

Damian Bereska¹

Silesian University of Technology

ORCID: 0000-0003-1110-8212

Karol Jędrasiak²

WSB University

ORCID: 0000-0002-2254-1030

Robert Wolański³

School of Aspirants of the State Fire Service in Krakow

ORCID: 0000-0002-5625-0936

Algorithm for automatic analysis and evaluation of communication skills of participants in VR public speaking exercises

Algorytm automatycznej analizy i oceny
umiejętności komunikacyjnych uczestników ćwiczeń VR
z zakresu wystąpień publicznych

Introduction

Virtual reality (VR) technology has opened new vistas in numerous domains, one of which is the realm of communication skills training. This

¹ Damian Bereska: dr inż. Politechnika Śląska, e-mail: Damian.Bereska@polsl.pl

² Karol Jędrasiak: dr inż., Akademia WSB w Dąbrowie-Górnicy,
e-mail: kjedrasiak@wsb.edu.pl

³ Robert Wolański: st. bryg. w st. spocz. dr inż., Szkoła Aspirantów Państwowej Straży
Pożarnej w Krakowie, email: rwolanski@sapsp.pl

research paper aims to explore and establish an algorithm for the automatic analysis of communication skills in public speaking exercises within a VR environment. Our study delves into not only the development of such an algorithm but also its implications and potential in enhancing the efficacy of public speaking training.

To set the context for our research, it is crucial to understand the current landscape of communication skills training methods and the role of technology therein. Traditional approaches, such as in-person coaching and video recording analysis, have been the mainstay for years. However, these methods come with inherent limitations, such as the requirement of human resources for evaluation, the subjective nature of feedback, and the lack of immersive, real-time environments for practice.

Virtual reality, as a technological innovation, addresses many of these limitations by providing an immersive environment for users to practice and hone their public speaking skills. VR allows for a controlled yet realistic setting where various scenarios and audience types can be simulated. This not only aids in reducing the logistical challenges of traditional training methods but also provides a safe space for learners to practice and receive instant feedback.

However, the utilization of VR in communication skills training is not without its challenges. The primary issue lies in the development of an effective algorithm that can accurately analyze and provide feedback on a participant's performance. This involves the intricate process of collecting and interpreting a wide array of data, ranging from speech patterns to non-verbal cues like gestures and eye contact. The complexity is further amplified when considering the personalization aspect of the training, catering to individual strengths and weaknesses, and the inclusion of participants with disabilities.

Our research is positioned at this juncture, aiming to bridge the gap by developing an algorithm that not only assesses a participant's communication skills in a VR setting but also provides personalized, objective, and comprehensive feedback. This paper outlines the algorithm's framework, its foundational principles, and its application in the context of VR-based public speaking exercises. We believe that the successful implementation of this algorithm can revolutionize the way communication skills are taught and

learned, making the process more efficient, accessible, and adaptable to a wide range of learners.

The method proposed in this study was subjected to a rigorous evaluation using a dataset comprising 20 recordings derived from various public speaking events in VR. This dataset was utilized to assess the efficacy of the method. Additionally, a comparative analysis was conducted aligning the outcomes of the method with assessments made by subject matter experts in the field. The observed high degree of congruence between the results obtained from the method and the expert evaluations provides a strong impetus for the authors to pursue further research and development in this particular domain.

In summary, this research contributes to the growing field of VR in education and training, specifically focusing on the nuanced area of communication skills development. By exploring the potential of VR in this domain, we aim to provide a comprehensive solution that overcomes the limitations of traditional training methods, paving the way for a new era in public speaking training.

Methods of Automated Analysis and Assessment of Communication Skills in VR

This chapter provides an overview of methods for the automatic analysis of communication skills. The authors focused on identifying factors and methods used in existing solutions for the automatic assessment of communication abilities.

Extensive scholarly work has delved into the development and refinement of research prototypes that harness audio and visual signals for the appraisal of presentation skills. A notable study (Scherer et al., 2012, p. 1114–1120) executed an integrated audiovisual analysis on a corpus comprising political speeches, uncovering various indicators pertinent to the assessment of speaking proficiency. Another significant contribution (Kurihara et al., 2007, p. 358–365) involved the creation of a system designed for presentation coaching, utilizing an array of sophisticated technologies including Automatic Speech Recognition (ASR), prosodic analysis, and image processing techniques. In

