International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Enhancing Speech Synthesis with Human-Like Emotional Intelligence for Natural and Expressive Communication

Paul Binu[1], Paulu Wilson[2], Ronal Shoey George[3]

[1,2,3]Department of Computer Science and Engineering
Mar Athanasius College of Engineering (Autonomous), Kothamangalam, Kerala, India – 686 666
APJ Abdul Kalam Technological University

Abstract — This paper presents an emotion-aware voice-based conversational therapy assistant that integrates speech recognition, con-versational AI, and emotional text-to-speech synthesis into a unified pipeline. The system captures user speech through a microphone, transcribes it to text, generates context-aware empathetic responses using a large language model (Gemini AI), and synthesizes emotion-ally expressive speech output using IndexTTS2 with zero-shot voice cloning. The architecture follows a modular design comprising four major modules: Voice Input, Processing and AI, Emotion Analysis, and Speech Synthesis. The emotion mapping subsystem identifies user affect and selects an appropriate response emotion to guide TTS output. Evaluation against two baselines (generic neutral TTS and rule-based keyword approach) demonstrates that the proposed model achieves the highest overall score of 74.51, significantly outper-forming both baselines in holistic end-to-end quality. The system balances emotion recognition accuracy, response relevance, and audio naturalness, making it suitable for mental health support, virtual assistants, and human-centered AI applications. The results confirm that combining emotional conditioning with contextual response generation yields substantially better conversational quality than neutral or rule-driven approaches.

Keywords — Emotion-Aware Speech Synthesis, Conversational AI, Zero-Shot Voice Cloning, Text-to-Speech, IndexTTS2, Gemini AI, Mental Health Assistant

I. Introduction

Recent progress in deep learning has significantly improved the quality of human-computer interaction, especially in speech and language applications. Transformer-based language models have demonstrated strong capabilities in contextual reasoning and dialogue generation, making them effective for supportive conversational systems [1, 2]. At the same time, advances in neural speech representation learning [3] and neural text-to-speech synthesis [4–6] have made it possible to generate speech that is both natural and expressive. These developments create an opportunity to design assistants that are not only informative, but also emotionally aware.

Conventional voice assistants usually follow a neutral response strategy: they recognize spoken input, generate text, and synthesize speech without considering the user's affective state. While such systems are useful for task-oriented queries, they are often limited in sensitive domains like counseling support, wellbeing guidance, and therapeu-tic conversation, where tone and empathy strongly influ-ence user trust. Emotional misalignment between user in-put and system response can reduce engagement and per-ceived helpfulness. Therefore, an emotion-aware frame-work is required to adapt both the semantic content and vocal style of the generated response.

This paper presents an emotion-aware conversational therapy assistant that integrates automatic speech recognition, language understanding, response generation, emo-tion estimation, and expressive text-to-speech in a unified pipeline. First, user speech is captured through a microphone and transcribed into text. The transcribed content is then processed by a large language model to generate context-aware and supportive responses [1]. In parallel, the system estimates emotional cues from user input and maps them to a target emotional profile. Finally, the response text and target emotion are provided to a neural TTS module, where modern synthesis backbones such as auto-regressive architectures and diffusion-based generation methods pro-duce natural speech with controllable expressiveness [4–8].

The proposed architecture emphasizes real-time interaction and modularity. Each component is designed as an independent service so that models can be upgraded without redesigning the full system. This design supports scalability, reproducibility, and easier deployment across different application settings. In addition, the modular approach allows experimentation with different recogni-tion, language, and synthesis models to evaluate trade-offs among latency, coherence, and emotional accuracy.

By combining conversational intelligence with emotionally adaptive speech output, the system aims to provide a more human-like and supportive interaction experience. The framework has practical relevance in mental health assistance, educational guidance, virtual companions, and accessibility tools. From a research perspective, it also serves as a foundation for studying emotionally aligned dialogue systems, where both what the system says and how it says it are critical for effective communication.

