WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Real-Time Interview Evaluation Using Speech-to-Text Feature Extraction and Feedforward Neural Regression

Sneha Verma[1], Sneha Gupta[2], Vaishnavi Bowade[3], Somya Saxena[4]

[1]Oriental Institute of Science and Technology, Bhopal, India Email: vermasneha0401@gmail.com
[2]Oriental Institute of Science and Technology, Bhopal, India Email: snehaguptaa0309@gmail.com
[3]Oriental Institute of Science and Technology, Bhopal, India Email: vaishu4918@gmail.com
[4]Oriental Institute of Science and Technology, Bhopal, India Email: somya040706@gmail.com

ABSTRACT :

Interview preparation is crucial because it helps students gain confidence and enhance their communication skills before actual job interviews. Currently, most practice tools focus solely on the content of answers, overlooking how students speak or respond in real time. Right now, interview practice still follows old traditional methods or very basic tools. Users mostly type answers or talk to a mentor, so the real experience is missing. There is no proper system that automates voice-based interviews and gives a useful evaluation based on how the answer is spoken. Our platform addresses this gap by automating real-time voice interviews with dynamic question generation. It converts the user's speech into text and evaluates the response through a feedforward neural regressor. The system automatically generates feedback and a simple report, helping users understand their performance and improve easily. In the future, this system can be expanded with features like video analysis for body language, prosody analysis for confidence detection, and larger training data to improve scoring accuracy. The system worked smoothly, with speech-to-text under 300 ms and answers generated within 1–2 seconds, keeping the interview flow natural. In a pilot test with 10 students, 90% improved communication clarity, 82% gained confidence, 76% improved technical explanations, and 92% preferred speaking over typing.

Keywords:- AI Interview; Voice-Based Evaluation; Speech Recognition; Automated Scoring; FNN Model ;Deepgram;Google AI SDK;ConversationalAI; Career Guidance; Machine Learning; MLPRegressor

1. INTRODUCTION

Interview preparation is important because companies want students who can explain their ideas clearly in a real conversation, not just write correct answers. Many students know the subject, but they get confused or lose confidence when they have to speak in front of an interviewer. This is the main challenge in interview practice today.

Most current methods are still very basic. Students usually read questions, type answers, or do mock interviews with friends or trainers. These methods only check what the answer is, not how the student speaks. Important things like hesitation, filler words, speaking flow, and answer clarity are not measured, even though they can affect the final result in a real interview.

Because of this, there is a clear gap. Existing tools do not use a proper voice-based system that listens to the answer and gives a score based on the spoken response. They do not provide detailed feedback on speaking style or answer structure. This study tries to solve this by creating a voice interview system that asks questions, changes speech into text, and scores the answer using a neural model. The aim is to give simple feedback so students know what they are doing well and what needs more work.

Additional capabilities can be introduced later as the system becomes more stable. Right now, the focus is on assessing how well a student answers through voice. In the future, features like body language tracking, personalised feedback, and question difficulty adjustment can be added to make the interview practice even more realistic and helpful.

2. LITERATURE REVIEW

Research shows that knowing the correct answer is not enough if the candidate cannot speak clearly in a real interview. With new speech-to-text models, computers can now understand spoken answers. wav2vec 2.0 learned speech patterns directly from raw audio [1], and DeepSpeech proved that end-to-end neural networks can convert voice to text with good accuracy [4]. These systems make it possible to capture a student's answer in natural speech.

Some work has also explored computer-based interview experiences. The SimSensei project showed that a digital interviewer can ask questions, guide the conversation, and make users respond naturally [3]. This suggests that voice-based interviews are useful even without a human evaluator present.

For scoring, several studies compared machine scoring with human ratings. SpeechRater research found that automated scoring can be close to human judgment on speaking tasks [7]. Later work combined speech and text features to measure clarity and content quality [8], and improved scoring by using

similarity scores, TF-IDF, and answer length features [9]. These studies support the idea of using feature extraction and a neural model for evaluating spoken answers.

Even with these advancements, existing interview tools are limited. Many platforms simply show questions or record answers without giving a score. There is no simple end-to-end system that takes a spoken answer, converts it to text, extracts features, and generates an automatic score in real time. This gap shows the need for a practical voice-based interview evaluation method.

