Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Exploring User Reception of Speech-Controlled Virtual Reality Environment for Voice and Public Speaking Training

Patryk Bartyzel[a], Magdalena Igras-Cybulska[a], Daniela Hekiert[b], Magdalena Majdak[c], Grzegorz Łukawski[d], Thomas Bohné[e], Sławomir Tadeja[e]

[a] AGH University of Science and Technology, Kraków, Poland
[b] SWPS University, Warsaw, Poland
[c] Polish Academy of Science, Kraków, Poland
[d] Kielce University of Technology, Kielce, Poland
[e] University of Cambridge, Cambridge, UK

ARTICLE INFO

Article history:
Received January 6, 2025

Keywords: Virtual Reality, VR, Voice Training, Voice User Interface, Public Speaking Training, Speech processing

ABSTRACT

In this paper, we explore the development and assessment of a virtual reality (VR) system designed to enhance public speaking and vocal skills among professional and non-professional speech users alike. The system's foundation lies in a speech recordings corpus of 529 utterances given during presentations by a total of 15 students. From these data, we extracted voice parameters such as pitch, timbre, and speech rate using speech processing methods. We also asked six expert annotators to evaluate the stress levels present within each presentation. This multi-faceted analysis facilitated the selection of specific parameters for real-time animation control of virtual characters responding dynamically to the change in the speaker's voice. Through these mechanics, we could cultivate user proficiency in voice modulation, thereby improving overall speaking abilities and confidence. Furthermore, the system fosters self-awareness of vocal quality, promoting proper utilization of the voice in professional settings. Our VR system offers a dual-mode environment that combines traditional public speaking scenarios in front of a virtual audience with a relaxing forest setting, where users can control weather conditions with their voice. To assess the system's efficacy, we conducted a pilot study with five participants. Additionally, we provide preliminary design guidelines informed by our user study to support the development of future VR-based speech trainers.

1. Introduction

Many professions require proper voice emission as a crucial skill in their day-to-day work activities, including singers, teachers and lecturers, call center workers, and movie and dubbing actors [2, 3, 4, 5, 6]. These and many other occupational voice users are susceptible to developing various vocal cord conditions, such as voice degradation or functional voice disorders caused by misuse or abuse of their voices.

The list of potentially resulting vocal pathologies [2, 7] includes vocal nodules, characterized by benign epithelial proliferations often resulting from excessive vocal fold adduction, and vocal polyps, fluid-filled lesions typically associated with vocal trauma. Moreover, laryngeal inflammation, often termed laryngitis, and muscle tension dysphonia may also contribute to vocal

⋆This article is an extended version of the paper presented at *the 3rd International Workshop on eXtended Reality for Industrial and Occupational Supports (XRIOS)*, at the 2024 IEEE Conference on Virtual Reality and 3D User Interfaces (IEEE VR), Bartyzel et al. [1]

e-mail: skt@eng.cam.ac.uk (Sławomir Tadeja)

dysfunction. In consequence, a diverse spectrum of resulting vocal pathologies can compromise occupational efficacy and precipitate enduring, more severe health aftereffects [2, 7].

Occupational voice disorders constitute a substantial portion of occupational diseases prevalent in developed economies. One such example is the Republic of Poland, which is a member state of the European Union (EU) since 2004. The government-backed report published in 2022 places voice-related health conditions among the three most commonly occurring occupational diseases responsible for 11.7 % of all such cases [8].

A more recent nationwide United States (US) survey indicated that 20.6% participants experienced a form of voice disorders during their lifetime [9]. This metric was significantly elevated ($p \leq 0.0001$) among employed individuals, particularly those in vocally demanding occupations such as singing and teaching. Moreover, the report revealed a pronounced association between voice disorder prevalence and the use of voice assistant and teleconferencing software (both $p \leq 0.0001$).

All these data suggest the existing and growing need for providing professional voice users with additional and adequate training to help them mitigate risks associated with their daily work-related activities. Moreover, due to the broad diversity of voice occupations and a large spectrum of related voice-related conditions [2, 7], such training must be tailored to the individual needs and wants of the workforce. One possible tool occupational and non-professional voice users can leverage for speech training is *Virtual Reality* (VR) environment. The VR environment allows for generating bespoke, entirely virtual and controllable environments that can offer unique opportunities for training across various professions and a broad range of use case scenarios. For instance, such immersive, VR-based systems can be developed to mimic medical facilities to train both patients and personnel about the intricacies of medical procedures such as medical imaging [10, 11] or support engineering design of aeroengine components [12].

Currently, we can observe ongoing attempts to use VR in voice and speech training in commercial and academia alike. In the former context, the VR speech training market is primarily dominated by *Ovation* [13] and *Virtual Speech* [14] applications. Both these systems allow for VR-based training scenarios in public speaking, sales, or interviewing and claim to utilize artificial (AI) intelligence for training progress tracking and analysis, as well as powering virtual avatars and audiences. The latter is also a burgeoning area of academic research that focuses on creating dynamic, interactive virtual audience simulations [15].

In this context, the work presented in this paper contributes to a broader project exploring the application of interactive VR interfaces for effective voice and speech training [16, 17].

We designed our prototype VR system utilizing the Unity game engine, adhering to standards developed by *Meta* and *OpenXR* to maximize compatibility across a wide range of VR apparatus. A video demonstration of our prototype system is available online[1]. Notably, our system leverages virtual non-player characters (NPCs) to provide trainees with real-time visual feedback. What is more, the behaviors and facial expressions of these avatars are based on voice quality features in the form of input parameters.

This paper presents an elaborated version of the short paper by Bartyzel et al. [1]. We revised all figures, adding visualizations of the VR environment (Figs. 1 and Fig. 3) and results (Fig. 5 and Fig.6). The text was also significantly edited and expanded. Specifically, We provided a detailed analysis of the speech corpus, describing data collection, annotation, and parameter selection. We expanded the literature review and provided detailed descriptions of the VR system, hardware, and software. Additionally, we reanalyzed user study data to inform preliminary design guidelines for future VR-based voice and speech training systems.

2. Related Work

Using immersive interfaces such as VR in speech and voice training is not new and has been attempted before [13, 14, 15, 18, 19, 20, 21, 22, 23, 24]. For example, the pioneering efforts to employ immersive environments to treat glossophobia can be dated to approximately twenty years ago [21, 22]. The VR technology was also deployed to mitigate the fear of children for public speaking [25]. The proof-of-concept system developed by the authors demonstrated that such an approach could lead to an observable decrease in speaking anxiety among children in the short term after receiving the treatment. Another VR system was used for preparing oral presentations by English language learners, resulting in reducing the effects of public speaking anxiety [24]

More recently, investigations into the efficacy and acceptance of employing VR to address the need for speech delivery training could have been observed [18, 26]. Palmas et al. [18] reported a positive tendency among 44 participants in the study towards effectiveness and acceptance of VR training for public speaking training. Similarly, Bachmann et al. [26] reported a significant increase in the effectiveness of speech delivery and high acceptance levels post-VR training. Public speaking skills are fundamental for effectively conveying ideas and concepts to groups of individuals. However, the fear of public speaking, a manifestation of social phobia, can significantly impair an individual's ability to communicate effectively in personal and professional settings alike [21, 23].

In contrast, a study by Buchta et al. [27], which examined users' comfort levels and willingness to use voice in VR environments, found that some users were reluctant to engage in voice communication within these settings. Their findings revealed a wide range of user responses, influenced by factors such as personality traits, and individual adaptability to emerging technologies as well as previous experience with voice interfaces. Despite the prevalence of digital literacy among younger demographics, some users exhibited reluctance to engage in voice communication within VR settings [28], suggesting that factors beyond technological familiarity play a significant role in shaping user behavior.

When designing immersive systems for public speaking training, we can observe a plethora of methods such as speech and body posture or gesticulation analysis used to provide the

[1] https://www.youtube.com/watch?v=9GMDqd2uS-o

training with detailed information or their progress and results [18, 23]. These can be further enhanced by capturing and analyzing physiological signals such as the heart rate that can be used to investigate the participant's nervousness [23, 29].

The second strand of research within VR usage for public speaking concerns the influence of the virtual audience [19, 20] as modulated by factors such as exhibiting the human-like traits (e.g. realism and behaviour) of the virtual characters, which has been the subject of past research [19]. For instance, Remacle et al. investigated the potential of VR in enhancing teachers' vocal skills within a simulated environment characterized by adjustable constraints, including background noise and restless children. Their results revealed enhancements in vocal frequency, intensity, intonation, and extended pause lengths compared to a control trial. Furthermore, the potential application of VR in stuttering therapy has also been examined [30]. These studies highlight the versatility of VR as a training tool capable of effectively addressing a wide range of professional development needs and wants.

The new advancements in virtual character animation for VR environments have significantly enhanced the responsiveness and realism of these virtual avatars (see Fig. 1(a)). Here, the machine learning algorithms turned out to be instrumental in creating more realistic and adaptive digital avatars that can exhibit real-time behavioral traits (e.g. reactions and movements), adapting to the environment and user actions [31]. Additionally, prior research has explored the integration of emotional intelligence into virtual characters, enabling them to exhibit and respond to emotional cues [32]. These developments highlight the substantial progress made in the field, advocating for characters that can seamlessly adjust to different social situations, promptly react to user actions, and exhibit emotional intelligence to enhance the overall VR training experience. Furthermore, ongoing research contributes to improving virtual characters' physical realism, such as their interactions with objects and other characters, to create even more immersive and believable experiences (see Fig. 1(a)).

In summary, three main differences between our system and previous approaches that can be found in the literature.

First, unlike traditional VR public speaking trainers that only simulate a real-world audience, our system incorporates both a public speaking scene and an innovative natural forest setting. The forest scene introduces supernatural interaction elements, such as changing weather conditions based on the speaker's vocal parameters and is designed to have a relaxational impact. The scene was prepared based on the previous qualitative investigation of relaxation environments [33].

Second, our solution provides immediate feedback by analyzing specific vocal parameters, including pitch, timbre, and speech rate. At present, our system stands out as it enables real-time evaluation of the user's vocal timbre and adjusts the virtual environment in response to captured changes in speech delivery. These parameters dynamically control the reactions of a virtual audience or alter environmental factors, giving users instant visual cues to adjust their delivery. This continuous feedback loop is crucial for public speaking skill improvement and is not typically found in other VR speech trainers.

Based on the exploratory user study, we also establish several design suggestions in the form of guidelines to inform future VR-based speech trainers. These address dimensions such as immersive audience behavior, supernatural interactions, safety, and progressive challenges.

