International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3541-3544
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0460

# AI-Based Monitoring Of Mental Health and Work Performance

*Bharti. P. Ahuja [1] , Ashutosh Bhagat [2], Riya Dhawale [3] , Rohan Ghute [4] , Sagar Pati [5]*
*[1]Associate professor, Dept. of CSE, Guru Gobind Singh College of Engineering and Research Centre, Maharashtra, India.*
*[2,3,4,5] UG, Dept. of CO, Guru Gobind Singh College of Engineering and Research Centre, Maharashtra, India.*
***Email ID:*** *bharti.ahuja@ggsf.edu.in[1], ashbhagat123d@gmail.com[2], riyadhawale75@gmail.com[3], rohanghute3379@gmail.com [4], sagarlpatil23@gmai.com[5]*

## Abstract
*Emotion aware intelligent systems are increasingly using techniques based on spoken interaction data, but two major issues arise for practical deployments: Inter speaker interference in shared conversations and a lack of user specific adaptation. In this work, we introduce a production level multi-tenant backend architecture for audio emotion recognition with retrieval-augmented generation (RAG) and extend it with an incremental speaker aware module that isolates the owner's voice to personalize predictions, allowing us to preserve the contextual entirety of conversation when retrieving similar tasks. The baseline pipeline consists of audio validation, preprocessing, Wav2Vec2 based embedding extraction, dual-head emotion inference (its global head plus user head), Whisper transcription where data is stored in MongoDB and Qdrant. And you apply a feedback loop and adaptive alpha blending to control personalization at a baseline the model starts out with global performance but goes user specific as corrected feedback is provided. For mixed-speaker sessions, we propose an Incremental V1 module, including VAD segmentation, segment-level speaker embeddings, clustering, owner verification and the construction of the owner only emotion path. This new enrolment structure exposes new endpoints and additive response metadata without breaking existing API contracts. The project presents an accuracy of 72.69% emotion (216 test samples) with practical low latency inference behavior, and places significant validation on speaker aware API and safety improvements based on extensive focused privacy tests covering 20 passing tests discrete to the set target environment. The resulting design axes deployability, personalization reliability and backward compatibility for real-world conversational AI systems.*

***Keywords:*** *Audio Emotion Recognition, Wav2Vec2, Dual Head Personalization, Speaker Verification, Diarization, RAG, FastAPI, MongoDB, Qdrant, Whisper.*

## 1. Introduction
Human-centered intelligent systems are increasingly required to be capable of understanding what users say and how they say it. Emotion recognition from speech is thus an integral component in the future of digital wellness, journaling, productivity assistants, and spirit-memento conversational memory apps. However, all existing applied systems degrade in the presence of multi-speaker conditions (overlap or interruptions within audio). In such cases, the direct end-to-end emotion inference system over mixed audio tends to misattribute affective signals and contaminates user-specific personalization. This paper discusses a full-stack system with an HTTP api for audio processing on the back-end and a web interface in which users can interact directly. The backend combines authentication, audio upload, emotion analysis (performed by neural embeddings), transcription, vector indexing and natural-language retrival of past conversations. Our system was implemented as a dual-head personalization approach comprising of a global emotion head that delivers strong baseline predictions and a user head that can adjust to corrected user feedback over time. This structure does enhance the quality of customization, even it still believes that any uploaded audio is only from the target user. The main technical challenge we address in this paper is the design of a practical production backend that can facilitate conversation-level context retention for retrieval, while also ensuring that emotion estimation and personalization

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3541-3544
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0460

attach only to the owner's voice if multi-speaker audio exists We answer with an extension of speaker-aware optimization that is compatible with existing APIs and which fits within CPU constraints. The Extension consists of owner enrollment, segment-level speaker analysis, owner-only emotion routing and personalization safety controls. The key contributions are threefold. This article introduces integrated architecture for multi-tenant service including emotion recognition, transcription, retrieval and adaptive personalization. Second, we present an Incremental V1 speaker-aware pipeline which separates the owner-sensitive emotion paths from the entire conversation retrieval paths. Third, we present implementation-level evidence from test outcomes and project indicators that show feasibility of the method, backward compatibility with pre-basic governance systems, as well signs of preparedness for incremental improvement.

### 1.1.System Overview

Speech signals are analysed here to identify emotional states and derive insights for mental well-being and performance at work, which is where this proposed system comes into play. This framework performs a pre-operation on the raw audio data in one go, then extracts relevant features and transits them through trained networks. We first preprocess the speech signal to eliminate noise, and sort out audio quality. The output signal is then framed to extract acoustic features from it. Then, useful acoustic features are extracted and integrated with pretrained speech embeddings to derive an informative representation of the emotion in audio. The emotion classification module uses the fused feature representation to classify the speaker's emotions. Moreover, speech transcription extracts linguistic details from the audio material. Lastly, the ensemble emotional fingerprints and session data are persisted in a database for long-term behavioural analysis and monitoring.

