Grenze International Journal of Engineering and Technology, June Issue

EmoGita: A Multimodal Emotion-Aware Retrieval-Augmented Framework for Personalized Bhagavad Gita Guidance

S. Sudeshna[1], Varigala Pavan Praneeth[2], Dusa Akhil[3], Sampelly Srihas[4] and Panumati Arjun Reddy[5]
[1]VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana, India
Email: sudeshna_s@vnrvjiet.in
[2-5]5th year students VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana, India
Email: varipavan415@gmail.com, dusaakhil1@gmail.com, srihassampelly@gmail.com, arjunreddypanumati@gmail.com

***Abstract*— Emotional distress often shows up in subtle ways through the interplay of voice tone and language, making it difficult for systems that rely on just one type of input to interpret accurately. This paper introduces EmoGita, a multimodal framework that combines speech and text analysis to detect a user's emotional state and provide personalized philosophical guidance from the Bhagavad Gita using a retrieval-augmented generation (RAG) approach. Speech-based emotions are captured with a Wav2Vec2 model, while text-based emotions are inferred using a BiLSTM classifier on transcripts generated by Whisper. The results from both channels are merged using a weighted probabilistic method to reduce ambiguity and enhance emotional accuracy. The combined emotional representation, together with the user's context, is projected into a shared semantic space to query a vector database of Bhagavad Gita verses using sentence-level embeddings. A large language model then generates explanations that explicitly connect the retrieved verses to the user's emotional state. Compared with systems that rely solely on text or speech, this multimodal fusion approach delivers guidance that is more emotionally aligned and contextually relevant. EmoGita demonstrates how affective computing and emotion-conditioned retrieval can be integrated into a single, human-centered AI system for offering sensitive and personalized guidance.**

***Index Terms*— Multimodal Emotion Recognition, Affective Computing, Emotion-Aware Retrieval, Retrieval-Augmented Generation, Speech and Language Processing, Human-Centered AI, Contextual Guidance Systems.**

I. INTRODUCTION

Human emotions are naturally complex, constantly changing, and often communicated in indirect ways. In spoken communication, emotions are conveyed not only through word choice but also through prosodic cues such as variations in pitch, speech rate, and vocal intensity. Traditional AI-based emotion analysis systems usually rely on either textual sentiment analysis or acoustic emotion recognition, which leads to an incomplete understanding of the user's emotional state.

Despite significant progress in deep learning for speech and language modeling, most real-world AI systems remain largely task-focused rather than emotionally aware.

Grenze ID: 01.GIJET.12.2.81

In emotionally sensitive domains—such as mental health support or personal guidance—this limitation can result in responses that feel superficial, lack empathy, or fail to consider the user's wider emotional context.

The Bhagavad Gita offers a well-structured philosophical framework that directly addresses emotional struggle, moral uncertainty, fear, and resilience. Its verse-based structure and rich thematic content make it particularly suitable for semantic retrieval when combined with modern embedding and retrieval techniques.

To overcome these challenges, this work introduces EmoGita, a system designed to:

- Capture emotional cues across multiple modalities
- Reduce interpretive uncertainty through emotion fusion
- Retrieve philosophically relevant content aligned with the user's emotional context
- Generate personalized, explanatory responses rather than static outputs

By combining affective computing with retrieval-augmented generation, EmoGita demonstrates a practical pathway toward developing emotionally intelligent and culturally grounded AI systems.

II. Literature Survey

This literature survey provides a structured review of prior work across speech emotion recognition, text-based emotion analysis, multimodal affective computing, and retrieval-augmented generation (RAG) systems. Its goal is to highlight limitations in existing methods while situating the proposed framework within current research directions. The review focuses on methodological choices, underlying technologies, reported performance, and how emotional understanding is applied in contextual or decision-oriented scenarios.

