2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) | 979-8-3503-2891-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISMAR-Adjunct60411.2023.00102

Manifest: Public Speaking Training Using Virtual Reality

Haania Siddiqui* Hafsa Irfan† Aliza Saleem Lakhani‡ Batool Ahmed§ Sadaf Shaikh¶
Muhammad Mobeen Movania ‖ Muhammad Farhan **

Habib University

Abstract

Public speaking is pivotal in personal and professional realms, evaluating communication skills. Employing VR training offers realistic practice, identifying weaknesses, improving speaking, and fostering confidence to overcome the fear of public speaking. This paper introduces a novel VR-based public speaking training system that comprehensively analyzes the speaker's speech, body language, and nervousness levels. The system requires the use of a VR headset that immerses the user in a virtual classroom setting where they deliver speeches, while simultaneously recording the speech, capturing video through an external camera, and monitoring the user's heart rate using a dedicated sensor. To evaluate the speech quality, the recorded speech is processed to calculate the scores of speaking rate, clarity, listenability, pronunciation, and number of pauses during the speech. The recorded video is further utilized to analyze body language, employing a pose estimation framework-YOLOv7. Furthermore, heart sensor data is employed to calculate the user's level of nervousness. The system provides detailed analysis for each training session, providing users with performance feedback that can be accessed through the website. By integrating immersive VR technology, speech analysis, body language assessment, and heart rate monitoring, this novel public speaking training application provides a comprehensive way to evaluate public speaking skills and receive feedback and thus is a viable solution for coaching in the field of public speaking.

Keywords: Virtual Reality Application, User Experience, Human-Computer Interaction, Public Speaking Training, Speech Analysis, Motion Tracking, Virtual Reality Development, User Interface

Index Terms: Human Computer Interaction—Public Speaking Training system

1 Introduction

In the modern era, effective communication skills, particularly public speaking, have become more crucial than ever before. It has become an essential requirement in various domains of professional life including law, banking and finance, counseling, teaching, public relations, politics, and sales. Public speaking, while vital for effective communication, remains one of the most significant and simultaneously feared forms of communication; the anxiety and apprehension associated with addressing an audience are often visible in subtle ways during presentations. This can include poor body posture, avoiding eye contact, and speaking without proper breath control [23]. In more severe instances, the anxiety surrounding public speaking can make these signs more apparent to the audience, such as mental blocks or loss of words, trembling hands or shaky legs, and excessive sweating.

Public speaking is an art and just like any other art one needs practice or training to master it [22]. Recent research has explored the use of Virtual Reality (VR) technology for training in public speaking. It offers a significant advantage in creating a strong sense of presence and embodiment within a virtual yet realistic environment [45]. That allows trainees to make mistakes within a safe learning environment, without facing real-world repercussions. It also eliminates the limitations of finite practice opportunities and enables individuals to confront and overcome their fears. VR applications, specifically in the field of public speaking, have shown promising results in reducing the fear of speaking in front of an audience. Research [32, 43] indicates that the increased confidence gained through VR-based interventions can be transferred to real-world scenarios and maintained even after the therapy concludes. Some existing solutions in the space are Ovation VR [9] which gives voice and text analysis and real-time feedback, Virtual Reality Speech Training (VR-ST) with direct feedback [39], a system built using Virtual Reality technology, Video 360, and Arduino heart sensors to overcome stage fright [28].

However, current Virtual Reality (VR) applications for public speaking training predominantly focus on speech or body analysis. Additionally, users are often required to remain in the VR environment to access their feedback after delivering their speech. This paper proposes a public speaking training application that utilizes VR to create a virtual audience and offers analysis and feedback on the trainee's speech, body language (using pose estimation), and nervousness through various algorithms. The main motivation of the paper is to show how a distributed system can be used to create an effective VR system for public speaking training, based on a comprehensive analysis of the speaker's speech, body language, and levels of nervousness. Our goal was to develop a solution prioritizing user-friendliness and practicality while avoiding unnecessary complexity. This approach is designed to operate seamlessly in diverse settings, considering the potential affordability limitations of Oculus Rift (or other headsets). This consideration prompted us to envision scenarios where multiple users could share a single headset, prompting us to explore alternatives to wearing the headset continuously. This is also of particular significance due to the potential discomfort and even VR sickness that may arise from prolonged VR headset use [19]. In light of this, we propose a solution where scores can be accessed outside the VR environment, offering a more comfortable and accessible approach.

Our contributions are as follows:

- An end-to-end system with a virtual environment and feedback on speech, body language, and the nervousness of a user.
- Automated analysis of each user on received audio, video, and heart rate using different algorithms to give feedback on the website.
- Integration of various hardware and software; using a headset to record audio, website camera to record video, and Arduino sensor integrated into a glove to measure heart rate specifically for public speaking training.

