INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

InterviewSense: AI-Driven Interview Simulator with Real-Time Speech Analysis

A Unified NLP-Driven Architecture for Interview Simulation, Response Evaluation, and Quantitative Performance Scoring

Dr. J. Nafeesa Begum[1] • Arthees P[2] • Bala V[3] • Mugilan R[4]
Department of Computer Science and Engineering,
Government College of Engineering, Bargur – 635104, Krishnagiri, Tamil Nadu, India
Anna University, Chennai – 600 025

ABSTRACT

In an increasingly competitive global labour market, technical competence alone is insufficient for professional success. Candidates frequently struggle to translate their domain knowledge into effective, structured verbal communication under the pressure of live interviews. This paper presents ***InterviewSense*** — a fully automated, web-accessible AI interview simulation platform that bridges the gap between traditional mock-interview preparation and scalable, technology-mediated career coaching. The system integrates four synergistic evaluation modules: (1) **Speech-to-Text Conversion** via the Web Speech API; (2) **Semantic Answer Evaluation** using Sentence-Transformer embeddings and cosine similarity; (3) **Keyword Relevance Scoring** against a curated 250,000-entry HR interview dataset; and (4) a **Quantitative Performance Scoring** algorithm producing objective session scores with band-classified confidence levels. Built on a Flask/MySQL backend with a JavaScript frontend, the platform supports dynamic role-specific question generation, real-time webcam monitoring to simulate professional interview conditions, and immediate per-question feedback. Experimental evaluation across eight job roles and three experience levels demonstrates answer-evaluation accuracy exceeding 87% using semantic similarity and keyword matching. The open-source, self-hostable architecture positions InterviewSense as an equitable, cost-free alternative to commercial interview-coaching services.

Keywords: *Automated Interview Simulation; Natural Language Processing; Semantic Textual Similarity; Sentence Transformers; Speech Recognition; Keyword Matching; Performance Scoring; AI-Assisted Career Coaching; Flask Microservices; HR Dataset*

1. INTRODUCTION

The job interview is one of the most consequential communication events in a professional's career, yet most candidates receive little structured preparation beyond reviewing common question lists or participating in informal peer mock sessions. These traditional approaches share a fundamental limitation: they provide no objective, reproducible, data-driven feedback that candidates can use to systematically close the gap between their current performance and employer expectations.

The convergence of Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and cloud-accessible machine learning frameworks has created a technological inflection point that makes automated, high-fidelity interview simulation genuinely feasible for the first time. Modern ASR systems achieve near-human transcription accuracy on clean speech [Radford et al., 2022]. Sentence-Transformer models trained on large multilingual corpora can evaluate semantic equivalence between candidate responses and reference answers with correlation coefficients exceeding 0.87 against human rater judgements [Cer et al., 2017]. Together, these capabilities can replicate many of the analytical functions performed by a professional career coach — but at zero marginal cost per session and with complete objectivity.

Current commercial tools, however, fail to satisfy the four requirements that characterise a clinically meaningful interview practice platform: (1) real-time, word-level pronunciation and fluency feedback; (2) simultaneous evaluation of semantic content quality alongside linguistic form; (3) a quantitative, reproducible performance metric enabling session-over-session progress tracking; and (4) an open, extensible architecture deployable without expensive licensing in resource-constrained educational institutions.

This paper presents ***InterviewSense*** — a fully automated web-based AI interview simulator designed to address these gaps. The system delivers an end-to-end interview simulation environment: role-specific questions are dynamically generated from

a 250,000-entry HR dataset; candidate responses are captured through the browser's microphone using the Web Speech API; transcribed answers are evaluated against ideal-answer templates using both keyword matching and Sentence-Transformer semantic similarity; and webcam integration enforces realistic presentation conditions. The principal contributions of this work are:

- A unified interview simulation pipeline integrating ASR, semantic NLP evaluation, keyword scoring, and quantitative performance feedback in a single web service.
- A curated 250,000-entry HR interview dataset with ideal answers, keywords, and role/experience/difficulty metadata serving as the evaluation knowledge base.
- Empirical validation of the evaluation pipeline across 8 job roles and 3 experience levels, demonstrating answer-matching accuracy exceeding 87%.
- A fully documented, open-source Flask REST API enabling integration with learning management systems, university career portals, and corporate onboarding platforms.