this context, a system employing marker-based computer vision for head movement tracking was developed. Despite this advancement, it has been observed that methods relying solely on marker-based or video-based tracking for monitoring bodily movements tend to suffer from practical limitations and susceptibility to inaccuracies. The advent of 3D motion tracking technologies, exemplified by devices like the Microsoft Kinect, has markedly enhanced the scope and efficacy of multimodal research endeavors. In a pioneering initiative (Nguyen, Chen, Rauterberg, 2012), advocated for the establishment of a system dedicated to the evaluation of public speaking, illustrating its application by recording students in a scientific presentation course via a Kinect device. Similarly (Batrınca et al., 2013, p. 116–128), embarked on the development of a training system for public speaking skills, integrating advanced multimodal sensing and virtual human interaction technologies. This system featured a virtual audience that interacted dynamically with the speaker, providing real-time feedback and training opportunities based on the quality of the presentation delivered. Additionally, the field of affective computing has been increasingly applied to the domain of public speaking, with a particular focus on analyzing speakers' stress responses (Giraud, Soury, Hua, 2013, p. 417–422; Kleinsmith, Bianchi-Berthouze, 2013). Two primary trends emerge from the review of relevant literature. Firstly, there is a technological convergence towards the utilization of 3D sensors like Kinect, in synergy with audio/video recording, as a foundational approach for the evaluation of presentation performances. Secondly, there is a burgeoning interest in the design and implementation of interactive systems specifically for the training of presentation skills.

We posit that the integration of multimodal technology could revolutionize traditional methods of public speaking assessment, which are typically scored by humans (Schreiber, Paul, Shibley, 2012) enhancing their reliability and cost-effectiveness. Consequently, our primary research question centers on the feasibility of machine-generated multimodal features to accurately predict human scores, derived from a meticulously crafted rubric designed for evaluating public speaking performances. It is important to note that the multimodal research conducted thus far (Nguyen et al., 2012; Batrınca

et al., 2013, p. 116–128) has predominantly relied on non-standardized scoring rubrics and human scoring processes, which raises concerns regarding their reliability. Moreover, a critical aspect often overlooked in these studies is the content of the speech itself, which remains a central element in conventional assessments of public speaking. To address this gap, our approach includes the application of natural language processing techniques to speech transcripts as an initial step toward quantifying content-related features. This approach aims to integrate a more comprehensive analysis that encompasses not only the delivery but also the substance of the speeches, thereby providing a more holistic evaluation of public speaking competencies.

As of the present day, the field of automatic evaluation of public speaking based on machine analysis of data collected during the speech remains relatively underdeveloped. The authors of the publication (Chen et al., 2014, p. 200–203) have recognized the Public Speaking Competence Rubric (PSCR), (Schreiber et al., 2012) as a foundational benchmark for the automatic evaluation of public speaking quality. The PSCR, developed for assessments by experts in public speaking, comprises an 11-item descriptive rubric designed to be accessible and understandable to audiences both within and outside the communication discipline. This rubric serves as a pivotal reference point toward which automatic quality assessment of public speaking should aspire. The authors (Chen et al., 2014, p. 200–203) presented that a combination of basic multimodal features, encompassing aspects of speech content, speech delivery (including fluency, pronunciation, and prosody), and nonverbal behaviors (such as movements of the head, body, and hands) can collectively yield significant predictions of human scores on presentation performance. This finding is a promising stride in the journey toward the development of an automated system for assessing public speaking skills. It not only marks a significant initial achievement but also highlights the urgent need for continued research in this area. Further explorations and enhancements in this field necessitate the expansion of the multimodal corpus. This involves not only accumulating more diverse data but also refining the quality and complexity of the features analyzed. One such advanced feature that merits attention is gesture recognition, which can provide deeper insights

into the nonverbal aspect of communication. Gesture recognition, when effectively integrated into the analysis, could greatly enhance the precision and depth of the assessment. Moreover, there is a pressing need to develop more sophisticated and robust scoring models for multimodal presentation assessment. These models should be capable of not only capturing the diverse aspects of public speaking but also of providing nuanced evaluations that align closely with human judgments. The ultimate goal is to create an automated system that not only replicates but potentially surpasses human expertise in assessing public speaking skills. This pursuit entails a multidisciplinary approach, combining insights and methodologies from communication studies, computer science, linguistics, psychology, and data analytics. By leveraging advanced technologies such as machine learning algorithms, natural language processing, virtual reality, and computer vision, researchers can construct a system that offers a comprehensive, objective, and nuanced evaluation of public speaking performances.

The potential applications of such a system are vast and varied. It can serve as an invaluable tool in educational settings, providing students with immediate, detailed feedback on their presentation skills and guiding them toward improvement. In the professional realm, this technology can be instrumental in training employees and leaders, enhancing their communication abilities, and thus contributing to more effective and persuasive public speaking. Furthermore, the development of this technology can contribute significantly to the broader field of automated assessment tools. It can provide insights into the challenges and possibilities of automating complex human skills evaluations, paving the way for future innovations in this domain. In conclusion, the research and development of an automated system for evaluating public speaking skills based on multimodal data analysis is not only a challenging endeavor but also an exciting opportunity. It promises to revolutionize the way public speaking skills are assessed and trained, offering a more objective, comprehensive, and accessible tool for individuals across various fields. As this technology evolves, it has the potential to significantly impact both educational practices and professional development, ultimately enhancing the art of public speaking in the digital age.

