IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15457

Voice Interview Analyzer

Mirza Daniyal Baig[1], Alfaiz Samani[2], Shehzan Shaikh[3], Anas Nakade[4], Alfiya Mulla[5], Zeeshan Khan[6]

CSE (Data Science), Anjuman-I-Islam's Kalsekar Technical Campus, Panvel , India[1,2,3,4]

Assistant Professor, CSE (Data Science), Anjuman-I-Islam's Kalsekar Technical Campus, Panvel , India[5]

Head of Department, CSE (Data Science) , Anjuman-I-Islam's Kalsekar Technical Campus , Panvel , India[6]

Abstract: Our Voice Interview Analyzer, a project built using Python, is designed to make the initial hiring process smarter and more efficient. Think of it as an AI assistant for recruiters; it listens to a candidate's spoken answers to standard interview questions, quickly converts their speech to text, and then dives deep into the content. Using natural language processing, it analyzes not just what the candidate says, but how they say it gauging their confidence through their tone, the richness of their vocabulary, there emotions when they talk, and how relevant their answers are. This system generates an objective, data-driven report that helps hiring managers save a lot of time while ensuring a fair and consistent screening process that can uncover promising candidates who could otherwise be overlooked.

Keywords: Speaker diarization, speaker recognition, speech separation, Emotion detection, speech clarity, Confidence Score, overall confidence, actionable feedback.

I. INTRODUCTION

The increasing volume of job applicants has made initial screening more time-intensive and inconsistent. Traditional resume-based filtering fails to capture a candidate's communication ability, confidence, and behavioral traits soft skills that are among the strongest predictors of job performance. With the rise of remote hiring, evaluating these qualities in a standardized and scalable manner has become a pressing need.

Advances in Automatic Speech Recognition, Natural Language Processing, and speaker analysis have made it feasible to extract meaningful evaluation signals directly from spoken responses. Pre-trained models such as WhisperX for transcription, Wav2Vec 2.0 for acoustic feature extraction, and BERT-based transformers for emotion classification now enable robust voice analysis pipelines at scale.

This paper presents the Voice Interview Analyzer, an end-to-end AI system that evaluates candidates based on their spoken interview responses. The system transcribes audio, separates candidate speech from interviewer questions, and scores responses across multiple dimensions confidence, emotion, answer relevance, vocabulary, and speech clarity generating a structured report with a final Hire or Do Not Hire recommendation.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15457

II. LITERATURE SURVEY

Reference	Year	Focus Area	Key Techniques & Findings
AI-based Behavioural Analyser for Interviews/Viva	2021	Emotion & Confidence Analysis	Uses AI to analyze facial expressions and speech patterns to assess candidate behavior during interviews.
An AI-Driven Approach to Enhance Interview Performance through Voice and Response Analysis	2025	Voice & Response Analysis	Integrates voice analysis with response evaluation to enhance interview performance.
Feasibility of an automated interview grounded in multiple miniinterview methodology	2022	Interview Methodology	Develops an automated interview system based on multiple mini-interview (MMI) methodology.
Utility of artificial intelligence-based conversation voice analysis for cognitive decline detection	2025	Cognitive Decline Detection	Utilizes AI-based voice analysis to detect signs of cognitive decline through conversational voice samples
Evaluating Speech-to-Text x LLM x Text-to-Speech Combinations for AI Interview Systems	2025	AI Interview System Evaluation	Analyzes combinations of speech-to-text, large language models, and text-to-speech for AI interview systems.
Mic Drop or Data Flop? Evaluating the Fitness for Purpose of AI Voice Interviewers	2025	AI Interviewer Evaluation	Evaluates the effectiveness of AI voice interviewers in quantitative and qualitative research contexts.

TABLE I : Survey of Existing System

Conclusion of Literature Survey:

The reviewed studies show that AI techniques are widely used for interview analysis, focusing on areas such as emotion detection, voice processing, and structured evaluation. However, most approaches address only specific aspects and lack a comprehensive, real-world applicable solution. This highlights the need for an integrated system that combines multiple analysis techniques for effective candidate evaluation.

III. PROPOSED SYSTEM

A. PROBLEM STATEMENT:

Identifying promising candidates quickly and fairly based on clarity , said evaluating confidence and emotional tone

B. PROPOSED METHODOLOGY

1) Audio Acquisition & Preprocessing:
The candidate's voice is captured and converted into a standard .wav format with a sampling rate of 16 kHz to ensure consistency in processing. The audio signal is then processed using techniques such as Fast Fourier Transform (FFT) and Mel filter banks to enhance and normalize the signal. Additionally, noise reduction and silence trimming are performed using spectral subtraction and voice activity detection (VAD) to improve the overall quality of the audio input before further analysis

$$.X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}, \quad k = 0,1,2,\dots,N-1$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15457