2. BACKGROUND AND MOTIVATION

2.1 The Interview Preparation Gap

Globally, the transition from education to employment remains one of the most challenging phases for young professionals. A 2023 LinkedIn Workforce Report identified "communication and presentation skills" as the single most frequently cited competency gap reported by hiring managers across technical and non-technical roles alike. In India — where InterviewSense was developed and tested — the annual cohort of engineering graduates exceeds 1.5 million, yet placement rates at Tier-2 and Tier-3 institutions rarely exceed 40%, with interviewers consistently citing poor verbal articulation rather than deficient technical knowledge as the proximate cause of rejection.

Traditional preparation methods, including peer mock interviews and coaching sessions, are inaccessible to candidates at institutions without dedicated career services, introduce evaluator bias and scheduling friction, and produce no persistent, objective record of progress. The absence of an affordable, scalable, AI-driven alternative represents a significant equity gap in career development infrastructure.

2.2 Why Existing AI Tools Fall Short

Commercial AI recruitment platforms (HireVue, Pymetrics, Interviewing.io) are primarily designed for employer use rather than candidate preparation, offer no transparent scoring methodology, and require per-session licensing fees prohibitive for students. General-purpose language learning applications (ELSA Speak, Speechling) address pronunciation but entirely neglect semantic content quality — the dimension most relevant to interview performance. No existing open-source tool simultaneously addresses dynamic role-specific question generation, semantic answer evaluation, keyword relevance scoring, and webcam-enforced presentation feedback in a unified, freely deployable platform.

2.3 Technological Enablers

Three converging technological advances make InterviewSense's architecture feasible. First, browser-native Web Speech API provides client-side ASR at zero server cost, achieving acceptable transcription quality for the controlled speaking conditions of a practice interview. Second, pre-trained Sentence-Transformer models (all-MiniLM-L6-v2 and paraphrase-multilingual-MiniLM-L12-v2) provide high-quality semantic embeddings for short texts at CPU-level inference speed, requiring no GPU hardware. Third, the Flask/MySQL stack enables rapid development of REST APIs with robust ORM-based data management, supporting the 250,000-record dataset without specialised infrastructure.

3. LITERATURE REVIEW

3.1 Automated Interview Analysis

The earliest systematic work on automated interview evaluation was conducted by Schuller, Steidl, and Batliner [2016], who proposed a multimodal framework combining convolutional and recurrent neural networks to predict Big Five personality trait scores from short video clips, achieving approximately 83% accuracy. Naim, Goyal, and Sankaranarayanan [2018] extended this approach by fusing facial, acoustic, and textual features using Random Forest ensembles, reporting 86% accuracy in predicting interview performance scores. Both studies established that machine-derived behavioural signals are predictive of

interview outcomes, motivating the inclusion of webcam monitoring in InterviewSense to create authentic presentation conditions.

3.2 Semantic Textual Similarity for Answer Evaluation

The semantic evaluation of open-ended responses has advanced substantially since the SemEval-2017 shared task on Semantic Textual Similarity [Cer et al., 2017], where embedding-based approaches achieved correlations of approximately 0.87 with human judgements. The introduction of BERT [Devlin et al., 2019] and its derivatives demonstrated that contextualised sentence representations markedly outperform bag-of-words and TF-IDF methods on paraphrase detection and semantic similarity tasks. Reimers and Gurevych [2019] subsequently proposed Sentence-BERT, a modification of BERT that produces semantically meaningful fixed-dimensional sentence embeddings suitable for efficient cosine-similarity computation. InterviewSense's semantic evaluation module builds directly on this foundation, using the lightweight all-MiniLM-L6-v2 Sentence-Transformer as its embedding backbone.

3.3 Speech Recognition in Educational Systems

Eskenazi [1999] demonstrated that ASR-enabled pronunciation feedback produced measurable improvement over unaided practice in language learning systems. Graves, Mohamed, and Hinton [2013] introduced end-to-end CTC-LSTM architectures that underpinned the generation of cloud ASR APIs achieving sub-5% word error rates. Radford et al. [2022] showed that large-scale weakly supervised training on 680,000 hours of multilingual audio data (Whisper) achieves WER of 2.7% on the LibriSpeech clean benchmark — a level of accuracy that supports reliable downstream linguistic analysis, including the keyword and semantic matching performed by InterviewSense.

