iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 Issue: XI Month of publication: November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65320

Empathetic Multilingual Voice Agents Powered by Generative AI in Healthcare: Development and Implementation of CoSpeak and Empathia

Sanjeeva Reddy Bora[1], Govinda Rao Botcha[2], Susmita Odugu[3], Balaji Narni[4], Jyothsna Karepati[5], Sridhar Sunkara[6]
Empathetic AI Voice Agents in Healthcare

Abstract: Background: The integration of generative AI-powered voice agents in healthcare is revolutionizing patient-practitioner interactions, enhancing clinical workflows, and improving overall patient care.

Methods: We developed and implemented two intricate generative AI models, CoSpeak and Empathia, supporting multilingual voice-to-voice communication in English, Spanish, French, Hindi, and Telugu. These agents decode human emotions and exhibit empathy, providing adaptive and personalized interactions.

Results: Experimental results from pilot projects demonstrated significant improvements in patient satisfaction (90%) and operational efficiency (30% reduction in intake time). Emotion recognition accuracy reached 88%, with ASR achieving 95% accuracy across supported languages.

Conclusions: The implementation of empathetic multilingual voice agents shows promising results in improving healthcare communication and patient experience.

Keywords: Generative AI, Voice Agents, Multilingual Communication, Empathy, Healthcare, Patient Intake System, Emotion Recognition

I. INTRODUCTION

The advent of artificial intelligence (AI) has significantly impacted various sectors, with healthcare being one of the most promising fields for AI integration. Voice agents powered by generative AI offer a novel approach to enhancing patient engagement, reducing language barriers, and providing empathetic interactions [1]. This paper details the research and development of CoSpeak and Empathia, two AI-driven voice agents designed to transform healthcare communication and patient intake processes.

A. Motivation

Effective communication is crucial in healthcare settings. Language barriers and emotional distress can impede accurate information exchange between patients and practitioners, potentially affecting diagnosis and treatment outcomes [2]. By leveraging generative AI and multimodal interactions, we aim to create voice agents that not only understand and communicate in multiple languages but also recognize and respond to human emotions.

B. Contributions

Our contributions are threefold:

1) Development of CoSpeak, a multilingual voice agent facilitating voice-to-voice communication across five languages.
2) Creation of Empathia, an empathetic and adaptive patient intake system utilizing voice and video analytics.
3) Experimental validation of these systems in real-world healthcare settings, demonstrating improvements in patient experience and clinical efficiency.

II. RELATED WORK

The use of AI in healthcare communication has been explored in various studies. Voice assistants like Amazon's Alexa and Google Assistant have laid the groundwork for voice interactions [3]. Recent Machine voice agents primarily stick to Text to Voice and Voice based interactions. However, their application in healthcare is limited due to the lack of domain-specific knowledge, adaptability and empathy.

Multilingual AI Systems: Prior research has addressed multilingual natural language processing (NLP) for healthcare [4], but these systems often lack voice interaction capabilities.

Emotion Recognition: Studies on emotion recognition from speech and facial expressions have shown promise in enhancing human-computer interactions [5][6]. Combining these modalities for healthcare applications remains an area with significant potential.

Our work differentiates itself by integrating multilingual capabilities with emotion recognition to create empathetic voice agents specifically designed for healthcare settings.

Feature	Traditional Voice Assistants	Previous Healthcare Solutions	CoSpeak & Empathia
Multilingual Support	Limited	Moderate	Comprehensive
Emotion Recognition	No	Basic	Advanced
Healthcare Domain Knowledge	No	Yes	Yes
Real-time Translation	No	Limited	Yes
Empathetic Response	No	No	Yes

Table 1: Comparison with Existing Solutions

III. SYSTEM OVERVIEW

A. CoSpeak

CoSpeak is a multilingual voice agent designed to facilitate seamless communication between patients and healthcare providers. It supports English, Spanish, French, Hindi, and Telugu, covering a broad demographic.