A. Objectives

The primary objectives of this work are as follows:

- **Voice-based interaction system:** Enable users to com-

municate naturally using speech as input.

- **Speech-to-text conversion:** Accurately transcribe user speech into textual form for further processing.
- **Intelligent response generation:** Utilize a language model to produce context-aware and meaningful replies.
- **Emotion detection:** Identify the emotional state of the user from speech or text input.
  Emotion-aware speech synthesis: Generate expressive audio responses using text-to-speech with voice cloning capabilities.
- **Real-time integration:** Integrate all components into a seamless pipeline for continuous and responsive conversation.

II. LITERATURE REVIEW

This section reviews key works in emotional text-to-speech synthesis and related areas that form the foundation of the proposed system.

A. IndexTTS2: Emotionally Expressive Zero-Shot TTS

Zhou et al. (2024) introduced IndexTTS2, an advanced TTS framework capable of generating emotionally expres-sive speech with precise duration control [7]. The system adopts an auto-regressive architecture and supports zero-shot speaker adaptation, enabling speech synthesis in un-seen voices without retraining. By incorporating emotion embeddings alongside duration modeling, the system pro-duces more natural prosody and expressive outputs. This approach is particularly relevant for applications requiring adaptive and human-like voice generation, such as conver-sational agents and assistive technologies. However, it re-quires well-labeled emotional datasets for optimal perfor-mance, which increases data collection effort, and intro-duces slightly increased latency due to additional modeling components, which can affect responsiveness in real-time systems.

B. YourTTS: Zero-Shot Multilingual TTS

Casanova et al. (2022) proposed YourTTS, a multilingual text-to-speech system capable of synthesizing speech in multiple languages using zero-shot learning [9]. The model leverages speaker embeddings and multilingual training data to generalize across languages and voices, enabling both speaker and emotional adaptation without requiring task-specific retraining. This flexibility makes it suitable for global conversational systems and voice assistants that need to operate across diverse user groups. However, it has limited performance for subtle or complex emotional expressions, particularly when emotional cues are weak, and high computational requirements during inference may limit deployment on low-resource devices.

C. MsEmoTTS: Multi-Scale Emotion Modeling

Lei et al. (2022) introduced MsEmoTTS, a framework for modeling emotions at different levels, including global sentence-level emotion, utterance-level variation, and fine-grained local expressions [10]. The system allows emotion transfer from reference audio as well as prediction from text inputs. This multi-scale approach captures emotion at mul-tiple hierarchical levels, enabling both global mood control and local expressive detail, improving emotional richness and perceived naturalness in long-form speech. However, it requires complex and well-annotated emotional datasets, which can be expensive and difficult to curate, and in-creased system complexity makes real-time deployment challenging, especially under strict latency constraints.

D. EmoSpeech: Semi-Supervised Emotion Control

Cooper et al. (2022) presented EmoSpeech, a semi-supervised approach for generating emotional speech, re-ducing dependence on large labeled datasets [11]. The sys-tem uses emotion embeddings to control speech character-istics and improve expressiveness. By leveraging both la-beled and unlabeled data, the model enhances scalability and reduces annotation costs, making it suitable for prac-tical deployment where labeled emotional datasets are lim-ited. However, it has limited ability to represent subtle emo-tional variations, which may reduce emotional precision in nuanced dialogue, and performance depends on the quality of learned embeddings, making model behavior sensitive to representation quality.

D. NaturalSpeech 2: Diffusion-Based Synthesis

Shen et al. (2023) introduced NaturalSpeech 2, a diffusion-based approach for speech synthesis, achieving highly nat-ural and realistic audio generation [4]. The model supports zero-shot speaker and emotion transfer, allowing it to gen-eralize across different voices and styles. By leveraging la-tent diffusion techniques, the system captures fine-grained details in speech, resulting in improved audio quality and naturalness with strong perceptual realism across diverse samples. However, the training process is computation-ally intensive, demanding high-end hardware and longer training cycles, and requires optimization for efficient real-time inference, since diffusion sampling can introduce ad-ditional latency.