Available Data in VR for Analysis

As noted in the preceding chapter, the utilization of virtual reality (VR) technology for training in public speaking might be a groundbreaking approach, especially considering the role of data collection and analysis in assessing a speaker's communicative abilities. This chapter aims to present available data types in VR for analysis.

VR training environments are meticulously crafted to simulate real-life speaking scenarios (il. 1), offering an immersive experience that captures extensive data vital for appraising a speaker's performance. This data spans a wide array of variables, encompassing everything from the precise recording of specific events within the VR session to the detailed examination of the speaker's vocal attributes and body language. For example, event recording in VR not only covers the timing of speech initiation and conclusion but also includes interactions with the virtual environment and audience reactions. These interactions can yield valuable insights, providing a richer context for understanding how speakers adapt to various scenarios and engage with their audience. Moreover, temporal data, such as the overall duration of the speech, the length of pauses, and the distribution of time across different segments of the presentation, is scrupulously documented. Such information is indispensable for assessing the speech's pacing and rhythm, which are critical elements of effective public speaking. Furthermore, voice analysis in VR goes beyond basic metrics like volume and pitch; it incorporates also speech analysis techniques, such as Speech2Text conversions, allowing for a more nuanced evaluation of the verbal content, including factors like clarity, fluency, and emotional expression.

Equally important is the analysis of physical communication, encompassing both head movements and gestures made using VR controllers. Tracking the participant's head position and orientation within the VR environment provides insights into their posture, attention direction, and engagement level with the virtual audience. This spatial data is further enhanced by analyzing the movements of VR controllers, which reveals information about the speaker's use of hand gestures and overall body language.



Il. 1. Examples of public speeches conducted using virtual reality. First row from the left: Ovation VR scenario 1, Ovation VR scenario 2. Second row from the left: Virtual Speech, eZawody

These non-verbal elements play a crucial role in communication, often conveying subtle messages that supplement the spoken words. The overarching objective of this extensive data collection and analysis is to offer actionable insights that can significantly improve a participant's communication skills. The data presents a comprehensive picture of the speaker's performance, addressing both verbal and non-verbal communication facets. For instance, the evaluation of speech delivery focuses not just on content but also on its presentation, considering aspects such as clarity, fluency, and emotional impact. The analysis of non-verbal communication extends beyond simply quantifying gestures; it delves into the appropriateness and impact of body language, including factors like posture, hand movements, and eye contact.

By harnessing the immersive capabilities of VR and the detailed data it provides, trainers and speakers can identify specific areas for improvement. This approach allows for targeted training interventions, enabling speakers to refine both their verbal and non-verbal communication skills in a controlled yet realistic environment. Moreover, the VR setting offers a safe space for speakers to experiment with different styles and techniques, receiving

immediate feedback based on the comprehensive data analysis. In addition to individual skill enhancement, the data gathered from VR training sessions can contribute to broader research in the field of communication. By analyzing aggregate data from multiple participants, patterns, and trends in public speaking can be identified, potentially leading to new insights and best practices. This data-driven approach to public speaking training not only benefits the individual speaker but also contributes to the evolving understanding of effective communication strategies. Utilization of VR technology in public speaking training represents a significant leap forward in the field of communication. By providing a rich, multi-dimensional dataset covering both verbal and non-verbal aspects of speech delivery, VR enables a more holistic and effective approach to developing public speaking skills. This technology not only offers individual speakers the tools to enhance their performance but also contributes to the larger body of knowledge in communication studies, paving the way for future innovations in training and assessment methods.

In the context of utilizing virtual reality (VR) technology for the automated assessment of a user's communication skills, a variety of data types are collected and made available for algorithmic analysis. This chapter outlines these data types, underscoring their significance in evaluating communicative proficiency within a VR environment.

- **Event Registration within Scenario Logic:** This involves the meticulous recording of various events as they occur within the logic of the VR scenario. It includes the tracking of specific actions taken by the user, interactions with the virtual environment, and responses triggered within the VR narrative. This data is crucial for understanding how the user navigates and responds to the simulated public speaking scenario.
- **Time Elapsed:** The measurement of elapsed time is a fundamental aspect of data collection in VR. This includes the duration of the entire session, the length of speeches, the timing of pauses, and the distribution of time across different segments of the presentation. Time-related data provides insights into the pacing and rhythm of

the user's speech, which are vital components of effective communication.

- **Voice Data:** The collection of voice data is comprehensive, extending beyond basic parameters like volume and pitch. It encompasses detailed aspects such as speech clarity, fluency, tonal variations, and emotional inflections. Advanced voice analysis techniques, such as Speech2Text conversion, are employed to delve deeper into the verbal content and receive 'words', offering a more granular assessment of the speaker's vocal delivery.
- **Head Movements:** The tracking of head movements involves capturing the XYZ positional coordinates and RX, RY, and RZ rotational angles in the 3D scene setup of the VR scenario, per unit of time elapsed. This data provides insights into the speaker's posture, orientation, and level of engagement with the audience, which are critical for effective communication.
- **VR Controller Movements (Controller 1):** The movements of the first VR controller are tracked in terms of its XYZ positional coordinates and RX, RY, and RZ rotational angles within the 3D scene of the scenario, for each time unit. This data is essential for understanding the use of hand gestures and their contribution to the overall body language of the speaker.
- **VR Controller Movements (Controller 2):** Similarly, the movements of the second VR controller are recorded, capturing the XYZ position and RX, RY, and RZ rotation in the 3D scene per time unit. This adds another layer of depth to the analysis of non-verbal cues, particularly in terms of how both hands are used in conjunction to enhance communication.