*email: hs06188@st.habib.edu.pk
†email: hi05946@st.habib.edu.pk
‡email: al05435@st.habib.edu.pk
§email: ba06180@st.habib.edu.pk
¶email: ss06054@st.habib.edu.pk
‖email: mobeen.movania@sse.habib.edu.pk
**email: muhammad.farhan@sse.habib.edu.pk

2771-1110/23/$31.00 ©2023 IEEE
DOI 10.1109/ISMAR-Adjunct60411.2023.00102

- Calibration of scores using Gaussian distribution to align clarity and speaking rate scores with local accents, thus enhancing the interpretation of these audio characteristics.

The subsequent sections of the paper are structured in the following manner: Section 2 discusses the existing research conducted in the field, Section 3 explains the system and methodology employed in the different modules: speech, body, and nervousness, Section 4 highlights the limitations and our plans for future work, and finally Section 5 concludes the paper.

2 Related Work

In this section, we provide a comprehensive review of existing solutions for public speaking training. Additionally, we delve into the existing methodologies employed for automatic speech recognition, and analysis of body language and nervousness.

2.1 Public Speaking Training

Public speaking training solutions can be categorized into two groups: those with Virtual Reality and those without. The majority of studies focusing on public speaking training without Virtual Reality provide feedback on either speech content and delivery [5,11,12], or solely on body language [35]. Ummo and LikeSo are personalized speech coaching applications that aim to analyze speech and provide real-time feedback on the usage of filler words, word repetition, clarity, volume, length of pauses and pace [5,12]. Using speech analysis algorithms, it mainly applies text analysis after speech transcription to determine pace (words per minute), filler words spoken, and total words spoken. Speaker Coach, by Microsoft, helps users rehearse their presentation and gives feedback in real-time on their pace, pitch, filler words, usage of informal speech and culturally sensitive terms, and how wordy the speech is [11]. After each rehearsal, it generates an analysis report on their performance. By developing an automated scoring model for public speaking performance, Lei Chen et al [20] found that using multi-modal features i.e. speech content, speech delivery, and hand, body, and head movement, together can significantly predict human scores on the presentation performance. [35] presents an online feedback system to evaluate public speaking by detecting emotions through body language, including fear and nervousness. This system utilizes a classification method with gestures and posture as input, offering feedback using a five-point scale from poor to excellent performance.

The literature for public speaking training in virtual reality has shown a gradual progression from training in front of a projected screen displaying a life-size audience [18,22,47] to using headsets as a more immersive experience in Virtual Reality [9,14,28,38,39]. Ovation and VirtualSpeech are virtual-reality-based public speaking training platforms that focus on creating as realistic environments in Virtual Reality (VR) as possible. The former gives real-time feedback, detects gaze patterns, does voice analysis to extract monotone, pauses, and mic distance and text analysis to extract filler words and pace, and tracks hand movement; the latter gives post-session feedback, measures eye contact and speaking pace, and detects and counts filler words. In the context of seamless transitions between diverse realities within virtual environments, recent advancements have been notable. Notably, a comprehensive exploration comprising two consecutive studies was conducted. This research [24] delved into the intricate dimensions of a system that facilitates uninterrupted shifts between realities without necessitating users to remove their headsets. A qualitative study on the effectiveness and usefulness of VirtualSpeech showed that the students who reported high levels of anxiety had decreased levels of nervousness after the training session, appreciating the realism and fail-safe feature of the environment [17]. Feedback in Ovation is displayed as a grade sheet to identify areas of improvement. While it offers analytics as another feature, multiple settings on the screen might make the user feel overwhelmed. Another study uses SIRVIGLOSS as an application that allows users to wear a headset, measures heart rate using a heart rate sensor, and determines the attentiveness of the virtual audience based on the heart rate [28]. The higher the heart rate, the less attentive the audience is. The results showed that after each simulation run, the number of times the users faced an inattentive audience declined.

2.2 Speech, Body Language and Nervousness Analysis

In this subsection, a comprehensive overview of the literature that leverages state-of-the-art techniques for the automated analysis of speech, body language, and nervousness is presented.