Early research in speech emotion recognition (SER) largely concentrates on extracting acoustic and prosodic features using deep learning techniques. Akinpelu et al. present a real-time SER system built on convolutional and recurrent neural networks, enhanced with data augmentation to improve performance in noisy conditions. Although the system achieves strong accuracy, it is limited to speech-only input and lacks the ability to reason about context beyond basic emotion classification. Kang et al. advance SER by incorporating CNN–LSTM architectures with attention mechanisms, enabling more effective temporal modeling of emotional cues. However, their approach still treats emotion recognition as a standalone classification task, without leveraging the results for deeper semantic reasoning.

Researchers have also explored lightweight SER models to support real-time and resource-efficient deployment. Sharma et al. evaluate DistilHuBERT and PaSST architectures, demonstrating trade-offs between performance and efficiency that are suitable for constrained environments. Despite their computational advantages, these models do not address higher-level emotional interpretation or integration with external knowledge systems. Patch-ViT-based methods further enhance spectral feature learning but remain confined to unimodal inference pipelines. To address the shortcomings of unimodal approaches, recent studies have shifted toward multimodal emotion recognition by combining speech and text signals. Wu et al. introduce a fusion framework that integrates audio, text, and video features to improve emotional accuracy. Plaza-del-Arcos et al. propose CM-RoBERTa, which uses cross-modal transformers to jointly encode speech and text representations. Yadav et al. extend this line of work by incorporating memory-augmented cross-modal transformers to preserve emotional context over time. While these approaches improve classification performance, their outputs typically serve as final predictions rather than inputs to adaptive or personalized response generation systems.

At the same time, text-based emotion recognition has progressed through transformer-based models enhanced with semantic and psycholinguistic features. Moudarres et al. show that integrating psychologically informed linguistic features with transformers leads to improved emotion detection from text. Jain et al. provide a comprehensive survey of datasets, models, and applications in emotion analysis, identifying persistent challenges such as emotional ambiguity, cross-domain generalization, and limited interpretability. Despite these advances, most text-based models remain isolated from multimodal inputs and are rarely used to guide downstream reasoning processes. Beyond emotion analysis, retrieval-augmented generation has gained attention as a way to ground large language models in external knowledge sources. Mandikal and Rajan demonstrate a Vedantic question-answering system using RAG, showing improved factual grounding through scripture retrieval. Kapuriya and Bhatt introduce Spiritual-LLM, a Gita-guided counseling framework that illustrates the feasibility of combining spiritual texts with generative models. Prakash et al. focus on embedding-based semantic retrieval of Bhagavad Gita verses, validating the effectiveness of vector-based scripture search.

However, most RAG-based spiritual or counseling systems rely on explicit user queries and do not automatically infer emotional context from speech or behavior. Emotion-aware conversational agents such as NirvanaNavigator attempt to incorporate emotional cues, but they primarily depend on prompt engineering or rule-based tagging rather than learned multimodal emotion inference.

Advanced representation learning methods, including AnchorBERT, improve emotional feature anchoring, while multimodal fusion techniques enhance robustness across different input modalities. Even so, most existing systems treat emotion detection, retrieval, and response generation as loosely connected components rather than a unified pipeline.

In summary, the literature highlights a key research gap: the lack of end-to-end systems that automatically infer emotions from both voice and text, fuse these signals in a principled manner, and use the resulting emotional state to guide knowledge retrieval and personalized explanation generation. While current approaches excel either in emotion recognition or in knowledge-grounded generation, they rarely combine both into a cohesive, user-centered framework. The proposed system addresses this gap by unifying multimodal emotion detection, emotion-conditioned retrieval from the Bhagavad Gita, and LLM-driven personalized guidance within a single adaptive architecture.