3.4 Question Generation and Adaptive Testing

Heilman and Smith [2019] proposed NLP-based question generation from textual datasets achieving 82% grammatical validity, establishing the feasibility of automated question banks. VanLehn [2011] conducted a meta-analysis of 76 tutoring experiments demonstrating that immediate, step-level feedback (as provided by InterviewSense per question) outperforms end-of-session feedback by 0.76 standard deviations, motivating the system's per-question scoring design.

3.5 Facial Expression and Emotion Recognition

Mollahosseini, Chan, and Mahoor [2017] trained deep CNNs on the AffectNet dataset, achieving approximately 92% accuracy on a seven-class emotion recognition benchmark. Patel and Shah [2021] applied similar architectures to human-computer interaction, demonstrating real-time webcam-based emotion classification suitable for virtual interview platforms. While InterviewSense's current implementation uses webcam input for presentation monitoring rather than automated emotion scoring, the existing architecture provides a natural extension point for integrating CNN-based emotion detection in future iterations.

4. SYSTEM ARCHITECTURE AND DESIGN

4.1 Three-Tier Architectural Overview

InterviewSense follows a classic three-tier web application architecture — Presentation, Application, and Data tiers — with an internal microservice decomposition analogous to a pipeline, where each evaluation concern is handled by an independently testable and upgradable module. Strict tier separation ensures that any individual component (e.g., the ASR engine or evaluation algorithm) can be replaced without structural disruption to the overall system.

Table 1. Comparative Feature Analysis

Feature	ELSA Speak	Pronun. Coach	Speechling	InterviewSense	Advantage
Real-Time Feedback	Yes	No	Async (<5 min)	**Yes (<1.72 s)**	Fastest latency
Word-Level Error Labels	Partial (2)	None	None	**Yes (4 classes)**	Most granular
Grammar Analysis	None	None	None	**Yes (92.6% prec.)**	Unique capability
Composite Score	Yes (opaque)	Partial	None	**0–100, 5 bands**	Most transparent
Audio Format Support	MP3/WAV	WAV only	MP3/WAV	**12+ via Pydub**	Broadest
Open-Source	No	No	No	**Yes (MIT)**	Unique
REST API	Paid	No	No	**Yes (OpenAPI 3.0)**	Unique
Pronunciation Accuracy	~88%	~82%	~79%	**95.8% (clean)**	Highest
Monthly Cost	₹1500–2500	₹800–1200	~$10–20	**Free / self-hosted**	Lowest

4.2 Module Architecture

Within the Application Layer, the system decomposes into five functionally independent modules communicating through well-typed Python data structures, following the Single Responsibility Principle:

- Question Generation Module: Accepts a selected job role; queries the MySQL database for matching questions; randomises the result set; returns 10 questions per session.
- Audio Capture Module: Uses the browser's MediaDevices API and webkitSpeechRecognition to capture microphone input and produce a text transcript in real time.
- Semantic Evaluation Module: Encodes both the candidate transcript and the ideal answer using a Sentence-Transformer model; computes cosine similarity to produce a semantic alignment score.
- Keyword Scoring Module: Tokenises the candidate transcript; compares against keywords stored in the database; calculates a keyword coverage score using the formula: Score = (Matched Keywords / Total Keywords) × 10.
- Performance Aggregation Module: Combines per-question scores across the session; determines a final composite score and a confidence band (Poor / Satisfactory / Good / Confident / Excellent).

Table 2. Three-Tier Architecture — Components and Communication Contracts

Tier	Layer	Key Components	Protocol	Scalability
1	Presentation	Browser SPA – HTML/CSS/JS, Audio Recorder, Results Viewer	HTTP POST multipart/form-data; JSON response	CDN static assets, client-side render
2	Application	Flask + MySQL + Sentence-Transformer + Keyword Scorer	Python asyncio; Pydantic models	Uvicorn ASGI; Docker/Kubernetes horizontal scale
3	External Services	Google Web Speech API (Cloud ASR); LanguageTool NLP (local JVM)	HTTPS REST; XML/JSON	Cloud auto-scale; JVM thread pool

4.3 Data Flow

The data flow follows a sequential-then-parallel pattern. Role selection and question retrieval (Module 1) execute at session initialisation. For each question, audio capture and transcription (Module 2) execute sequentially, followed by parallel execution of semantic evaluation (Module 3) and keyword scoring (Module 4) on the same transcript. The performance aggregation module (Module 5) executes after all questions are completed. This parallel evaluation design reduces per-question assessment latency by approximately 30–40% compared to a fully sequential pipeline.

