



# IJRASET

International Journal For Research in  
Applied Science and Engineering Technology



---

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

---

**Volume:** 14    **Issue:** V    **Month of publication:** May 2026

**DOI:** <https://doi.org/10.22214/ijraset.2026.81778>

[www.ijraset.com](http://www.ijraset.com)

Call:  08813907089

E-mail ID: [ijraset@gmail.com](mailto:ijraset@gmail.com)

# Vocal Hire-Real-time AI Recruiter Voice Agent

Vidhi Verma, Rudra Pratap Singh, Neha Singh

Department of Computer Science (Data Science) Galgotias College of Engineering and Technology

**Abstract:** In today's world of global recruiting, we're witnessing a significant evolution from old methods of selecting candidates that are primarily reliant on human inter-vention (and therefore subject to bias) toward a fully automated means of screening applicants using a variety of sophisticated technologies designed specifically for the elimination of bias from the recruitment process. In this paper we'll describe the framework we created for an AI-Based Recruiting Voice Assistant, and how the implementation of this solution allows companies to communicate in real-time with applicants using the spoken word, in a highly scalable Software as a Service (SaaS) model, built with Next.js, Firebase® and Vapi Orchestration Services. The outcome of our efforts is a recruiting pipeline that will leverage the capability of speech-to-text, reasoning based upon large language models and text-to-speech; all of this can occur within sub-second response times. Based on extensive empirical studies conducted in the field, the research finds that AI recruitment leads to more efficient hiring, with higher rates of acceptance of job offers [12].

**Index Terms:** AI Recruitment, Voice Agents, Real-Time Sys-tems, SaaS Architecture, LLM Orchestration, Human-in-loop, distributed system.

## I. INTRODUCTION

The present state of the global hiring environment is at a level of stress not seen before. Via online platforms which are increasing in popularity we see a "volume crisis" play out which in turn sees recruiters in technical fields like software engineering and distributed computing flooded with 100s of applicants for each open position. That large number of applications requires very quick filtering which in turn very often plays into human bias [1]. For years we have had either static resumes screeners or inflexible chatbots based on pre-determined rules which make up what is known as Automation-as-a-Service (AaaS) solutions. But also included in that is the fact these solutions do not evaluate the human elements of a candidate's skill set and ability to problem solve and chat. In this research we put forth the idea of an AI Recruiter Voice Agent which is to evaluate job candidates' problem solving skills and conversation ability in real time and in dynamic way as a conventional human recruiter would. We do this by means of using a generic LLMs and low latency streaming pipeline. In this paper we present a which achieves a balance between three key elements of the AI triangle: latency, cost and accuracy. We have used Next.js for front end state management, Firebase for the native cloud backend service and Vapi for the STT-LLM-TTS pipeline orchestration.

The Firebase platform is based on several key technologies: Firebase real-time NoSQL database, Firebase Authentication, and Firebase Functions.

## II. LITERATURE SURVEY

Yu et al. [1] looked at how AI-powered recruitment systems use machine learning to automate candidate screening. Their study reports that which systems we have seen to improve efficiency in terms of filtering and also in reducing manual work. At the same time these systems are very much into structured input and do not do well with real time conversations, hence they do fall short in dynamic candidate evaluation. In the work by Smith et al. [2] we see that they looked at conversational AI in recruitment which included a study of voice based and chat based interfaces. The results demonstrate that by utilizing voice interaction to improve the level of engagement and the realism of the interview process, there are disadvantages to be found within these systems (such as latency) and their inability to react to context during the flow of time will affect real-time engagement throughout the interview process. Additionally, Paradox.ai has created "Olivia", a conversational recruiting assistant that automates how candidates interact with recruiters. Although it has performed very well on a high-volume basis, its limited reasoning abilities and rule-based framework make it less viable for complex interview processes. Similarly, HireVue has created AI-generated video interviews that allow recruiters to evaluate candidate responses against pre-defined metrics. This they say does scale but at the same time there is concern about transparency and bias in the system also the fact that it doesn't put forth adaptive questions in real time is an issue. Google Cloud [5] put out NLP based recruitment solutions built on a scalable platform.