3. Stress-Related Speech Parameters in Public Speaking

Public speaking often triggers stress and anxiety, which can significantly impact a speaker's vocal performance for both professional and non-professional voice users [34]. Understanding how stress manifests in speech is crucial for developing practical training tools that help individuals manage and improve their public speaking skills. Here, we explore the specific speech parameters affected by stress based on new data collected from real-life students' oral presentations. We also explore how these parameters can be used to provide targeted feedback in a virtual reality training environment.

3.1. Data Collection and Preprocessing

As the first phase of our project, we conducted recordings and analysis of a dataset focusing on stress-related speech parameters, which consisted of 529 real-life public speaking recordings. Participants were informed about the assumptions, objectives, procedure, and duration of the study, as well as the methods used, potential risks and hazards, and any inconveniences associated with participating, together with their rights and responsibilities. They provided informed and voluntary consent to partake in the study. Each volunteer provided informed consent and signed the necessary documents and confirmation forms. Participants were also required to complete questionnaires immediately before and after their presentations.

The recordings were conducted in the university lecture room using the Zoom H4N recorder and microphones from the Boya BY-WM4 K2 set, consisting of two microphones, two microphone clips, two wireless signal transmitters, a wireless signal receiver, and a mini TRRS to mini TRS cable. The microphone was clipped to the participant's collar, with the cable leading to a wireless transmitter attached to the waist at the back. The transmitter sent the signal to the wireless receiver, connected to the Zoom H4N recorder via the mini TRRS to mini TRS cable. The recordings were saved on the recorder. Two microphones were used because the presentations were conducted in pairs, but each participant was recorded individually with only one microphone. The sampling frequency was $44,100$ $[Hz]$, and the bit depth was 24 *bits* per sample. The files were saved in WAV format.

Each recording underwent basic editing using Reaper software. This process included cutting out longer pauses and removing disturbances such as the sounds of drawing on a board, noise from material rubbing against the microphone, and persistent background noise from the room. We also split a single presentation conducted by two people into separate files for each person and divided files into short segments. For files lasting between 8 and 23 seconds, labels corresponding to the level of audible stress in the recording were created. These labeled files were then used to test the stress detection module. As a result of the basic processing, the following dataset was created:

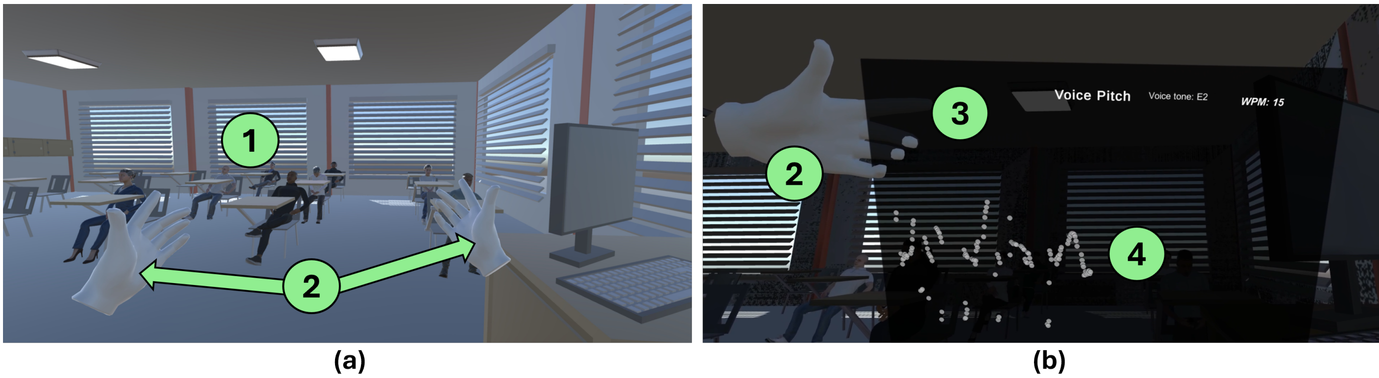


Fig. 1. (a-b) Two screen captures from within our immersive environment showcasing the *Mode 1* of the prototype controlled with bimanual gestural input. (a) A virtual audience seated in a classroom with their behavior programmed to respond to the speaker's voice. The figure also includes (2) hand gestures the user makes while practicing public speaking. (b) A virtual representation of the (2) user's hand is visible, indicating an interactive element. On the right side of the image, a (3) display panel shows `Voice Pitch` with a graphic equalizer and metrics such as `Voice tone: CH3` and `WPM: 109`, reflecting the speaker's vocal characteristics in real-time. Meanwhile, (4) shows the intonation contour of the utterance, i.e., a progression of the fundamental F0 frequency over time.

- Dataset Name: *Glosso Stress Dataset*
- Number of Recordings: *529*
- Number of Participants: *15*
- Minimum Recording Length: 8 [*sec*]
- Maximum Recording Length: 23 [*sec*]
- Sampling Frequency: $44,100$ [*Hz*]
- Bit Depth: 24 *bits per sample*
- Number of Channels: *1*
- File Format: *lossless WAV*

3.2. Data Annotation

This dataset was created using audio samples of student presentations delivered during seminar classes. Recording lengths ranged from 8 to 23 seconds, with a median of 15.5 seconds and a standard deviation of 4.6 seconds. Six dedicated annotators evaluated the samples. The annotators group included two males and four females and comprised a speech therapist, a speech trainer, three students of acoustic engineering and a psychologist.

Expert annotators rated the perceived stress levels on a Likert-like scale ranging from 1 to 5. The primary objective of this study was to identify which speech parameters might lead listeners to perceive a speaker as stressed during auditory reception. To determine the samples suitable for testing, we calculated the mean and standard deviation of the labels for each recording. Files were then categorized into two subsets: (i) neutral speech and (ii) stressed speech. The neutral speech subset included samples with a mean label value less than or equal to 2 and a standard deviation not exceeding 1. Conversely, the stressed speech subset comprised files with a mean label value greater than or equal to 4, with a standard deviation not exceeding 1.

This sieving process ensured that the chosen samples represented instances where experts were highly agreed upon on the presence or absence of audible stress. The resulting test set consisted of 96 files, with 30 files categorized as stressed speech and 66 as neutral speech, representing approximately 18% of the entire dataset.

3.3. Manifestations of Stress in Speech

In the reviewed content, stress was evident in several facets of vocal and speech production, especially in breathing regulation, breath support, resonance, pitch variation, speech pace, and articulation. These signs indicate both physiological and psychological strain, which can notably impact speech quality [35]. They could also imply an insufficient knowledge of effective vocal strategies in formal speech delivery, further supporting the notion that the way one speaks conveys a message, aligning with prior research [36].

One of the most noticeable stress indicators detected in the recordings was associated with breathing patterns, which directly influence sound quality and are characteristic of spontaneous speech. [37, 38]. Moreover, audible, shallow breathing, which gives the impression of insufficient air, often leads to a voice that is unsteady and fatigued, causing phrases to trail off quickly and final consonants to lack resonance. Conversely, poor resonance and inadequate breath support can result in a voice that sounds suppressed, faint, and internally focused, which is contrary to the desired outward sound projection. During the analysis of the audio samples, the professional speech coach observed that the speakers' body posture (often rigid and ungrounded due to stress) also needed adjustment to restore flexibility and ease.

3.4. Additional Symptoms and Psychological Implications

In addition to the common issues related to voice production, other stress-related symptoms observed in the analyzed material included a rapid speech pace and some instances of involuntary laughter, likely beyond the speaker's control. These behaviors may arise from self-awareness, the expectations of the audience, the circumstances of the task, previous public speaking experiences and the speaker's perceived self-image during the presentation [39]. Furthermore, these behavioral patterns could signify insecurities in social interactions, a fear of being judged, and concerns about criticism [40]. Such reactions are often motivated by a desire to conclude the speech swiftly, either to avoid potential confrontation or due to a sense of pressure related to occupying too much time. Additionally, these symptoms might be triggered by self-reflection, doubt, or uncertainty during the presentation, leading to hesitations, uncertainty and fluctuations in the emotional state [34].

3.5. *Consideration of Individual Speech Characteristics*

It is essential to acknowledge that specific individual speech traits may give the impression of stress, even when the speaker is not genuinely stressed [41]. For example, habitual reflectiveness or a certain academic detachment in student presentations could be mistaken for signs of stress. Additionally, a trembling voice might be due to organic factors [42], while the absence of stress markers in speech could reflect a high level of technical proficiency [34]. Thus, it is essential to remember that stress does not always manifest in perceptible changes in speech sound. Conversely, certain intonation patterns might create the illusion of stress.

3.6. *Implications for Training and Education*

The analysis of the collected samples indicates that most speakers likely lacked prior training in voice projection and emission. This absence of knowledge regarding healthy voice use and an understanding of how emission quality affects message delivery and audience perception underscores the need for training in voice and speech techniques.

4. Selection of Speech Parameters

Our goal was to identify voice and speech features crucial for improving speaker-audience communication's efficacy. After reviewing relevant literature on the fundamentals of the topic [34, 43, 44] as well as previous approaches [45, 26, 23, 14, 13] and analyzing the collected dataset, we selected voice timbre, pitch, and speech rate as the key parameters.

These parameters include voice source characteristics, spectral features (which pertain to vocal tract parameters), and temporal features such as prosody. Each of these elements plays a significant role in shaping how the speaker's voice is perceived and, consequently, how effectively the speaker can engage with the audience.

Voice timbre is the speech attribute central to distinguishing between sounds that are otherwise similar in loudness, pitch and duration [46]. Timbre allows listeners to perceive two sounds as different, even when presented in an identical manner with the same intensity and frequency [47]. For instance, even minor variations in timbre can lead to distinct perceptions and emotional responses from the audience [48, 49]. For instance, a warmer timbre may be perceived as more engaging or trustworthy, while a harsher timbre might evoke discomfort or tension. Understanding and manipulating timbre can, therefore, be a powerful tool in ensuring that the speaker's message is received as intended.

Voice pitch is determined by the perceived frequency of the sound waves generated by the vocal folds during voice production [50]. Pitch allows listeners to discern the intonation of speech, which is vital for conveying meaning and emotion. For example, a higher pitch might be associated with excitement, nervousness, or urgency, while a lower pitch might convey calmness, authority, or even boredom [51, 52, 53]. By adjusting pitch, speakers can modulate how their message is interpreted, thereby influencing the emotional tone of the interaction.

Speech rate, defined as the number of words or syllables spoken per minute, is another critical parameter that affects listener perception [54]. A slow speech rate can be interpreted as monotonous or dull, potentially losing the audience's attention, while an excessively high speech rate can lead to confusion and difficulty in reception; a moderate speech rate is generally perceived as the most engaging, while deviations towards slower or faster rates can potentially diminish the speaker's perceived reliability and assertiveness [54, 55]. Therefore, being aware of and appropriately adjusting one's speech rate is essential for maintaining effective communication, particularly in a virtual setting where immediate feedback can be used to fine-tune performance.