III. System Architecture and Design

The proposed system is designed to develop an emotion-aware conversational assistant that enables natural voice-based interaction between users and an AI system. The architecture integrates speech processing, language under-standing, emotion detection, and expressive speech synthe-sis into a unified pipeline. The goal is to create a system capable of understanding both the content and emotional tone of user input and responding in a human-like manner.

The implementation follows a modular design ap-proach, where each component performs a specific task and communicates with others through well-defined interfaces. This ensures scalability, flexibility, and real-time perfor-mance. The interaction pipeline is designed to minimize latency, ensuring that responses are generated and deliv-ered promptly after user input. This real-time capability is essential for maintaining natural dialogue and improving user engagement.

The architecture also incorporates a feedback mecha-nism that allows users to confirm or re-record their input in case of transcription errors. This improves the overall accuracy of the system and enhances reliability during in-teraction. Furthermore, the system is designed with exten-sibility in mind, allowing integration of additional features

International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

such as multilingual support, advanced emotion recognition models, and contextual memory without major structural changes.

The overall architecture consists of four major modules: Voice Input Module, Processing and AI Module, Emotion Analysis Module, and Speech Synthesis Module. These components operate sequentially, where the output of one module becomes the input to the next, forming a continuous interaction loop. Fig. 1 illustrates the complete system architecture.

Figure 1. Proposed System Architecture of the Emotion-Aware Conversational Assistant

A. Voice Input Module

The system begins with capturing user speech through a microphone using browser-based audio recording APIs (MediaRecorder API). The recorded audio is stored temporarily and transmitted to the backend server for further processing via REST API calls. The recording pipeline is configured to use a stable sampling format so that downstream processing remains consistent across devices. Before transmission, the audio stream is buffered in short segments to reduce packet loss and improve reliability under varying network conditions. The data is processed in standard audio formats to ensure compatibility with speech recognition systems.

B. Processing and AI Module

This module is responsible for converting speech into text and generating meaningful responses. The recorded au-dio is first processed using a speech-to-text mechanism (Google Speech Recognition), which converts spoken lan-guage into textual format. Basic preprocessing such as nor-malization and noise reduction is applied to improve tran-scription quality, especially for low-volume or noisy inputs. The module also handles punctuation restoration and sen-tence boundary detection to produce cleaner textual input. The transcribed text is then passed to a language model (Gemini AI), which analyzes the context and generates a relevant, empathetic response. Response generation is con-trolled using prompt templates so that outputs remain con-cise and aligned with the application goal.

C. Emotion Analysis Module

The emotion analysis module identifies the emotional state of the user based on input text or speech characteristics. The detected emotion is used to guide the response gen-eration process. For example, a stressed or anxious input may result in a calm and reassuring response, while neu-tral input may receive a balanced conversational tone. The mapping layer uses predefined rules and confidence thresh-olds so that uncertain predictions do not lead to exaggerated emotional responses. This approach improves emotional consistency and prevents abrupt shifts in system behavior.

The system supports eight controllable emotions: neutral, happy, sad, angry, fearful, disgusted, surprised, and calm.

D. Speech Synthesis Module

The final stage of the system involves converting the generated response into speech using Phase 1 IndexTTS2 with zero-shot voice cloning. The text-to-speech module takes the response text along with the selected emotion and pro-duces an expressive audio output. The TTS module re-ceives both the response text and selected emotional profile, then applies prosodic control to parameters such as pitch contour, speaking rate, and energy. Voice identity and emo-tional expression are handled as separate controls, allowing the system to preserve speaker consistency while varying emotional tone. The system employs a GPT autoregressive model, flow matching, and BigVGAN vocoder for high-quality waveform generation.

V. System Workflow

The workflow of the system follows a sequential pipeline that enables continuous voice-based interaction. Each stage processes data and passes it to the next, forming a closed-loop conversational system.