The following processed raw input data, derived from VR-based interactions and analyses was used by the authors:

- **Facial Expressions (Mouth Movements):** This input is derived from the analysis of the 'Words' variable. It focuses on the movements of the mouth, providing insights into the nuances of facial expressions, which are crucial for effective communication.

- **Gestures:** Utilizing the XYZ positional data and RXRYRZ rotational data from VR controllers 1 and 2, gestures are analyzed. This algorithm encompasses the movement and positioning of the hands and arms, offering a detailed understanding of how gestures contribute to the overall communicative process.
- **Spatial Movement (Territory):** Based on the XYZ position of the head, this input parameter assesses the speaker's movement within the VR space. It offers insights into how the speaker navigates and utilizes physical space during a presentation, which can be indicative of confidence and engagement.
- **Voice Tone:** The tone of the voice is analyzed using Sound data. This aspect of the analysis evaluates the nuances in the speaker's voice tone, which can convey various emotional states and levels of confidence.
- **Voice Timbre:** Also derived from Sound analysis, this parameter examines the quality or color of the voice. Voice timbre can affect the perception of the speaker's message and is an essential component of effective verbal communication.
- **Speech Tempo:** The pace at which words are spoken is another critical aspect derived from Sound analysis. Speech tempo can influence the clarity of the message and the audience's ability to follow the speaker's thoughts.
- **Words:** This involves mapping the 'Words' variable to analyze the actual verbal content. It includes the recognition of spoken words, their arrangement, and their relevance to the topic.

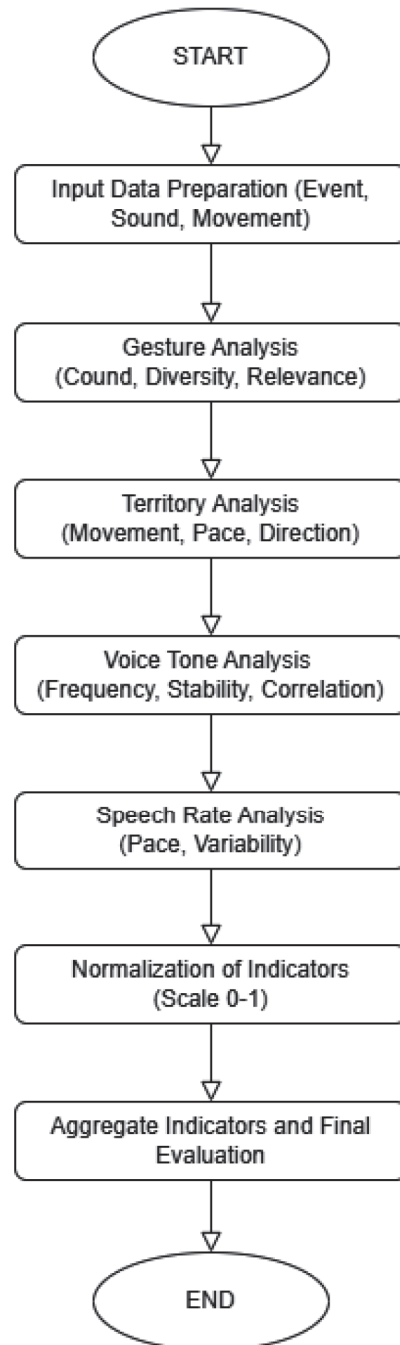
Further processing these variables enables the extraction of various additional information. For example, voice analysis can reveal not just the words being spoken but also their pacing and other characteristics initially discussed in the chapter. The subsequent chapter will elaborate on how the authors of the article decided to process the available data for automated analysis and evaluation of public speaking performances.

Algorithm for automatic assessment of the participant's communication skills in public speaking exercises

Considering the unique requirements of users with disabilities, as discussed with their caregivers and instructors, a decision was made to develop an algorithm that is both comprehensible for the instructor and the exercise participant and crucially, also repeatable. This algorithm for automated analysis of communication skills comprises the following steps (il. 2).