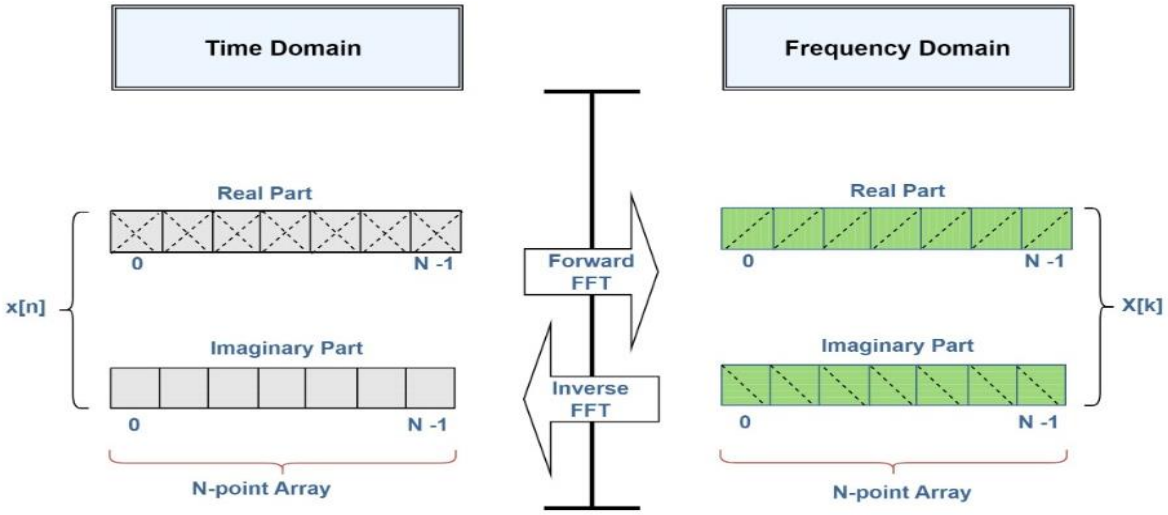

Fig 1. Fast Fourier Transform (FFT)

2) Speech-to-Text (Transcription):
The transcription process utilizes WhisperX, an advanced extension of the Whisper model, designed for robust automatic speech recognition. It incorporates a Conformer-based encoder–decoder architecture to effectively process audio signals and generate accurate transcriptions. The model employs Connectionist Temporal Classification (CTC) loss along with cross-attention mechanisms to achieve precise alignment between audio input and textual output. Additionally, it leverages activation functions such as ReLU and GELU, along with multi-head self-attention layers, to enhance contextual understanding and improve overall transcription performance.

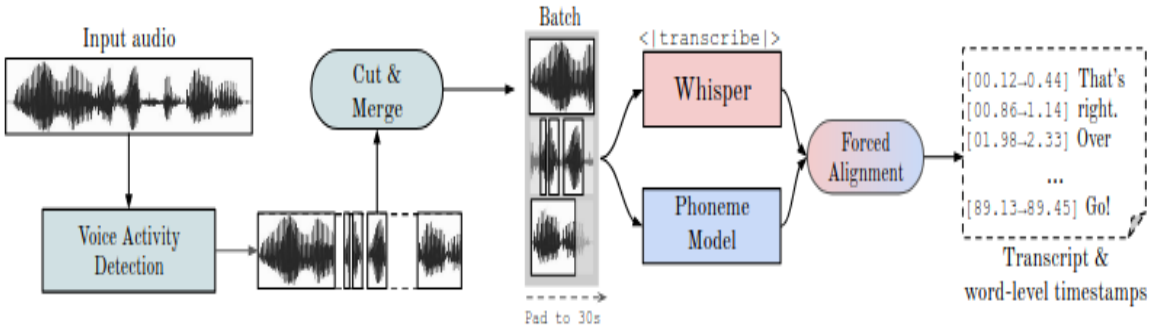

Fig. 2: Diariazation

3) Feature Extraction:
Wav2Vec2.0 is utilized to extract contextualized speech embeddings directly from raw audio signals. The model is trained using a contrastive self-supervised learning approach, enabling it to differentiate between true latent representations and negative samples. It combines temporal convolutional filters, such as one-dimensional CNNs, with Transformer-based architectures to effectively capture both short-term and long-range dependencies in speech data.

$$L_{\text{wav2vec2}} = -\log \frac{\exp\left(\frac{\text{sim}(c_t, q_t^+)}{\kappa}\right)}{\sum_{q \in Q} \exp\left(\frac{\text{sim}(c_t, q)}{\kappa}\right)}$$

4) Speaker Identification:
Pyannote.audio 3.1 is utilized for speaker diarization and identification through a neural clustering-based pipeline. The system generates embeddings for voice segments using pretrained speaker encoders and applies agglomerative