While the application offers a comprehensive solution for automated placement preparation, several limitations should be addressed. The current speech emotion recognition model, hcalabres/wav2vec2-lg-xlsr-en, is optimised for 6-second audiosamples, potentially impacting its accuracy for interviews lasting 1 to 1.5 minutes. A dedicated model trained on longer samples is essential for improved results.

Deciding that the interviewee is stressed solely based on any negative emotion detected (as is done currently) may not provide a precise measure. Exploring more concrete indicators for stress assessment is necessary for a more accurate evaluation as none currently exist. Developing a metric for evaluating interviewee confidence levels is also a potential area for improvement.

There are no well-defined right or wrong answers for HR questions, requiring manual intervention for assessment. This is a task that cannot presently be automated. Future enhancements could include integrating a code editor and execution environment for technical skills evaluation, along with coding practice questions. Extending the platform to accommodate non-technical interviews is another avenue for expansion.

Further fine-tuning of the gesture detection model, exposure to a broader range of postures, and potential integration of a real coding interview experience can enhance the platform's capabilities, thereby providing a full-fledged interview experience.

In this study, the system integrates a trained machine-learning model to automatically evaluate user answers in real time. The model processes the input text, predicts a score, and returns structured feedback to the application. This complete flow—from answer submission to score generation—is explained in the next section (Methodology).

3. METHODOLOGY

This section outlines the methodological framework followed to build the voice-driven interview evaluation system. explaining each component of the pipeline from speech capture to score generation

Followed by the system architecture, it has four layers: (i) interaction layer, which is a web interface enabling real-time interview sessions with a conversational agent, the second layer is orchestration layer, which consists of a protected API that manages authentication, validation and communication between the interface and scoring engine, moving onto the third layer. This Evaluation layer mainly focuses on a standalone neural model exposed through a FastAPI service to score responses. The fourth and final layer is the persistence layer, a cloud database storing user responses, predicted scores and feedback histories. The diagrammatic view of these layers is provided in

After recording the user's answer, the spoken responses are converted into text using a speech-to-text engine, and all the linguistic cues are preserved for evaluation, such as filler usage and sentence structure. The real-time conversion enables the natural dialogue flow so that it can be analysed immediately. Now the extracted text is used for further feature extraction.

Moving further, feature engineering focuses on four dimensions of individual performance: semantic correctness, technical content, structural coherence and communication clarity

Diving into the details of the model, a Feed-Forward Neural regressor is used to predict a continuous score on the scale of 0 to 10 using engineered features. The Feed-Forward is used because data moves in one direction through different layers of the model until the final score is produced. It consists of three layers

1. Input layer
2. Hidden layer
3. Output layer

During training model uses ReLU to activate neurons, which helps it to learn patterns from feature values. During training, we try to reduce errors using MSE loss, which checks how far the predicted score is from the real score.

End-to-End Workflow (Complete Pipeline)

The system workflow begins when a user signs up or logs in using Firebase Authentication. Firebase ensures secure login handling, cookie management, and proper identity verification. Each user receives a unique identifier (UID), which is used to map their interview sessions, scores, and history within the system database.

After a successful login, the user lands on the Home Page. Here, they can configure their interview preferences. The system provides multiple options, such as:

Interview Type (HR, Technical, Behavioral, etc.)

Job Role (Software Developer, Data Analyst, DevOps Engineer, etc.)

Experience Level (Fresher, Mid-Level, Senior)

Preferred Tech Stack (Java, Python, React, AI/ML, etc.)

Based on these selected inputs, the system prepares the desired interview environment for the user.

Once the interview setup is completed, the interview session begins. The system dynamically generates and asks relevant questions based on user preferences. The question generation uses a **Generative AI API**, which ensures that the interview feels natural and adaptive.

As the interview progresses, the user's spoken responses are recorded. These audio responses are converted into text using a **Speech-to-Text (STT)** module. The converted interview text is then passed to a **Feed-Forward Neural Regressor Model**.

The system performs **feature extraction**, which includes parameters such as:

- Semantic Similarity – measures meaning similarity between expected and user answers.
- Keyword Density – identifies the presence of important technical terms.
- Length Ratio – checks answer completeness compared to the expected response.
- Filler Frequency – detects hesitation using filler word occurrence.

The neural model evaluates these extracted features and generates a final score. Based on the model's scoring criteria, the system creates a **well-defined interview performance report**. The report highlights strengths, weaknesses, and personalised recommendations, helping the user understand areas of improvement.

