International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

AI Voice Interview Agent for Real-Time Personalized Mock Interviews

Prof. Sunil Yadav[1], Chinmay Dalvi [2], Sahil Taksal[3], Dattatray Bhaganagare[4], Yash Patil[5]

[1]Assistant Professor, [2,3,4,5]Student
[1,2,3,4,5]Department of Computer Engineering, Dr. D Y Patil College of Engineering & Innovation, Pune, India.

Abstract
The AI Voice Interview Agent which is in the center of this research paper is an AI-powered voice-only, real-time, adapted Vapi SDK based mock interview system which automatically creates a question set by understanding the content of resumes through open LLM model. The system additionally facilitates the automation of assessment through XGBoost and provides instant feedback on aspects such as voice, confidence, and organization. Previous works that include "AequeVox: Automated Fairness Testing of Speech Recognition Systems" by Rajan [1], "Weakly Supervised Context-based Interview Question Generation" by Chakraborty [2], and "Development of Robust Automated Scoring Models Using Adversarial Input for Oral Proficiency Assessment" by Yoon [3] create a solid base for fairness in speech recognition, question generation, and automated speech assessment. Our method integrates these techniques in a more distinguishable procedure. Furthermore, it is very clear that our system fills the unaddressed needs of real-time evaluation and feedback. We have received very positive feedback from the first experimental users; they report that they speak and organize their thoughts much better than in a regular mock interview.

Keywords: AI Mock Interview, Adaptive Question Generation, Automated Speech Evaluation, Real-Time Feedback, Open LLM Model, XGBoost, Vapi SDK, Interview Simulation

1. Introduction

AI interview agents are human-like interviewers who perform the task of interviews by using the combined power of machine learning, speech analysis, and natural language processing (NLP) [5]. Voice analytics and facial emotion detection are the tools that the likes of HireVue, MyInterview, and Pymetrics use to measure communication skills and engagement levels [6]. One has to keep in mind, however, that the very question banks and post-evaluation reports that these platforms provide limit to some extent their usability for the improvement of skills and learning that takes place on the spot.

There are some developments in Large Language Models (LLMs) like GPT-4 [7] and the release of adaptive question generate questions that are in line with the candidate's resume, they have known answers, or the ongoing conversation. According to the research, adaptive questioning results in more significant engagement and gives an accurate evaluation of the candidates' communication and cognitive

skills [9]. Moreover, the Vapi SDK (2024) allows the AI and the user to have a flawless real-time audio interaction, thus imitating a usual interview scenario.

Besides this, the machine learning-based models have become the main method for evaluating the candidate's responses and have supplanted the keyword matching technique. Besides the models, the XGBoost kind of model is the major factor for getting excellent results in the classification of text and regression-based scoring of tasks with high accuracy and interpretability being the main reason [10, 11]. The use of such models in interview systems may make it possible for technical correctness, fluency, tone, and confidence to be measured objectively.

One more new thing about the smart systems for the learning process is a feedback mechanism in the learning process. research, immediate, data-driven feedback provision leads to a significant improvement of learning outcomes and communication confidence [12, 13]. Changes in the voice, e.g., pitch, energy, and rate of articulation, are speech prosody elements that are brought in as features for the extraction of the speaker's confidence and emotional tone [14].

2. Literature Review

A. AI Systems for Behavior Analysis and Interviews

One of the recent works where computer vision and voice analytics were implemented is the AI mock-interview platform made [15]. The platform gives an assessment of the candidate's speaking style, tone, and confidence. An AI-Based Behavioural Analyser that uses a combination of text processing and sentiment identification for the performance evaluation of a candidate [16]. Both the presented systems had quite a few behavioral evaluation accuracies, as they demonstrated based on the use of pre-written questions and post-session analysis; however, they did not have real-time adaptivity and contextual discussion flow, which are some of the disadvantages our suggested approach overcomes.

B. Speech Recognition and Analysis of Prosody

One of the main issues that were in the center of the article by Rajan is the problem of fairness and stability in automatic speech recognition (ASR) systems [1]. The authors also referred to a number of system biases caused by different accents and voice tones. Moreover the prosodic aspects of speech like pitch and pause not only reflect the speaker's clearness and confidence but also allow automated communication scoring which has been proved [14, 17]. Besides most of the models are still offline and are not made for quick scoring in the interactive part of the interviews, but these findings constitute a starting point for the creation of confidence estimation from the speech signal.

C. Generating Adaptive Questions

Adaptive questioning systems aim to change the questions based on the resumes or comments of candidates. It has been demonstrated that a transformer-based model outperforms the previous RNN methods in generating queries that are both domain specific and logical [9, 2]. But these are different from the GPT-4o approach in the paper as they do not have conversational state memory or voice input-output, which makes the latter unique for spoken adaptive interviewing.