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15457

hierarchical clustering to distinguish between different speakers. This approach enables accurate identification and separation of individuals, ensuring reliable attribution of responses in multi-speaker interview scenarios

$$y_t = \sigma(W x_t + b)$$

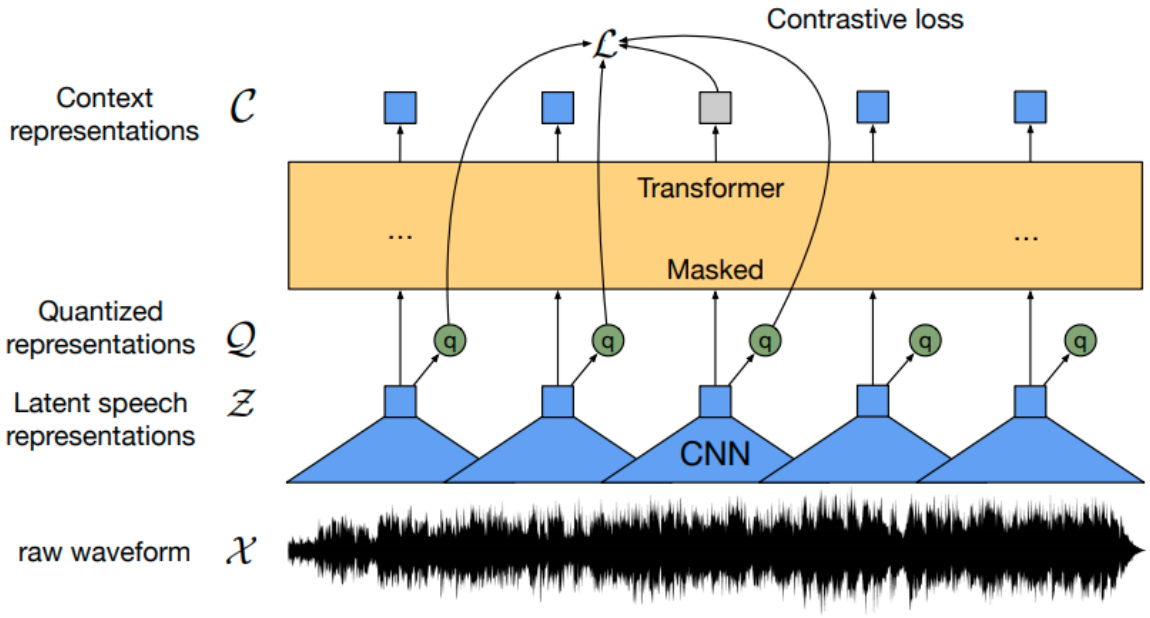

Fig. 3: Wave2vec2

5) Text Understanding & Analysis:
The transcribed and refined text is processed using MythoMax (13B), a large language model fine-tuned for tasks such as semantic understanding, sentiment analysis, and emotional tone detection. The model utilizes causal attention mechanisms to preserve contextual continuity and maintain coherence across sentences. Additionally, it employs activation functions like Swish (SiLU), which contribute to improved learning efficiency and enhanced performance in text-based inference tasks.

6) Sentiment, Confidence, and Clarity Evaluation:
The system performs sentiment classification using MythoMax, where responses are categorized into multiple classes such as confident, nervous, or neutral through attention-based feature analysis. Confidence scoring is derived from vocal characteristics, including intensity, pitch variation, and fluency, utilizing embeddings generated by Wav2Vec2. Additionally, clarity and relevance are evaluated by comparing the transcribed responses with job-specific keywords using techniques such as cosine similarity and TF-IDF weighting.

$$P(y_t \mid y_{<t}, x) = \text{softmax}(W_o h_t)$$

$$\text{cosine similarity}(A, B) = \frac{A \cdot B}{|A||B|}$$

IV. SYSTEM DESIGN

The Voice Interview Analyzer is designed as a modular pipeline that processes candidate responses from audio input to final evaluation. The system begins with capturing real-time or uploaded audio files in standard format, which are then segmented into smaller chunks using voice activity detection or frame-based methods. The audio undergoes preprocessing to remove noise and improve quality before being passed to the speech-to-text module, where it is transcribed into text with timestamps and confidence scores. Speaker diarization is performed to distinguish between different speakers, followed by role classification to label segments as either applicant or interviewer. The transcribed content is further refined through grammar and punctuation correction. The system then extracts both acoustic features, such as pitch and tone, and textual features related to semantic meaning. These features are analyzed to evaluate parameters including confidence, emotional tone, clarity, and response relevance, while also identifying filler words and generating feedback. Additionally, interviewer responses are classified based on question type. The final outputs are

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15457

aggregated into structured dashboards and downloadable reports. The backend architecture supports asynchronous processing using task queues, with separate services for speech recognition, diarization, and language analysis, and is deployed using scalable infrastructure.