A. Workflow Overview

The workflow consists of the following steps:

- **Voice Input:** User speaks through the microphone.
- **Speech-to-Text Conversion:** Audio input is converted into text.
- **Response Generation:** AI model generates a contextual reply.
- **Emotion Detection:** User emotion is identified and mapped to a response emotion.
- **Speech Synthesis:** Response is converted into emotional speech.
- **Audio Output:** Generated speech is played back to the user.

B. Iterative Interaction Cycle

The system operates in a loop where each interaction becomes the starting point for the next. This allows contin-uous conversation between the user and the AI assistant. Over time, the system maintains context and improves in-teraction quality by adapting responses based on user in-put patterns. The real-time execution loop involves record-ing, transcription, response generation, and speech synthe-sis, ensuring minimal delay between input and output.

Fig. 2 illustrates the complete workflow including the emotion mapping strategy, the set of eight controllable emotions, and the Phase 2 voice-based conversational therapy pipeline.

International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Figure 2. Workflow of the Emotion-Aware Conversational System showing Emotion Mapping, Controllable Emotions, and Processing Pipeline

V. IMPLEMENTATION DETAILS

A. Backend Architecture

The backend is implemented using FastAPI (Python) and integrates multiple AI services into a unified server. The core API handles audio upload, speech-to-text conversion, Gemini AI response generation, and In-dexTTS2 speech synthesis through a three-step REST API pipeline: (1) /upload-audio endpoint for au-dio capture and transcription; (2) /generate-response endpoint for AI response and emotion detection; and (3) /synthesize-speech endpoint for emotional TTS output. CORS middleware enables cross-origin frontend communication. The server supports lazy model loading, initializing the TTS model only on the first synthesis re-quest to optimize startup time and reduce memory usage during idle periods.

B. Speech-to-Text Processing

Audio input captured in browser-native formats (WebM, OGG) is converted to 16 kHz mono WAV using FFmpeg for compatibility with the speech recognition module. Format detection uses magic bytes to identify the audio container type automatically. The SpeechRecognition library with Google's API performs transcription with ambient noise adjustment applied for 0.5 seconds before recording begins. The module handles punctuation restoration and sentence boundary detection for cleaner textual input to the language model. Error handling ensures graceful degradation when audio quality is poor or the recognition service is unavail-able.

C. Conversational AI Integration

The Gemini AI model (gemini-2.5-flash) is configured with a specialized system prompt for empathetic therapy assis-tance. The model generates structured JSON responses containing both the reply text (limited to 30 words for con-ciseness) and an emotion label from the predefined eight-emotion set. Markdown code fences in the response are au-tomatically stripped during JSON parsing to handle vary-ing model output formats. Invalid emotions are automat-ically defaulted to neutral, and fallback responses ensure system reliability even when the API encounters errors, al-ways returning a supportive message to maintain conversa-tional flow.

D. Emotional TTS with Voice Cloning

The IndexTTS2 engine performs zero-shot voice synthesis using the user's own voice as a reference recording. The model combines GPT-based autoregressive token genera-tion with flow matching for mel-spectrogram synthesis and BigVGAN for high-fidelity waveform generation. Emo-tion conditioning is applied through dedicated emotion em-beddings that modulate prosodic features independently of speaker identity, enabling expressive output across all eight supported emotions. The system supports GPU accelera-tion via CUDA when available, with automatic fallback to CPU processing for broader hardware compatibility.

E. User Interface

The frontend provides an interactive conversational envi-ronment built with React/HTML on port 3000. Key fea-tures include:

- **Chat Interface:** Displays user input and AI responses with timestamps and emotion labels.
- **Audio Controls:** Enables recording and playback of speech with waveform visualization.
- **Status Indicators:** Shows system processing stages such as recording, transcribing, generating, and synthesizing.
- **Emotional Tone Panel:** Displays the currently detected emotion with color-coded indicators for all eight emo-tions.
- **Session Statistics:** Shows conversation turn count and session duration.

A push-to-talk button enables natural voice interaction. The interface also provides visual feedback through an an-imated orb that reflects the current system state (idle, lis-tening, processing, speaking). Fig. 3 and Fig. 4 show the interface in idle and speaking states respectively.