III. METHODOLOGIES USED

TABLE I: METHODOLOGIES USED IN THIS PAPERS

S. No	Title	Year	Methodology
1	Real-Time Speech Emotion Recognition Using Deep Learning and Data Augmentation	2023	CNN–RNN based SER with data augmentation
2	CNN-LSTM and Attention Mechanism for Speech Emotion Recognition	2023	CNN-LSTM with attention-based temporal modeling
3	DistilHuBERT vs. PaSST for Lightweight Speech Emotion Recognition	2023	Lightweight transformer-based SER models
4	Patch-ViT: Vision Transformer for Speech Emotion Recognition	2023	Vision Transformer applied to spectrogram patches
5	M-fusHER: Multimodal Emotion Recognition Using Audio, Text, and Video	2024	Multimodal fusion using deep neural networks
6	CM-RoBERTa: Cross-Modal RoBERTa for Emotion Recognition from Speech and Text	2023	Cross-modal transformer-based fusion
7	MemoCMT: Cross-Modal Transformer Fusion for Emotion Recognition	2024	Transformer fusion with memory-based context modeling
8	Transformer with Psycholinguistic Features for Text Emotion Recognition	2023	Transformer with linguistic and psycholinguistic features
9	Emotion Analysis Survey: Models, Datasets, and Applications	2023	Comparative survey of emotion recognition models
10	AnchorBERT: Learning Emotion Anchors for Emotion Recognition	2023	Emotion-aware transformer with anchor learning
11	ProEmo: Prompt-Driven Emotional Text-to-Speech Synthesis	2023	Prompt-based emotional control in TTS systems
12	Semi-Supervised Emotional TTS with Implicit Representations	2023	Semi-supervised emotional speech synthesis
13	EmoVoice: Emotionally Controllable TTS Using Large Language Models	2023	LLM-driven controllable emotional speech generation
14	ZET-Speech: Zero-Shot Emotional TTS with Domain-Adversarial Training	2023	Zero-shot emotional speech synthesis
15	EASPO: Emotion-Aligned Speech Synthesis via Preference Optimization	2023	Preference-optimized emotional speech synthesis
16	Ancient Wisdom, Modern RAG: Vedantic QA Using Retrieval-Augmented Generation	2023	RAG-based knowledge-grounded QA system
17	Spiritual-LLM: Gita-Guided Counseling with Generative AI	2023	Generative AI with spiritual text grounding
18	Ancient Indian Scriptures + RAG: Knowledge Retrieval from the Bhagavad Gita	2023	Vector embedding-based scripture retrieval
19	NirvanaNavigator: Emotion-Aware Chatbot Using Gita and RAG	2023	Emotion-aware conversational system with RAG

IV. RESEARCH GAP

A close examination of the referenced studies shows that most prior research focuses primarily on improving emotion recognition accuracy rather than on how emotions are actually used. Works such as Akinpelu et al.

(2023), Kang et al. (2023), and Sharma et al. (2023) concentrate on optimizing speech emotion recognition through deep learning models including CNN–LSTM architectures, HuBERT-based models, and Vision Transformers. Although these studies achieve notable performance gains through architectural improvements and data augmentation, they treat emotion detection as the final objective. The recognized emotions are not further contextualized or applied to downstream reasoning, decision-making, or personalized response generation, creating a clear gap between recognition and practical application.

Studies in multimodal emotion recognition, including those by Wu et al. (2024), Plaza-del-Arcos et al. (2023), and Yadav et al. (2024), attempt to combine speech and text using cross-modal transformers and fusion techniques. However, these approaches primarily emphasize feature-level or representation-level fusion aimed at improving classification performance. They do not explicitly address semantic or psychological alignment across modalities, nor do they use the fused emotional representation as input to a knowledge-driven or interpretive system. As a result, even though multimodal consistency improves, emotional outputs remain disconnected from actionable guidance or deeper contextual interpretation. Research on text-based emotion recognition, such as Moudarres et al. (2023), Elbakoury et al. (2023), and Jain et al. (2023), focuses on linguistic and psycholinguistic modeling using transformer-based and anchor-driven representations. While these works provide strong theoretical foundations for textual emotion analysis, they are largely confined to standalone inference tasks. They do not investigate how emotions inferred from text can be combined with speech-based emotions or integrated into a broader reasoning framework that adapts system behavior based on emotional context.