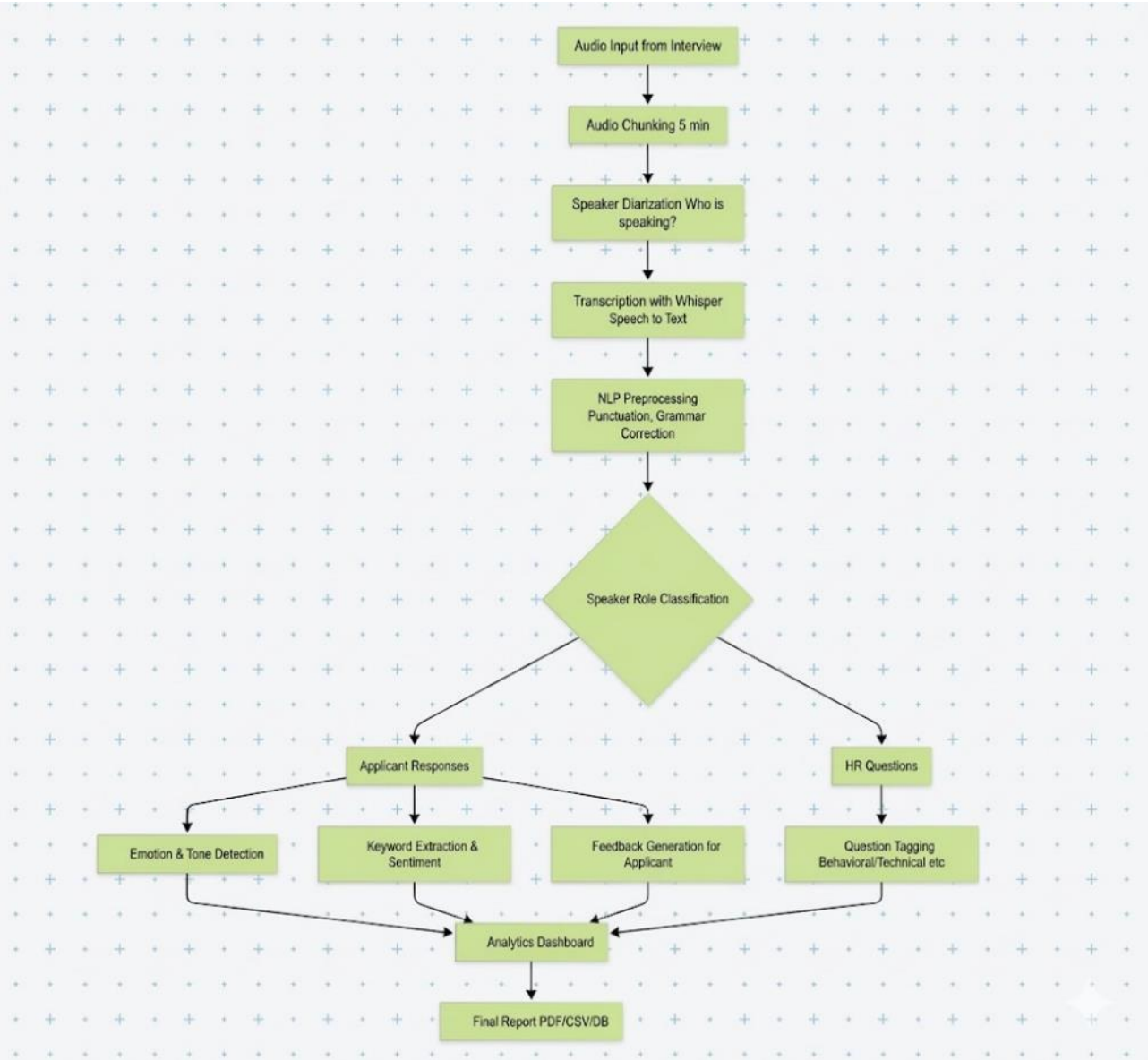

Fig. 4: System Design

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15457

V.RESULTS

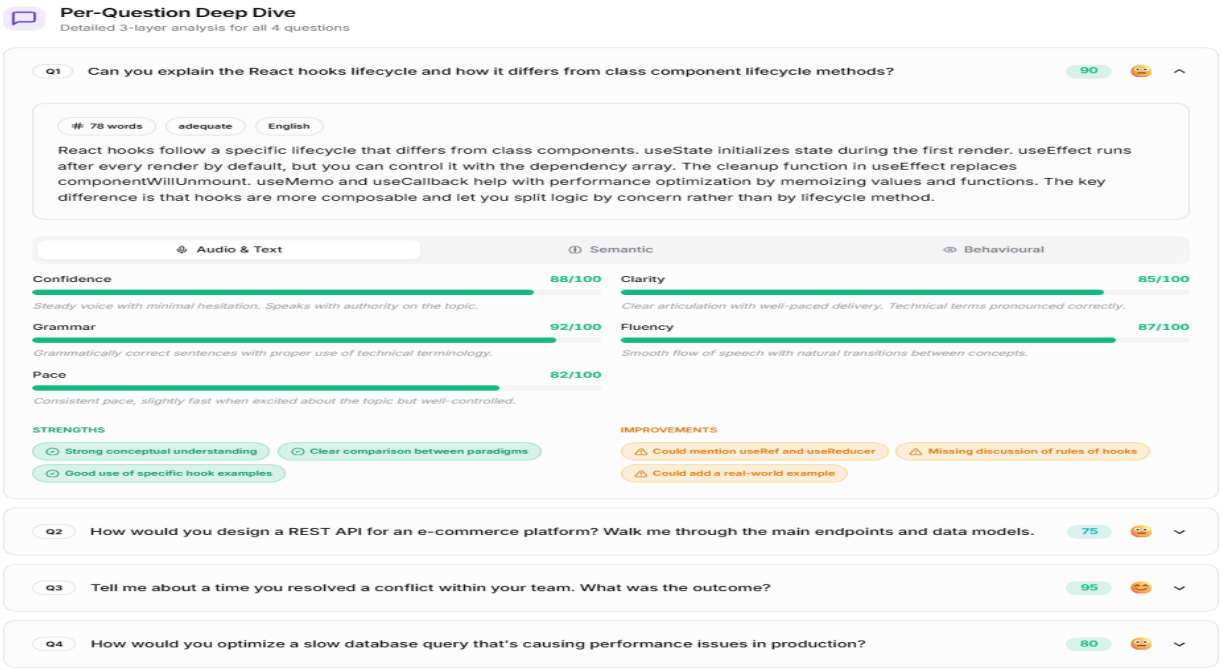

Figure 4.1: Per Question analysis

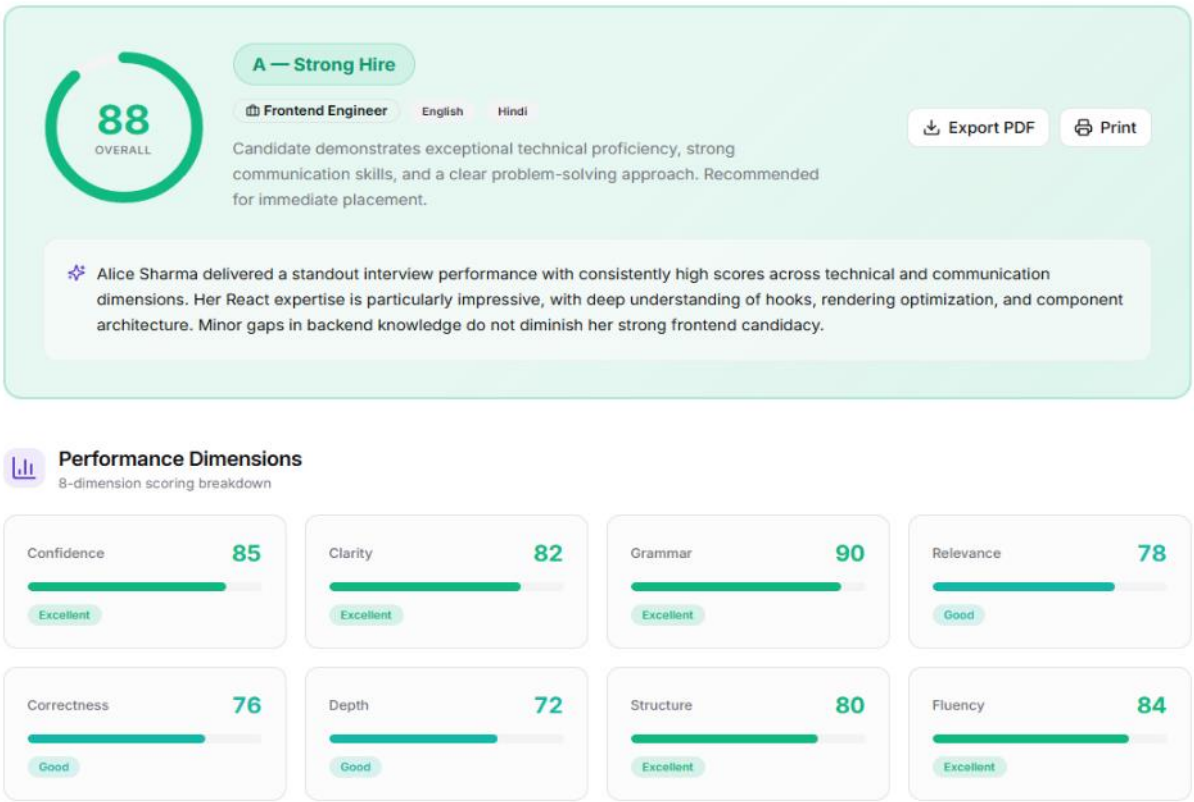

Figure 4.2: Overall Feedback

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15457

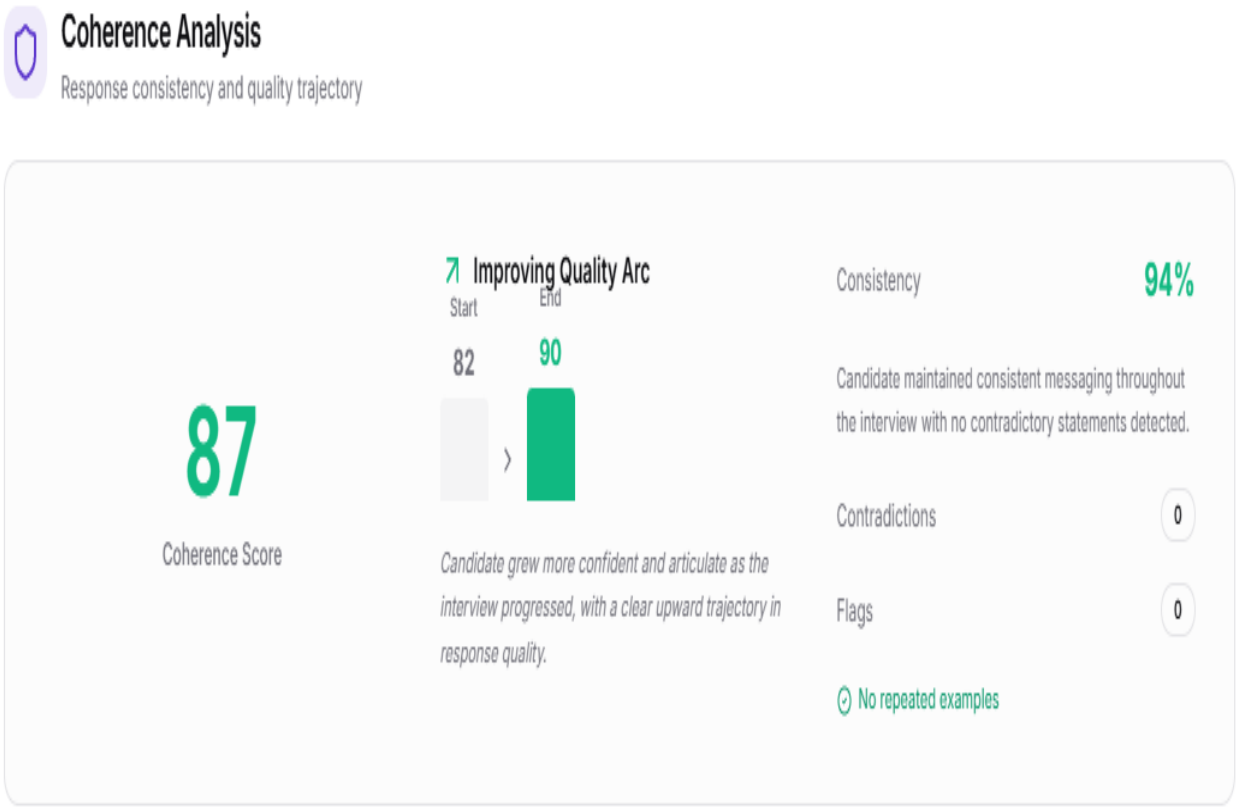

Figure 4.3: Detection of contradiction

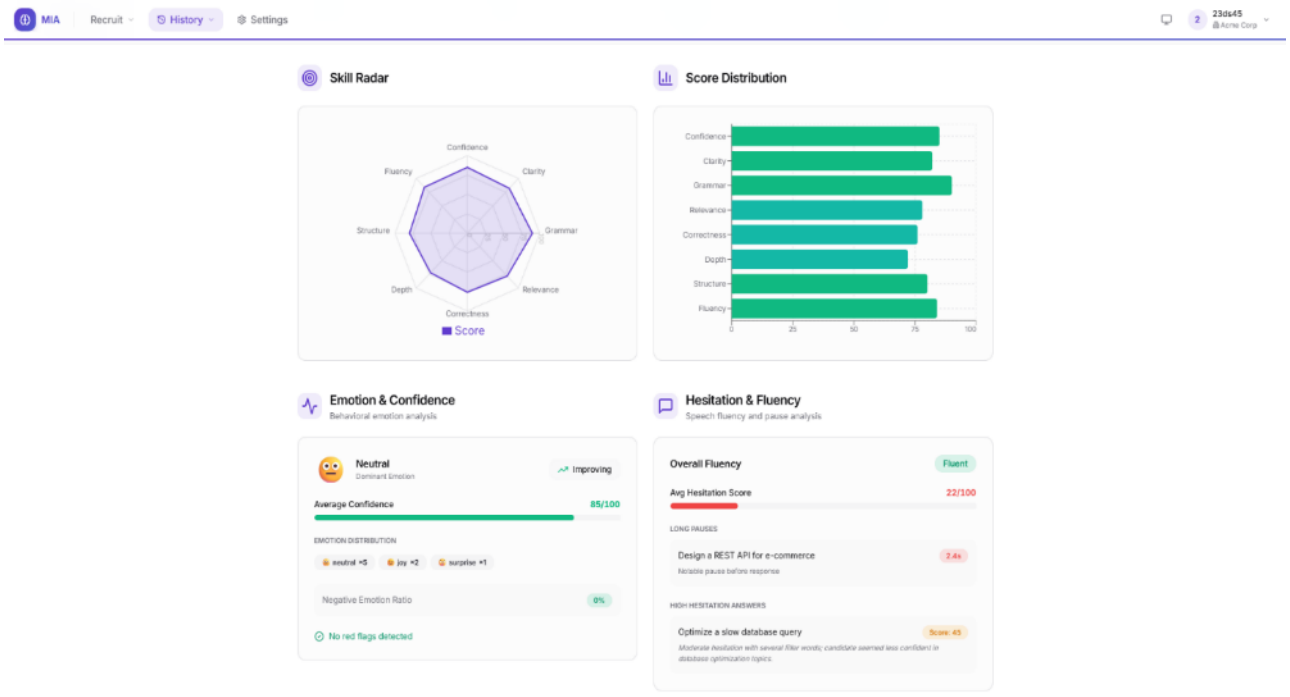

Figure 4.4: Graph explaining capability of candidate

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15457