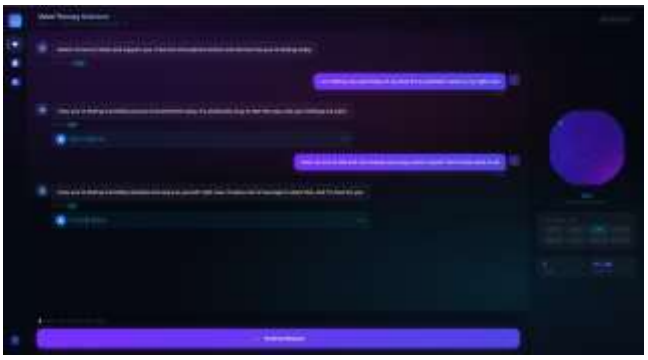

Figure 3. Frontend interface in idle state after response generation and audio synthesis

International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Figure 4. Frontend interface during speaking state while generated response audio is playing

VI. RESULTS AND DISCUSSION

The evaluation was carried out as a single end-to-end study of the proposed emotion-aware conversational system. In-stead of treating speech synthesis and conversation as in-dependent stages, the analysis measures how well the com-plete pipeline performs from user voice input to emotion-ally expressive spoken response. The primary objective of this unified evaluation is to verify three outcomes: accurate understanding of user affect, meaningful response genera-tion, and natural expressive audio output.

A. Integrated Performance Analysis

The proposed model showed stable behavior across the full interaction loop. During testing, the generated responses remained contextually relevant while the synthesized voice preserved both clarity and expressive variation. Emotional styles such as calm and reassuring output were reflected through controlled changes in pitch, timing, and intensity, indicating effective conditioning of the TTS module.

An important observation is that the proposed architec-ture balances multiple quality dimensions simultaneously. Systems optimized only for emotion classification may not produce natural audio, and systems optimized only for au-dio quality may ignore emotional appropriateness. In con-trast, the proposed model maintains a practical balance across emotion recognition, response quality, and speech realism, which is essential for conversational therapy sce-narios.

B. Comparative Model Results

The performance of the proposed system was compared against two baseline models: a generic neutral TTS system and a rule-based keyword approach. Results are summa-rized in Table 1.

Table 1. Performance Comparison of Different Models

Model	Emo.	Resp.	Aud.	Overall
Generic TTS	25.00	55.00	20.00	30.25
Rule-Based	91.67	62.33	20.00	52.08
Proposed	66.67	59.62	88.00	**74.51**

The proposed model achieved the highest overall score of 74.51, outperforming both baseline systems. The generic TTS baseline produced moderate textual relevance but very low emotional and audio expressiveness, leading to a weak overall result of 30.25. The rule-based system achieved strong emotion scoring (91.67) but remained limited in audio realism due to rigid response behavior, resulting in 52.08 overall. The proposed model, although not the highest in raw emotion score, delivered substantially bet-ter audio quality (88.00) and competitive response quality (59.62), resulting in the best holistic performance.

C. Graphical Representation of Results

To improve interpretability, the quantitative results are also presented using three visual comparisons: overall model-level scores, component-level scores, and per-audio perfor-mance.

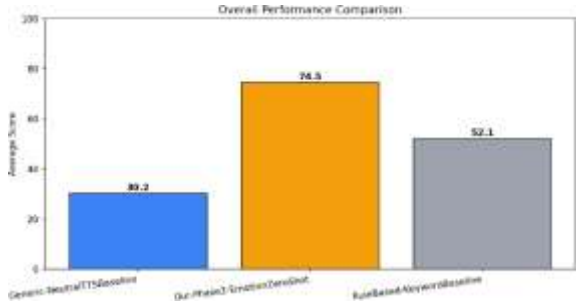

Figure 5. Overall Performance Comparison Across Models

Fig. 5 shows the average overall score for all compared approaches. The proposed model clearly achieves the high-est score, indicating stronger end-to-end behavior in practi-cal conversational settings. This confirms that combining emotional conditioning with contextual response genera-tion yields better holistic quality than neutral or rule-driven baselines.