Recent advancements, as discussed in [31], have introduced End-to-end (E2E) systems to the realm of ASR. The work [31] focused on three prominent E2E techniques for ASR: Connectionist Temporal Classification(CTC), Attention-based Encoder-Decoder (AED), and Recurrent Neural Network Transducer (RNN-T) where RNN-T outperformed the other two. Another work [40] introduced a new model 'Branchformer' utilizing convolution, Multi-headed self-attention, and a multi-layer perceptron neural network. The model gave better results than the standard transformer and cgMLP (MLP module with convolutional gating) and performed relatively similarly to the Conformer model. In [51], Attention-Based CTC is used which primarily solves two main end-to-end LVCSR problems: data alignment problem and directly outputting the target transcription. RNN-transducer can map the input sequence to an output sequence of arbitrary length regardless of the length of the input. On the other hand, a model based on attention can indirectly learn the soft alignment between the input sequence and the output sequence. Lastly, a study highlighted by [41] showcased a robust approach in automatic speech recognition (ASR). Previous models focused on unsupervised pre-training, using larger unlabeled datasets for testing. However, they often neglected decoder optimization, limiting their utility. In contrast, ASR systems with supervised pre-training across domains demonstrated better generalization. The study bridged this gap, scaling weakly supervised ASR to 680,000 hours of labeled audio data, pre-training multilingual and multitask components. The model, named Whisper, consistently outperformed state-of-the-art models across various evaluations, approaching human accuracy. The study suggests further research to improve performance in languages with limited resources.

The review for body analysis focused on approaches and cutting-edge models for body analysis (tracking) and Pose estimation. Human Pose Estimation (HPE) from monocular images classify pose estimation techniques into two different categories [21,25]: generative and discriminative. The former starts with a pose and projects it onto the image while the latter starts from the image evidence and typically learns a mechanism modeling the relations between image evidence and human poses based on training data. Other surveys classify pose estimation into bottom-up and top-down approaches [25,34]. Top-down approaches first use a bounding box object detector to detect a human body instance and then focus on detecting their pose while bottom-up first detects the body parts and then groups them into human body instances [34]. The surveys [25,34,36] discussed deep learning-based methods for pose estimation. Two of the methods discussed were DeepPose and ConvNet Pose. DeepPose is the first significant research paper that utilized deep learning methods for human pose estimation by using AlexNet as backbone architecture [49]. In ConvNet, CNNs are used to generate heatmaps to predict the position of joints [48]. Authors in [52] introduced BlazePose which used a top-down approach that was further improved in [27]. The model's performance was compared to other state-of-the-art approaches and it outperformed those. However, being a top-down approach it cannot detect the pose of the human body if the person's eyes are occluded [34]. Another pose estimation model YOLOv7 [50], employs a bottom-up approach (unlike BlazePose), detecting landmarks across the entire image and then identifying the person based on these landmarks. Previous studies have measured nervousness as a basis to present different scenarios of audience attentiveness [28] or tracked body language to extract emotional states such as fear and nervousness [35]. In [28], Arduino Rev3 is used to send heart rate readings via Bluetooth module to the SIRVIGLOSS application. After averaging values across 10 seconds, if the heart rate sensor, touching either the tip of the ear or the fingertip, detects a normal heartbeat (60-100 bpm) then the first scenario is run where the audience is paying attention and is silent. If the heart rate sensor detects an abnormal heart rate where the user is labeled as nervous (greater than 100 bpm), the second scenario in the application is run where a display of an audience appears that does not pay attention to the user, looks sleepy, or is busy. If the user experiences nervousness for more than 30 seconds, the third scenario is run where the virtual audience is very crowded and cheers "boos" to the presenter. These scenarios then repeat themselves based on the heart rate detected unless the simulation time is ended. In [35], Microsoft Kinect is used to offer both colored video and motion information to generate emotional signals from body language.