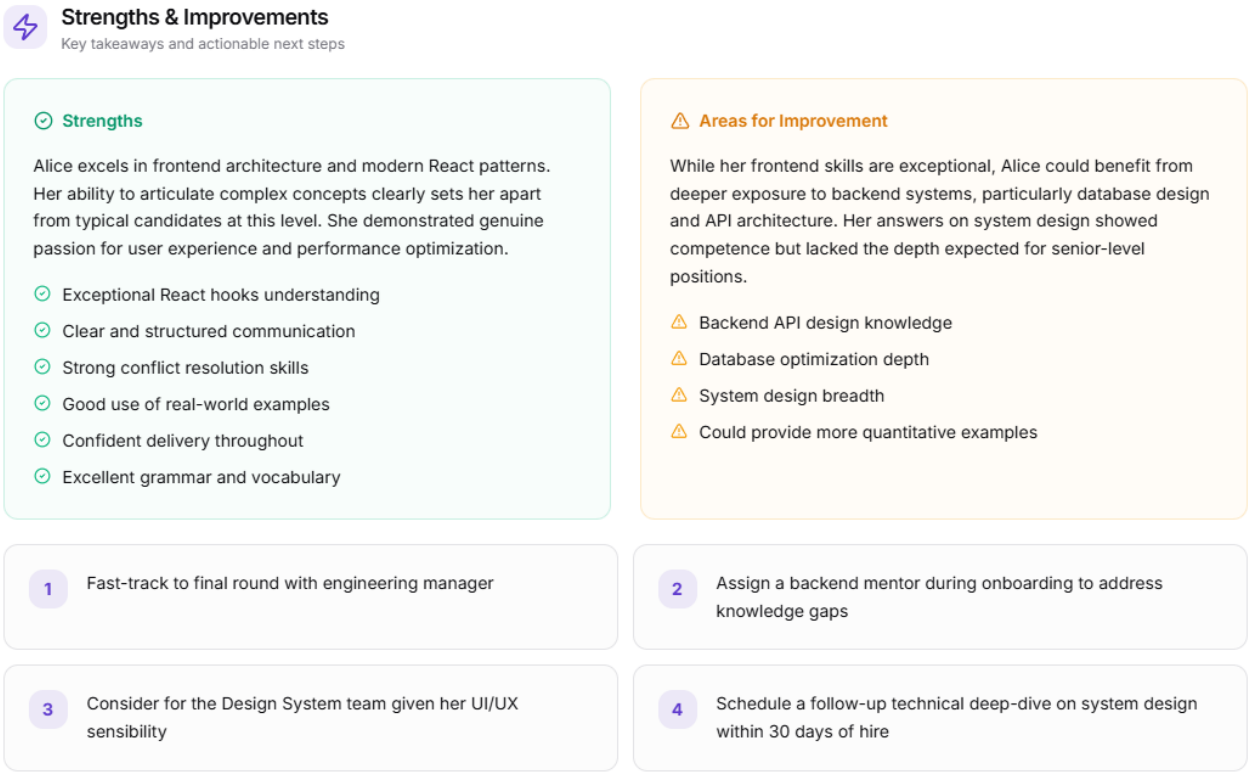

Figure 4.5: Strengths and Weakness

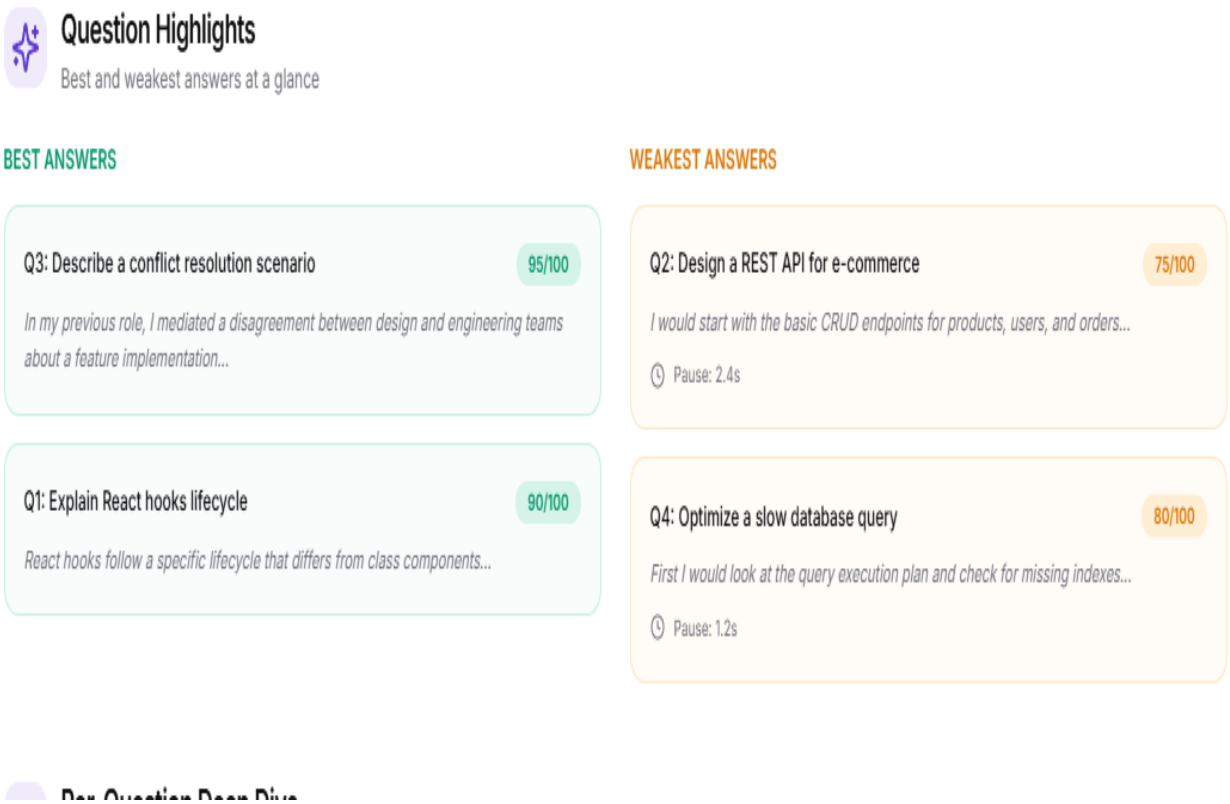

Figure 4.6: Overall question Analysis

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.15457

IV. CONCLUSION

The Voice Interview Analyzer is an AI-driven system that processes interview recordings to produce accurate transcripts, identify speakers, and classify their roles. It evaluates applicant responses for emotion, confidence, and content relevance, detects filler words, and generates actionable feedback. By providing aggregated analytics, visual