4.1. *Timbre, Pitch and Speech Rate in VR Speech Training*

The three key parameters, i.e., *timbre*, *pitch*, and *speech rate*, offer a rich source of data that can be utilized to develop an immersive and responsive digital environment. By interpreting these parameters and converting them into corresponding animations or reactions in virtual characters (see Fig. 1(a)), users can gain real-time feedback on their vocal delivery, helping them fine-tune their public speaking abilities and voice modulation.

For example, if a person speaks too rapidly, the virtual audience might show signs of confusion, prompting the need to slow down (see Fig. 1). Conversely, the characters could appear disinterested or bored if the speech is overly slow. This immediate feedback system allows users to adapt their speaking style on the fly, leading to more impactful communication.

Furthermore, examining vocal timbre through metrics such as pitch and *Mel Frequency Cepstral Coefficients* (MFCC) [56] can yield a deeper understanding of the speaker's emotional condition, vital in public speaking. A voice that communicates calmness and self-assurance is likely to trigger positive reactions from the virtual audience, whereas a voice that sounds tense or anxious may lead the virtual characters to display concern or unease. By embedding these feedback mechanisms, the virtual setting becomes an invaluable resource for practicing and honing public speaking skills, empowering users to achieve greater mastery over their vocal delivery and overall performance.

5. Speech Processing

Following the selection of the stress-related speech parameters, we focused on identifying the most effective algorithms for their extraction and analysis.

5.1. *Speech Timbre and MFCC*

Our system utilizes MFCC to analyze voice timbre [56]. This analysis includes spectral analysis of energy distribution across

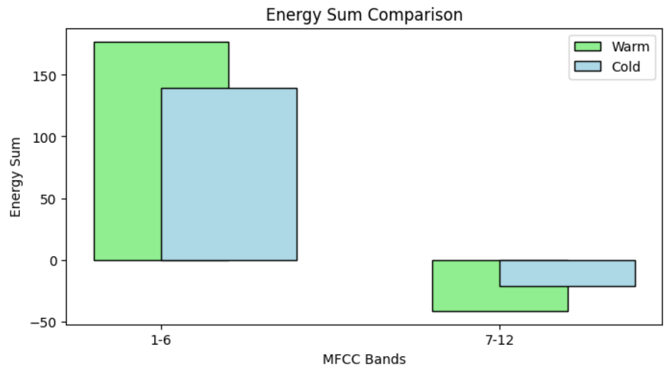


Fig. 2. The bar plot compares the energy sum across two frequency bands (1-6 and 7-12) for warm and cold voice timbres of the example recording of the same speaker [57]. The warm timbre exhibits significantly higher energy in the lower frequency band (1-6), while the cold timbre has slightly more energy concentrated in the higher frequency band (7-12).

various frequency bands to assess the perceived warmth or coldness of the voice. Particularly, our system determines whether the energy is primarily concentrated in lower or higher frequency bands [57, 58]. When the energy is concentrated in the higher frequency range, indexed between 7 and 12 in the 12-point MFCC array, the voice is identified as having a "cold" timbre [57]. In contrast, if the energy is skewed towards lower frequencies, indexed between 1 and 6, the voice is labeled as "warm" [57]. We illustrate this fundamental difference in Fig. 2. This distinction is essential, as it directly impacts how virtual characters react to the speaker's voice within the system (see Fig. 1(a)).

5.2. F0, Pitch and Intonation

Voice pitch, predominantly modulated by the vocal signal's fundamental frequency (F0), plays a crucial role in speech processing [50]. F0, which corresponds to the lowest frequency of a periodic waveform, reflects the vibration rate of the vocal folds during speech. Typically, the average F0 ranges from $100 - 146$ $[Hz]$ for men and $188 - 221$ $[Hz]$ for women [59]. This frequency arises from the vibration of the vocal folds within the larynx, driven by lung airflow. Notably, F0 is only present in voiced sounds and is absent in unvoiced ones, such as hissing or whispering. Prior research has demonstrated that pitch alone can reliably indicate emotions [60, 61]. In addition, studies on voice quality modifications suggest that professional vocal training can result in a lowered F0 [62, 40].

In the case of our system, we used the *Robust Algorithm for Pitch Tracking* (RAPT) [63] to track the fundamental frequency. RAPT is a sophisticated algorithm recognized for its high accuracy and resilience to noise, making it particularly effective in challenging acoustic environments [64].

To measure voice monotony, we use the standard deviation of pitch as a crucial indicator. This allows the virtual environment to adapt dynamically based on whether the user's voice is rich or monotonous. Adjusting the pitch variation parameters allows the environment to be easily customized to meet specific requirements or preferences. If the system identifies the voice as monotonous, as signified by the standard deviation falling below a predetermined threshold, it reduces a numerical value from the overall engagement score. This decline mirrors a perceived loss of interest from the virtual audience, as a monotonous voice can be seen as less engaging.

Nevertheless, when pitch variation exceeds the threshold, indicating a more dynamic and expressive voice, the system raises the threshold, reflecting increased audience engagement. This approach effectively replicates real-life audience reactions, providing users with valuable feedback that can help them improve their public speaking skills, especially in voice modulation and maintaining audience interest.

5.3. Speech Rate

The prototype assesses the speaker's verbal communication speed, expressed in words per minute (WPM). The optimal range for clear and effective speech delivery typically lies between $80 - 120$ $[WPM]$ [65, 54]. When the user's speech rate falls within this desired range, the system incrementally raises the engagement threshold, simulating increased interest from the virtual audience due to the speaker's clarity and pacing. This adjustment triggers animations that depict the audience as more attentive, thereby enhancing the realism of the virtual environment and offering the user positive reinforcement. At present, the system categorizes speech rates into several ranges, each corresponding to different deviations from the optimal pace:

Slight Deviation. The values fall within the range of $60 - 80$ $[WPM] \wedge 120 - 140$ $[WPM]$, indicating that the speech rate is slightly faster or slower than the ideal range but still within tolerable limits. The system detects this minor deviation and subtly lowers the engagement threshold.

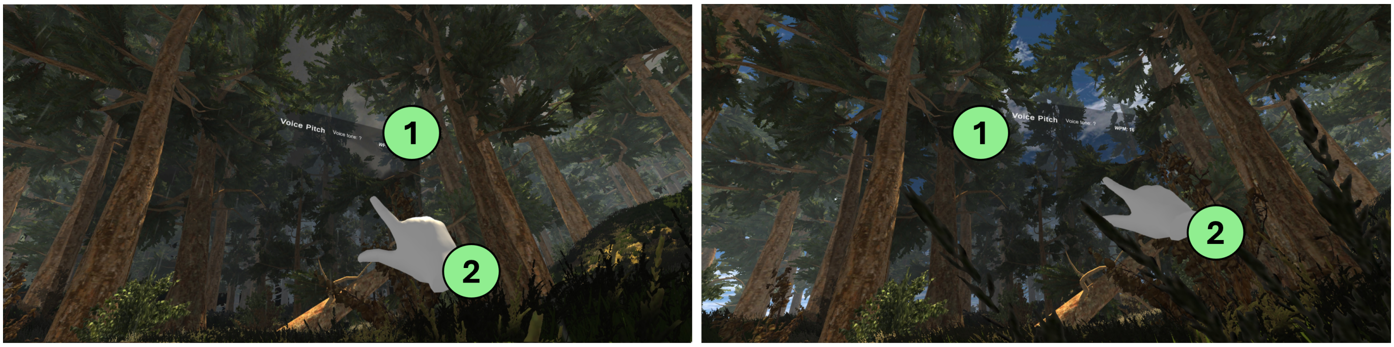
Moderate Deviation. In the range of $40 - 60$ $[WPM] \wedge 140 - 160$ $[WPM]$, the speech rate is noticeably slower or faster than optimal, potentially compromising clarity and effectiveness. As a result, our virtual audience reacts more strongly to these deviations (see Fig. 1(a)).

Severe Deviation. If the speech rate is below 40 $[WPM]$ or above 160 $[WPM]$, we consider this value as an extreme departure from the ideal speech rate, where the pace is either too slow or too fast to support effective communication. The system responds with pronounced animations (e.g., *Male Sitting Pose Not Interested Yawning*, *Sitting Looking Around* or *Sitting Head Down* from *Mixamo*[2] library composed of thousands of full-body character animations, captured from professional motion actors), reflecting a significant drop in virtual audience engagement.

6. Virtual Reality Training Environment

Here, we describe the VR environment, including the used software and hardware apparatus, as well as a detailed analysis of tasks that our system needs to support to provide its user with an effective training tool.

[1]Mixamo library: `https://www.mixamo.com/`. [Accessed: 30-Aug-2024]



(a) (b)

Fig. 3. The screenshot from within the VR environment showing *Mode 2* of our prototype. This scene is set in a dense, natural forest environment with various trees and foliage under a hazy sky, suggesting a serene atmosphere. At the top of the image, a translucent (1) heads-up display (HUD) shows the text `Voice Pitch`, with graphic equalizer and metrics such as `Voice tone` and `WPM`, reflecting the speaker's vocal characteristics in real-time. Within this VR experience, the weather conditions might fluctuate in response to the speaker's vocal attributes, altering the virtual environment's ambience in real-time. The HUD position can be manipulated by the user with the (2) hand gestures .

6.1. Apparatus

Our immersive environment was developed using the *Unity* game engine, specifically the 2020.3 LTS version. Unity is a widely adopted platform for prototyping and evaluating VR systems deployed on various head-mounted display devices [10, 11, 12, 66], making it a suitable choice for our experiments. We developed and tested the system using a desktop PC, which features an Intel i5-11400F CPU, 32 GB of DDR4 RAM, and an Nvidia RTX 3060 GPU with 32 GB of VRAM, operating under Windows 11 OS.

The VR experience was facilitated using the Meta Quest 2 headset. This device provides a high field of view (FoV) of approximately 100° and a display resolution of 1832×1920 pixels per eye. Interaction within the VR environment was primarily conducted using the headset's standard controllers and laser point (see Fig. 4), which were equipped with precise motion tracking capabilities and enabled accurate and responsive input within the virtual space.

Fig. 4. The system menu allows easy selection of (right-hand side) the scene and (left-hand side) the audio recording device.

6.2. System Requirements

We identified four key tasks of the voice-controlled VR environment that the system needs to support in order to provide its users with an effective training solution. These are as follows.