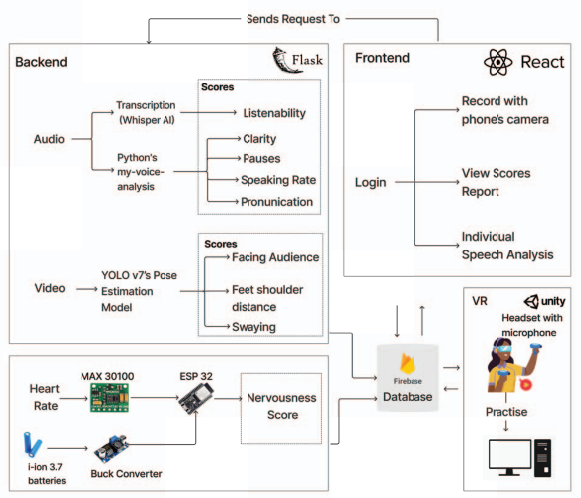


Figure 1: System Flow Diagram of the VR Application

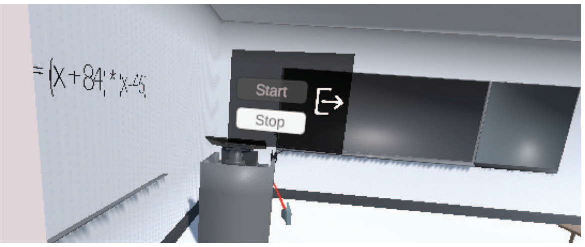


Figure 2: Start/Stop buttons for a training session

3 Manifest System Overview

This section entails a systematic explanation of the various system components, their interaction, and the techniques utilized. Section 3.1 gives an overview of the system architecture, and the inflow, and outflow of data between the components. Sections 3.2 and 3.3 provide us with the methods employed to gather useful data from the user and technologies used to set up the VR environment and the website component, respectively. Sections 3.4, 3.5, and 3.6 cover the various data analysis techniques for evaluating the quality of the speaker's speech, body language that is then applied to the incoming data to provide feedback for the training of the user. The prototype is developed specifically for use with the Oculus Rift headset [8] connected to a PC, with NVIDIA GeForce GTX 790 graphics card, operated on the Microsoft Windows 10 Operating System.

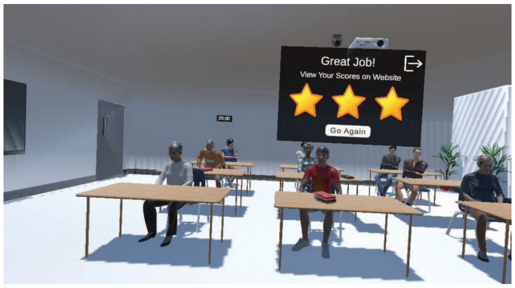


Figure 3: Classroom view in VR

3.1 System Design

Figure 1 provides an overview of our application design and the flow of data for each component. The system is a combination of a web interface and Virtual Reality (VR). Within the VR environment, a realistic classroom setting is simulated, collecting data from multiple sources: speech via the Oculus Rift headset's microphone [8], body movements analyzed from recorded video through image processing, and user heart rate from a glove sensor. Subsequently, the website component analyses the acquired training data using selected techniques and presents an insightful analysis. Details regarding the hardware and software employed for constructing each component are discussed within their corresponding subsections.

3.2 Virtual Reality (VR) Environment

Figure 3 shows our VR environment. The environment was developed using Unity3D [13] game engine. A classroom environment is created containing nine human models, each performing different actions (clapping, head down, rubbing arms). For environment assets such as classrooms, shelves, and books, freely available 3D models from CGTrader [1] and SketchFab [44] are used. The human models were created using MakeHuman [33], and Mixamo [6] was employed to incorporate lifelike animations.

The VR application commences with an initial non-VR login screen. After a successful login, users are transported to a virtual classroom setting that demands the use of a VR headset for complete immersion and exploration. Inside this virtual environment, a pair of buttons labeled 'Start' and 'Stop,' are displayed on the left of the user, as depicted in Figure 2. Clicking the 'Start' button triggers the start of the session and recording of the user's speech, body language, and heart rate, while the 'Stop' button, once clicked, concludes the recording and saves it. The recorded speech is then securely transmitted to the cloud storage platform, Firebase [3], where it is stored in the database.

3.3 Website Component

The web component of the VR application, as shown in depicted in Figure 1, is constructed using React [10] and Flask [4]. After logging in, users are guided to the homepage, presenting an overview of the application's functionalities. Once the device's camera is activated through the website, video recording will automatically initiate, capturing the user's body language, when the user commences a session within the VR application. To monitor progress and evaluate performance, users can navigate to the Report section, which offers a comprehensive overview of their scores for the latest completed training session. Additionally, users can track their progress across multiple training sessions via graphs accessible on the website, showcasing each evaluation metric, as illustrated for the pronunciation metric in Figure 4. An additional feature to get an analysis of their audio is also available on the website for those with no access to the VR headset. This feature ensures flexibility and convenience in utilizing the speech training capabilities of the application.