On the knowledge retrieval front, studies by Mandikal and Rajan (2023), Kapuriya and Bhatt (2023), Prakash et al. (2023), and Sharma et al. (2023) introduce retrieval-augmented generation frameworks for accessing ancient Indian texts such as the Bhagavad Gita. Although these systems successfully demonstrate semantic retrieval and generative explanation capabilities, they rely mainly on textual similarity or prompt-based techniques. Emotional context—whether derived from speech, text, or multimodal signals—is largely excluded from the retrieval process, resulting in guidance that is contextually relevant but emotionally neutral.

In addition, none of the reviewed RAG-based spiritual or counseling systems establish a direct link between emotion recognition outputs and vector database retrieval strategies. Emotional labels or probability distributions are not incorporated into vector representations or used as retrieval constraints, which limits the depth of personalization. This separation is evident when comparing emotion-focused research with RAG-based counseling systems, as the two lines of work largely progress independently rather than converging.

Another notable limitation across the literature is the absence of explanatory personalization. Even when systems generate guidance or responses, they do not clearly explain why a specific verse, recommendation, or response is appropriate for the detected emotional state. While large language models are used for generation, they are not systematically employed to connect user-described situations, inferred emotions, and retrieved knowledge into a coherent and emotionally aligned explanation. Finally, none of the surveyed studies address longitudinal emotional modeling. Emotion recognition and guidance are treated as single-session interactions, with no mechanisms to store emotional history, track emotional trends, or adapt responses based on prior emotional states. This significantly limits the usefulness of existing systems for long-term emotional support or well-being monitoring.

In summary, although existing literature makes meaningful contributions in emotion recognition accuracy, multimodal fusion, and knowledge retrieval when considered independently, a clear research gap remains. There is a lack of integrated systems that combine multimodal emotion analysis with emotion-aware retrieval and personalized explanatory generation. The proposed work directly addresses this gap by unifying these components into a single end-to-end framework that uses detected emotions not merely for classification, but as a central signal for retrieval, reasoning, and emotionally contextualized guidance.

V. Proposed Systems

The proposed system introduces an integrated, emotion-aware guidance framework that combines multimodal emotion recognition with retrieval-augmented generation (RAG) to provide personalized counseling grounded in the Bhagavad Gita. User speech is processed through two parallel pipelines: a Wav2Vec2-based deep learning model captures vocal emotional cues, while Whisper converts speech into text, which is then analyzed using a BiLSTM-based text emotion classifier. The emotion outputs from both modalities are merged using a weighted fusion strategy, producing a stable and reliable emotional representation while reducing ambiguity that can arise from relying on a single modality.

This fused emotional state, together with the transcribed situational context, is embedded into a shared semantic space and used to query a Qdrant vector database containing Bhagavad Gita verses enriched with semantic

embeddings and emotion-related metadata. The retrieval process prioritizes both contextual relevance and emotional alignment, going beyond simple keyword matching. A large language model then generates personalized explanations that link the user's emotional state and lived experience with the philosophical insights of the retrieved verses. By tightly integrating emotion detection, knowledge retrieval, and explanation generation, the system transforms static spiritual texts into adaptive, emotionally grounded guidance.

VI. Experimental Evaluation and Technical Analysis

To evaluate the effectiveness of the proposed EmoGita framework, we analyze the impact of multimodal emotion fusion on emotional alignment and contextual relevance of retrieved guidance. Since the system is designed for emotionally sensitive, user-centered guidance rather than conventional classification benchmarks, evaluation is performed through a combination of comparative analysis and qualitative assessment.