(R1) Replicate Realistic Virtual Environment: The system should accurately replicate real-world environments within the VR space, such as classrooms or auditoriums (see Fig. 1(a)). This replication of potential, real-world conditions of public speaking (e.g. a classroom or other auditorium) is crucial for creating an immersive experience that closely resembles actual public speaking scenarios [10, 11], thereby increasing the effectiveness of the training provided to users.

(R2) Enable User Navigation and Interaction: The system should support intuitive user navigation and interaction within the VR environment. The users must be able to move freely within the virtual space and interact with characters [10, 11], which is essential for simulating real-life public speaking situations and enhancing the user's sense of presence [15].

(R3) Analyze Voice Parameters in Real-Time: To provide meaningful feedback, the system must analyze voice parameters in real-time. This analysis enables the virtual environment to react dynamically to the user's speech, allowing the system to simulate audience reactions based on the user's vocal performance (see Fig. 1(a-b)) [15].

(R4) Ensure User Safety and Comfort: The system must include safety measures to prevent accidents because VR headsets can disconnect users from their physical surroundings [67, 68]. Additionally, the system should be designed to minimize discomfort, such as motion sickness, by optimizing frame rates and ensuring smooth navigation [69].

All these four key requirements (R1-R4) are supported by the implemented functionalities within the VR system.

6.3. Immersive Scenes and Their Purpose

Our VR-based public speaking system offers a dual-mode training environment (see Fig. 1–4). Namely, aside from more

No.	Age	Gender	Field of Study	Gaming Frequency	VR Usage	PRCS Score	SUS Score
P1	23	M	Engineering	Occasionally	Never	24	87.5
P2	23	F	Social sciences	Occasionally	Occasionally	44	95.0
P3	23	M	Engineering	Daily or almost daily	Occasionally	29	97.5
P4	20	F	Social sciences	A few times a month	Occasionally	28	92.0
P5	22	F	Engineering	A few times a month	Never	45	90.0

Table 1. Characteristics of Study Participants with PRCS and SUS Scores

typical VR simulations within a classroom or lecture theater in front of a responsive virtual audience (see Fig. 1), we introduce a natural setting with voice-based interaction and supernatural elements (see Fig. 3). Users can dynamically influence their environment, enhancing immersion and providing a more comprehensive training experience. This hybrid approach provides a more comprehensive and engaging training experience than traditional VR-based systems.

Public speaking scene. In the virtual classroom setup (see Fig. 1), each character is carefully configured with three essential parameters: *Threshold Multiplier*, *Animation Delay*, and *Threshold and Animation Ranges*. These components are crucial for developing a dynamic and believable audience interaction.

Threshold multiplier is key to generating a range of audience reactions. The system adjusts the threshold value based on the user's vocal features, applying a specific multiplier for each character, which results in different levels of responsiveness. This variation adds to the authenticity of the virtual audience, as some characters will react more strongly to changes in speech rate, simulating the natural diversity of responses found in real audiences.

Animation delay adds variety to the timing of character responses by setting each one to evaluate its threshold and trigger animations at different moments. This staggered timing prevents the characters from reacting simultaneously, which would seem unnatural. Instead, it mirrors the natural variation in how people in an actual audience react, with characters showing different levels of engagement and response times.

Threshold and animation ranges for each character are scaled between 0 and 100 and divided into four specific animation ranges $[0-25]$, $[25-50]$, $[50-75]$, and $[75-100]$. The current threshold, influenced by the user's vocal attributes and the multiplier applied, determines which animation is triggered within these ranges.

Together, these features create a rich and realistic virtual environment. Animations were applied from the Mixamo library. By using animation delays and threshold multipliers, each character in the virtual audience presents a distinct and authentic reaction, closely reflecting the diverse responses typical of a real-world audience. In the following, we present the mapping of the segments with specific examples:

- Range of $[0-25]$ (disengaged behavior): virtual characters display signs of disinterest, such as looking away, checking their phones, or fidgeting. For example, characters might perform animations like "looking around the room" or "crossing arms," suggesting a lack of interest in the speaker's performance.
- Range of $[25-50]$ (neutral behavior): the audience's engagement is moderate, indicating a neutral response to the user's speech. Characters might appear attentive but not particularly enthusiastic, with animations like "sitting upright" or "nodding occasionally." These behaviors suggest that the audience is paying attention but is not highly engaged.
- Range of $[50-75]$ (positive engagement): as the threshold increases, virtual characters begin to display more positive reactions. In this range, characters show active listening behaviors, such as "leaning forward" or "smiling." These animations signal that the audience is engaged and interested in the content being delivered.
- Range $[75-100]$ (high engagement): this range represents the highest level of engagement from the virtual audience when characters react with enthusiastic gestures, like "applauding lightly" or "nodding more frequently." These responses simulate an audience that is highly impressed or moved by the speaker's performance.

Forest scene. This relaxation scene (see Fig. 3) offers feedback through environmental shifts, particularly in weather patterns that respond to the user's voice characteristics. This scene serves as a practice space, allowing users to refine their speaking skills in a serene, informal environment, and is based on earlier research into calming VR settings [33].

The weather responds dynamically to the user's voice threshold, which is categorized into four distinct levels: *very rainy* ($[0-25]$ threshold), representing heavy rainfall and indicating a less favorable reaction to the user's speech; *rainy* ($[25-50]$ threshold), reflecting moderate rainfall and a neutral response; *drizzle* ($[50-75]$ threshold), symbolizing light rain as a sign of positive feedback; and *sunny* ($[75-100]$ threshold), where clear weather signifies the best reception, highlighting effective communication and strong engagement.

7. System Assessment with Target Users

We carried out a small-scale exploration to assess the potential of our immersive VR environment for vocal training and public speaking. During this exploratory study, five participants interacted with the VR prototype, seated or standing, using a Meta Quest 2 headset and controllers. The main aim of the study was to assess the VR system's capacity to create a realistic and engaging immersive environment, as well as its ability to facilitate effective learning.

7.1. Participants, Experimental Protocol and Task

Five Polish volunteers [70], two males and three females, averaging 22.2 years old, as shown in Tab. 1, engaged with the VR environment and provided insights on the system functionality, the user interface design, as well as an overall experience. The group included individuals from both technical and social sciences educational backgrounds. Their gaming habits ranged from occasional to daily, and their familiarity with VR varied, with three participants having previous experience using VR technology.

After signing the informed consent, participants were initially asked to fill out the Polish version (prepared using back-translation method [71]) of a shortened version of the *Personal Report of Confidence as a Speaker* (PRCS) [72]. This 12-item version was created to provide a more concise and efficient evaluation, targeting the critical aspects of anxiety related to public speaking [73, 74]. The purpose was to gauge their experience and anxiety levels, specifically in the context of public speaking. A 5-point Likert scale was used for responses. The example items were: *1. My hands tremble when I try to handle objects on the podium. 2. I constantly worry that I will forget my speech. 3. I am in a constant state of anxiety while preparing my speech.*

Participants were introduced to the VR system through a brief orientation provided by the experimenter. The experimenter assisted participants with wearing the head-mounted display (HMD) and explained the basic functionality of the VR environment, including how to use the controllers and navigate the virtual space. Once participants were familiar with the equipment, they began their interactions with the VR system.

Next, participants completed two main tasks within the VR environment, presented in the following sequence, to ensure a structured experience. First, they engaged with the public speaking scene, where they delivered a short, spontaneous speech while observing real-time feedback from the virtual audience based on their vocal performance (see Fig. 1). In the second task, participants explored the forest scene, where they could observe dynamic environmental changes in response to their voice attributes, offering a more relaxed and reflective interaction experience (see Fig. 3). The experimenter provided prompts if necessary, including assistance with the VR controls and answering any technical questions.

After interacting with the VR tool, participants were asked to fill out the widely recognized 5-points *System Usability Scale* (SUS) [75] to assess the system's usability. The example items were: *1. I think that I would like to use this system frequently. 2. I found the system unnecessarily complex. 3. I thought the system was easy to use.* At the end of the experiment, we conducted semi-structured interviews to delve into the participants' overall experience. The example questions were: *1. Can you describe any challenges or difficulties you encountered while using the system? 2. Which features of the system did you find most and less useful, and why? 3. Were there any features that you felt were missing or could be improved?* These interviews sought to gather immediate impressions and more considered reflections, offering crucial insights into the effectiveness and influence of the VR experience.

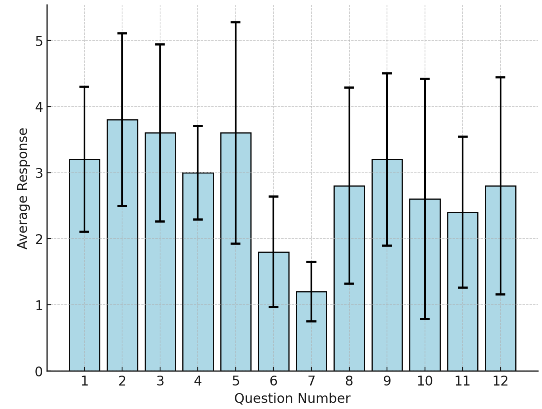


Fig. 5. Distribution of PRCS responses by question (mean values and standard deviation) illustrates a wide range of public speaking anxiety levels among participants, with significant variability in responses, especially on questions related to specific fears and self-perception like 6 and 7, indicating the need for personalized training approaches. Here, some participants reported a manageable level of fear, with generally lower scores observed for these questions.

8. Study Results

Overall, the findings from our exploratory study suggest that, despite being in the early stages of development, our VR prototype holds significant potential as an interactive learning platform for public speaking. The combination of high usability and positive user experiences indicates that the foundation has been laid for a tool that could be highly effective in helping individuals improve their public speaking abilities. As the system continues to evolve, further research with more extensive and diverse participant groups will be essential to validate these initial findings and refine the system to better meet the needs of its users.

8.1. Speaker Confidence

The PRCS [72] results revealed a wide range of public speaking anxiety levels among the participants, underscoring the limitations of a one-size-fits-all approach to training (see Fig. 5). The shortened version of the PRCS does not have a formally established scoring scheme with cut-off levels, meaning that the interpretation of results can vary depending on the study. In our case, we adopted an approximate interpretation of PRCS scores based on psychometric guidelines [76]: *low anxiety* level: $12-23$, *moderate anxiety* level: $24-42$, and *high anxiety* level: $43-60$. The Cronbach's alpha for the PRCS scale ($\alpha = 0.86$) indicates a high level of internal consistency, suggesting that the items reliably measure the construct of public speaking anxiety.