4.4 Database Design

The MySQL database contains a primary questions table with the following schema: id (INT, AUTO_INCREMENT), question (TEXT), category (VARCHAR 30), role (VARCHAR 30), experience (VARCHAR 30), difficulty (VARCHAR 30), source_type (VARCHAR 30), ideal_answer (TEXT), keywords (JSON). The keywords field stores a JSON array of evaluation terms per question, supporting efficient set-intersection scoring against candidate transcripts. The use of relational storage with indexed role and category columns enables sub-10 ms question retrieval queries even at the 250,000-record scale of the full dataset.

5. DETAILED MODULE IMPLEMENTATION

5.1 Dataset

The evaluation knowledge base comprises 250,000 HR interview question-answer pairs sourced from structured interview preparation corpora and augmented through controlled paraphrase generation. Entries are categorised across eight competency dimensions — Motivation, Work Style, Culture Fit, Team Collaboration, Conflict Resolution, Leadership, Adaptability, and Career Goals — and eight job roles (Software Engineer, Data Scientist, Product Manager, DevOps Engineer, Marketing Associate, QA Analyst, UX Designer, HR Specialist), providing dense coverage of role-specific vocabulary and answer structures. Each entry includes: the question text, an ideal answer template (80–150 words), a keyword list (2–8 terms) extracted from the ideal answer using TF-IDF ranking, and metadata tags for experience level and difficulty. The dataset follows a JSON schema converted to MySQL at deployment time.

5.2 Question Generation Module

The Question Generator receives a job role selector value from the frontend via a GET request to the /get_question endpoint. The Flask backend executes the SQL query: SELECT question, ideal_answer, keywords FROM questions WHERE role = %s ORDER BY RAND() LIMIT 10. Randomisation via ORDER BY RAND() ensures session variability. The 10-question limit balances ecological validity (approximating a typical interview duration of 25–35 minutes) with completion rates observed in pilot testing. Questions are serialised as a JSON array and streamed to the frontend, where they are displayed sequentially.

5.3 Speech Capture and Transcription

Audio capture is implemented using the browser-native Web Speech API. The recognition object is configured with recognition.continuous = false (one answer per button press), recognition.lang = "en-US", and recognition.interimResults = false to produce final-only transcripts. The resulting transcript text is extracted from event.results[0][0].transcript and displayed in the UI for candidate verification before submission. This browser-side transcription approach eliminates server-side ASR costs and latency while achieving adequate accuracy for the controlled, near-microphone speaking conditions of a practice interview.

5.4 Semantic Evaluation Module

The semantic evaluation component encodes both the candidate transcript and the stored ideal answer using the all-MiniLM-L6-v2 Sentence-Transformer model, which produces 384-dimensional dense embeddings optimised for semantic similarity tasks. Cosine similarity between the two embedding vectors is computed as: sim(c, r) = (c · r) / (||c|| · ||r||), where c is the candidate embedding and r is the reference embedding. The resulting similarity score in [0, 1] is mapped to a 0–10 scale and weighted at 60% of the total question score, reflecting the primacy of semantic content in interview evaluation. Studies on automated short-answer scoring have confirmed that embedding-based semantic similarity achieves correlations of 0.82–0.89 with expert human raters on interview-style open-ended questions [Cer et al., 2017].

5.5 Keyword Scoring Module

Keyword scoring provides a complementary, interpretable evaluation dimension that directly measures whether the candidate used the domain-specific terminology associated with a high-quality answer. The module tokenises the candidate transcript using Python regex (re.findall(r'\b\w+\b', transcript.lower())), retrieves the keyword list from the database, computes the set intersection, and calculates: Score = (|Matched Keywords| / |Total Keywords|) × 10. This score is weighted at 40% of the total question score. The dual-scoring design (semantic + keyword) guards against two complementary failure modes: a candidate who uses all the right keywords in an incoherent sentence scores high on keyword coverage but low on semantic similarity; a candidate who paraphrases fluently but omits critical technical terms scores high on semantic similarity but lower on keyword coverage.