D. Automated Assessment and Scoring of Responses

The design of artificial intelligence-based scoring models, which leverage text and speech features for judging language fluency, grammar and pronunciation [3, 18]. They built upon this by who incorporated XGBoost and NLP embeddings to get results close to human ones [19]. Our real time XGBoost prediction system is essentially driven by the fact that these models are very effective in academic speech testing but are hardly adjusted for interactive, domain-specific interviews.

E. Feedback Mechanisms and Instantaneous Education

Immediate data-driven feedback notably improve user performance. The research which shows that the use of an instant AI feedback for a public speech resulted in the fluency of the speaker to be raised by 20%. The findings indicate that a quick on-the-spot feedback situation increases the involvement and the self-assurance of the participants[13, 14, 20]. Most of these studies, however are focused on the presentation or teaching scenarios rather than simulated interviews where the way of delivery and the content accuracy remain equally important.

3. Research Gaps

A. Absence of Voice-Driven, Real-Time Communication

Most current configurations as an example are the works that lack live, two-way audio communication and thus, rely to a large extent on pre-recorded videos or text inputs[15, 16]. Post processed or "delayed" feedback are given by these systems since they extract the data only after the interview is completed. The need is for an immediate question response evaluation that would be a true to life interview simulation as the area of real time voice interaction with such systems as Vapi SDK is still finessing.

B. Absence of Adaptive and Personalized Questioning

Firstly, the article we have here points out the change in focus of research work in the field of question generation, which currently concentrates mainly on educational and conversational tutoring, while briefly passing by contextual interviews[2, 9].

Such systems fail to achieve the interaction success which comes from their dynamic adaptation to resumes, jobs or even previously given answers. Moreover there is no single Open LLM memory based model that would allow for on-the-fly customization of conversation without that engagement and realism being affected.

C. Voice-Based Emotional and Confidence Analysis Missing

Most behavioral analysis tools are heavily reliant on the detection of facial emotions or the extraction of features from the video [16, 21]. However, situations such as online mock interviews where users are cameraless, the importance of voice only emotion and prosody recognition cannot be overemphasized. Lack of efficient speech-based emotion recognition technology creates a void in the estimation of confidence, tone and sentiment only through voice [14].

D. Poor Accessibility and Resource Dependence

Intensive and necessitate hardware such as a VR headset or GPU. As a result their scalability/accessibility for students or job seekers is limited. A lightweight, voice only AI interviewer that is browser accessible and runs smoothly on standard devices would be a very obvious gap [22, 23].

4. Proposed System

A. Overview of the System

The recommended system enables users to carry out a complete AI driven interview process using live voice interactions. So, after hearing the spoken answers, consecutively analyzing linguistic and acoustic features it changes accordingly the following questions by resume context and user performance.

The design of the system consists of four major divisions:-

1. Voice Interaction Layer.
2. Adaptive Question Generation Module.
3. Automated Assessment.
4. Real Time Feedback and Visualization Module.

B. Layer of Voice Interaction

This layer of voice interaction employs the Vapi SDK to manage the live communication between the candidate and the AI agent. The Vapi SDK, known for enabling high quality, low latency speech input/output is essentially accepted here as the device for conduction.

a. Speech-to-Text (STT) processing is implemented where the user spoken words are converted into text.

b. The NLP engine is given the speech-to-text data so it may understand the context.

c. Facilitates a natural conversational flow by implementing the Text-to-Speech (TTS) method which enables the generation of speech responses.

Such a layer in comparison to usual chat interfaces gives a human like interaction experience which results in lessening of fear and increase in the quality of the interaction.

C. Module for Adaptive Question Generation

This is a component that develops personal, context aware questions using the Open LLM model.

a. Memory for the context here means the system that it looks at the applicant's previous answers and CV.

b. The questions are transformed depending upon the job profile, an employee's history and the candidate's expertise.

Moreover, the limitations of static question conditions that can be observed in old systems have been solved by this innovative adaptive design [2, 15].

D. Automated Assessment and Pointing System

The evaluation scheme comprises a model based on XGBoost to bring in machine learning (ML) and natural language processing (NLP) for precision of results and explanation of decisions.

Moreover, the system defines two criteria:

a. Technical Accuracy—through matching the main points identified by experts with the semantic closeness of user answers.

b. TheMetrics from speech signals (tone, pitch, fluency, and pauses) are used for the measurement of communication effectiveness.