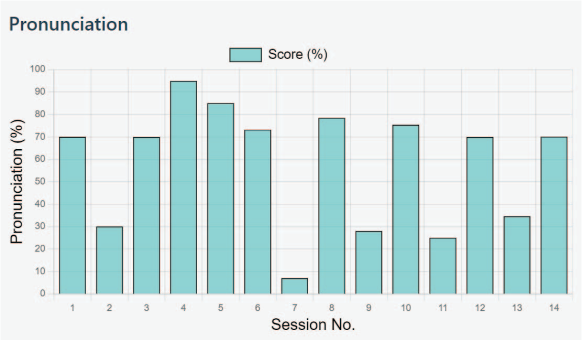
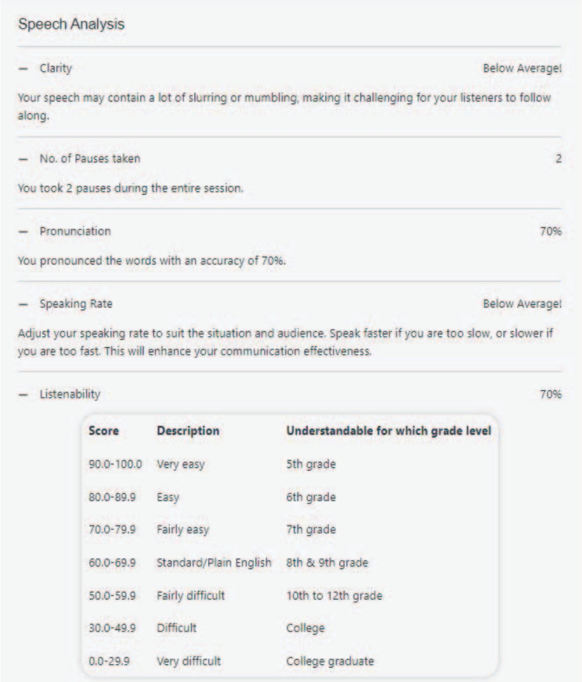


Figure 4: Visualization of pronunciation scores over four training sessions



Speech Analysis

— Clarity Below Average!

Your speech may contain a lot of slurring or mumbling, making it challenging for your listeners to follow along.

— No. of Pauses taken 2

You took 2 pauses during the entire session.

— Pronunciation 70%

You pronounced the words with an accuracy of 70%.

— Speaking Rate Below Average!

Adjust your speaking rate to suit the situation and audience. Speak faster if you are too slow, or slower if you are too fast. This will enhance your communication effectiveness.

— Listenability 70%

Score	Description	Understandable for which grade level
90.0-100.0	Very easy	5th grade
80.0-89.9	Easy	6th grade
70.0-79.9	Fairly easy	7th grade
60.0-69.9	Standard/Plain English	8th & 9th grade
50.0-59.9	Fairly difficult	10th to 12th grade
30.0-49.9	Difficult	College
0.0-29.9	Very difficult	College graduate

Figure 5: Speech Results on Website

3.4 Speech Analysis

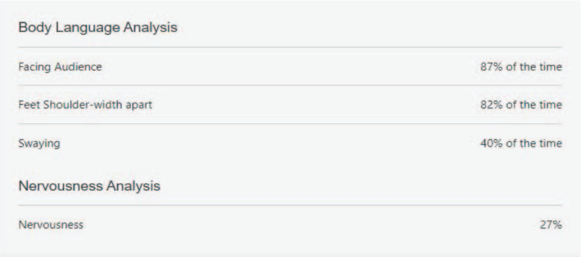
The user's voice is recorded using the integrated microphone in the Oculus Rift [8] headset. Figure 5 illustrates the utilization of five evaluation metrics (or scores) to analyze speech delivery. The clarity score measures the articulation rate of an audio signal by calculating the number of pronounced syllables per second, excluding pauses. This score offers insights into the speaker's clarity of speech. The "speaking rate" quantifies the number of syllables pronounced per second in an audio signal, providing information about the pace of speech delivery. The "pronunciation" evaluates the number of pronounced syllables in an audio signal by detecting vocalic segments. The "number of pauses" quantifies the occurrences of pauses (moments of silence) exceeding a specific threshold duration within the audio signal. The "listenability" assesses the comprehensibility of speech using the Flesch Reading Ease algorithm [29], which determines the ease of reading and comprehension. The algorithm utilizes the transcribed text to determine the level of comprehensibility of the speech. This is accomplished by calculating the count of sentence length and syllables per word. To analyze the recorded voice, the Whisper [42] model was used to transcribe the speech to text, enabling subsequent text analysis. We chose Whisper for two reasons. Firstly, it provides state-of-the-art performance with high accuracy. Secondly, its training encompassed multilingual data, aligning with our utilization of data from local English speakers with distinct local accents. The transcribed text from Whisper is then evaluated using the Flesch Readability Ease [29] algorithm to calculate the listenability score, providing insights into the text's comprehensibility and the ease with which the speaker's speech can be understood. For the calculation of clarity, speaking Rate, pronunciation, and pauses, the Python library 'my-voice-analysis' [7] is employed. This library includes specific functions and algorithms tailored for voice analysis, facilitating the computation of the scores for our system.