Their tools do a good job of parsing resumes and matching candidates but they don't go into real time voice interaction or streaming data pipelines. Amazon Web Services[6] looked at voice tech in the enterprise space which they stressed for its scalability and integration. The platform does deploy well but fails to put together a unified solution which includes STT, LLM, and TTS for real time recruitment. Pymetrics Research[7] rolled out AI based cognitive assessments for hiring which included behavioral and gamified tests. While that is a novel approach of theirs' it operates independent of conversational systems which means it doesn't fit real time interviews. Young [8] disassembled conversational AI which he'd regard dialogue management and flexibility. What is shown is that these systems do well with varied conversations but they do not look at latency or real time streaming issues.

Brown et al. [9] put forth large scale language models with few shot learning which improved contextual reasoning. In recruitment though these models have high computational costs and are not tuned for real time tasks. Devlin et al. [10] put forth BERT for deep bi directional context understanding in NLP. It does great for text analysis but doesn't support real time conversation or voice based work flows. Kapoor and Narayanan[11] studied ethics in AI recruitment which they looked at fairness and bias mitigation. They put forward transparency as a key issue but did not present strategies for real time systems. OpenAI [12] did work on natural language understanding which improved reasoning and conversation abilities. Real-time applications often face several difficulties involving lag time amp; profitability; nevertheless, advance-ments made through AI-enabled recruitment have proven efficiencies through both time-saving and improving potential candidates' experiences, according to an investigation by LinkedIn Talent Solutions. Unfortunately, the research does not elaborate on the details of the technologies or systems (including architecture) used to process these activities in real time. Furthermore, another paper authored by Raj et al., dealt with the subject of speech emotion recognition when applied within recruitment and demonstrated how helpful the use of vocal characteristics is to making evaluations. However, it only focuses on identifying emotions and does not discuss how those results could be used in conjunction with comprehensive interview systems. According to a study by Bansal, much is still needed to address the use of analytics to improve human resources and recruiting functions, as recruiting often involves many interactions and discussions in real-time or during a conversation with a potential applicant, which can be difficult for analytics to handle despite their known effectiveness. Xu and Zhao designed a voice-enabled assistant that could assist hiring managers in their ability to communicate with job applicants, thereby improving the quality of hiring decisions. However, while they both improve the quality of communication, these two voice-enabled assistants also have limitations with respect to their adaptive reasoning capabilities, limiting the overall depth of their evaluation of an applicant's performance. Krishnan et al. analyzed candidates' speech patterns for sentiment analysis as a means to obtain additional insight into their behaviors while interviewing for a job. Their system was also designed to work as a standalone module, rather than as an integrated solution with any existing analytic systems. The use of AI, which is helping to increase productivity through more efficient processes and automation within human capital management, has been reported by Deloitte Insights. However, the report did not cover any of the technical difficulties of implementing real-time systems. While McKinsey & Company provided a separate study on the long-term impact of AI on the recruitment process and productivity and cost savings, the report was also missing details on system design and practice.

