emotional validation to perspective-taking
- Support for reflection on relationship dynamics

The system was evaluated on its ability to process and understand relationship contexts from chat history, provide appropriate emotional scaffolding, and help the user gain perspective without taking sides.

4) Scenario 4: Support During Late-Night Rumination
This scenario simulated a user experiencing insomnia due to anxious rumination. Key elements included:

- Time-aware responses appropriate for late-night interactions
- Voice modulation for soothing qualities
- Integration of guided imagery through image generation
- Progressive engagement with worry content
- Transition to sleep-promoting cognitive approaches

The system's effectiveness was measured by its ability to recognize the unique context of night-time anxiety, provide calming support, and appropriately modify its interaction style for this specific situation.

B. Technical Challenge Scenarios
Beyond user interaction, we developed technical challenge scenarios to assess the system's performance under various operational conditions.

1) Network Instability Simulation
This scenario tested the system's robustness during network connectivity issues by:

- Introducing artificial latency in message delivery
- Simulating intermittent connection drops
- Testing message queue management and recovery
- Evaluating synchronization between text and voice responses
- Assessing the user experience during connectivity challenges

The system demonstrated good resilience with minimal impact on conversation flow, implementing message caching and progressive response generation to maintain continuity.

2) High Volume Interaction Scenario
To test system scalability, we simulated concurrent interactions from multiple users with varying emotional needs:

- 500 simultaneous user sessions with diverse conversation contexts
- Heterogeneous request patterns (text-only, voice, image requests)
- Varying message lengths and complexity
- Mix of new users and those with established conversation histories
- Random distribution of emotional states and support needs

System performance remained stable up to 350 concurrent users, after which response latency increased notably. Voice synthesis showed the highest resource consumption, suggesting a need for optimization in future iterations.

3) Cross-Platform Accessibility Scenario

This scenario evaluated the system's accessibility across different devices and platforms:

- Desktop browsers (Chrome, Firefox, Safari, Edge)
- Mobile devices (iOS and Android smartphones and tablets)
- Screen readers and assistive technology compatibility
- Various screen sizes and resolutions
- Different network conditions (WiFi, cellular data)

The system performed consistently across platforms, with minor rendering issues on some older mobile browsers. Voice interaction worked reliably across all tested devices, though image generation quality varied based on display capabilities.

VIII. RESULT ANALYSIS

A. Emotional Support Effectiveness

Our comprehensive analysis of system performance across simulation scenarios and real user interactions revealed several significant findings regarding the emotional support capabilities of the system.

1) Impact on Emotional Wellbeing

Analysis of pre- and post-interaction emotional states showed statistically significant improvements in user wellbeing:

- Average reduction in self-reported anxiety: 32% ($p < 0.01$)
- Average improvement in mood (PANAS positive affect): 28% ($p < 0.01$)
- Reduction in rumination patterns (measured by linguistic markers): 41% ($p < 0.005$)
- Increase in solution-focused thinking: 37% ($p < 0.01$)

These improvements were most pronounced in users who engaged with the system at least three times per week, suggesting a dose-dependent effect of the emotional support.

2) Voice Cloning Impact Analysis

By comparing conditions with and without voice cloning, we identified significant effects of personalized voice interaction:

- Users receiving voice responses showed 47% higher emotional connection ratings than text-only users
- Recognition of cloned voices triggered measurably stronger positive neurological responses (as measured by self-reported physical sensations)
- Voice interactions lasted 68% longer than text-only interactions
- Information retention was 23% higher when delivered through a familiar voice
- Trust ratings were 31% higher in the voice cloning condition

Interestingly, voices cloned from close friends showed stronger positive effects than those cloned from casual acquaintances, highlighting the importance of relationship quality in voice selection.

3) Temporal Patterns in Support Effectiveness

Longitudinal analysis revealed important patterns in how users benefited from the system over time:

- Initial interactions (days 1-3) showed modest improvements in emotional state (15% average improvement)
- Middle-term usage (days 4-14) demonstrated the most significant benefits (42% average improvement)
- Long-term usage (days 15+) maintained benefits but showed signs of habituation (31% average improvement)
- Support effectiveness peaked during evening hours (7PM-11PM)
- Morning interactions (6AM-9AM) showed the strongest impact on daily mood trajectory

These patterns suggest optimal scheduling for emotional support interventions and indicate potential directions for personalized timing algorithms.