### 1.2.Audio Preprocessing

- Validate file format and size
- Store in user-specific directory
- Resample to standard 16kHz
- Apply voice activity detection trimming

### 1.3.Speaker-Aware Processing (Core Contribution)

**Segmentation:** Audio split into speech segments using VAD

**Speaker Embedding:** Each segment converted into a speaker embedding

**Clustering:** Segments grouped by speaker similarity

**Owner Verification:** Compare clusters with enrolled owner profile

**Output**

- Owner segments identified
- Other speaker segments separated

### 1.4. Dual Path Processing

Path 1: Emotion Recognition (Owner Only)

- Combine only owner speech segments
- Extract Wav2Vec2 embeddings
- Pass through dual-head classifier: Global Head
- Pre-trained general model
- User Head
- Personalized model trained on feedback

**Adaptive Blending (Key Method)**

$$\alpha = \alpha_{data} \times \alpha_{conf}$$

Where:

- α_data → depends on feedback count
- α_conf → depends on global confidence

Final prediction:

- Weighted combination of global + user model

Path 2: RAG Retrieval (Full Conversation)

- Full audio is transcribed using Whisper
- Speaker tags added (Speaker 1, Speaker 2, Owner)
- Stored in Qdrant as embeddings
- Used for:
  - Question answering
  - Context retrieval

### 1.5. Personalization Mechanism

**Feedback Loop**

- User corrects wrong emotion predictions
- Data stored in MongoDB

**Training Policy**

- Training starts after 20 feedback samples
- Updates every 10 samples

**Safety Control**

- Training allowed only if:
  - Owner speech ratio ≥ 25%

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3541-3544
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0460

- Owner is verified

## 2. Results And Discussion

### 2.1.Results

The system has been successfully proven to work well for owner-centric personalization and conversation-complete retrieval. This step includes the separation of pipeline at speaker-analysis stage which preserves the richness of RAG and mitigate personalization drift due to non-owner speech. This becomes even more rewarding if you are in situations where there are multiple speakers in a conversation (e.g. group discussions, meetings etc.).

Incremental design—optimizing for practical deployment over max diarization accuracy This approach gives decent performance for a first production roll out, using lightweight segment embeddings and clustering methods. The system architecture also allows future incremental improvements in the form of plugging in more sophisticated diarization models without having to change the contracts for external APIs.

### 2.2.Discussion

From a software-engineering standpoint, the risk of such deployments is made trivial in deployment through backward-compatible schema extensions and feature-flag fallbacks. Existing clients using the service are not affected, while new clients can take advantage of speaker-based metadata that aids a richer end-user experience and better diagnostics. While this clustering method works in most cases, clustering purity can still be sub-optimal when speaker-overlap or background-noise is heavy. These limitations highlight future work possibilities of better diarization and noise-robust speaker embedding. Yet due to the modularity of the system, such improvements can be made with a lot of the underlying structure unabated and stagnancy at greatly higher scalability.

## Conclusion

This paper proposed a full-stack, multi-tenant audio intelligence backend that fuses emotion recognition with personalization and RAG retrieval With the baseline dual-head architecture backstage for solid cold-start performance and feedback-based training, Incremental V1 is aimed directly at a large real-time problem: contamination of an owner personalization due to mixed speaker. The outcome is a system that forces owner-only emotion paths and full-conversation retrieval paths into one upload workflow, creates enrollment & verification APIs as well as safety guards on training without breaking existing clients. Through knowledge gathered from visible project evidence (i.e, baseline model performance and extension tests), we can say that such a design is feasible for practical deployment.

## Acknowledgements

The authors express sincere gratitude to Dr. Bharti P. Ahuja, Associate Professor, Department of Computer Science and Engineering, Guru Gobind Singh College of Engineering and Research Centre, for her invaluable guidance, mentorship and continuous support throughout this research work. We also thank the Department of Computer Engineering for providing the necessary infrastructure and resources. Special appreciation to our peers for their encouragement during the development of this AI-based mental health and work performance monitoring system.

## References

[1]. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460. DOI: 10.48550/arXiv.2006.11477

[2]. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. (2015). A review of depression and suicide risk assessment using speech analysis. Speech Communication, 71, 10–49 DOI: 10.1016/j.specom.2015.03.00.

[3]. Zhang, Z., Zhang, Z., & Deng, J. (2018). Speech emotion recognition using deep learning. IEEE Signal Processing Letters, 25(10), 1510–1514. DOI: 10.1109/LSP.2018.2867159

[4]. Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing, 20(2), 356–370. DOI: 10.1109/TASL.2011.2125954

[5]. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3541-3544
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0460

Information Fusion, 37, 98–125. DOI : 10.1016/j.inffus.2017.02.003