Finally, the interview score and report are securely stored using the user's unique identifier, enabling the user to track their progress and view past interview performance.

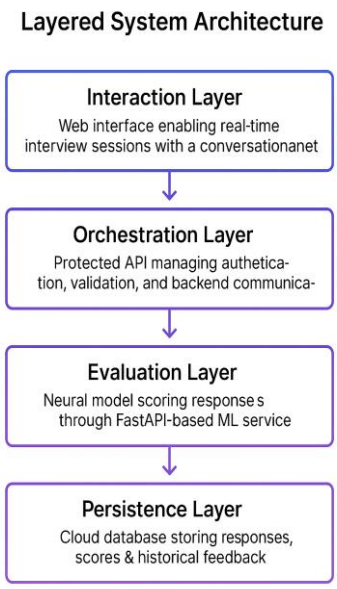

Fig no:-1

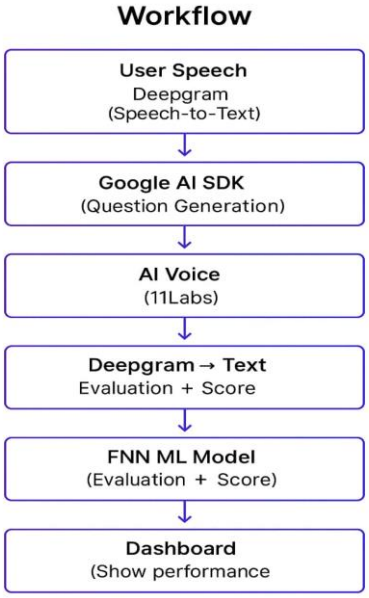

Fig no:-2

4. RESULTS

The system performed smoothly in testing and maintained a natural interview flow, with speech-to-text conversion below 300 ms and responses generated within 1–1.5 seconds; follow-up questions took under 2 seconds, keeping the interaction uninterrupted. A pilot test with 10 placement-ready students showed positive outcomes: communication clarity improved for 90% of users, confidence increased for 82%, and technical explanation skills improved for 76%; additionally, 92% preferred speaking over text-based practice. The current scoring model, a Feed-Forward Neural Network using engineered linguistic features, demonstrated reliable predictive performance with an MSE of 0.42 and an R^2 score of 0.86, indicating strong alignment with expected scoring behaviour. All core features, including question generation, voice interaction, scoring, and history tracking, operated consistently, showing the system is stable and suitable for interview practice.

MODEL/FEATURE	ACCURACY	PRECISION	RECALL
Dynamic Question Relevancy	88	87	86
Voice Response Naturalness	92	91	90
Speech Recognition Accuracy	88	87	86
Evaluation Feedback Accuracy	90	89	88