Step 1: Preparation of Input Data

1. Processing data from event registration, time elapsed, sound, words, and movements of the head and VR controllers.
2. Normalization of the sound signal.
3. Conversion of sound to WAV format with a frequency of 44.1 KHz.
4. Conversion of sound to text using Mozilla DeepSpeech (Speech2Text).
5. Sound analysis to obtain information about the tone of voice by determining the fundamental frequency (F0) values for successive sound segments and observing changes over time. F0 is calculated using the YIN method.
6. Sound analysis to obtain information about the timbre of the voice, which is associated with the individual characteristics of the speaker's voice. This involves calculating the Mel-frequency cepstral coefficients (MFCC).
7. Identification of syllables in the sound sample, which includes:
 - a) Calculating the energy of the sound signal for subsequent time frames in a 30 ms time window with a 15 ms overlap.
 - b) Calculating F0 values for subsequent time frames using the YIN method.
 - c) Selecting energy and F0 thresholds that differentiate syllables from intersyllabic pauses.
 - d) Detecting syllables based on exceeding energy and F0 thresholds, as a syllable is typically characterized by higher energy and a



II. 2. Flowchart illustrating the process of multimodal data analysis for behavioral evaluation. The process begins with input data preparation, including events, sounds, and movements. Subsequent steps involve gesture analysis (count, diversity, relevance), territory analysis (movement, pace, direction), voice tone analysis (frequency, stability, correlation), and speech rate analysis (pace, variability). The data is then normalized to a scale of 0–1, leading to the aggregation of indicators and the final evaluation.

distinct F0 value. It allows for the necessity of “calibration” before beginning automatic analysis.

8. Analysis of sound to obtain information about the speaking rate, understood as the number of syllables per unit of time.

This comprehensive step involves multiple layers of sound processing and analysis, aiming to accurately capture and interpret various aspects of the speaker’s voice and speech patterns. The normalization and conversion of the sound signal ensure that the data is in a consistent and analyzable format. The use of advanced techniques like Mozilla DeepSpeech for Speech2Text conversion and the YIN method for F0 analysis, coupled with MFCC calculations, facilitates a detailed examination of the voice’s tone and timbre. Syllable identification, incorporating sound energy and frequency analysis, allows for a nuanced understanding of speech pacing and rhythm. The need for initial calibration underscores the importance of tailoring the analysis to individual speaker characteristics, ensuring that the assessment is as accurate and personalized as possible. This step lays a crucial foundation for the automated analysis of communication skills, as it meticulously processes and prepares the diverse range of auditory and verbal data collected during VR exercises.

Step 2: Gesture analysis based on XYZ, RXRYRZ data of VR controllers 1 and 2

The analysis of gestures in the context of virtual reality (VR) training exercises involves a detailed examination of the data obtained from VR controllers 1 and 2, particularly focusing on their XYZ positioning and RXRYRZ rotational information. This process includes several key aspects:

1. **Gesture Detection:** Identifying gestures based on the movements of the VR controllers, employing machine learning techniques to recognize specific gestures.
2. **Number of Gestures:** Calculating the total number of gestures made by the participant during their presentation.

3. Diversity of Gestures: Assessing the variety of gestures used, taking into account different types of gestures such as illustrative, accentuating, and symbolic gestures.
4. Relevance of Gestures to Speech Content: Evaluating the extent to which the participant's gestures are related to the content of their speech and support verbal communication. This evaluation is based on the analysis using deep neural networks to be described in another article.
5. Gesticulation Index:
 - The formula for the Gesticulation Index is given as: $\text{Gesticulation_Index} = w_1 * \text{number_of_gestures} + w_2 * \text{diversity_of_gestures} + w_3 * \text{relevance_of_gestures_to_content}$, where w_1 , w_2 , w_3 represent the weights corresponding to each factor. Interpretation of the Index:
 - » If the Gesticulation_Index is below a lower threshold, it indicates insufficient gesticulation by the participant.
 - » If the Gesticulation_Index exceeds an upper threshold, it suggests excessive gesticulation by the participant.
 - » If the Gesticulation_Index falls between the lower and upper thresholds, the participant's gesticulation level is considered appropriate.

This comprehensive approach to gesture analysis in VR public speaking training incorporates advanced data processing and machine learning algorithms to offer a nuanced understanding of non-verbal communication skills. By quantifying and assessing gestures in relation to the spoken content, the algorithm provides valuable insights into the effectiveness and appropriateness of the participant's body language, thereby contributing to a more holistic evaluation of their communication proficiency. The Gesticulation Index serves as a critical metric, guiding participants towards achieving a balanced and effective use of gestures, which is essential for impactful and engaging public speaking.

Step 3: Territory analysis (movement) based on XYZ head position data

In the analysis of participant movement during public speaking exercises in a virtual reality (VR) environment, several key metrics are calculated to assess the physical dynamics of the presentation:

1. Calculation of Distance Covered: This involves measuring the total distance traveled by the participant throughout the presentation. It is determined by summing the lengths of each movement segment between consecutive time points.
2. Calculation of Movement Pace: The pace of movement is quantified by dividing the total distance covered by the duration of the presentation.
3. Analysis of Directional Changes: This is defined as a change in direction exceeding a threshold angular degree of 30 degrees.
4. Territory Indicator:
 - The formula for the Territory Indicator is: $\text{Territory_Indicator} = w1 * \text{distance} + w2 * \text{pace_of_movement} + w3 * \text{number_of_directional_changes}$, where $w1, w2, w3$ are weights corresponding to each factor. Interpretation of the Indicator:
 - » If the Territory_Indicator is below a lower threshold, it suggests that the participant is overly static.
 - » If the Territory_Indicator exceeds an upper threshold, it indicates excessive movement by the participant.
 - » If the Territory_Indicator falls between the lower and upper thresholds, the participant's movement is considered at an appropriate level.