A calibration process was performed to align the clarity and speaking rate scores with local accents. Initially, these scores were computed in syllables per second. However, recognizing the influence of local accents on speech characteristics, the scores are adjusted to a more culturally relevant and understandable scale. This calibration process ensures that the scores accurately reflect the clarity and speaking rate within the context of local accents, enabling better interpretation and analysis of the audio characteristics. For the calibration process, a total of 93 audio files were used. To collect the data, we obtained 63 audio files through a form created using Google Forms [26]. Participants were asked to record their voices while reading passages of varying difficulty levels: easy, medium, and hard. Each difficulty level had three passages. Additionally, we curated 30 audio samples from YouTube [15] of local accents.

Provided that the Python 'my-voice-analysis' library calculated clarity and speaking rate in syllables per second, we deemed it necessary that the values generated be calibrated into metrics that are more easily understandable by the users on the website. As such, the metrics were chosen to be either Below Average, Average, or Above Average. The calibration process consisted of 5 steps. Firstly, audio files were categorized and labeled as either Above Average, Below Average, or Average. For each labeled audio file, both clarity and speaking rate was calculated using the Python library. Thirdly, mean and standard deviation was found using the Gaussian distribution for each audio file for the three categories. For each training session, clarity and speaking rate are measured for the audio file generated. Using the mean and standard deviation found earlier, a z-score is generated for each metric. These z-scores represent the deviation from the mean in terms of standard deviations. Using the z-table, probabilities are computed for all three categories; the metric (either Above Average, Below Average, or Average) with the highest probability is assigned to that measure. Thus, as shown by Figure 5, the two scores (clarity and speaking rate) are evaluated as categories (Below Average, Average, Above Average). Other scores, such as pronunciation and listenability, are presented as percentages.

3.5 Body Analysis

Body language is an equally important aspect of public speaking for capturing and engaging the audience's attention. Previous research on effective body language [16, 37], influenced the selection



Body Language Analysis

Facing Audience	87% of the time
Feet Shoulder-width apart	82% of the time
Swaying	40% of the time

Nervousness Analysis

Nervousness	27%

Figure 6: Body Language and Nervousness Results on Website

of three distinct metrics (as shown in Figure 6): swaying, facing the audience, and standing feet-shoulder distance apart. Swaying assesses whether the user exhibited swaying or rocking motions in a stationary position, as swaying is discouraged in public speaking. The score, facing the audience, analyzes the user's alignment toward the audience to evaluate their ability to maintain an appropriate direction. It is crucial to face the audience as it is considered a metric for confidence. Lastly, feet-shoulder distance estimates the user's stance width, specifically measuring the distance between their feet. For performing analysis, our optimal choice was YOLOv7 [50], a pose estimation model, as it offers enhanced visibility, improved handling of occlusions, and superior tracking of rapid movements [30]. Moreover, even with constraints like wire tethering of the Oculus Rift and the limitations of the phone's camera, it could still accurately detect the correct body key points by relying on the shoulder key points, even if the entire body was not visible.

We employed pose estimation and retrieved the key points of joints of the body. The key points for the left and right ear were used to estimate the facing the audience score. A bounding box of width 1 in the *x* and *y* direction around the right ear key point is created. If the left ear key point exists within that bounding box, the person is not facing the audience, and the score initialized at 0, is incremented by 1. For swaying, a bounding box of width 1 in the *x* and *y* direction is created around the right hip joint key point. If the left hip joint key point exists within the bounding box, the person is considered as swaying, and the score initialized at 0, is incremented by 1. The distance between the left and right shoulder key points is computed and the distance between the left and right foot key points is computed. If the difference between both distances is greater than 2 units in the *x* direction, the score of feet-shoulder distance is incremented by 1.

3.6 Nervousness Analysis

Nervousness, the feeling of unease or anxiety that is experienced prior to or while being in the spotlight, can hinder the speaker's ability to deliver the desired content, in the form of a shortage of breadth, mumbled-up words, and a shaking of the voice. This feeling of unease is directly linked with the heart rate of the speaker and therefore is the deciding factor for this analysis. If the heart rate lies between 60 beats per minute and 100 beats per minute, the nervousness score is unchanged, however, any heart rate value less than 60 and above 100 increments the nervousness score by 1. We measure heart rate using the MAX30100 heart rate sensor [46], ESP32 [2] as a microcontroller, and a DC Buck Converter LM2596 to convert voltage coming from the LiPo cells to 5V.