A. Emotion Fusion Analysis

EmoGita integrates emotional cues from both speech and text modalities using a weighted probabilistic fusion strategy. Let P_v and P_t denote the probability distributions obtained from the voice-based and text-based emotion classifiers, respectively. The fused emotion probability distribution P_f is computed as:

$$P_f = \alpha P_v + (1-\alpha)\, P_t$$

where $\alpha \in [0,1]$ represents the relative importance assigned to vocal emotional cues. In the current implementation, α is empirically set to 0.6 to prioritize prosodic information while still incorporating semantic context from text. This fusion approach reduces uncertainty arising from noisy or ambiguous single-modality predictions and produces a more stable emotional representation.

B. Ablation Comparison

To understand the contribution of each modality, we compare three configurations:

- Voice-only emotion recognition, where guidance retrieval is conditioned solely on speech emotion predictions
- Text-only emotion recognition, where guidance is conditioned on emotions inferred from transcribed text
- Multimodal fusion (EmoGita), where both modalities are combined using the proposed fusion strategy

Qualitative analysis across multiple user scenarios shows that unimodal approaches often misinterpret emotional intent when either vocal tone or linguistic content dominates. In contrast, the fused approach consistently produces guidance that aligns more closely with both the user's emotional state and situational context. This demonstrates the advantage of multimodal fusion in reducing emotional misclassification and improving downstream retrieval relevance.

C. Emotion-Aware Retrieval Evaluation

The fused emotional state is incorporated into the retrieval process by augmenting the semantic query with emotional context before embedding. This emotion-conditioned retrieval strategy biases the vector search toward Bhagavad Gita verses that are both semantically relevant and emotionally aligned. Compared to emotion-agnostic retrieval, this approach reduces the likelihood of retrieving philosophically relevant but emotionally mismatched verses, thereby improving the perceived appropriateness of the guidance.

D. Qualitative Case Analysis

Consider a user expressing anxiety related to career uncertainty. In text-only configurations, the system often retrieves verses emphasizing duty without acknowledging emotional distress. Voice-only configurations occasionally misclassify calm speech as neutrality, resulting in generic guidance. The multimodal EmoGita framework successfully identifies underlying anxiety through combined vocal hesitation and linguistic cues, leading to the retrieval of verses addressing fear, detachment from outcomes, and inner resilience. This qualitative improvement highlights the importance of emotion-aware retrieval for meaningful personalization.

E. Discussion on Limitations

While the proposed framework demonstrates improved emotional alignment, it currently evaluates emotional interactions on a per-session basis and does not model long-term emotional trends. Additionally, the fusion weights are empirically chosen rather than learned dynamically. These limitations suggest opportunities for future work, including adaptive fusion strategies and longitudinal emotional modeling.

VII. Conclusion

The proposed system introduces an integrated, emotion-aware guidance framework that combines multimodal emotion recognition with retrieval-augmented generation (RAG) to offer personalized counseling based on the Bhagavad Gita. User speech is processed through two parallel pipelines: a Wav2Vec2-based deep learning model is used to extract emotional cues from vocal signals, while Whisper converts speech into text, which is then analyzed by a BiLSTM-based text emotion classifier. The emotional outputs from both modalities are combined using a weighted fusion strategy to produce a stable and reliable final emotional state, reducing the ambiguity that often arises from single-modality analysis.

This fused emotional representation, together with the transcribed situational context, is embedded into a shared semantic space and used to query a Qdrant vector database that stores Bhagavad Gita verses enriched with semantic embeddings and emotion-related metadata. The retrieval process prioritizes both contextual relevance and emotional alignment, moving beyond simple keyword-based matching. A large language model then generates personalized explanations that link the user's emotional state and lived experience with the philosophical essence of the retrieved verses. By closely integrating emotion detection, knowledge retrieval, and explanation generation, the system transforms static spiritual texts into adaptive, emotionally grounded guidance.