The observed scores varied significantly, with some participants (P1, P3, and P4) reporting moderate anxiety, reflected in scores in the lower part of the moderate anxiety level as 24 out of 60 (see Tab. 1). These participants exhibited a manageable level of fear, particularly in response to questions like Q6 ("*The faces of my audience become blurred when I look at them*") and Q7 ("*I feel disgusted with myself after attempting to speak before a group of people.*"), where lower scores indicated that these specific anxieties were less pronounced.

In contrast, some P2 and P5 scored as high as 45, indicating a much more intense level of anxiety that could severely impact their public speaking performance. These higher scores, particularly on questions Q6 and Q7, suggest a more profound struggle with the visual and emotional aspects of audience engagement. This variability in responses highlights the need for personalized training programs tailored to address each individual's distinct challenges, particularly in overcoming specific fears related to audience perception and self-criticism during public speaking [74].

8.2. System Usability

On the other hand, the SUS [77] scores provided a clear indication of the system's usability, with participants reporting an average score of 92.5 out of 100 (see Fig. 6). This very high score suggests that our VR system was well-received and considered intuitive and easy to use, which is critical for encouraging consistent use and engagement.

Notably, Cronbach's alpha for the SUS scale is approximately equal to 0.76, which indicates a good level of internal consistency among the items in this scale. Responses to the fourth question (Q4) of the questionnaire, which assessed the need for technical support, were relatively high (2.4) in comparison with typical patterns observed in SUS responses for other even-numbered questions ($1 - 1.4$). In the SUS, odd-numbered questions correlate positively with usability, while even-numbered questions correlate negatively. For a high usability rating, we expect high scores on odd questions and low scores on even ones. However, Q4, an even-numbered question, stood out with a higher score, indicating that users felt a greater need for technical support than expected in a highly usable system. The skewness of the SUS Q4 data is approximately 0.20, indicating a slight positive skew. The kurtosis is approximately -0.74, which suggests the distribution is slightly platykurtic, flatter than a normal distribution.

This suggests that while the system is generally user-friendly, users may find certain aspects or functionalities complex enough to require additional assistance, especially the VR setup. Here, the assistance provided to participants may have impacted their responses to the SUS Q4 item, suggesting a need for additional technical support. Therefore, we plan to revise our system to incorporate a more intuitive onboarding tutorial to further reduce reliance on external guidance.

The positive overall feedback indicates that the application has largely met the requirements for a user-friendly interface, which is a promising sign for its future development and deployment. However, it is essential to note that these findings are based on small sample size, and as such, the numerical values should be interpreted with caution to avoid overgeneralization.

8.3. Semi-Structured Interviews

The qualitative data from post-interaction interviews further enriched our understanding of how participants approached public speaking. Participants indicated that they prepare for speeches by repeating the content in their minds or out loud at home, noting keywords, or preparing a full script. A single participant reported practicing with a close person to refine speech delivery. These methods highlight the importance of practice and feedback in building confidence for public speaking, which aligns with the features they value most in our application.

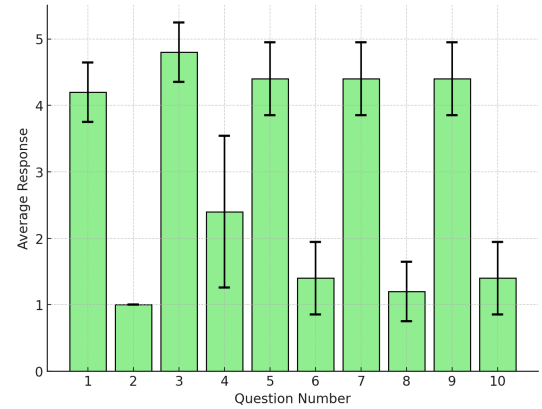


Fig. 6. Distribution of SUS responses by question (mean values and standard deviation) indicates high usability, with participants rating the system favorably. At the same time, the higher-than-expected score on question 4 suggests that some users may require additional technical support when using the VR system.

The ability to adjust the virtual world's reactions based on what is being said and the simulation of an audience was particularly appreciated (see Fig. 1(a)), as they allowed users to test their skills in a relaxed and controlled environment. Furthermore, real-time voice tone and speech rate display (see Fig. 1(b)) were crucial for users to monitor and improve their performance, offering immediate, actionable insights.

Participants identified the potential target audience for the application as individuals who frequently engage in public speaking but struggle with confidence and those who aspire to become better speakers, such as students, schoolchildren, and professionals. They noted that the application's ability to help users become more accustomed to public speaking and practice in an intuitive way would be a significant motivator for use.

However, participants also pointed out potential barriers, such as the entry challenges related to VR technology and the cost of the equipment. Despite these concerns, participants appreciated features like the clarity of the menu, the presence of an audience, the ability to see voice parameters, as well as the calming effect of the virtual forest, along with the ability to move around the space and receive visual feedback. They also expressed a desire for further enhancements, such as greater interaction with the audience, responses to more stimuli beyond just voice, the inclusion of an in-app assistant, and an option to save feedback, which could further enhance the application's usability and appeal.

9. Preliminary Design Guidelines for VR Speech Trainers

Based on our experience developing and testing the VR system and utilizing participants' quantitative and qualitative feedback, we arrive at nine early stage design guidelines to help inform future iterations of VR-based speech training systems. The purpose of these early guidelines is to lay the groundwork for further research based on both our observations and insights from initial user interactions with the system. Formulating preliminary design guidelines aligns with the previously postulated approach, which advocates that the path toward new ubiquitous conversational voice interactions must be based on several evolutionary steps [78]. At the same time, we advise caution when following these guidelines' suggestions.

1. **Real-life replication.** Create immersive VR environments that replicate real-world public speaking scenarios [10, 11] (to support **R1**), featuring virtual audiences that dynamically respond to changes in the user's vocal delivery [21, 79] (to support **R2**). Use voice parameters to trigger audience reactions that reflect real-life feedback (to support **R1–R3**). During the study, several participants noted the impact of audience responses on their training experience, with one stating, "*The audience reactions and voice parameter dashboard were the most helpful*" (P2). Another added, "*The visual representation showing if people are listening or not was great*" (P4).

2. **Real-time feedback.** Implement real-time feedback mechanisms that respond to the user's vocal parameters (to support **R1–R2**). This feedback loop should include both auditory and visual cues to help users adjust their speech in real-time and improve their overall delivery (to support **R3**) [80, 81]. P4 highlighted that "*tracking the speed of my speech helped me the most*", while P4 mentioned, "*I liked that it showed my tone because when I am stressed, I speak higher.*"

3. **Personalization.** Offer personalized training modules that adapt to the user's public speaking anxiety levels and specific vocal characteristics. Use voice analysis data (such as pitch, timbre, and speech rate) to tailor the training experience to individual needs (to support **R3**) [15]. As P1 remarked, "*the system adapts to your voice, which is crucial.*"

4. **Progressive challenges.** Structure the training program to introduce users to adaptive and challenging tasks [82] concerning public speaking conditioned on their vocal performance metrics. As users improve, they gradually increase the complexity of the scenarios, audience size, and feedback intensity. Implement adaptive algorithms that continuously monitor and adjust the training scenarios based on the user's vocal performance [82]. Use detailed analytics of voice parameters to fine-tune the difficulty and feedback. P1 commented, "*This app could encourage people by helping them get comfortable with public speaking*", while P4 noted that "*it really can help improve public speaking skills.*"

5. **Performance Tracking.** Offer detailed performance tracking and analytics, focusing on key voice metrics such as pitch, timbre, speech rate, and stress levels (to support **R1–R3**). Use visual dashboards to display progress over time [83], allowing users to set goals and track their improvement in both vocal quality and public speaking skills (to support **R2–R3**). P5 valued the system's ability to track vocal adjustments and commented that "*the tone tracking was the most helpful part.*"

6. **Supernatural Interaction.** Integrate "supernatural" interactions where the user's voice can manipulate the environment in ways that are impossible in the real world [84], such as changing the weather or altering the landscape (to support **R1–R2**). These supernatural elements enhance the immersive experience and may allow users to see the powerful impact of their voice in a visually engaging and memorable way [84]. Users expressed that these features added a unique layer to the training experience, making it more enjoyable and memorable. They noted how the environment's dynamic response to their voices added depth to the training experience. For instance, P1 emphasized, "*the most important feature was the environment's ability to adapt to your voice*," which underscored the engagement and responsiveness this feature provided. While P2 shared that the forest setting, with its calming and adaptive elements, "*had a soothing effect*", suggesting that such supernatural features not only enhance immersion but also help reduce anxiety. Participants valued the safe and flexible training environment enabled by these features (P4: "*The most important aspect was the ability to test myself in safe conditions and a relaxed atmosphere.*" These comments suggest that including supernatural interactions can enhance user engagement with the training, making the experience more immersive and potentially supporting meaningful learning outcomes.

7. **Safety and Comfort.** Given that VR headsets can disconnect users from their physical surroundings [67, 68], ensure that the system includes safety measures to prevent accidents (to support **R4**). Design the system to minimize discomfort, such as motion sickness, by optimizing frame rates and ensuring smooth navigation (to support **R1–R2**) [69]. Implement virtual boundaries and prompts to keep users within safe physical spaces during use (to support **R4**). As P4 expressed, "*the chance to test in a safe and relaxed atmosphere was essential.*"

8. **Minimalistic interface.** Keep the VR interface design clean and focused, presenting only the essential elements and information needed for the task at hand (to support **R1–R2**) (see Fig. 4) [85, 86]. Avoid unnecessary visual clutter that could distract or overwhelm the user (see Fig. 1 and Fig. 3). P1 and P3 highlighted the simplicity and clarity of the interface as helpful and appreciated the streamlined design.

9. **Help and documentation.** Ensure the VR system includes an intuitive interface that offers tutorials and contextual help [85, 86] (see Fig. 4). During our study, participants indicated the value of clear guidance and support within the VR system to facilitate ease of use (P1, P2, P4). For instance, they noted that the initial assistance provided by the experimenter helped them feel comfortable with the interface. At the same time, responses to the SUS Q4 item revealed a perceived need for further support in navigating

VR functionalities. Based on this feedback, we recommend incorporating an onboarding tutorial to reduce reliance on external help and promote a more seamless, user-driven experience.

10. Limitations

This study has several limitations that should be considered when interpreting the findings. First, the small sample size of only five participants limits the generalizability of the results. With such a limited pool of users, it is possible to capture the majority of usability issues [87] but it remains difficult to draw definitive conclusions about the overall usability and effectiveness of the VR system across diverse populations. Additionally, the study was conducted in a controlled environment, which may not fully replicate the complexities and challenges of real-world public speaking scenarios. Furthermore, the presence of the experimenter during the tests, offering assistance and guidance, may have influenced the participants' experience and perceptions of usability, potentially reducing the ecological validity of the results. This controlled setting might have contributed to the system's high usability scores, but future studies should explore its performance in more varied and dynamic environments to ensure its applicability in real-life situations.