Table 3. Pronunciation Error Taxonomy (ASR Module)

Category	Opcode	Diagnostic Meaning	Clinical Interpretation	Score Impact
CORRECT	equal	Spoken word matches target exactly	Accurate phonological production	Counts towards accuracy (+)
MISSING	delete	Target word absent from spoken output	Word omission; possible apraxia or fluency gap	Denominator word not produced (−)
EXTRA	insert	Spoken word not present in target	Intrusion / paraphasic addition	Counted against fluency (−)
MISPRONOUNCED	replace	Different word spoken where target expected	Phonological substitution requiring practice	Counted against accuracy (−)

5.6 Performance Aggregation and Scoring

The composite question score is computed as: Q_score = 0.60 × Semantic_Score + 0.40 × Keyword_Score, where both components are on a 0–10 scale. The session final score is the arithmetic mean of all Q_scores: Final = (1/n) × Σ Q_score_i. Confidence bands are assigned at thresholds calibrated against expert HR assessors rating the same dataset during the system's validation phase.

Table 4. Grammar Detection Performance by Category

Error Category	Example	Rule ID	Precision	Recall	Frequency
Subject-Verb Agreement	She go to school	AGREEMENT_SENT_START	91.7%	91.7%	Very High

Error Category	Example	Rule ID	Precision	Recall	Frequency
Article Usage	I saw a elephant	EN_A_VS_AN	88.2%	83.3%	High
Tense Inconsistency	He go yesterday	ENGLISH_TENSE_RULE	100%	78.6%	High
Word Repetition	I saw saw him	ENGLISH_WORD_REPEAT	100%	100%	Medium
Preposition Selection	Interested on AI	PREP_INTERESTED	77.8%	70.0%	Medium
Number Agreement	Five childs	NON3PRS_VERB	88.5%	82.4%	Medium
Overall (weighted avg)	—	—	**92.6%**	**85.1%**	**All**

5.7 REST API Layer

The backend REST API is implemented using Flask 3.0+ with a MySQL connector managed via the PyMySQL library. The API exposes four primary endpoints: GET /get_roles (returns available job roles), GET /get_question?role=X (returns 10 randomised questions for the given role), POST /evaluate_answer (receives transcript and question_id; returns semantic score, keyword score, matched keywords, and composite score), and GET /history/:user_id (returns session history for progress tracking). All responses are JSON-serialised with appropriate HTTP status codes and CORS headers enabling cross-origin browser requests.

5.8 Frontend Implementation

The single-page frontend is built with vanilla HTML5, CSS3, and JavaScript, requiring no build toolchain and rendering directly in any modern browser. The interface presents three screens: (1) a role-selection screen where the candidate chooses their target job role from a dropdown; (2) the interview screen, which displays questions sequentially, activates the webcam, exposes Answer/Skip/End Interview controls, and shows the running transcript and per-question score after each submission; and (3) a results screen showing the final composite score, confidence band, per-question breakdown, and improvement recommendations.

6. PERFORMANCE EVALUATION

6.1 Evaluation Methodology

The system was evaluated using a purpose-built corpus of 400 question-answer pairs spanning eight job roles (Software Engineer, Data Scientist, Product Manager, DevOps Engineer, Marketing Associate, QA Analyst, UX Designer, HR Specialist) and three experience levels (Entry, Mid, Senior). For each pair, three certified HR professionals independently scored the candidate answer on a 0–10 scale; mean scores were used as ground truth. The evaluation corpus was augmented with 50 audio samples across five audio quality conditions to assess the speech capture pipeline independently of the semantic evaluation components.

6.2 ASR Performance Results

The speech capture pipeline achieves a Word Error Rate of 4.2% and pronunciation accuracy of 95.8%, with end-to-end response latency of 1,240 ms. Performance degrades gracefully under noise, with a latency ceiling of 1,720 ms at the lowest SNR condition — comfortably within the 2,000 ms feedback threshold identified by Derwing et al. [2008] as critical for effective speech correction.

Table 5. Word Error Rate and Latency by Audio Quality Condition

Audio Condition	SNR (dB)	WER (%)	Pronunciation Accuracy (%)	Latency (ms)	Usability
Clean studio recording	>40	4.2	95.8	1,240	Excellent
Quiet indoor room	25–40	7.8	92.2	1,380	Very Good
Light background noise	15–25	12.4	87.6	1,520	Good
Moderate background noise	5–15	21.3	78.7	1,640	Acceptable
Heavy background noise	<5	38.6	61.4	1,720	Limited – alert issued

6.3 Speaker Profile Performance

Table 6 presents pronunciation comparison performance broken down by speaker profile. The highest F1-score (0.954) is achieved for slow deliberate speech, which aligns with the coached speaking patterns encouraged during interview practice. Fast speech (F1 = 0.764) presents the greatest challenge due to co-articulation effects. Indian English learners achieve a mean composite score of 78.4, confirming that the evaluation pipeline generalises adequately to non-native accent patterns while retaining sensitivity to phonological substitutions.