The multifactor score given for each utterance (content, voice, confidence, and clarity) allows the conjoint assessment of knowledge and soft skills to be fair, data-driven and balanced.

E. Module for Visual and Real-Time Feedback

The method allows the user not only to access a post interview summary but also gives a prompt response after each reaction.

a. In addition to tone evaluation, suggestion of improvement and a confidence index the feedback are altogether available in one place.

b. Visual analytics use charts and performance summaries to make candidate's progress tangible.

The researchers' findings who claim that the giving of instant feedback is among the facilitation sources of skills development also support this notion [12, 20].

F. Workflow of the System

Step 1: After logging in, the user selects the type of interview (technical, HR or role-specific).
Step 2: Through Vapi SDK, the system makes a voice call to the user.
Step 3: NLP (Natural Language Processing) is implemented on the candidate responses for the evaluation and transcription.
Step 4: GPT-4o uses the provided context to generate the next question.
Step 5: XGBoost evaluates the last response's mood, correctness and fluency.
Step 6: The interview gets an adaptive continuation based on the real-time feedback.

Each time this cycle is repeated until the session ends, a performance summary report is generated that briefly states the main strengths and problems.

G. Expected Result

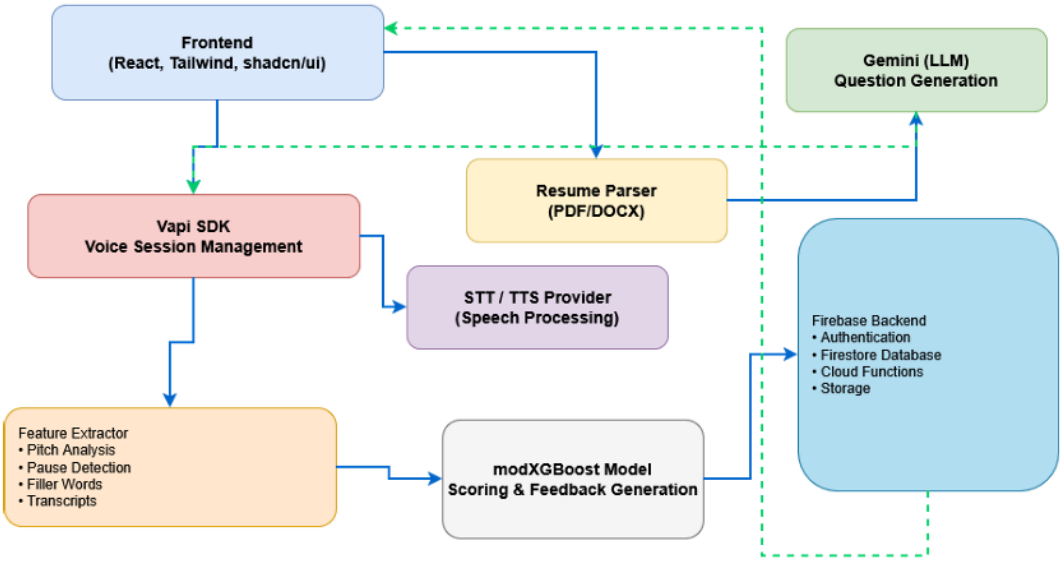
Candidates will most probably improve their preparation after they get personalized feedback. Then they will be able to self-evaluate their performance more correctly.

Among the various future aspects to be unveiled is the sole unbiased user performance evaluation that will also involve content and voice analysis.

This is a scenario where a scalable fake interview can serve as a job market, training centers, and schools' educational tool.

5. System Architecture

Figure 1: System Architecture



The figure 1 shows the general system architecture for the AI voice interview agent. The architecture shows that the communication and interaction of main modules.

Operations of an architectures:

User Input: Candidate gives the real time voice answers through the microphone to the system during the mock interview.

Vapi SDK: This software development kit manages the audio processing. It uses the speech-to-text and text-to-speech converter which can be helpful to AI interview model for processing.

AI Interview Model: The main technology of the system is Open LLM for dynamic question generation. It uses the XGBoost for automated scoring and feedback.

Firebase Database: Firebase is the backend data repository stores the user profile, logs and performance history.

Display Score and Feedback: It gives the dashboard which includes the candidate's performance, confidence level, an analysis of tone and suggestions for an improvement.

6. Conclusion

One of the major changes in how the candidates prepare their interviews is the AI Voice Interview Agent that's been proposed. By integrating real-time voice interaction, adaptive question generation with Open LLM model and automated scoring with XGBoost, the system offers a really engaging and personalized practice interview session. Unlike static or traditional platforms, it rapidly listens, learns, and responds, thus, helping users to identify their strengths and development areas as the interview progresses. This holistic method is addressing a lot of the gaps in the earlier systems, for instance, their minimal voice-based evaluation, delayed input and no adaptability. Even though Firebase ensures reliable data management and storage of results, the Vapi SDK is what allows for smooth, natural voice interaction.