All the scores are subsequently normalized based on the total length of the user's recording. This normalization process allows for the expression of scores as a percentage of 100, providing a standardized and comparative measure. By normalizing the scores, it is ensured that variations in recording duration do not influence the overall interpretation and assessment of the user's performance. This normalization technique enables fair and consistent evaluation across different recordings, facilitating meaningful comparisons and comprehensive analysis.

3.7 Technical Challenges

The Oculus Rift is officially compatible with an NVIDIA GTX 1050 Ti graphics card. However, due to inherent resource constraints, we opted for an NVIDIA GTX 790 Ti graphics card for our implementation. Our approach involved decoupling the system to effectively optimize hardware utilization. The VR system was solely for the immersive environment, only incorporating the functionalities mentioned in 3.2. After retrieving data from the database, the processing of speech and video was seamlessly executed on the backend of the website. External sources were employed for recording users' videos and measuring their heart rates. By obtaining input from external sources and processing it independently of the VR application, we ensured that there was no compromise on the quality of the environment and the VR experience. We made sure to avoid intricate functionalities that might introduce lag or disruptions to the system's performance.

4 Limitations And Future Work

The system encompasses several potential areas for future improvement. Firstly, it can introduce more realistic, immersive, and diverse environments simulating various public speaking venues, enriching realism and effectiveness. This would enable users to practice in different environments. Secondly, enhancing feedback by specifically identifying instances of metric violations would offer targeted guidance for user improvement, thereby facilitating their growth as speakers. Thirdly, incorporating improved visualizations could aid users in tracking their progress over time. Due to resource constraints, certain features like slide presentations and real-time audience feedback were omitted; their inclusion would elevate effectiveness. Another intriguing prospect involves measuring nervousness through self-touch behaviors by utilizing a touch sensor.
Furthermore, a valuable avenue for future research could involve conducting a comparative user study that directly contrasts the outcomes of a group using a completely VR-based training approach with those using our application. This comparative analysis, along with the addition of comprehensive statistical analysis, would provide robust insights into the distinct advantages and benefits of our application in public speaking training. These enhancements and future research prospects contribute to the project's ongoing development and its impact on public speaking training.

5 Conclusion

This work introduces a novel public speaking training application that utilizes Virtual Reality (VR) to simulate an audience, providing comprehensive feedback on speech, body language, and nervousness, which is accessible through a website. Our work advances knowledge in distributed systems by integrating non-VR and VR methods in public speaking training, highlighting the potential for improved user experience. This study contributes to the integration of various software and hardware components and highlights the effectiveness of a multi-modal public speaking VR application, incorporating audio, video, and nervousness components, in addition to a user-friendly website for viewing scores and feedback. Collectively it aims to enhance the training experience and provide a comprehensive approach to improving public speaking skills.

Acknowledgments

We extend our profound gratitude to Dr. Mark Billinghurst for his invaluable guidance and unwavering support for helping us with this work.

References

[1] CGTrader. https://www.cgtrader.com/.

[2] ESP32-DevKitC. https://www.espressif.com/en/products/devkits/esp32-devkitc. Accessed: 2023-08-14.

[3] Firebase. https://firebase.google.com/.

[4] Flask. https://flask.palletsprojects.com/en/2.3.x/.

[5] LikeSo App. https://sayitlikeso.com/.

[6] Mixamo. https://www.mixamo.com/.

[7] my-voice-analysis. https://pypi.org/project/my-voice-analysis/.

[8] Oculus Rift. https://www.oculus.com/rift/.

[9] OvationVR App. https://www.ovationvr.com/.

[10] React. https://reactjs.org/.

[11] SpeakerCoach App by Microsoft. https://support.microsoft.com/en-gb/office/rehearse-your-slide-show-with-speaker-coach-cd7fc941-5c3b-498c-a225-83ef3f64f07b.

[12] Ummo App. http://www.ummoapp.com/.

[13] Unity 3D. https://unity.com/.

[14] VirtualSpeech App. https://virtualspeech.com/.

[15] Youtube. https://www.youtube.com/.

[16] Five ways to improve your body language during a speech. *Columbia University School of Professional Studies*, 2018.

[17] M. J. Alsaffar. Virtual reality software as preparation tools for oral presentations: Perceptions from the classroom. *Theory and Practice in Language Studies*, 11(10):1146–1160, 2021.

[18] L. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer. Cicero-towards a multimodal virtual audience platform for public speaking training. In *Intelligent Virtual Agents: 13th International Conference, IVA 2013, Edinburgh, UK, August 29-31, 2013. Proceedings 13*, pp. 116–128. Springer, 2013.

[19] E. Chang, H. T. Kim, and B. Yoo. Virtual reality sickness: A review of causes and measurements. *International Journal of Human–Computer Interaction*, 36(17):1658–1682, 2020.