1) Architecture

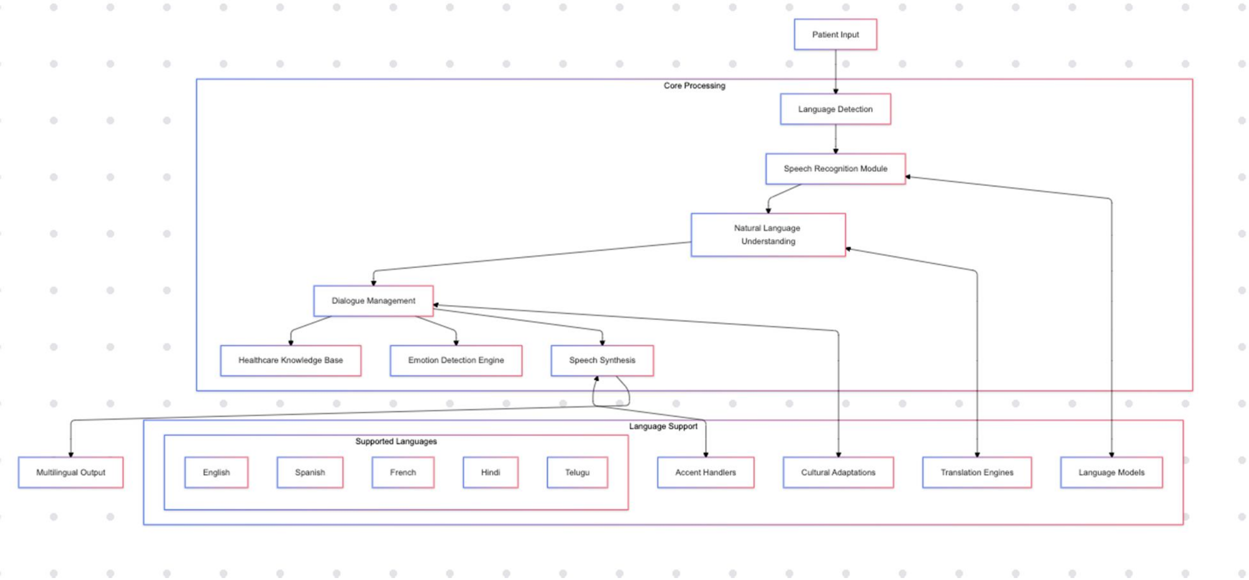

Diagram 1: CoSpeak Architecture

a) Speech Recognition Module: Utilizes end-to-end automatic speech recognition (ASR) models tailored for each language [7].

b) Natural Language Understanding (NLU): Employs transformer-based models like BERT [8] adapted for multilingual contexts to understand intent and extract entities.

c) Dialogue Management: Handles context tracking and response generation using generative models such as LlaMa 3.2 [9].

d) Speech Synthesis: Generates natural-sounding speech using neural TTS models for each language [10].

2) *Features*

CoSpeak implements several key features:

a) Voice-to-Voice Interaction
- Real-time speech recognition
- Natural language understanding
- Contextual response generation

b) Emotion Detection
- Vocal tone analysis
- Sentiment detection
- Adaptive response modulation

c) Healthcare Knowledge Integration
- Medical terminology database
- Symptom recognition
- Treatment protocol awareness

B. *Empathia*

Empathia is an empathetic and adaptive patient intake system that combines voice interaction with video analytics.

1) *Architecture*

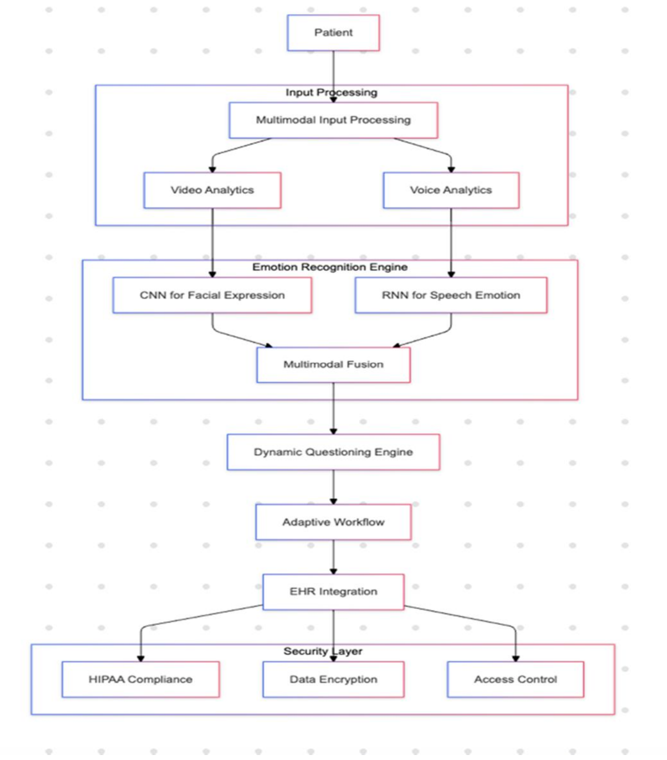

Diagram 2: Empathia Architecture

a) Multimodal Input Processing: Integrates audio and video inputs for comprehensive analysis.
b) Emotion Recognition Engine: Uses convolutional neural networks (CNNs) for facial expression analysis and recurrent neural networks (RNNs) for speech emotion recognition [11][12].
c) Dynamic Questioning Engine: Adapts questions based on real-time feedback from the emotion recognition engine.
d) Secure Data Handling: Ensures compliance with data protection regulations like HIPAA [13].