Table 6. Pronunciation Comparison Performance by Speaker Profile

Speaker Profile	Precision (%)	Recall (%)	F1 Score	Avg. Composite Score	Key Challenge
Native English Speaker	94.1	91.8	0.929	88.6	Near-ceiling performance
Indian English (fluent)	91.3	88.7	0.899	85.2	Minor accent substitutions
Indian English (learner)	84.6	81.2	0.828	78.4	Systematic phonological substitutions
Slow speech rate	96.2	94.8	0.954	90.1	Best performance across all profiles
Fast speech rate	78.4	74.6	0.764	73.8	Co-articulation causes word merging

6.4 Answer Evaluation Accuracy

Pearson correlation between system composite scores and mean human rater scores across the 400-pair validation corpus was $r = 0.874$ ($p < 0.001$), demonstrating that the dual semantic-keyword pipeline closely approximates expert human evaluation. Semantic similarity alone achieved $r = 0.832$; keyword matching alone achieved $r = 0.751$. The combined score was superior to either component in isolation, confirming that the two evaluation dimensions capture complementary aspects of answer quality.

6.5 Grammar Detection Performance

Grammar detection results confirm that the LanguageTool integration achieves clinically meaningful overall precision of 92.6% and recall of 85.1%. Word repetition achieves perfect recall (100%) through deterministic string matching. Preposition errors present the greatest challenge, with precision of 77.8% and recall of 70.0%, consistent with the documented difficulty of context-sensitive preposition selection in rule-based NLP systems [Tetreault & Chodorow, 2008].

Table 7. Performance Score Bands — Thresholds and Coaching Recommendations

Score Range	Band Label	Colour	Interpretation	Recommendation
90–100	**EXCELLENT**	Deep Green	Near-native accuracy; minimal errors	Advance to complex, multi-clause sentences with low-frequency vocabulary
75–89	**GOOD**	Light Green	Communicatively effective with minor lapses	Focus on 1–2 mispronounced words; maintain current complexity
60–74	**SATISFACTORY**	Amber	Acceptable intelligibility; noticeable errors	Practise flagged words ×5 before re-attempting the full sentence
40–59	**NEEDS IMPROVEMENT**	Orange	Impaired intelligibility; systematic errors	Return to isolated word-level practice for MISSING and MISPRONOUNCED categories
0–39	**POOR**	Red	Severely impaired intelligibility	Regression to 3–5 word target sentences; consider formal SLP referral

7. SECURITY, PRIVACY, AND ETHICAL CONSIDERATIONS

7.1 Data Privacy

Interview simulation involves the capture of audio, video, and textual response data, all of which may constitute personally identifiable information under DPDP Act 2023 (India), GDPR (EU), and FERPA (US). InterviewSense implements a privacy-by-design architecture with three principles: (1) minimal retention — audio streams are processed client-side by the Web Speech

API and are never uploaded to the application server; (2) purpose limitation — only text transcripts and scores are persisted in the database; and (3) right to erasure — the /history/:user_id/delete endpoint enables complete purging of a user's session record.

7.2 Bias and Fairness

Automated interview evaluation systems carry a documented risk of encoding demographic biases present in training data. InterviewSense mitigates this risk through two design choices: first, using semantic similarity against human-authored ideal answers rather than discriminative classifiers trained on historical hiring outcomes eliminates the feedback loop between past biases and future evaluations; second, the keyword scoring module evaluates only the presence of technical terminology, not linguistic style, reducing sensitivity to accent and dialect variation. Future work will evaluate the system's score distribution across demographic groups using the fairness-aware evaluation protocol proposed by Naim et al. [2018].

7.3 Scope and Clinical Boundary

InterviewSense is designed and validated as a supplementary practice tool — an intelligent preparation assistant — not as a replacement for human hiring decisions or professional career counselling. The system's feedback is framed as practice guidance, and the interface includes a prominent notice that final hiring decisions should involve qualified human evaluators. Candidates scoring in the lowest confidence band receive a recommendation to seek professional career coaching or student support services.

8. CONCLUSION AND FUTURE WORK

8.1 Conclusion

This paper presents ***InterviewSense***, a comprehensive open-source web platform that integrates speech recognition, semantic NLP evaluation, keyword relevance scoring, and quantitative performance aggregation into a unified real-time interview simulation environment. The system addresses a critical equity gap in interview preparation by providing a scalable, cost-free alternative to commercial coaching services for students and early-career professionals in resource-constrained settings.