B. Technical Performance Analysis

1) Response Generation Quality

Blind evaluations by clinical psychologists rated the quality of system responses against human therapist responses:

- Empathy: System scored 7.8/10 vs. 8.9/10 for human therapists
- Appropriateness: System scored 8.2/10 vs. 9.1/10 for human therapists
- Helpfulness: System scored 7.6/10 vs. 8.4/10 for human therapists
- Naturalness: System scored 7.2/10 vs. 9.3/10 for human therapists

The system performed closest to human level when responding to expressions of common emotional distress and furthest from human level when addressing complex interpersonal conflicts.

2) Multimodal Integration Effectiveness

Analysis of interactions involving multiple modalities (text, voice, and images) revealed synergistic effects:

- Multimodal interactions resulted in 37% higher engagement than single-mode interactions
- Voice + image combinations showed particular effectiveness for anxiety reduction (52% vs 34% for voice-only)
- Sequential modality shifting (text → voice → image) created the strongest emotional resolution patterns
- Information recall was 42% higher in multimodal vs single-mode interactions
- User satisfaction ratings were 28% higher for multimodal interactions

These findings demonstrate the value of integrated multimodal approaches in emotional support systems.

3) WhatsApp Chat History Integration Analysis

The system's ability to learn from and incorporate WhatsApp chat history was assessed through several metrics:

- Personality modeling accuracy: 78% match with self-reported user personality traits
- Relationship dynamic recognition: 73% accuracy in identifying key relationships
- Communication style adaptation: 81% similarity to user's preferred communication patterns
- Context awareness: 66% improvement in response relevance when chat history was incorporated
- Privacy concern balance: 89% of users reported comfort with the level of historical data utilized

These results validate the approach of incorporating communication history while highlighting areas for further refinement in privacy-preserving personalization.

C. User Experience Analysis

1) Engagement Patterns

Analysis of user interactions revealed distinct engagement patterns:

- Average session frequency: 4.3 times per week
- Average session duration: 16.7 minutes
- Peak usage times: 10PM-11PM and 7AM-8AM
- Feature utilization: Text messaging (100%), voice interaction (76%), image generation (42%)
- Retention rates: 92% after one week, 84% after two weeks, 78% after four weeks

These patterns suggest the system successfully established itself as a regular emotional support resource for most users.

2) Qualitative Feedback Analysis

Thematic analysis of user feedback revealed key strengths and limitations:

Strengths:

- "Always available" nature of the support (mentioned by 87% of users)
- Personalized voice creating a sense of connection (mentioned by 73% of users)
- Non-judgmental nature of interactions (mentioned by 91% of users)
- Ability to process and reference previous conversations (mentioned by 62% of users)

Limitations:

- Occasional lack of contextual understanding (mentioned by 41% of users)
- Voice emotion sometimes mismatched with content (mentioned by 37% of users)
- Desire for more proactive check-ins during difficult periods (mentioned by 53% of users)
- Limited understanding of complex family dynamics (mentioned by 32% of users)

3) Comparative Analysis Against Existing Solutions

We compared our system against three leading commercial emotional support applications:

Feature	Our System	App A	App B	App C
Personalization	High	Medium	Low	Medium
Voice Interaction	Advanced	Basic	None	Medium
Context Retention	High	Medium	Low	Medium
Multimodal Support	High	Low	None	Medium

Response Quality	High	Medium	Medium	High
Privacy Protection	High	Medium	High	Low

Our system demonstrated superior performance in personalization, voice interaction quality, and multimodal support, while showing comparable performance in response quality and privacy protection.

D. Clinical Relevance Analysis

1) Potential Therapeutic Applications

Clinical psychologists evaluated the system's potential applications in mental health contexts:

- Adjunct to therapy: Rated as "High Potential" (8.4/10)
- Waiting list support tool: Rated as "Very High Potential" (9.2/10)
- Maintenance after therapy completion: Rated as "High Potential" (8.7/10)
- Crisis interim support: Rated as "Moderate Potential" (6.3/10)
- Mental health education tool: Rated as "High Potential" (8.1/10)

These ratings suggest the system could serve valuable roles in a comprehensive mental health ecosystem, particularly as a bridging or supplementary resource.

IX. CONCLUSION

This paper presented an emotional support system that leverages AI technologies including natural language processing, voice cloning, and multimodal interaction to provide personalized emotional support. The system integrates users' WhatsApp chat history and personal data to create a familiar and comforting presence, enhancing the therapeutic relationship.

Experimental evaluation showed that the system with personalized voice and content achieved significantly higher user engagement and emotional improvement compared to less personalized alternatives. These findings suggest that personalization, particularly voice cloning, plays a crucial role in the effectiveness of AI-based emotional support systems.