[20] L. Chen, G. Feng, J. Joe, C. W. Leong, C. Kitchen, and C. M. Lee. Towards automated assessment of public speaking skills using multimodal cues. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 200–203, 2014.

[21] Y. Chen, Y. Tian, and M. He. Monocular human pose estimation: A survey of deep learning-based methods. *CoRR*, abs/2006.01423, 2020.

[22] M. Chollet, T. Wörtwein, L.-P. Morency, A. Shapiro, and S. Scherer. Exploring feedback strategies to improve public speaking: an interactive virtual audience framework. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 1143–1154, 2015.

[23] M. Dall. *Sicher prasentieren: Wirksam vortragen*. Redline Wirtschaft, 2014.

[24] C. George, A. N. Tien, and H. Hussmann. Seamless, bi-directional transitions along the reality-virtuality continuum: A conceptualization and prototype exploration. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 2–7, 2020. doi: 10.1109/ISMAR50242.2020.00067

[25] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, and E.-h. Zahzah. Human pose estimation from monocular images: A comprehensive survey. *Sensors*, 16(12), 2016.

[26] Google LLC. Google forms. https://www.google.com/forms/.

[27] I. Grishchenko, V. Bazarevsky, A. Zanfir, E. G. Bazavan, M. Zanfir, R. Yee, K. Raveendran, M. Zhdanovich, M. Grundmann, and C. Sminchisescu. Blazepose ghum holistic: Real-time 3d human landmarks and pose estimation, 2022.

[28] D. Herumurti, A. Yuniarti, P. Rimawan, and A. A. Yunanto. Overcoming glossophobia based on virtual reality and heart rate sensors. In *2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pp. 139–144. IEEE, 2019.

[29] K. J. Peter, F. J. Robert P., R. Richard L., and S. C. Brad. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training, University of Central Florida*, 1975.

[30] K. Kukil and V. Gupta. Yolov7 pose vs mediapipe in human pose estimation, May 2023.

[31] J. Li. Recent advances in end-to-end automatic speech recognition, 2022.

[32] J. Luiselli and A. Fischer. *Computer-Assisted and Web-Based Innovations in Psychology, Special Education, and Health*. 01 2016.

[33] MakeHuman Community. MakeHuman. https://www.makehumancommunity.org/.

[34] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang. The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. *IEEE Access*, 8:133330–133348, 2020.

[35] A.-T. Nguyen, W. Chen, and M. Rauterberg. Online feedback system for public speakers. In *2012 IEEE Symposium on E-Learning, E-Management and E-Services*, pp. 1–5. IEEE, 2012.

[36] T. D. Nguyen and M. Kresovic. A survey of top-down approaches for human pose estimation. *CoRR*, abs/2202.02656, 2022.

[37] S. G. S. of Business. Make body language your superpower, May 2014.

[38] F. Palmas, J. Cichor, D. A. Plecher, and G. Klinker. Acceptance and effectiveness of a virtual reality public speaking training. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 363–371, 2019.

[39] F. Palmas, R. Reinelt, J. E. Cichor, D. A. Plecher, and G. Klinker. Virtual reality public speaking training: Experimental evaluation of direct feedback technology acceptance. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 463–472, 2021.

[40] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding, 2022.

[41] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[42] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[43] M. Safir, H. Wallach, and M. Bar-Zvi. Virtual reality cognitive-behavior therapy for public speaking anxiety: One-year follow-up. *Behavior modification*, 36:235–46, 12 2011. doi: 10.1177/0145445511429999

[44] Sketchfab. Sketchfab. https://sketchfab.com/.

[45] M. Slater, C. Guger, G. Edlinger, R. Leeb, G. Pfurtscheller, A. Antley, M. Garau, A. Brogni, and D. Friedman. Analysis of physiological responses to a social situation in an immersive virtual environment. *Presence*, 15:553–569, 10 2006.

[46] R. Strogonovs. Implementing pulse oximeter using max30100. *MorfCoding and Engineering*, 2017.

[47] S. Stupar-Rutenfrans, L. E. Ketelaars, and M. S. van Gisbergen. Beat the fear of public speaking: Mobile 360 video virtual reality exposure training in home environment reduces public speaking anxiety. *Cyberpsychology, Behavior, and Social Networking*, 20(10):624–633, 2017.

[48] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks, 2015.

[49] A. Toshev and C. Szegedy. DeepPose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2014. doi: 10.1109/cvpr.2014.214

[50] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022.

[51] D. Wang, X. Wang, and S. Lv. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8), 2019.

[52] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann. Mediapipe hands: On-device real-time hand tracking, 2020.