Through this detailed analysis, a comprehensive understanding of the participant's spatial utilization and movement dynamics during a presentation is developed. The measurement of the distance traveled and the pace of movement, combined with the analysis of directional changes, provides insights into how effectively the participant navigates and occupies the speaking area. The Territory Indicator serves as a critical metric in evaluating the balance and appropriateness of physical movement, contributing to the

overall assessment of public speaking skills. This metric is particularly useful in VR training environments, where spatial awareness and effective use of physical space are key components of successful communication.

Step 4: Analyze the tone of voice per unit of time – it should vary accordingly

In the process of analyzing the voice qualities of participants during public speaking exercises, several key metrics related to the fundamental frequency (F0) are calculated:

5. **Fundamental Frequency Range (F0):** This involves calculating the range of the participant's fundamental frequency, taking into account the minimum, maximum, and average F0 values throughout the presentation. This range needs to be calibrated to the participant's gender, age, and individual voice characteristics.
6. **Voice Tone Stability:** The stability of the participant's voice tone is analyzed by examining the variations in the fundamental frequency. Stability is measured as the standard deviation of F0.
7. **Correlation of Voice Tone with Emotions and Speech Content:** This assessment is conducted using deep neural networks to evaluate the extent to which the participant's voice tone correlates with the content of their speech and supports verbal communication.
8. **Voice Tone Correctness Index:**
 - The formula for the Voice Tone Correctness Index is: $\text{Voice_Tone_Index} = w_1 * \text{F0_range} + w_2 * \text{tone_stability} + w_3 * \text{tone_content_correlation}$, where w_1 , w_2 , w_3 are the weights corresponding to each factor. Interpretation of the Index:
 - » If the Voice_Tone_Index is below a lower threshold, it indicates that the participant's voice tone is too monotonous.
 - » If the Voice_Tone_Index exceeds an upper threshold, it suggests that the participant's voice tone is excessively variable.
 - » If the Voice_Tone_Index falls between the lower and upper thresholds, the participant's voice tone is considered appropriate.

Through this comprehensive voice analysis, the participant's ability to modulate their voice by the content and emotional context of their speech is evaluated. The assessment of the fundamental frequency range, tone stability, and the correlation of voice tone with speech content offers insights into the effectiveness and appropriateness of the participant's vocal delivery. The Voice Tone Correctness Index serves as a crucial metric in determining the quality of voice modulation, contributing significantly to the overall evaluation of public speaking skills. This metric is particularly valuable in VR training environments where vocal expression plays a key role in successful communication.

Step 5: Analyze the timbre of the voice – analyze whether the timbre of the voice is in a range that is pleasing to the listener

In evaluating the vocal qualities of participants in public speaking exercises, the following key aspects are analyzed to assess the timbre of the voice:

1. **Clarity:** The clarity of the participant's voice is assessed by measuring the ratio of the signal's energy in the higher formants to that in the lower formants. This analysis helps in understanding how clear and distinct the voice sounds.
2. **Resonance:** The resonance of the participant's voice is evaluated by analyzing the harmonic characteristics of the sound signal, including the amplitude of the harmonic series and the harmonic-to-noise ratio (HNR).
3. **Projection:** The projection of the participant's voice is assessed by analyzing characteristics of the sound signal related to loudness and carry, such as the sound signal's power level.
4. **Voice Expression:** The expression of the participant's voice is evaluated by analyzing the sound signal's features related to intonation, articulation, and the dynamics of speech.
5. **Voice Timbre Index:**
 - The formula for the Voice Timbre Index is: $\text{Voice_Timbre_Index} = w_1 * \text{clarity} + w_2 * \text{resonance} + w_3 * \text{projection} + w_4 * \text{voice_}$

expression, where w_1 , w_2 , w_3 , w_4 represent the weights corresponding to each factor. Interpretation of the Index:

- » If the Voice_Timbre_Index is below a lower threshold, it indicates that the participant's voice timbre is insufficiently distinctive.
- » If the Voice_Timbre_Index exceeds an upper threshold, it suggests that the participant's voice timbre is overly distinctive.
- » If the Voice_Timbre_Index falls between the lower and upper thresholds, the participant's voice timbre is considered at an appropriate level.

This detailed assessment of voice timbre encompasses various dimensions of vocal quality, from basic clarity and resonance to the more complex aspects of projection and expression. The Voice Timbre Index serves as a comprehensive metric for evaluating the effectiveness and appropriateness of the participant's vocal characteristics, contributing significantly to the overall assessment of their public speaking skills. This index is particularly useful in VR training environments, where the nuances of voice play a crucial role in successful communication and audience engagement.