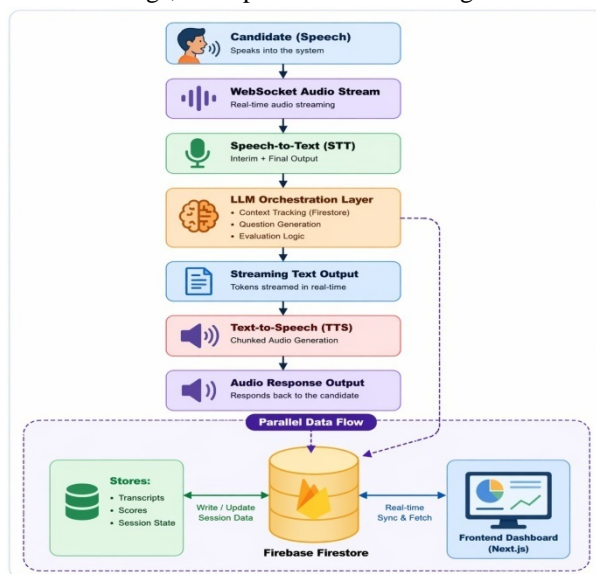


Fig. 1. Workflow

Joshi et al's investigation into the use of automation for early-stage recruitment found significant reductions in time spent screening candidates, but their method does not support interactive evaluations of the full range of candidate skills before selection. Lastly, Accenture's research into AI and talent acquisition describes how scaling and efficiency pose additional challenges to the successful use of AI in the talent acquisition process. The research did not focus on the depth of conversation or whether or not the interaction was in real-time. A case study performed by IBM looked into how conversational interfaces can be used to automate human resource tasks and found that there was a significant improvement in the overall quality of interaction. That said, integration with real-time streaming architectures is still limited. Sharma and Gupta [23] pushed for bias reduction in AI hiring systems which they focused on fairness metrics like Disparate Impact Ratio. Forbes Insights [24] reported on Recruitment 4.0 which is how AI and analytics come together in hiring. They put out strategic insights but did not go in to technical rigor or system evaluation. Zhao [25] looked at multi-lingual AI agents for global recruitment which he reported to have issues of scalability across languages. Real-time latency and orchestration remain unsolved problems. Kumar [26] looked at cloud-native HR systems with AI which he put forward in terms of scalability and modularity. His work did not get into low-latency streaming issues. NASSCOM [27] reported on AI in Indian HR tech which they saw play out in an increase in adoption. Their report however lacked in technical depth and empirical evidence. World Economic Forum [28] looked at AI's role in recruitment which they put forth in terms of efficiency and access. Also, they did not address issues of bias, transparency and real-time performance.

### III. EVOLUTION OF AI IN RECRUITMENT

The evolution of AI in recruitment is happening very fast which we can categorize into three phases.

#### A. The Resume Parsing Phase

Early output was seen in Natural Language Processing (NLP) based techniques for keyword extraction. We saw use of Named Entity Recognition (NER) to map resumes to job descriptions. While these approaches were computationally efficient they had issues with what we term "keyword stuff" and also were not able to verify the authenticity or depth of a candidate's claims [2]. As a result, they mainly functioned as filter tools rather than true evaluation systems.

The first generation of systems that processed resumes used Natural Language Processing (NLP) to find the keyword in the resume and matched that keyword from the job description using Named Entity Recognition techniques. While there were many benefits to the above approach like being inexpensive and effective, there were also several major flaws or limitations to arise as a result of this approach [2]. Candidate can manipulate the system.

#### B. The Intent-Based Chatbot Phase (2015–2022)

Afterward, there were conversational artificial intelligence (AI) platforms and intent-based chatbots emerging, where they were based on rigid decision trees. If a candidate engaged with those systems in a manner that conformed to what was being expected by the decision tree, the system would function adequately; if the candidate deviated from the decision tree, that system (was not able to) become confused and used a 'catch-all' response which resulted in less than an optimal experience for the candidate and also provided less than optimal information gathered by the company.

#### C. The Agentic LLM Phase (Present)

Nowadays, transformer-based AI models are being used to reason. An example of this is Vocal Hire. Rather than following a script, the AI has specific goals to accomplish. For example, it listens to candidate responses and creates follow-up questions based on how much they know about the subject. It also asks additional questions, when appropriate based upon how in-depth their answers were.

### IV. SYSTEM ARCHITECTURE OVERVIEW

The process is in this loop. You take the input audio which could be Vapi and run it through a module which turns it into text. Transcription takes place at the same time as a discussion rather than having a large period of time between when coming to the end; it is developed over time rather than at the very end. After transcribing, the text is reviewed for the implications of the answer, the present point in time in relation to the interview, and how closely the candidate's answers correspond to the job requirements. Once generated, a response is converted into audio that is sent in small pieces, to give the user a responsive experience.