dashboards, and downloadable reports, the system streamlines the interview evaluation process, ensuring efficiency, consistency, and objective insights for HR decision-making.

REFERENCES

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Advances in Neural Information Processing Systems, 2020.

[2] G. Drakopoulos et al., "Speaker Diarization: A Review of Objectives and Methods," Applied Sciences, vol. 15, no. 4, 2025. [Online]. Available:
https://www.mdpi.com/2076-3417/15/4/2002

[3] K. Kinoshita et al., "Microsoft Speaker Diarization System for the VoxCeleb Challenge 2020," arXiv, 2020. [Online]. Available:
https://arxiv.org/abs/2010.11458

[4] X. Y. Fu et al., "Improving Punctuation Restoration for Speech Transcripts," WNUT / ACL, 2021. [Online]. Available:
https://aclanthology.org/2021.wnut-1.19.pdf

[5] T. Alam et al., "Punctuation Restoration Using Transformer Models," WNUT / ACL, 2020. [Online]. Available:
https://aclanthology.org/2020.wnut-1.18/

[6] G. Drakopoulos, K. Dimitropoulos, and I. Pitas, "Emotion Recognition from Speech: A Survey," Science of Computer Programming, 2019. [Online]. Available:
https://www.scitepress.org/Papers/2019/84950/84950.pdf

[7] P. Jafarzadeh et al., "Speaker Emotion Recognition: Leveraging Self-Supervised Models," arXiv, 2024. [Online]. Available:
https://arxiv.org/abs/2411.02964

[8] D. Snyder et al., "Speaker Recognition for Multi-Speaker Conversations," arXiv, 2021. [Online]. Available:
https://arxiv.org/abs/2110.00560

[9] H. Hamza et al., "Speaker Diarization and Emotion Identification from Speech," arXiv, 2023.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," ICASSP, 2018. [Online]. Available:
https://www.danielpovey.com/files/2018icasspxvectors.pdf