Step 6: Speech rate analysis – analyze whether the speaking rate is too fast or too slow

In assessing the speech pace of participants in public speaking exercises, the following metrics are analyzed:

1. Variability in Speech Pace: This is calculated by determining the standard deviation of the Speech Per Minute (SPM) rate for each segment of the speech. This measurement provides insights into the consistency of the speech pace throughout the presentation.
2. Speech Pace Correctness Index:
 - The formula for this index is: $\text{Speech_Pace_Index} = w_1 * \text{SPM} + w_2 * \text{variability_in_pace}$, where w_1 and w_2 are the weights corresponding to each factor. Interpretation of the Index:
 - » If the Speech_Pace_Index is below a lower threshold, it indicates that the participant is speaking too slowly.

- » If the `Speech_Pace_Index` exceeds an upper threshold, it suggests that the participant is speaking too quickly.
- » If the `Speech_Pace_Index` falls between the lower and upper thresholds, the participant's speech pace is considered appropriate.

This comprehensive approach to evaluating speech pace involves not only measuring the average rate of speech but also examining its variability throughout the presentation. The Speech Pace Correctness Index serves as a crucial metric for determining the effectiveness and appropriateness of the participant's speech rate. This index is particularly important in VR training environments, where pacing can significantly impact the clarity and effectiveness of communication. By assessing both the average speech pace and its consistency, trainers, and speakers can gain valuable insights into how to adjust their speaking rate for optimal communication and audience engagement.

Step 7: Normalize indicators to obtain values from 0 to 1

Each of the aforementioned indicators possesses the potential to be effectively utilized for analyzing public speaking skills individually. However, when these indicators are combined, they provide a more comprehensive expression of the analyzed individual's abilities. To facilitate their amalgamation, it is essential to first normalize the values of these individual indicators. This process of normalization ensures that each indicator contributes equally to the overall assessment, allowing for a balanced and holistic evaluation of public speaking skills. The integration of these diverse metrics enables a nuanced understanding of the speaker's strengths and areas for improvement, encompassing various aspects of communication such as voice quality, speech pace, and gesture use. By synthesizing these indicators into a unified analysis, a more detailed and accurate profile of the speaker's public speaking competence is obtained, which is crucial for targeted training and development in this field.

Step 8: Aggregate indicators and evaluate communication

The process involves the calculation of a weighted average of the communication indicators, taking into account the significance of each characteristic. This method ensures that more critical aspects of communication have a proportionally greater impact on the overall evaluation. Subsequently, the resulting scores are rescaled to a more comprehensible range, such as 1-10. This rescaling is done to simplify the interpretation of the results for both the instructor and the exercise participant. By adopting this approach, the assessment becomes more user-friendly while maintaining its analytical rigor. This system not only aids in the immediate understanding of performance levels but also facilitates more effective communication between instructors and participants regarding areas of strength and those requiring improvement in public speaking skills.

Tests

The developed method for automated assessment of public speaking using virtual reality (VR) was subjected to a comparative analysis against traditional human assessment. For this purpose, a collection of 20 public speaking recordings conducted in VR was prepared. This dataset encompassed all the VR data types discussed in this publication. Subsequently, five experts were requested to evaluate these recordings using the Public Speaking Competence Rubric (PSCR), (Schreiber and contributors, 2012) using 10 points scale. The median of the experts' scores was adopted as a reference value for comparison with the algorithm's outcomes. The algorithm then performed its analysis, and its results were juxtaposed with the reference scores to estimate its effectiveness. The algorithm was awarded a point whenever its score, rounded to the nearest whole number, matched the experts' assessment. The research equipment used for measurements during the tests consisted of Meta Quest 2 VR goggles (il. 3). Visual presentation of sample results obtained using this apparatus (il. 4). The final score achieved by the algorithm was 70% similar to expert scores, with detailed results presented in table 1.

II. 3. Meta Quest 2 VR goggles and controllers used as research equipment for measurements during the tests



II. 4. Visual presentation of sample results obtained using Meta Quest 2 VR goggles, illustrating head movements along the k (up), l (forward), and j (right) axes during a public speech. The need for noise filtering in the measurement data is evident



Table 1. The results of the expert evaluation and the developed algorithm. Comparison of the results of the algorithm with the median of the experts' evaluations. Recordings are evaluated by experts and algorithms on a scale from 1 (lowest quality) to 10 (highest quality)

Recordings No.	Expert						Algorithm	Alg. score
	1	2	3	4	5	Median		
1	7	8	8	9	7	8	7	0
2	4	5	6	4	5	5	5	1
3	9	9	9	9	8	9	9	1
4	2	1	2	1	3	2	3	0
5	6	6	5	5	5	5	5	1
6	5	5	6	6	7	6	6	1
7	4	4	4	3	2	4	4	1
8	8	7	6	6	7	7	7	1
9	6	5	6	5	5	5	6	0
10	9	8	7	7	8	8	7	0
11	8	7	6	6	7	7	6	0
12	3	3	2	2	1	2	2	1
13	3	3	3	4	4	3	3	1
14	5	6	6	5	4	5	5	1
15	3	3	3	2	3	3	3	1
16	4	6	5	6	4	5	6	0
17	5	5	6	5	4	5	5	1
18	1	2	1	2	1	1	1	1
19	4	4	5	4	5	4	4	1
20	3	5	4	4	3	4	4	1
							Total score:	70%
							No. Errors	6

Source: own study.