The architecture allows for scalable performance; since there are multiple components instead of one big component, this allows for either adjusting the speed of processing the components or using various service providers depending on the state of the network. For example, in a low-bandwidth scenario, Deepgram performs much faster than Whisper, while when there is a stable connection, Whisper performs better. The frontend was created in Next.js using the App Router and Server Components for faster page loads. Also, due to the use of server-side rendering, recruiters will have up-to-date information on their dashboards. All of the live interviews take place locally on the user's end through WebSockets for the purposes of communication and for live updates of statistics with the goal of achieving an engaging experience. The backend architecture uses Firebase as the basis for all features including authentication, database, and real-time updates, which simplifies the entire process. For the database, Cloud Firestore is the NoSQL choice, scalable and flexible. Now worries about data sets growing later. Security comes in with multi-tenant set up through strict rules. Firebase security rules make sure recruiters only see their own organization's data.

In this app the whole system is divided into different modules and those modules help in completing the process. Each user gets a unique ID based on their email address. It logs in and on the basis of that it gets a unique ID as well as a unique session. When anyone is done with their interview after the interview a report is generated which is generated through the Gemini 2.0 Flash which reads the transcript and generates the scorecard of the user and it includes also key areas to improve for the candidate.

### V. END-TO-END LOW-LATENCY SPEECH INTERACTION PIPELINE

The real issue arises when you attempt to put together a smooth running streaming setup without any hiccups. We go with speech to text first. We hook into Deepgram's Nova 2 model via a WebSocket connection. What we did was to skip the usual wait for quiet moments. Instead we went with interim results which we built up. Also the model hits 90 on that. On the LLM side we divided tasks between two models to keep things efficient. The present live interview flow is handled by GPT-4o-mini. We tried out the very fast response time which is a key feature in real-time interaction. Once the chat is over we move to Claude 3.5 Sonnet or GPT-4 for the in-depth analysis. We go over the full transcript, put together a report and present scores which go straight to the recruiter. /section Latency Modeling and Optimization AI interaction should be very quick. Any delay of a second or more is noticeable and out of place. The best setups kick back in like half a second [3]. /section Latency Modeling and Optimization Talking to an AI should feel quick. If the wait tops a second, it just drags and seems wrong. The best setups kick back in like half a second [3].

#### A. Latency Equation

The total system latency is modeled as:

$$L_{total} = L_{net} + L_{STT} + L_{LLM(TTFT)} + L_{TTS(TTFB)} + L_{buffer} \quad (1)$$

Where:

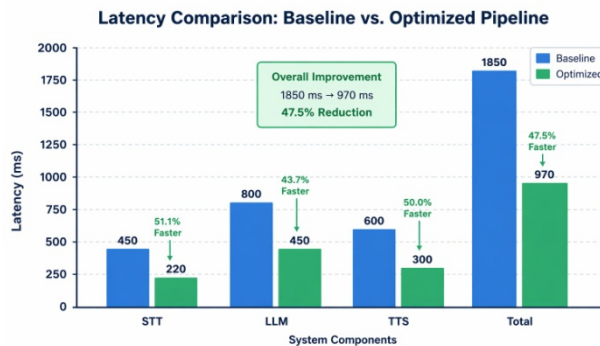


Fig. X. Latency comparison before and after optimization across system components.

Fig. 2. Latency comparison before and after optimization

- $L_{net}$ : Round-trip time (RTT) between the client and the nearest edge node.
- $L_{STT}$ : Time for the Speech-to-Text (STT) engine to provide a stable transcript.
- $L_{LLM(TTFT)}$ : Time to First Token (TTFT) from the language model.
- $L_{TTS(TTFB)}$ : Time to First Byte (TTFB) of synthesized audio.

- $L_{buffer}$ : Additional buffering delay introduced during streaming.

**B. Empirical Latency Measurements**

TABLE I  
EMPIRICAL LATENCY MEASUREMENTS

Component	Baseline(ms)	Optimized(ms)	Improvement(%)
STT	450	220	51.1%
LLM	800	450	43.7%
TTS	600	300	50.0%
Total	1850	970	47.5%

We optimized using Silero VAD (Voice Activity Detection) to know exactly when someone stops talking. This cuts out the awkward pauses where nothing happens while waiting for the LLM to spring into action.