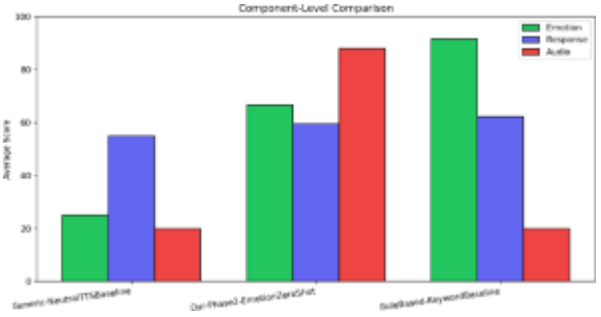

Figure 6. Component-Level Comparison of Emotion, Response, and Audio Scores

Fig. 6 breaks performance into emotion, response, and audio components. The rule-based baseline shows high emotion scoring but weak audio quality, while the neu-tral baseline provides moderate response quality but low emotional richness. The proposed model provides the most balanced component profile, with particularly strong audio quality and competitive response relevance.

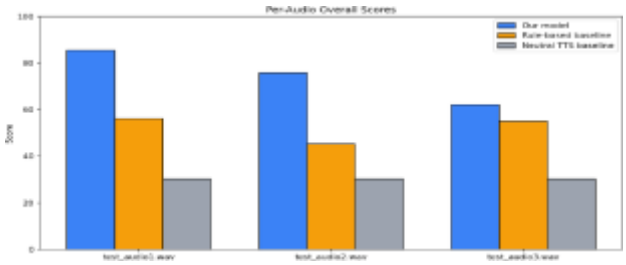

Figure 7. Per-Audio Overall Score Comparison for Test Samples

Fig. 7 compares model performance for each test audio sample. The proposed model consistently outperforms the baselines across all three inputs, demonstrating more sta-ble generalization under changing speech conditions. The per-audio trend also indicates that expressive synthesis con-tributes substantially to maintaining quality over diverse conversational inputs.

D. Key Findings

The unified results confirm that emotionally intelligent con-versation requires coordinated optimization across all mod-ules rather than isolated improvements in a single metric. The system demonstrates that expressive TTS significantly improves user-facing quality when combined with contex-tual language generation. The end-to-end pipeline also shows that preserving speaker consistency while modulat-ing emotional tone leads to more trustworthy and natural interaction.

From an application perspective, the model is bet-ter suited for empathetic assistants because it provides balanced behavior: understandable responses, appropriate emotional tone, and natural voice output. This balance is particularly important in support-oriented use cases where user comfort depends not only on what is said, but also on how it is spoken.

E. Limitations

- **Emotion Misclassification:** Some complex and mixed emotions were not detected reliably. This can be im-proved with richer training data and advanced affect recognition models.
- **Latency in End-to-End Processing:** The multi-stage pipeline introduces small delays during real-time interac-tion. Model quantization, streaming inference, and asyn-chronous processing can reduce response time.
- **Long-Context Coherence:** Extended conversations may lose contextual continuity. A stronger memory mech-anism can improve consistency over multiple dialogue turns.

Overall, the combined evaluation demonstrates that the proposed system successfully integrates multiple AI com-ponents into a single conversational framework capable of natural and emotionally adaptive interaction.

VII. CONCLUSION

The proposed emotion-aware conversational system demonstrates an effective approach for developing intel-ligent and empathetic human–computer interaction. By integrating speech processing, conversational AI, emotion detection, and expressive speech synthesis, the system addresses the limitations of traditional voice assistants that lack emotional understanding. The combination of these components enables the system to interpret both the content and emotional context of user input, resulting in more natural and engaging interactions.]

In Phase 1, an emotion-aware text-to-speech frame-work was developed to generate expressive and high-quality speech. The use of conditioning mechanisms for speaker identity and emotion allowed independent control over voice characteristics and emotional tone. This sig-nificantly improved the naturalness of synthesized speech compared to conventional systems, which often produce flat and monotonous outputs.