11. Discussion

This study introduces the design of a VR environment that dynamically responds to alterations in voice and speech parameters, offering real-time visualization of these changes within a responsive virtual setting. The system is specifically designed to help users develop proficiency in critical aspects of public speaking, such as maintaining an optimal speech rate, timbre and voice modulation.

11.1. System Assessment

Our initial assessment of the system hints at its usability and effectiveness in providing immediate feedback, which is crucial for users aiming to refine their public speaking skills. The feedback and insights gathered from this initial user study have proven valuable for guiding the future development and enhancement of the application. For example, participants emphasized the importance of adjusting the virtual environment and audience reactions in real-time based on their speech. This dynamic responsiveness was identified as a key factor in creating an immersive training experience that closely mirrors real-life public speaking scenarios critical for building users' confidence and competence [88, 21].

In addition, participants highly valued the real-time feedback on voice parameters like tone and speech rate, allowing them to make immediate, informed adjustments to their performance. This suggests that the system is effective not only as a training tool but also as a self-assessment platform, enabling users to continuously improve their public speaking abilities.

11.2. Barriers for System Adoption

During the post-experimental interviews, participants reflected on several potential barriers to the widespread adoption of the system, notably the technological entry threshold associated with VR and the cost of the necessary equipment.

Furthermore, as it was commented on, for users who are not already familiar with VR technology, the learning curve can be steep, potentially discouraging them from fully engaging with the system. The financial investment required to purchase VR hardware could be prohibitive, especially for individuals or institutions with limited budgets. At the same time, HMDs and other devices facilitating the VR interface are becoming more affordable for each new generation, increasing their availability and prior familiarization with this technology.

Moreover, we built our system utilising the well-known and free for-to-use version of the Unity game engine that allows for prompt deployment to multiple hardware platforms of varying vendors. We also adhere to open and royalty-free standards such as OpenXR when developing our system to warrant potential cross-platform deployment.

11.3. Improvement Opportunities

A promising avenue for improvement is the integration of machine learning models to enhance the system's ability to perform more elaborate and refined analyses of voice timbre and other vocal characteristics like pause patterns [37] or speech emotion recognition [89]. This would allow the system to offer more personalized and precise feedback, improving its effectiveness as a training tool. In addition, expanding the range of voice features analyzed—such as incorporating elements like emotional tone or vocal stress—could provide a more comprehensive assessment of the user's public speaking performance.

Furthermore, advanced visualization techniques could lead to a richer, more immersive user experience. For instance, developing more complex and varied audience reactions or creating scenarios that simulate a broader range of public speaking contexts could better prepare users for the diversity of situations they might encounter in real life. Also ameliorating the feedback with additional modalities, e.g., gestures [90, 23], posture, eye movement and psychophysiology [91] tracking, would provide more complex feedback to the users.

A novel approach to reducing training stress could involve designing the virtual environment to react inversely to the quality of the user's speech. Specifically, when the system detects higher levels of vocal stress or anxiety—such as increased pitch variability, rapid speech rate, or erratic vocal timbre—the environment could respond in a calming manner rather than mirroring the user's anxiety. For example, if the user's voice becomes tense, the virtual environment could gradually dim the lights, soften the background noise, or display more neutral and supportive audience reactions. This inverse feedback loop could help mitigate the pressure of real-time performance, creating a more supportive and less stressful training atmosphere which would stay consistent with behavioral therapy of anxiety [92, 93]. By reducing the intensity of the virtual audience's reactions when users struggle, the system could help users regain composure, encouraging them to continue without the added pressure of feeling judged by the virtual audience.

11.4. System Design Guidelines

The proposed design guidelines for VR-based speech trainers are well-aligned with well-established usability heuristics [85, 86] while being geared explicitly towards speech training. For example, the guideline on **real-time feedback** is consistent with the original *visibility of system status* heuristic [85, 86], as it ensures that users are constantly informed about how the system is reacting to their vocal inputs, thus maintaining transparency and user control. Similarly, our **personalization** guideline overlaps with *match between system and the real world* by using analogous concepts tailored to individual users, making the system more intuitive and easier to navigate. Furthermore, **progressive challenges** guideline also aligns with usability heuristics by reflecting on a scalable learning experience that adapts to the user's skill level, ensuring that both beginners and advanced users can engage effectively with the system.

11.5. Supernatural Elements of Interaction Methods

Supernatural interaction or out-of-reality interaction in the context of VR refers to actions that go beyond the physical limitations of the real world, allowing users to engage in experiences that are impossible in everyday reality [94, 95]. These interactions are often described as fantastical interactions or immersive interactions, especially when they involve the ability to manipulate the environment in extraordinary ways, such as using voice commands to change weather conditions, moving objects with a mere gesture, or even altering time and space within the virtual environment. These interactions are designed to showcase the creative and emotionally engaging potential of VR, offering experiences that cannot be replicated in the physical world. For example, allowing users to summon or move objects with a gesture or control natural elements like wind or rain with their voice adds a layer of immersion that enhances the overall experience, making it more memorable and impactful.

11.6. Dual-Mode Training Environment

One of the key novelties of our approach lies in its dual-mode training environment, which significantly expands the scope of traditional VR-based public speaking training systems (see Fig. 1 and Fig. 3). While most existing approaches primarily focus on simulating the experience of delivering a presentation in front of a virtual audience [19, 20], our system goes beyond this by introducing an additional mode that allows users to practice voice emission in a natural setting (see Fig. 3).

This natural environment is enhanced with voice-based interaction and *supernatural* interaction, where the user's voice can dynamically influence the surroundings, such as changing the weather or altering the landscape (see Fig. 3). This innovative feature offers a unique blend of realism with features that could only be delivered via an immersive interface, providing users with a more comprehensive and engaging training experience. Consequently, our system offers a unique and underexplored approach to VR-based public speaking and voice training by occupational and non-professional voice users alike.

12. Conclusion and Future Work

This project explores the user reception of a speech-controlled, highly responsive VR-based tool for public speaking and voice training. The system's capability to offer real-time feedback based on ongoing monitoring of voice parameters, such as pitch, timbre, and speech rate (see Fig. 1(b)), provides users with insights into their vocal characteristics and how these may influence their communication effectiveness. The immediate, interactive feedback loop offers users a means to engage with real-time adjustments in voice modulation and delivery, providing valuable insights for potential confidence enhancement in public speaking.

The system's innovative approach provides a distinct environment where users can explore simulated public speaking scenarios, with responsive feedback intended to encourage and support skill development. By simulating the pressures and dynamics of real-world speaking scenarios (see Fig. 1(a)), the tool provides an engaging platform for speakers at various experience levels to explore different speaking strategies. As users engage with the system, they can better understand the subtleties of their vocal performance, such as the impact of stress or nervousness on their speech, and learn to make adjustments that lead to more effective communication.

As the system continues to evolve, with potential enhancements like machine learning integration and expanded voice feature analysis, it could become an effective resource for anyone looking to improve their vocal skills and overall communication proficiency.

We plan for future iterations of our system to focus on expanding the capabilities of the current voice timbre analysis. First, we intend to transition from a simplistic binary classification of vocal tone to a more nuanced, multi-dimensional spectrum of possible human voices. This will enable the system to accurately identify and respond to a broader range of vocal qualities, further enhancing the quality of system output.

Another strand of the future development direction of our approach is to investigate the influence of ambient, background audio effects (e.g., equipment sounds, white noise, or crowd whispers) in helping to facilitate more realistic public voice usage scenarios. At the same time, some of these can be considered a form of distraction that could be present during public speech delivery. Thus, their inclusion would help us explore the degree of their influence on the speaker while immersed in VR.

What is more, to address the need for additional technical support, as indicated in the usability assessment, we plan to implement a more intuitive onboarding tutorial to minimize the need for external guidance.

In future work, we will also test whether supernatural interactions enhance learning by providing immediate feedback that clarifies the impact of vocal modulation on engagement. We plan a between-subject experimental study with a quantitative approach, hypothesizing that these interactions will accelerate learning by making vocal qualities more tangible. Additionally, a longitudinal study with pre- and post-assessments will assess the sustained impact on skill retention, tracking users over time to see if immediate feedback leads to long-term gains in vocal modulation, audience engagement, and confidence.

Finally, present-day VR interfaces offer unique capabilities, such as the possibility of tracking body parts, such as head and hand tracking [96]. These can be either supported directly by the selected VR headset [11, 10] or be facilitated with additional sensor [66, 97]. Such a real-time analysis of human posture and movement may provide a method for a more complex understanding of nonverbal communication, which is an essential and frequently underestimated component of effective public address [98]. Moreover, the role of multimodal feedback to improve user engagement was previously explored [99], together with the impact of gestural cues on communicative effectiveness [90].

We plan to evaluate each of these highlighted streams of research in a series of controlled user studies with a balanced and diverse group of participants to validate our initial findings and ensure that the system is effective across a broader range of users.

We will also incorporate pre- and post-assessment procedures to capture sustained changes and refine the training process based on observed skill retention. Additionally, we will conduct a longitudinal study to assess the long-term effectiveness of our training system in improving public speaking skills.

As both the usability and effectiveness of our training system are important considerations, we plan to deploy a range of subjective (e.g. interviews and questionnaires) and objective (e.g. task completion times) metrics utilizing a mixed-methods approach. The former will include cognitive load with *NASA Task Load Index* (NASA TLX) [100] experienced by experiment participants as well as their opinions on the interface through SUS [75]. Furthermore, to reason about the participant engagement and immersion, we will capture the “feel of presence” [101] levels experienced by the VR system users undergoing public speaking training.

Acknowledgments

This work was supported by grant no LIDER/55/0230/L-11/19/NCBR/2020 by the National Center for Research and Development, Poland. All user studies conducted as part of the project were carried out in accordance with the approval of the Ethics Committee for Research at the Institute of Psychology, Jagiellonian University (decision no. KE/52_2021).

CRediT authorship contribution statement

Patryk Bartyzel: Conceptualization, Methodology, Software, Visualization. **Magdalena Igras-Cybulska**: Conceptualization, Methodology, Investigation, Project administration, Funding acquisition, Writing – original draft. **Daniela Hekiert**: Methodology (psychology), Data Investigation, **Magdalena Majdak**: Methodology (voice analysis), Voice Data Investigation. **Grzegorz Łukawski**: Methodology, Formal analysis, Data curation. **Thomas Bohné**: Methodology, Writing – review. **Sławomir Tadeja**: Supervision, Conceptualization, Methodology, Writing – original draft, Writing – review.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

[1] Bartyzel, P, Igras-Cybulska, M, Tadeja, SK, Majdak, M, Łukawski, G, Hekiert, D. Voice-Responsive Virtual Reality Training Environment for Occupational and Non-Professional Voice Users. In: 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). 2024, p. 208–213. doi:10.1109/VRW62533.2024.00042.