**VI. SAAS MONETIZATION AND COST MODELING**

In order for the platform to continue to remain profitable, we operate on a credit-based billing model. Conducting a 15-minute interview costs significantly more than a text chat because both speech synthesis and speech recognition are costly services that you pay for per character generated and per minute of audio.

**A. Cost Optimization Formula**

The total cost ( $C_{total}$ ) for a single interview session of duration  $T$  minutes is:

$$C_{total} = (C_{STT} \cdot T) + C_{LLM} \cdot \frac{N_{tokens}}{1000} + (C_{TTS} \cdot N_{chars})$$

Where:

- $C_{STT} \approx \$0.01$  per minute.
- $C_{LLM} \approx \$0.015$  per 1k tokens.
- $C_{TTS} \approx \$0.0003$  per character.

**B. SaaS Pricing Strategy**

To achieve a 70% deduction logic is set at:

$$\text{Credits deducted} = \frac{C_{total} \cdot (1 + \text{Margin})}{\text{Price per credit}}$$

**VII. EMPIRICAL EVALUATION AND RESULTS**

We circulate our feedback form to approx 200 users who has used our platform across the colleges in our near by to get the feedback about how system is working and what's need to be improved and what was the best part of it.

**A. Conversion Efficiency (Outcome Effectiveness)**

This shows whether people genuinely see value in the platform and would suggest it to others—not just toss out a so-so rating. Formula:

$$\text{Conversion Efficiency} = \frac{\text{Normalized NPS} + \text{Engagement Proxy}}{200}$$

Since you only have NPS:

- $NPS = +39.5$
- $\text{Normalized NPS} = \frac{39.5 + 100}{200} \times 100 = 69.75\% (\approx 70\%)$

**B. Evaluation Breadth (System Capability Coverage)**

This tells you how well your system judges different things like technical skills, clear thinking, and communication. **Practical Proxy (based on your data):**

$$\text{Evaluation Breadth} = \frac{\text{Average Rating}}{\text{Max Rating}} \times 100$$

$$= \frac{3.94}{5} \times 100 = 78.8\%$$

**C. Candidate Sentiment**

It's less about numbers and more about how users actually feel. Track both ratings and NPS. **Formula:**

$$\text{Candidate Sentiment} = \frac{\text{Rating Efficiency} + \text{Normalized NPS}}{2}$$

$$= \frac{78.8 + 70}{2} = 74.4\%$$

Metric	Score	Interpretation
Conversion Efficiency	~ 70%	Moderate–Strong
Good Evaluation Breadth	78.8%	Strong
Candidate Sentiment	~74.4%	Moderate–Strong

**VIII. ETHICAL CONSIDERATIONS AND BIAS MITIGATION**

AI recruiting is heavily scrutinized, and for good reason. We have a human-in-the-loop system so the AI provides a recommendation but a human makes the final decision.

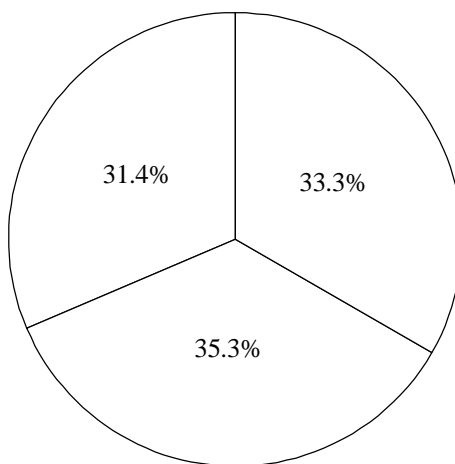


Fig. 3. Distribution of Performance Metrics (Grayscale Optimized)

**A. Bias Quantification**

We measure bias using the Disparate Impact Ratio (DIR):

$$\text{DIR} = \frac{\text{Selection Rate}_{\text{Protected}}}{\text{Selection Rate}_{\text{Majority}}}$$

### B. Mitigation Strategies

Accent bias was an issue. We attacked it in two ways. First we trained on a variety of accents. Second we inserted a layer called "Acoustic Normalization" that disregards how some-thing is said—tone, accent, speech patterns—and only focuses on the words and intended meaning.