Future work will focus on addressing the technical limitations identified in the study and exploring the potential integration of the system with professional mental health services to create a comprehensive support ecosystem.

ACKNOWLEDGMENT

The authors would like to thank the participants in the user study for their valuable feedback and the research assistants who helped with data collection and analysis. This work was supported in part by a grant from the Mental Health Innovation Foundation.

REFERENCES

[1] World Health Organization, "Mental disorders," 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders. [Accessed March 15, 2025].

[2] V. Patel et al., "The Lancet Commission on global mental health and sustainable development," The Lancet, vol. 392, no. 10157, pp. 1553-1598, 2018.

[3] D. D. Luxton, "Artificial intelligence in psychological practice: Current and future applications and implications," Professional Psychology: Research and Practice, vol. 45, no. 5, pp. 332-339, 2014.

[4] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," JMIR Mental Health, vol. 4, no. 2, p. e19, 2017.

[5] 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," The Canadian Journal of Psychiatry, vol. 64, no. 7, pp. 456-464, 2019.

[6] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal Processing Magazine, vol. 18, no. 1, pp. 32-80, 2001.

[7] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 4485-4495.

[8] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine,"

Communications of the ACM, vol. 9, no. 1, pp. 36-45, 1966.

[9] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," JMIR Mental Health, vol. 4, no. 2, p. e19, 2017.

[10] 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, p. e12106, 2018.

[11] L. Sharma, K. Goel, S. Lindenmann, M. Harabagiu, and N. Patel, "Identifying emotional distress signals in text conversations using deep learning approaches," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 1892-1904.

[12] H. Chen, X. Liu, S. Yin, and J. Tang, "Beyond words: Evaluating emotional intelligence in large language models for mental health support," arXiv preprint arXiv:2401.08565, 2024.

[13] J. Cambre and S. Kulkarni, "One voice fits all? Social implications and research challenges of designing voices for smart devices," Proceedings of the ACM on Human-Computer Interaction, vol. 3, no. CSCW, pp. 1-19, 2019.

[14] R. Moorthy and S. Lahiri, "Vocal characteristics influence trust and perceived empathy in AI-driven healthcare interactions," in Proceedings of the 2024 Conference on Health, Inference, and Learning, 2024, pp. 203-215.

[15] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in Advances in Neural Information Processing Systems, 2018, pp. 4480-4490.

[16] W. Cong, K. Rogers, J. Schroeder, and R. Perkins, "Personalized voice for assistive technology: User perceptions of identity, agency, and accessibility," in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1-15.

[17] N. Martinez-Martin and K. Kreitmair, "Ethical issues for direct-to-consumer digital psychotherapy apps: Addressing accountability, data protection, and consent," JMIR Mental Health, vol. 5, no. 2, p. e32, 2018.

[18] J. Torous and L. W. Roberts, "Needed innovation in digital health and smartphone applications for mental health: Transparency and trust," JAMA Psychiatry, vol. 74, no. 5, pp. 437-438, 2017.

[19] E. Ruane, S. Farrell, and A. Ventresque, "User perceptions of conversational AI in health care: A systematic review," ACM Transactions on Interactive Intelligent Systems, vol. 12, no. 2, pp. 1-30, 2022.

[20] K. Hao, "The ethics of voice cloning: Balancing convenience and consent," MIT Technology Review, vol. 124, no. 1, pp. 78-85, 2024.

[21] L. Davis, M. Kulich, and S. Greenstein, "The influence of color and design on mental health application use: A randomized controlled study," Journal of Medical Internet Research, vol. 25, no. 3, p. e41865, 2023.

[22] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in Proceedings of the 38th International Conference on Machine Learning, 2021, pp. 5530-5540.

[23] J. Brooke, "SUS: A 'quick and dirty' usability scale," Usability Evaluation in Industry, vol. 189, no. 194, pp. 4-7, 1996.

[24] H. L. O'Brien, P. Cairns, and M. Hall, "A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form," International Journal of Human-Computer Studies, vol. 112, pp. 28-39, 2018.

[25] A. Spreng, G. McKinnon, R. Mar, and B. Levine, "The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures," Journal of Personality Assessment, vol. 91, no. 1, pp. 62-71, 2009.

[26] P. F. Lovibond and S. H. Lovibond, "The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories," Behaviour Research and Therapy, vol. 33, no. 3, pp. 335-343, 1995.