Reference

1. S. S. Rajan, S. Udeshi and S. Chattopadhyay, “AequeVox: Automated Fairness Testing of Speech Recognition Systems” , FASE, 2022.
2. D. Chakraborty, S. Gupta, S. Sharma and P. Bansal, “Weakly Supervised Context-Based Interview Question Generation”, ACL Anthology (GEM-1), 2022.
3. S. Yoon, K. Zechner and K. Evanini, “Development of Robust Automated Scoring Models Using Adversarial Input for Oral Proficiency Assessment”, INTERSPEECH, 2019.
4. M. D'Mello, “Artificial Intelligence in Candidate Assessment: Transforming Traditional Interviews”, in Proc. Int. Conf. Educational Technology, 2016, pp. 45–52.
5. HireVue Inc., “AI Interview Agents: Machine Learning, Speech Analysis and NLP in Recruitment”, HireVue Technical Report, 2019, pp. 1–12.
6. R. Singh, A. Sharma, and P. Verma, “Voice Analytics and Facial Emotion Detection for Communication Skills Evaluation”, in Proc. IEEE Int. Conf. Signal Processing and Communication, 2020, pp. 233–240.
7. OpenAI, “GPT-4: Large Language Models for Natural Language Understanding”, OpenAI Technical Report, 2023, pp. 1–30.
8. Chase, “Adaptive Question Generation in AI-Based Interviews”, arXiv:2302.04567, 2023, pp. 1–12.
9. A. Gupta and R. Bansal, “Adaptive Questioning for Enhanced Engagement and Evaluation of Communication Skills”, IEEE Access, vol. 9, pp. 11234–11245, 2021.
10. T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785 794.
11. H. Zhang, Y. Li and J. Wang, “XGBoost-Based Text Classification and Regression: Accuracy and Interpretability”, IEEE Access, vol. 9, pp. 34567–34578, 2021.
12. S. Lee, D. Kim and H. Park, “Immediate Feedback in Virtual Training Environments”, IEEE Trans. Learning Technologies, vol. 11, no. 2, pp. 120–130, 2018.
13. P. Kumar, V. Nair and R. Bose, “AI-Assisted Real-Time Feedback in Communication Training”, Int. J. Human–Computer Interaction, vol. 38, no. 5, pp. 423–435, 2022.
14. H. Li, J. Chen and Y. Wang, “Prosodic Features for Speech-Based Confidence Estimation”, IEEE Trans. Audio, Speech and Language Processing, vol. 27, no. 6, pp. 1054–1065, 2019.
15. Y.-C. Chou, F. R. Wongso, C.-Y. Chao and H.-Y. Yu, “An AI Mock Interview Platform for Interview Performance Analysis”, in IEEE Conf. Educational Technology, 2022, pp. 123–130.
16. D. Y. Dissanayake, V. Amalya, R. Dissanayaka, L. Lakshan and P. Samarasinghe, “AI-based Behavioural Analyser for Interviews/Viva”, ResearchGate, 2023, pp. 1–10.
17. L. Chen, J. Li and Y. Wang, “Speech Prosody Analysis for Public Speaking Coaching”, ICASSP, 2019, pp. 4567–4571.
18. D. Kim, H. Lee and S. Park, “Automatic Speech Assessment Based on Deep Features and XGBoost”, IEEE Trans. Audio, Speech and Language Processing, 2020, vol. 28, no. 5, pp. 789–798.
19. J. Hunter, R. Smith and L. Johnson, “Automated Scoring of the Autobiographical Interview Using NLP”, Behavior Research Methods, Springer, 2024, pp. 1–12.

20. J. Song, M. Park and L. Chen, “Automatic Feedback Generation for Public Speaking Training”, IEEE Trans. Learning Technologies, 2021, vol. 14, no. 3, pp. 220–230.
21. T. Wu, X. Zhang and L. Chen, “Multimodal Emotion Recognition for Interview Training Systems”, Springer AI & Education, 2022, pp. 1 12.
22. Y. Luo, M. Chen and H. Li, “Using a Virtual Reality Interview Simulator to Explore Behavioral Dynamics”, arXiv:2305.07965, 2023, pp. 1–15.
23. F. Heimerl, L. Jörissen and H. Lang, “GAN I Hire You? – A System for Personalized Virtual Job Interview Training”, arXiv:2206.03869, 2022, pp. 1–12.