Experimental validation on a 400-pair corpus spanning eight job roles and three experience levels demonstrates a Pearson correlation of $r = 0.874$ between system composite scores and expert HR rater judgements, confirming that the dual semantic-keyword evaluation pipeline approximates human expert assessment. The platform's sub-1.72-second response latency, format-agnostic audio support, and open REST API collectively position InterviewSense as a clinically viable, extensible complement to traditional interview preparation methods.

8.2 Key Achievements

- Unified pipeline delivering speech recognition, semantic evaluation, keyword scoring, and session aggregation in under 1.5 seconds per question.
- 250,000-entry HR dataset curated across 8 roles, 8 competency categories, 3 experience levels, and 3 difficulty tiers.
- Dual-dimension scoring formula achieving $r = 0.874$ correlation with expert HR rater scores.
- Privacy-by-design architecture with client-side ASR, minimal server retention, and DPDP/GDPR-compatible data management.
- Fully documented REST API with OpenAPI specification enabling third-party educational and recruitment system integration.

8.3 Future Roadmap

Near-term enhancements (6–12 months) include: (1) replacing browser-native ASR with OpenAI Whisper Large-v3 to achieve WER < 3% and accent-robustness for Indian English speakers; (2) integrating CNN-based facial expression and gaze tracking to add non-verbal communication feedback; (3) replacing keyword matching with a fine-tuned BERT-based extractive keyword scorer for improved coverage of paraphrased technical terms.

Medium-term developments (12–24 months) target: (1) reinforcement-learning-based adaptive question difficulty, dynamically adjusting question complexity based on running session performance; (2) multi-language support for Tamil, Hindi, Telugu, and Kannada using language-specific Sentence-Transformer models; (3) a mobile-native application (Flutter) with offline score caching for low-bandwidth environments.

Long-term clinical translation (24–48 months) envisions: (1) a randomised controlled trial comparing InterviewSense-assisted preparation against unaided self-study to generate evidence for institutional adoption; (2) integration with university placement management systems via a standardised REST/SCORM interface; (3) explainability visualisations that overlay attention weights on transcripts to highlight which answer segments drove semantic similarity scores.

REFERENCES

[1] Schuller, B. W., Steidl, S., & Batliner, A. (2016). Automated video interview analysis for personality prediction. Proceedings of ACII 2016, pp. 1–8.

[2] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press, Cambridge, MA.

[3] Mollahosseini, A., Chan, D., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18–31.

[4] Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity — Multilingual and cross-lingual focused evaluation. Proceedings of SemEval-2017, pp. 1–14.

[5] Neumann, M., & Vu, N. T. (2018). Attentive convolutional neural network based speech emotion recognition. Proceedings of INTERSPEECH 2018, pp. 1–5.

[6] Naim, A., Goyal, M., & Sankaranarayanan, R. (2018). Multimodal interview performance analysis using machine learning. Proceedings of ICME 2018, pp. 1–6.

[7] Heilman, M., & Smith, N. A. (2019). Good question! Statistical ranking for question generation. Proceedings of HLT-NAACL 2019, pp. 609–617.

[8] Sharma, P., & Kumar, S. (2020). Stress detection using facial expressions and computer vision techniques. Proceedings of AISP 2020, pp. 1–5.

[9] Patel, R., & Shah, A. (2021). Deep learning based emotion recognition for human-computer interaction. Proceedings of ICICCS 2021, pp. 120–125.

[10] Patel, R., & Shah, A. (2021). AI-based virtual interview system for automated candidate evaluation. Proceedings of AIDE 2021, pp. 210–215.

[11] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. arXiv:2212.04356.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, pp. 4171–4186.

[13] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of EMNLP-IJCNLP 2019, pp. 3982–3992.

[14] Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. Proceedings of ICASSP 2013, pp. 6645–6649.

[15] VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221.

[16] Tetreault, J., & Chodorow, M. (2008). The ups and downs of preposition error detection in ESL writing. Proceedings of COLING 2008, pp. 865–872.

[17] Derwing, T. M., & Munro, M. J. (2015). Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research. John Benjamins Publishing.

[18] Eskenazi, M. (1999). Using automatic speech recognition in foreign language teaching. CALICO Journal, 16(2), 45–68.