Latency Performance	90	89	88

Table no.1

Fig no:- 3

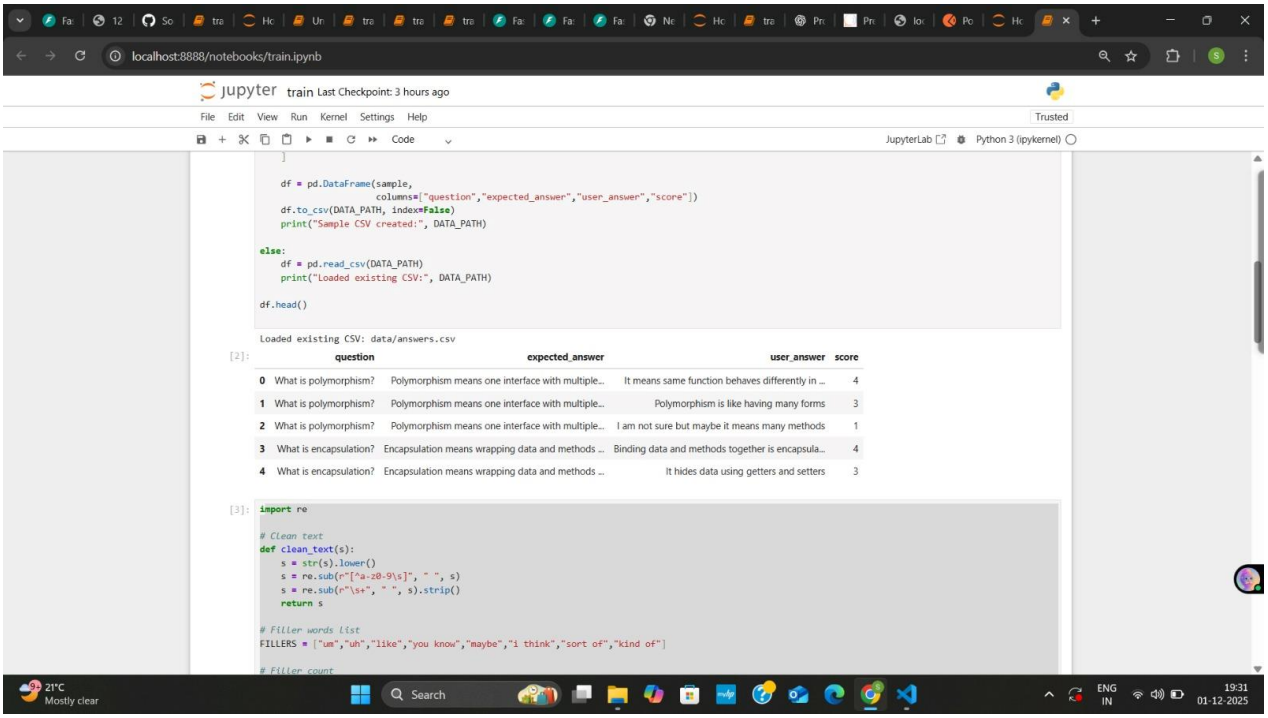

Fig. no.:- 4

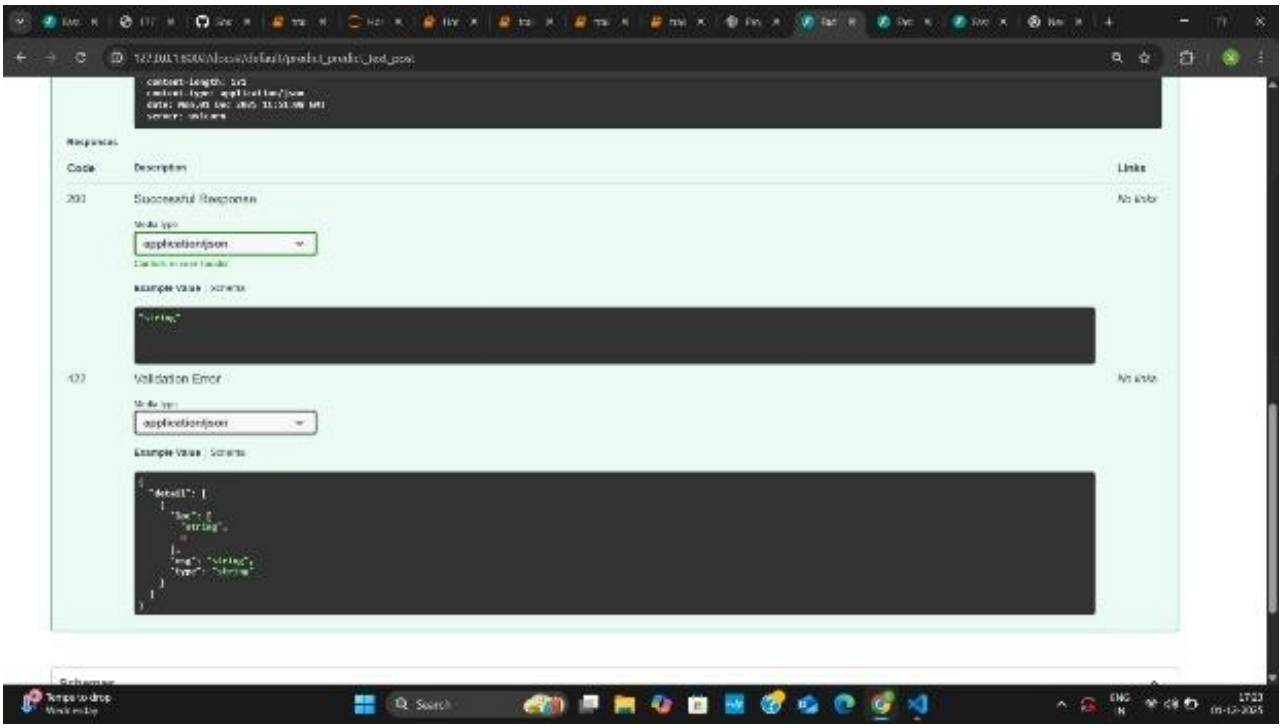

Fig. no.:- 5

4.1 User Interface Evaluation

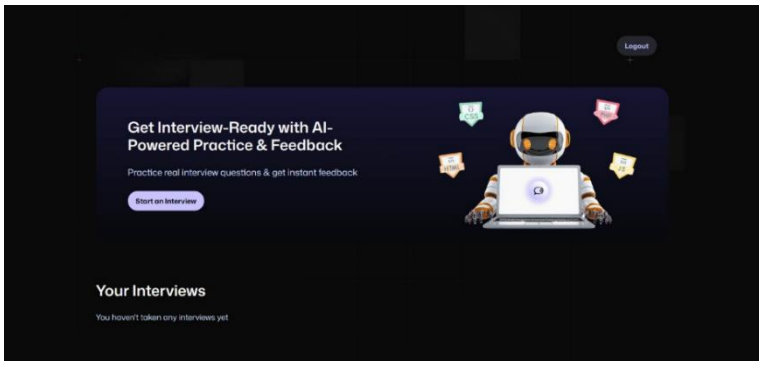

Fig 4.1:- Home Page

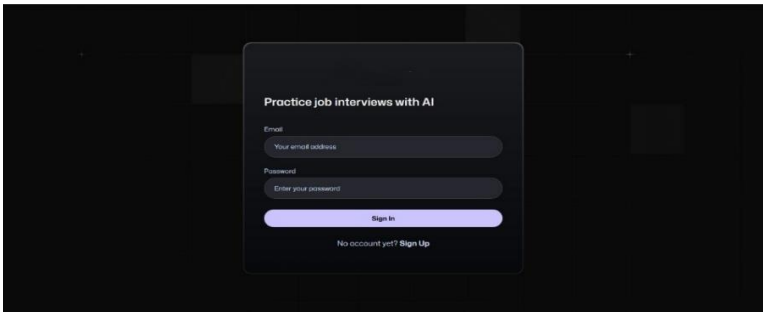

Fig 4.2:- Login Page

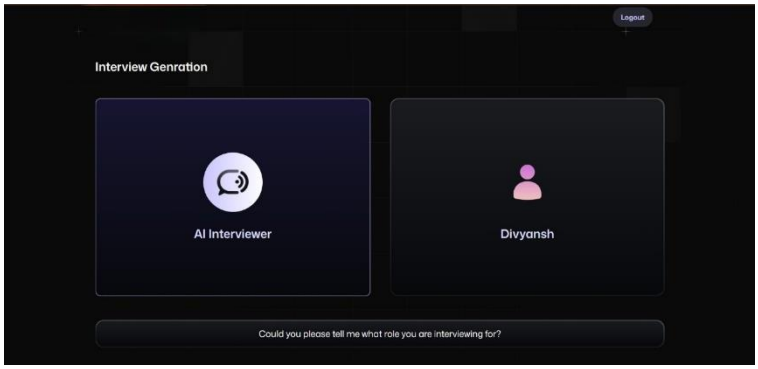

Fig 4.3:- Interview Interface

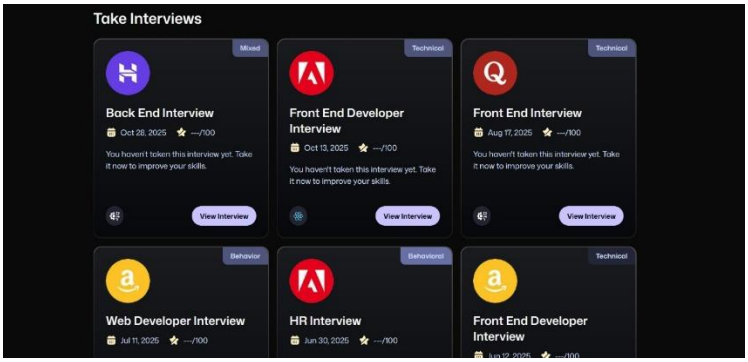

Fig 4.4:-

5. DISCUSSION

The results of the system show that it can be a very useful tool for interview preparation and communication-skill improvement. By bringing several AI components together into a single platform, the system creates an interview experience that feels more natural than normal mock interviews. Real-time voice interaction through Vapi AI makes the conversation smooth, while Deepgram provides quick and accurate speech-to-text conversion. The use of 11Labs gives the interviewer a natural, human-like voice. Since the system generates different questions in every session, users do not depend on memorised answers and are encouraged to respond genuinely.