References

[1] Akinpelu, Adetola, et al. (2023). Real-Time Speech Emotion Recognition Using Deep Learning and Data Augmentation. Applied Sciences, 13(5), 3138. https://doi.org/10.3390/app13053138

[2] Kang, Yoonjoo, et al. (2023). CNN-LSTM and Attention Mechanism for Speech Emotion Recognition. Information, 14(2), 82. https://doi.org/10.3390/info14020082

[3] Sharma, Deepak, et al. (2023). DistilHuBERT vs. PaSST for Lightweight Speech Emotion Recognition. Applied Sciences, 13(2), 1092. https://doi.org/10.3390/app13021092

[4] Wu, Xiaoqing, et al. (2024). M-fusHER: Multimodal Emotion Recognition Using Audio, Text, and Video. Journal of Big Data, 12(1). https://journalofbigdata.springeropen.com/articles/10.1186/s40537-025-0460-2

[5] Plaza-del-Arcos, Faustino, et al. (2023). CM-RoBERTa: Cross-Modal RoBERTa for Emotion Recognition from Speech and Text. Electronics, 12(5), 1143. https://doi.org/10.3390/electronics12051143

[6] Yadav, Harshit, et al. (2024). MemoCMT: Cross-Modal Transformer Fusion for Emotion Recognition. Scientific Reports, 14. https://doi.org/10.1038/s41598-025-89202-x

[7] Moudarres, Soudabeh, et al. (2023). Transformer with Psycholinguistic Features for Text Emotion Recognition. Sensors, 23(4), 2077. https://doi.org/10.3390/s23042077

[8] Jain, R., et al. (2023). Emotion Analysis Survey: Models, Datasets, and Applications. arXiv preprint. https://arxiv.org/abs/2306.12271

[9] Mandikal, J. P., & Rajan, K. S. (2023). Ancient Wisdom, Modern RAG: Vedantic QA Using Retrieval-Augmented Generation. arXiv preprint. https://arxiv.org/abs/2307.06615

[10] Kapuriya, Raj, & Bhatt, Riddhi. (2023). Spiritual-LLM: Gita-Guided Counseling with Generative AI. arXiv preprint. https://arxiv.org/abs/2311.11892

[11] Prakash, Sumit, et al. (2023). Ancient Indian Scriptures + RAG: Knowledge Retrieval from the Bhagavad Gita. International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIT). https://www.ijariit.com/manuscript/271689

[12] Sharma, Deepak, et al. (2023). NirvanaNavigator: Emotion-Aware Chatbot Using Gita and RAG. IRJMETS, 6(8). https://www.irjmets.com/Volume6Issue8/PID708.pdf

[13] Elbakoury, Mohamed, et al. (2023). AnchorBERT: Learning Emotion Anchors for Emotion Recognition. SSRN. https://www.ssrn.com/abstract=4396586

[14] Akinpelu, Adetola, et al. (2023). Patch-ViT: Vision Transformer for Speech Emotion Recognition. Electronics, 12(16), 3438. https://doi.org/10.3390/electronics12163438

[15] Lee, Seok-Woo, et al. (2023). ProEmo: Prompt-Driven Emotional Text-to-Speech Synthesis. arXiv preprint. https://arxiv.org/abs/2304.00426

[16] Liu, Feng, et al. (2023). Semi-Supervised Emotional TTS with Implicit Representations. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31. https://doi.org/10.1109/TASLP.2023.3287529

[17] Kang, Donghyun, et al. (2023). EmoVoice: Emotionally Controllable TTS Using Large Language Models. arXiv preprint. https://arxiv.org/abs/2307.08589

[18] Jang, Joohyun, et al. (2023). ZET-Speech: Zero-Shot Emotional TTS with Domain-Adversarial Training and Diffusion. IEEE Access, 11. https://doi.org/10.1109/ACCESS.2023.3274509

[19] Hu, Zhiyao, et al. (2023). EASPO: Emotion-Aligned Speech Synthesis via Preference Optimization. arXiv preprint. https://arxiv.org/abs/2305.12107