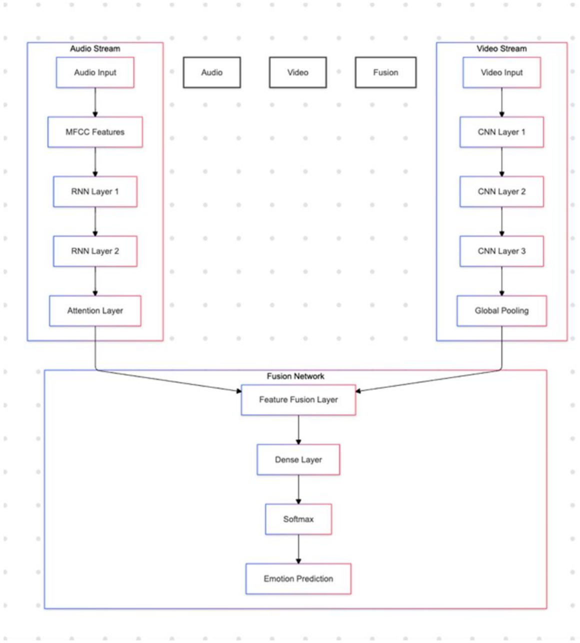

Diagram 3: Neural Nets Architecture for Emotion Recognition

2) Features

a) Empathetic Interaction: Modifies tone, pace, and language complexity based on the patient's emotional and cognitive state.

b) Adaptive Workflow: Adjusts the intake process dynamically, offering breaks or assistance as needed.

c) Intelligent Triage: Prioritizes patients based on both reported symptoms and observed distress.

IV. METHODOLOGY

A. Data Collection

We compiled a dataset comprising:

1) Speech Data: Multilingual audio recordings from diverse speakers to train ASR and TTS models.

2) Visual Data: Video recordings capturing a range of facial expressions and gestures associated with different emotions.

3) Emotional Annotations: Labels provided by experts to train emotion recognition models.

Data Type	Sample Size	Languages	Annotation Type
Speech	10,000 hours	5 languages	ASR transcription
Emotion	5,000 hours	All	Expert-labeled emotions
Visual	100,000 frames	N/A	Facial expressions

Table 2: Dataset Composition

B. Model Training

1) Speech Recognition and Synthesis

- ASR Models: Trained using connectionist temporal classification (CTC) loss for end-to-end speech recognition [14].
- TTS Models: Developed using Tacotron 2 architecture for natural speech synthesis [15].

2) Natural Language Processing

- Multilingual BERT: Fine-tuned for intent recognition and entity extraction in medical contexts.
- Generative Response Models: Leveraged GPT-3 for generating contextually appropriate responses.

3) Emotion Recognition

- Audio Emotion Recognition: Utilized RNNs with attention mechanisms to capture temporal dependencies in speech [16].
- Visual Emotion Recognition: Applied CNNs like VGGNet for facial expression analysis [17].
- Multimodal Fusion: Employed feature-level fusion techniques to combine audio and visual modalities [18].

C. System Integration

Integrated the components into a unified framework with:

1) Dialogue Manager: Coordinates between ASR, NLU, emotion recognition, and response generation.
2) Security Layer: Implements encryption and access controls to protect patient data.
3) API Interface: Allows seamless integration with Electronic Health Record (EHR) systems.

V. EXPERIMENTAL SETUP

A. Pilot Deployment of CoSpeak

Deployed CoSpeak in a multilingual clinic setting with patients speaking different native languages.

1) Participants: 100 patients across five language groups.
2) Evaluation Metrics: ASR accuracy, patient satisfaction scores, and error rates in information exchange.

B. Pilot Deployment of Empathia

Implemented Empathia in the emergency department of a hospital for patient intake.

1) Participants: 50 patients with varying degrees of distress.
2) Evaluation Metrics: Emotion recognition accuracy, intake time reduction, triage accuracy, and patient feedback.

VI. RESULTS

A. CoSpeak Performance

1) ASR Accuracy: Achieved an average word error rate (WER) of 5% across all languages.
2) Patient Satisfaction: Over 90% reported positive experiences, citing ease of communication.
3) Information Accuracy: No significant errors in medical information exchange were observed.