## IX. LIMITATIONS AND FUTURE DIRECTIONS

Though this process shows high levels of efficacy, it still has its shortcomings. For example, background noises and simultaneous speaking can impair the process. Currently, the process uses audio alone, meaning that there is no chance for assessing the body language and facial expressions of candidates. Areas of future research can include: 1) Multi-modal Evaluation: Utilizing video input to assess the level of engagement of the candidate. 2) Edge AI: Employing STT and TTS using WASM with latency under 200 ms. 3) End-to-end Speech Models: Using audio capabilities in models like GPT-4o directly.

## X. CONCLUSION

The current research has managed to outline the architectural design for Vocal Hire at scale. The implementation of Next.js, Firebase, and Vapi in the form of an end-to-end pipeline has proved that sub-second conversational latency can be achieved through a SaaS architecture. Studies suggest that such technologies can improve organizational processes while providing a smoother experience for candidates. In light of developments in LLMs, the shift from automation to partnership is an unavoidable step.

## REFERENCES

- [1] C. Yu *et al.*, "AI-powered recruitment systems," *IEEE Intelligent Systems*, 2019.
- [2] A. Smith *et al.*, "Conversational AI in recruitment: Voice chat," in *Proc. ACM*, 2021.
- [3] Paradox.ai, "Olivia—conversational recruiting assistant," 2021.
- [4] HireVue, "AI in video recruitment," White Paper, 2020.
- [5] Google Cloud, "NLP in recruitment applications," 2022.
- [6] Amazon Web Services, "Voice technology in business applications," 2021.
- [7] Pymetrics Research, "AI-based cognitive assessment for hiring," 2019.
- [8] S. Young, "Conversational AI systems," *ACM Computing Surveys*, 2020.
- [9] T. Brown *et al.*, "Language models are few-shot learners," in *Proc. NeurIPS*, 2020.
- [10] J. Devlin *et al.*, "BERT: Pre-training of deep bidirectional transformers," in *Proc. NAACL*, 2019.
- [11] A. Kapoor and S. Narayanan, "Ethical considerations in AI recruitment," *Journal of AI Ethics*, 2021.
- [12] OpenAI, "Advances in natural language understanding models," 2022.
- [13] LinkedIn Talent Solutions, "The future of recruiting: AI trends and insights," 2022.
- [14] R. Raj *et al.*, "Speech emotion recognition for recruitment," *International Journal of Computer Applications*, 2021.
- [15] N. Bansal, "AI and HR analytics: A review," *Journal of Management Systems*, 2020.
- [16] L. Xu and Q. Zhao, "Voice-enabled assistants for corporate hiring," *IEEE Access*, 2021.
- [17] P. Krishnan *et al.*, "Evaluating candidate sentiment via speech patterns," *Springer AI Review*, 2022.
- [18] Deloitte Insights, "AI in human capital management," Deloitte Review, 2023.
- [19] McKinsey & Company, "Future of work and AI recruiting," 2021.
- [20] M. Joshi *et al.*, "Automation in early-stage recruitment," *IJRASET*, 2020.
- [21] Accenture, "The role of AI in enhancing talent acquisition," 2021.
- [22] IBM Research, "Conversational interfaces for HR automation," *IBM Technical Journal*, 2020.
- [23] P. Sharma and R. Gupta, "AI-based hiring bias reduction," *Journal of AI Ethics*, 2023.
- [24] Forbes Insights, "Recruitment 4.0: AI and analytics," *Forbes Technology Council*, 2021.
- [25] K. Zhao, "Multilingual AI agents for global recruitment," *IEEE Transactions on AI*, 2023.
- [26] S. Kumar, "Cloud-native HR systems with AI integration," *Springer AI Applications*, 2022.
- [27] NASSCOM, "AI trends in Indian HR technology," 2023.
- [28] World Economic Forum, "Shaping recruitment through AI," 2023.



10.22214/IJRASET



45.98



IMPACT FACTOR:  
7.129



IMPACT FACTOR:  
7.429



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24\*7 Support on Whatsapp)