[2] Williams, NR. Occupational groups at risk of voice disorders: a review of the literature. Occupational medicine 2003;53(7):456–460.

[3] Phyland, D, Miles, A. Occupational voice is a work in progress: active risk management, habilitation and rehabilitation. Current opinion in otolaryngology & head and neck surgery 2019;27(6):439.

[4] Niebudek-Bogusz, E, Śliwińska-Kowalska, M. An overview of occupational voice disorders in poland. International journal of occupational medicine and environmental health 2013;26:659–669.

[5] Ziwei, S, Lee, PJ, Jung, H. Prevalence of voice handicap among nurses in intensive care units due to occupational noise during pandemic. Frontiers in Public Health 2023;11:1250512.

[6] Darius, S, Voigt-Zimmermann, S, Böckelmann, I. Effects of occupation-specific vocal stress on the mental health of day care teachers. Folia Phoniatrica et Logopaedica 2023;:1–1.

[7] Vilkman, E. Occupational risk factors and voice disorders. Logopedics Phoniatrics Vocology 1996;21(3-4):137–141.

[8] Świątkowska, B, Hanke, W. Occupational diseases in Poland in 2022, https://www.gov.pl/web/psse-przemysl/choroby-zawodowe-w-polsce-w-2022-roku. 2023.

[9] Huston, MN, Puka, I, Naunheim, MR. Prevalence of voice disorders in the united states: a national survey. The Laryngoscope 2023;.

[10] Tadeja, SK, Godula, K, Bohné, T, Woźniak, MM. Using VR to Present a Mobile MRI Unit in Confined Physical Space: Reporting Results of a Field-Deployment at a Radiology Exhibition. In: 2023 IEEE Intl Conf on Cyber Science and Technology Congress (CyberSciTech). 2023, p. 1051–1057. doi:10.1109/DASC/PiCom/CBDCom/Cy59711.2023.10361485.

[11] Tadeja, SK, Bohné, T, Godula, K, Cybulski, A, Woźniak, MM. Immersive presentations of real-world medical equipment through interactive VR environment populated with the high-fidelity 3D model of mobile MRI unit. Computers & Graphics 2024;120:103919. doi:10.1016/j.cag.2024.103919.

[12] Tadeja, SK, Kipouros, T, Lu, Y, Kristensson, PO. Supporting Decision Making in Engineering Design Using Parallel Coordinates in Virtual Reality. AIAA Journal 2021;59(12):5332–5346. URL: https://doi.org/10.2514/1.J060441. doi:10.2514/1.J060441.

[13] Ovation VR, https://www.ovationvr.com/. 2022.

[14] VirtualSpeech, . VirtualSpeech, https://virtualspeech.com/. 2022.

[15] El-Yamri, M, Romero-Hernandez, A, Gonzalez-Riojo, M, Manero, B. Emotions-responsive audiences for VR public speaking simulators based on the speakers' voice. In: 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT); vol. 2161-377X. 2019, p. 349–353. doi:10.1109/ICALT.2019.00108.

[16] Igras-Cybulska, M, Cybulski, A, Hekiert, D, Majdak, M, Salamon, R, Słomka, P, et al. glossoVR - voice emission and public speech training system. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). 2020, p. 832–833. doi:10.1109/VRW50115.2020.00267.

[17] Igras-Cybulska, M, Hekiert, D, Cybulski, A, Tadeja, S, Witkowski, M, Nakonieczny, K, et al. Towards multimodal vr trainer of voice emission and public speaking-work-in-progress. In: 2023 IEEE Conference on

Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). IEEE; 2023, p. 355–359.
[18] Palmas, F, Cichor, J, Plecher, DA, Klinker, G. Acceptance and effectiveness of a virtual reality public speaking training. In: 2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). 2019, p. 363–371. doi:10.1109/ISMAR.2019.00034.
[19] Sterna, R, Strojny, P, Rębilas, K. Can virtual observers affect our behavior? Social Psychological Bulletin 2019;14(3):1–18.
[20] Remacle, A, Bouchard, S, Etienne, AM, Rivard, MC, Morsomme, D. A virtual classroom can elicit teachers' speech characteristics: evidence from acoustic measurements during in vivo and in virtuo lessons, compared to a free speech control situation. Virtual Reality 2021;25(4):935–944.
[21] Pertaub, DP, Slater, M, Barker, C. An experiment on public speaking anxiety in response to three different types of virtual audience. Presence 2002;11(1):68–78.
[22] Slater, M, Pertaub, DP, Barker, C, Clark, DM. An experimental study on fear of public speaking using a virtual environment. CyberPsychology & Behavior 2006;9(5):627–633.
[23] Siddiqui, H, Irfan, H, Lakhani, AS, Ahmed, B, Shaikh, S, Movania, MM, et al. Manifest: Public speaking training using virtual reality. In: 2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). 2023, p. 468–473. doi:10.1109/ISMAR-Adjunct60411.2023.00102.
[24] Alsaffar, MJ. Virtual Reality Software as Preparation Tools for Oral Presentations: Perceptions From the Classroom. Theory and Practice in Language Studies ????;11(10):1146–1160. URL: https://tpls.academypublication.com/index.php/tpls/article/view/1581. doi:10.17507/tpls.1110.02; number: 10.
[25] Sülter, RE, Ketelaar, PE, Lange, WG. Speakapp-kids! virtual reality training to reduce fear of public speaking in children – a proof of concept. Computers & Education 2022;178:104384. URL: https://www.sciencedirect.com/science/article/pii/S036013152100261X. doi:https://doi.org/10.1016/j.compedu.2021.104384.
[26] Bachmann, M, Subramaniam, A, Born, J, Weibel, D. Virtual reality public speaking training: effectiveness and user technology acceptance. Frontiers in Virtual Reality 2023;4. URL: https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2023.1242544. doi:10.3389/frvir.2023.1242544.
[27] Buchta, K, Wójcik, P, Nakonieczny, K, Janicka, J, Gałuszka, D, Sterna, R, et al. Modeling and optimizing the voice assistant behavior in virtual reality. In: 2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). IEEE; 2022, p. 397–402.
[28] Buchta, K, Wójcik, P, Pelc, M, Górowska, A, Mota, D, Boichenko, K, et al. NUX IVE - a research tool for comparing voice user interface and graphical user interface in VR. In: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). 2022, p. 982–983. doi:10.1109/VRW55335.2022.00342.
[29] Herumurti, D, Yuniarti, A, Rimawan, P, Yunanto, AA. 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). 2019, p. 139–144. doi:10.1109/ICIAICT.2019.8784846.
[30] Leclercq, AL, Ménard, L, Bouchard, S, Moïse-Richard, A. Virtual reality-a promising tool in youths who stutter. In: Oxford Dysfluency Conference. 2021,.
[31] Liu, B. Research on virtual interactive animation design system based on deep learning. Computational Intelligence and Neuroscience 2022;2022.
[32] Wang, H, Gaddy, V, Beveridge, JR, Ortega, FR. Building an emotionally responsive avatar with dynamic facial expressions in human—computer interactions. Multimodal Technologies and Interaction 2021;5(3):13.
[33] Wrońska, G, Sterna, R, Łukawski, G, Igras-Cybulska, M. Exploring user preferences of vr relaxation experiences: A comparative mixed-methods study. In: 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). IEEE; 2024, p. 877–878.
[34] Wellens, WA, Van Opstal, M. Performance stress in professional voice users. Occupational Voice: Care and Cure The Hague, Kugler 2001;:81–100.
[35] Kooijman, P, De Jong, F, Oudes, M, Huinck, W, Van Acht, H, Graamans, K. Muscular tension and body posture in relation to voice handicap and voice quality in teachers with persistent voice complaints. Folia phoniatrica et logopaedica 2005;57(3):134–147.
[36] Brisson, V, Fournier, C, Pelletier, A, Joyal, M, Defoy, L, Tremblay, P. Vocal health and vocal health knowledge among occupational voice users in the province of quebec. Journal of Voice 2022;.
[37] Igras-Cybulska, M, Ziółko, B, Żelasko, P, Witkowski, M. Structure of pauses in speech in the context of speaker verification and classification of speech type. EURASIP Journal on Audio, Speech, and Music Processing 2016;2016:1–16.
[38] Shriberg, E. Spontaneous speech: how people really talk and why engineers should care. In: INTERSPEECH. Citeseer; 2005, p. 1781–1784.
[39] Sugiyati, K, Indriani, L. Exploring the level and primary causes of public speaking anxiety among english department students. Journal of Research on Language Education 2021;2(1):57–66.
[40] Majdak, M. Voice self-assessment in the light of acoustic analysis–case studies. Logopaedica Lodziensia 2023;(8):65–78.
[41] Kramer, E. Judgment of personal characteristics and emotions from nonverbal properties of speech. Psychological Bulletin 1963;60(4):408.
[42] Milo, SK. The Voice Teacher's Guide to Vocal Health for Voice Students: Preventing, Detecting, and Addressing Symptoms. The Ohio State University; 2014.
[43] Scherer, KR, London, H, Wolf, JJ. The voice of confidence: Paralinguistic cues and audience evaluation. Journal of Research in Personality 1973;7(1):31–44.
[44] Lucas, SE, Stob, P. The art of public speaking. McGraw-Hill; 2020.
[45] El-Yamri, M, Romero-Hernandez, A, Gonzalez-Riojo, M, Manero, B. Designing a vr game for public speaking based on speakers features: a case study. Smart Learning Environments 2019;6(1):12.
[46] Melara, RD, Marks, LE. Interaction among auditory dimensions: Timbre, pitch, and loudness. Perception & psychophysics 1990;48(2):169–178.
[47] Oh, Y, Zuwala, JC, Salvagno, CM, Tilbrook, GA. The impact of pitch and timbre cues on auditory grouping and stream segregation. Frontiers in Neuroscience 2022;15:725093.
[48] Weiss, B, Burkhardt, F. Voice attributes affecting likability perception. In: Interspeech. 2010, p. 2014–2017.
[49] Yarnall, S. The 'real'me: Practical application of research into the perception of vocal timbre. Contemporary Music Review 2017;36(6):562–579.
[50] Hirst, DJ, de Looze, C. Measuring speech. fundamental frequency and pitch. Cambridge Handbook of Phonetics 2021;(1):336–361.
[51] Rodero, E. Effectiveness, attractiveness, and emotional response to voice pitch and hand gestures in public speaking. Frontiers in communication 2022;7:869084.
[52] Vaissière, J. Perception of intonation. The handbook of speech perception 2005;:236–263.
[53] Grichkovtsova, I, Lacheret, A, Morel, M. The role of intonation and voice quality in the affective speech perception. In: INTERSPEECH. 2007, p. 2245–2248.
[54] Rodero, E. A comparative analysis of speech rate and perception in radio bulletins. Text & Talk 2012;32(3):391–411.
[55] Feldstein, S, Dohm, FA, Crown, CL. Gender and speech rate in the perception of competence and social attractiveness. The Journal of social psychology 2001;141(6):785–806.
[56] Sahidullah, M, Saha, G. Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. Speech Communication 2012;54(4):543–565. URL: https://www.sciencedirect.com/science/article/pii/S0167639311001622. doi:https://doi.org/10.1016/j.specom.2011.11.004.
[57] Majdak, M, Igras, M. Metafory głosu–analiza akustyczna. Prace Filologiczne" nr 2014;66:179–199.
[58] Terasawa, H, Berger, J, Makino, S. In search of a perceptual metric for timbre: Dissimilarity judgments among synthetic sounds with mfcc-derived spectral envelopes. Journal of the Audio Engineering Society 2012;60(9):674–685.
[59] Bäckström, T, Räsänen, O, Zewoudie, A, Zarazaga, PP, Koivusalo, L, Das, S, et al. Introduction to Speech Processing. 2 ed.; 2022. URL: https://speechprocessingbook.aalto.fi. doi:10.5281/