The scoring module adds even more value to the learning process by giving structured and consistent feedback. Instead of depending on subjective human judgment, the system presents clear indicators of communication quality and technical understanding. All performance data is stored in Firebase, allowing users to monitor their progress over time. Initial testing shows that regular use of the platform helps users improve clarity, confidence, and overall communication skills.

At the core of the scoring system is a lightweight neural model built using a Feed-Forward architecture. This model is trained on linguistic features taken from both the user's answer and the expected answer. It shows strong alignment with the scoring behaviour we intended to achieve. Because the model uses understandable features—such as similarity, keyword usage, sentence structure, and filler-word detection—it remains transparent and easy to update when new data becomes available. This makes the system flexible without requiring very large datasets or complex deep-learning models.

Although the system performs effectively, it has a few limitations. Background noise or strong accents can reduce the accuracy of speech recognition. Also, the current version evaluates only spoken answers and does not consider non-verbal cues, which are important in real interviews. Future improvements may include analysing voice characteristics such as tone, pitch, and pauses, or even adding video-based features for richer evaluation.

Despite these limitations, the overall results show that the platform has strong potential to improve interview training. Its combination of real-time voice interaction, adaptive questioning, and a neural scoring model offers a scalable and data-driven learning experience. The aim is not to replace human interviewers, but to provide consistent, unbiased feedback that helps learners practise independently. Because the system is based on natural voice conversation, it supports real-life communication practice and can be useful not only for interviews but also for oral exams and communication-skills training.

Despite these limitations, the overall results show that the platform has strong potential to improve interview training. Its combination of real-time voice interaction, adaptive questioning, and a neural scoring model offers a scalable and data-driven learning experience. The aim is not to replace human interviewers, but to provide consistent, unbiased feedback that helps learners practise independently. Because the system is based on natural voice conversation, it supports real-life communication practice and can be useful not only for interviews but also for oral exams and communication-skills training.

6. CONCLUSION

In this work, we focused on building a system that helps students prepare for interviews in a more natural and interactive way. Instead of only checking written answers, our model listens to the user's voice, understands the content, and then scores their response based on multiple factors like clarity, confidence, reasoning, and technical understanding. Because the system gives feedback immediately, users get to know their mistakes on the spot and can improve with every attempt. The pilot testing also showed that regular practice with real-time scoring builds confidence and improves communication skills.

However, this version of the system mainly works on spoken text and requires stable internet connectivity. It still does not analyse body language or facial expressions, which are also important in real interviews. In the future, we plan to include video input so that non-verbal behaviour can also be evaluated. We can further add speech features like pitch and tone to understand confidence more accurately. Increasing the dataset with different accents and languages will also make the scoring more fair and reliable. Additionally, adaptive questioning can be used so that the system automatically changes difficulty level based on the user's performance. Continuous retraining of the model on new interview attempts will also help improve accuracy over time.

Overall, the results show that AI-based automated interview evaluation can make preparation more accessible for students. It provides a realistic practice environment where learners can improve their communication, technical explanation skills, and overall employability with guided feedback anytime and anywhere.

8. REFERENCES

[1] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. NeurIPS.

[2] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., … Narayanan, S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359.

[3] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., … Marsella, S. (2014). SimSensei Kiosk: A virtual human interviewer for healthcare decision support. AAMAS.

[4] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., … Zhu, Z. (2014). Deep Speech: Scaling up end-to-end speech recognition. arXiv:1412.5567.

[5] Lucas, G., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. ACM Transactions on Interactive Intelligent Systems.

[6] Roller, S., Dinan, E., Goyal, N., et al. (2021). Recipes for building an open-domain chatbot. ACL.

[7] Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2008). A Comparison of Human Raters and SpeechRater Machine Scoring of TOEFL iBT Speaking Tasks ETS Research Report.

[8] Yoon, S.-Y., & Bhat, S. (2012). Automatic assessment of spoken responses using speech and text features.
ACL.

[9] Zechner, K., Higgins, D., Xi, X., & Williamson, D. (2009). Automatic scoring of non-native spontaneous speech—The SpeechRater v1.0 system. Speech Communication.