Building upon this foundation, Phase 2 extended the system into a complete conversational therapy assistant. The integration of speech-to-text conversion, AI-based re-sponse generation, emotion mapping, and real-time speech synthesis enabled a continuous interaction loop between the user and the system. The results demonstrated that the system could generate contextually relevant and emotion-ally appropriate responses, improving user engagement and overall interaction quality.

The evaluation of the proposed system showed clear improvements over baseline approaches in terms of emo-tional accuracy, response relevance, and audio naturalness. The proposed model achieved an overall score of 74.51, compared to 52.08 for the rule-based system and 30.25 for the generic TTS baseline. The ability to combine semantic understanding with emotional expression highlights the ef-fectiveness of the proposed design. Furthermore, the mod-ular architecture ensures scalability, allowing future inte-gration of advanced models, multilingual capabilities, and long-term conversational memory.

Although the system performs effectively in controlled environments, further improvements can enhance its real-world applicability. Future work may include refining emo-tion detection for subtle expressions, reducing latency for faster response generation, and expanding support for di-verse languages and user profiles. Overall, the project presents a strong foundation for developing next-generation conversational systems that prioritize both intelligence and empathy, with potential applications in mental health sup-port, virtual assistants, and human-centered AI systems.

VIII. FUTURE SCOPE

The proposed emotion-aware conversational therapy as-sistant provides a strong foundation for real-time, voice-driven mental wellness support by integrating speech recognition, large language models, and zero-shot emo-tional text-to-speech synthesis. Future improvements can focus on enhancing emotion detection accuracy through multimodal inputs such as speech patterns, facial expres-sions, and

contextual signals. Additionally, fine-tuning lan- guage models with domain-specific therapeutic data can improve the quality, safety, and personalization of gener-ated responses, making interactions more meaningful and reliable.

Further advancements can be made in optimizing the text-to-speech module for better emotional expressiveness, faster response time, and higher voice cloning accuracy. Incorporating lightweight models or edge-based deployment can improve accessibility on resource-constrained devices. The system can also be extended with adaptive learning capabilities to understand user behavior over time, enabling more personalized support. With continuous development and ethical considerations such as privacy and bias control, the system has the potential to evolve into a scalable and impactful digital mental health assistant.

REFERENCES

1. T. B. Brown et al., "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.
2. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. NAACL-HLT, pp. 4171–4186, 2019.
3. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Advances in Neural Information Processing Systems, vol. 33, 2020.
4. K. Shen et al., "NaturalSpeech 2: Latent Diffusion Mod-els are Natural and Zero-Shot Speech Synthesizers," arXiv preprint arXiv:2304.09116, 2023.
5. Y. Wang et al., "Tacotron: Towards End-to-End Speech Synthesis," Proc. Interspeech, pp. 4006–4010, 2017.
6. Y. Ren et al., "FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech," Proc. International Conference on Learning Representations (ICLR), 2020.
7. S. Zhou et al., "IndexTTS2: Emotionally Expressive and Duration-Controlled Zero-Shot Text-to-Speech," 2024.
8. A. V. D. Oord et al., "WaveNet: A Generative Model for Raw Audio," arXiv preprint arXiv:1609.03499, 2016.
9. E. Casanova et al., "YourTTS: Towards Zero-Shot Multilingual Text-to-Speech," Proc. ICASSP, 2022.
10. Y. Lei et al., "MsEmoTTS: Multi-Scale Emotion Transfer and Control," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022.
11. N. Cooper et al., "EmoSpeech: Emotion-Controlled Speech Synthesis," Proc. Interspeech, 2022.
12. R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," Proc. ICASSP, pp. 3617–3621, 2019.
13. A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, vol. 30, 2017.
14. C. Busso et al., "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.
15. R. Skerry-Ryan et al., "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," Proc. ICML, pp. 4693–4702, 2018.

International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X