zenodo.6821775.

[60] Dimitrova-Grekow, T, Klis, A, Igras-Cybulska, M. Speech emotion recognition based on voice fundamental frequency. Archives of Acoustics 2019;:277–286.

[61] Hekiert, D, Igras-Cybulska, M. Capturing emotions in voice: A comparative analysis of methodologies in psychology and digital signal processing. Roczniki Psychologiczne 2019;22(1):15–34.

[62] Majdak, M, Igras, M, Domeracka-Kołodziej, A. Looking for natural voice—the effectiveness of the program of postgraduate studies of voice and speech training. In: XXII Annual Pacific Voice Conference (PVC). IEEE; 2014, p. 1–6.

[63] Talkin, D, Kleijn, WB. A robust algorithm for pitch tracking (rapt). Speech coding and synthesis 1995;495:518.

[64] Ferro, M, Tamburini, F. Using deep neural networks for smoothing pitch profiles in connected speech. IJCoL Italian Journal of Computational Linguistics 2019;5(5-2):33–48.

[65] Kobayashi, M, Morimoto, M, Sato, H, Sato, H. Optimum speech level to minimize listening difficulty in public spaces. The Journal of the Acoustical Society of America 2007;121(1):251–256.

[66] Tadeja, SK, Rydlewicz, W, Lu, Y, Bubas, T, Rydlewicz, M, Kristensson, PO. Measurement and Inspection of Photo-Realistic 3-D VR Models. IEEE Computer Graphics and Applications 2021;41(6):143–151. doi:10.1109/MCG.2021.3114955.

[67] Albrecht, M, Assländer, L, Reiterer, H, Streuber, S. MoPeDT: A Modular Head-Mounted Display Toolkit to Conduct Peripheral Vision Research. In: IEEE VR. 2023, p. 691–701. doi:10.1109/VR55154.2023.00084.

[68] McGill, M, Boland, D, Murray-Smith, R, Brewster, S. A Dose of Reality: Overcoming Usability Challenges in VR Head-Mounted Displays. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15; New York, NY, USA: ACM. ISBN 9781450331456; 2015, p. 2143–2152. URL: https://doi.org/10.1145/2702123.2702382. doi:10.1145/2702123.2702382.

[69] Zhang, C. Investigation on motion sickness in virtual reality environment from the perspective of user experience. In: 2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE). 2020, p. 393–396. doi:10.1109/ICISCAE51034.2020.9236907.

[70] Turner, CW, Lewis, JR, Nielsen, J. Determining usability test sample size. International encyclopedia of ergonomics and human factors 2006;3(2):3084–3088.

[71] Brislin, RW. Back-translation for cross-cultural research. Journal of cross-cultural psychology 1970;1(3):185–216.

[72] Paul, BG, Hart, EL. A personal report of confidence as a speaker. Behavioral Therapy 1974;5:628–638.

[73] Hook, JN, Smith, CA, Valentiner, DP. A short-form of the personal report of confidence as a speaker. Personality and individual differences 2008;44(6):1306–1313.

[74] Heeren, A, Ceschi, G, Valentiner, DP, Dethier, V, Philippot, P. Assessing public speaking fear with the short form of the personal report of confidence as a speaker scale: confirmatory factor analyses among a french-speaking community sample. Neuropsychiatric disease and treatment 2013;:609–618.

[75] Brooke, J. Sus: a "quick and dirty'usability. Usability evaluation in industry 1996;189(3):189–194.

[76] Kline, P. Handbook of psychological testing. Routledge; 2013.

[77] Brooke, J. Sus: a retrospective. J Usability Stud 2013;8:29–40.

[78] Murad, C, Munteanu, C, Cowan, BR, Clark, L. Revolution or evolution? speech interaction and hci design guidelines. IEEE Pervasive Computing 2019;18(2):33–45. doi:10.1109/MPRV.2019.2906991.

[79] Chollet, M, Wörtwein, T, Morency, LP, Shapiro, A, Scherer, S. 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. 2015, p. 1143–1154.

[80] Yang, J, Sasikumar, P, Bai, H, Barde, A, Sörös, G, Billinghurst, M. The effects of spatial auditory and visual cues on mixed reality remote collaboration. Journal on Multimodal User Interfaces 2020;14(4):337–352.

[81] Kawase, T, Sakamoto, S, Hori, Y, Maki, A, Suzuki, Y, Kobayashi, T. Bimodal audio–visual training enhances auditory adaptation process. Neuroreport 2009;20(14):1231–1234.

[82] Dey, A, Chatburn, A, Billinghurst, M. Exploration of an eeg-based cognitively adaptive training system in virtual reality. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). 2019, p. 220–226. doi:10.1109/VR.2019.8797840.

[83] Gallagher, T, Slof, B, van der Schaaf, M, Arztmann, M, Fracaro, SG, Kester, L. Learning analytics dashboard design: Workplace learner preferences for reference frames in immersive training in practice. Journal of Computer Assisted Learning 2024;.

[84] Shneiderman, B. Why not make interfaces better than 3d reality? IEEE Computer Graphics and Applications 2003;23(6):12–15. doi:10.1109/MCG.2003.1242376.

[85] Nielsen, J. Usability inspection methods. In: Conference Companion on Human Factors in Computing Systems. CHI '94; New York, NY, USA: Association for Computing Machinery. ISBN 0897916514; 1994, p. 413–414. doi:10.1145/259963.260531.

[86] Nielsen, J. Enhancing the explanatory power of usability heuristics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '94; New York, NY, USA: Association for Computing Machinery. ISBN 0897916506; 1994, p. 152–158. doi:10.1145/191666.191729.

[87] Nielsen, J. Why you only need to test with 5 users. 2000. URL: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/; accessed: 2024-11-08.

[88] Slater, M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philosophical Transactions of the Royal Society B: Biological Sciences 2009;364(1535):3549–3557.

[89] Jahangir, R, Teh, YW, Hanif, F, Mujtaba, G. Deep learning approaches for speech emotion recognition: State of the art and research challenges. Multimedia Tools and Applications 2021;80(16):23745–23812.

[90] Skrzek, T, Malawski, F, Tadeja, S, Hekiert, D, Hemmerling, D, Igras-Cybulska, M. Automatic recognition of gesture qualities in speech delivery in vr-work-in-progress. In: 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). 2024, p. 624–627. doi:10.1109/VRW62533.2024.00122.

[91] Sterna, R, Cybulski, A, Igras-Cybulska, M, Pilarczyk, J, Siry, A, Kuniecki, M. Psychophysiology, eye-tracking and vr: exemplary study design. In: 2021 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW). IEEE; 2021, p. 639–640.

[92] Morina, N, Ijntema, H, Meyerbröker, K, Emmelkamp, PM. Can virtual reality exposure therapy gains be generalized to real-life? a meta-analysis of studies applying behavioral assessments. Behaviour research and therapy 2015;74:18–24.

[93] Anderson, P, Jacobs, C, Rothbaum, BO. Computer-supported cognitive behavioral treatment of anxiety disorders. Journal of clinical psychology 2004;60(3):253–267.

[94] Slater, M, Usoh, M. Body centred interaction in immersive virtual environments. Artificial life and virtual reality 1994;1(1994):125–148.

[95] Piumsomboon, T, Lee, Y, Punpongsanon, P. Supernatural enhancements of perception, interaction, and collaboration in mixed reality. 2023.

[96] Pięta, P, Jegierski, H, Babiuch, P, Jegierski, M, Płaza, M, Łukawski, G, et al. Automated classification of virtual reality user motions using a motion atlas and machine learning approach. IEEE Access 2024;.

[97] Tadeja, SK, Lu, Y, Rydlewicz, M, Rydlewicz, W, Bubas, T, Kristensson, PO. Exploring Gestural Input for Engineering Surveys of Real-Life Structures in Virtual Reality using Photogrammetric 3D Models. Multimedia Tools and Applications 2021;80(20):31039–31058. URL: https://doi.org/10.1007/s11042-021-10520-z. doi:10.1007/s11042-021-10520-z.

[98] Dargue, N, Sweller, N, Jones, MP. When our hands help us understand: A meta-analysis into the effects of gesture on comprehension. Psychological Bulletin 2019;145(8):765.

[99] Igras-Cybulska, M, Hekiert, D, Cybulski, A, Tadeja, S, Witkowski, M, Nakonieczny, K, et al. Towards Multimodal VR Trainer of Voice Emission and Public Speaking -Work-in-Progress. In: 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). 2023, p. 355–359. doi:10.1109/VRW58643.2023.00079.

[100] Hart, S, Staveland, L. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. Advances in Psychology 1988;:139–183doi:10.1016/s0166-4115(08)62386-9.

[101] Schubert, T, Friedmann, F, Regenbrecht, H. The experience of presence: Factor analytic insights. Presence: Teleoper Virtual

Environ 2001;10(3):266–281. URL: https://doi.org/10.1162/105474601300343603. doi:10.1162/105474601300343603.