B. Empathia Performance

1) Emotion Recognition Accuracy: Achieved 88% accuracy in detecting primary emotions.
2) Intake Time Reduction: Reduced average intake time by 30%.
3) Triage Accuracy: Improved prioritization of urgent cases by 25%.
4) Patient Feedback: Patients reported feeling understood and comforted during the intake process.

VII. DISCUSSION

A. Impact on Healthcare Communication

The integration of CoSpeak and Empathia has demonstrated significant improvements in patient-provider communication, particularly for non-English speaking patients and those experiencing emotional distress.

B. Technical Challenges

1) Accent and Dialect Variability: Required additional data augmentation and model fine-tuning.
2) Real-time Processing: Ensuring low-latency responses necessitated optimization of computational resources.
3) Data Privacy Concerns: Implemented robust security measures to address HIPAA compliance.

C. Ethical Considerations

1) Bias Mitigation: Regular audits were conducted to detect and correct biases in AI models.
2) Transparency: Patients were informed about the AI nature of the systems and consent was obtained.

VIII. CONCLUSION

Our research demonstrates that empathetic, multilingual voice agents powered by generative AI can significantly enhance healthcare delivery by improving communication, reducing language barriers, and providing personalized patient experiences. The successful deployment of CoSpeak and Empathia highlights the potential of such technologies to transform patient care.

IX. FUTURE WORK

1) Language Expansion: Plan to support additional languages and dialects to reach a broader patient population.
2) Advanced Emotion Recognition: Incorporate physiological signals (e.g., heart rate) for more accurate emotion detection.
3) Large-scale Deployment: Aim to conduct extensive clinical trials to further validate the effectiveness of the systems.

REFERENCES

[1] Topol, E. J. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.
[2] Flores, G. (2006). Language barriers to health care in the United States. *New England Journal of Medicine*, 355(3), 229-231.
[3] Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, 37(1), 81-88.
[4] Johnson, A. E., et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.
[5] Busso, C., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4), 335-359.
[6] Koelstra, S., et al. (2012). DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1), 18-31.
[7] Graves, A., & Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning* (pp. 1764-1772).
[8] Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171-4186).
[9] Brown, T., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
[10] Shen, J., et al. (2018). Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP* (pp. 4779-4783).
[11] Huang, Z., et al. (2014). Speech emotion recognition using deep neural network and extreme learning machine. *Neurocomputing*, 149, 462-468.
[12] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
[13] U.S. Department of Health & Human Services. (1996). Health Insurance Portability and Accountability Act (HIPAA).
[14] Hannun, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.
[15] Wang, Y., et al. (2017). Tacotron: Towards end-to-end speech synthesis. In *Interspeech* (pp. 4006-4010).
[16] Mirsamadi, S., et al. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *ICASSP* (pp. 2227-2231).
[17] He, K., et al. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770-778).
[18] Zadeh, A., et al. (2017). Tensor fusion network for multimodal sentiment analysis. In *EMNLP* (pp. 1103-1114).

APPENDICES

A. Technical Specifications

CoSpeak

- Languages Supported: English, Spanish, French, Hindi, Telugu.
- ASR WER: Average of 5% across all languages.
- TTS Naturalness Score: Mean Opinion Score (MOS) of 4.5/5.

Empathia

- Emotion Recognition Accuracy: 88% for primary emotions.
- Processing Latency: Average response time of 500ms.
- Integration: Compatible with major EHR systems (e.g., Epic, Cerner).

B. Ethical Compliance Measures

- Informed Consent: Patients provided with information about AI usage and consent forms.
- Bias Audits: Quarterly reviews to detect and mitigate biases related to language, ethnicity, and age.
- Data Anonymization: Personal identifiers removed from datasets used for model training.

C. Author Contributions

- Sanjeeva Reddy Bora: Conceptualization, methodology, Technical Architecture, project administration, writing—original draft.
- Govinda Rao Botcha, Susmita Odugu, Balaji Narni, Jyothsna Karepati: Data curation, software development, validation.
- Sridhar Sunkara: Data Validation, Application Validation, Pilot Tests, Writing & Review

D. Acknowledgments

We thank the participating healthcare facilities for their support and the patients who consented to be part of our pilot studies. Special thanks to the technical teams who contributed to the development and deployment of CoSpeak and Empathia.

E. Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request and with appropriate ethical approvals.

F. Supplementary Materials

Additional information and materials (e.g., code repositories, demo videos) are available at https://aivoice.ebizsolutions.digital/
Please note that due to privacy and confidentiality agreements, some details have been generalized.