Table 2. Comparison of the time taken by experts to conduct the evaluation and the proposed algorithm

Recordings No.	Rec. Length [min]	Expert					Algorithm [s]
		1 [s]	2 [s]	3 [s]	4 [s]	5 [s]	
1	20	1200	1280	1310	1304	1524	43
2	27	1620	1733	1815	1770	1824	47
3	25	1500	1621	1683	2053	1732	51
4	18	1080	1340	1424	2105	2056	38
5	15	900	1104	1245	2053	2021	35
6	19	1140	1053	1352	1846	1754	39
7	33	1980	2134	2345	2567	2156	44
8	38	2280	2435	2703	2856	2673	51
9	29	1740	1843	1934	2135	2205	48
10	25	1500	1624	1573	1647	1580	44
11	23	1380	1413	1536	1735	1675	36
12	22	1320	1364	1424	1412	1594	33
13	17	1020	1067	1112	1447	1521	29
14	18	1080	1402	1585	1843	1945	35
15	20	1200	1426	1653	1742	1345	38
16	24	1440	1535	1745	1898	1953	41
17	22	1320	1320	1420	1520	1642	40
18	25	1500	1587	1624	1720	1683	48
19	31	1860	1890	1934	2104	2048	53
20	33	1980	2016	2144	2267	2074	56

Source: own study.

We consider this outcome to be highly promising and motivating for further research in this area. However, the analysis of just 20 recordings is merely the tip of the iceberg, indicating the need for more in-depth studies in this field. One significant advantage of using the algorithm, observed during the tests, is the time efficiency of the analysis. The proposed algorithm processed all presentations in under 60 seconds without optimization, whereas the experts' analysis, according to their statements, took at least 15 minutes per recording (table 2). Unlike human evaluators, the algorithm does not need

to watch the recorded data in real-time and is capable of performing calculations and entering various types of data rapidly.

This stark contrast in analysis time between the algorithm and human evaluators highlights the potential for significant time savings in the assessment process. Moreover, the algorithm's ability to process data quickly without the need for real-time observation or manual data entry underscores its efficiency and potential for scalability. The fact that the algorithm scored 7 out of 10 points compared to the experts' assessments further reinforces its validity and reliability in evaluating public speaking skills. The relatively high score achieved by the algorithm, despite its rapid processing time, suggests that it can serve as an effective tool for assessing public speaking competencies, particularly in settings where quick and efficient evaluation is required.

Furthermore, the use of VR technology in this assessment method adds an innovative dimension to public speaking training and evaluation. VR provides a controlled yet realistic environment for speakers to practice and hone their skills, while the algorithm offers an objective and efficient means of assessing these skills. The integration of VR and automated analysis could revolutionize the way public speaking is taught and assessed, offering new possibilities for immersive training experiences combined with precise and rapid feedback.

Conclusions

This research paper presents a significant advancement in the field of communication skills training, particularly in public speaking, through the application of virtual reality (VR) technology. It focuses on the development and implementation of an algorithm for the automatic analysis of communication skills within a VR environment, aimed at enhancing the effectiveness of public speaking training. Traditional training methods, while useful, have limitations such as the need for human resources, subjective feedback, and the absence of immersive practice environments. VR technology addresses these issues by providing a realistic and controlled setting for practice and instant feedback. The main challenge in utilizing VR for training is

developing an effective algorithm that can accurately analyze and provide feedback on a participant's performance. This task involves collecting and interpreting a wide range of data, from speech patterns to non-verbal cues like gestures and eye contact, while also personalizing the training to cater to individual needs, including those of participants with disabilities. The proposed algorithm in this research is designed to fill this gap, offering personalized, objective, and comprehensive feedback. It is poised to transform the way communication skills are taught and learned, making the process more efficient, accessible, and adaptable. The effectiveness of the algorithm was rigorously tested using a dataset of 20 VR public speaking recordings and compared against assessments made by subject matter experts using the Public Speaking Competence Rubric (PSCR). The results showed a high degree of congruence between the algorithm's outcomes and expert evaluations, indicating the potential of this method for further research and development in this domain. A notable finding from the testing phase is the time efficiency of the algorithm. It processed all presentations in under 60 seconds without optimization, significantly faster than the 30 minutes per recording required by human evaluators. This efficiency, coupled with the algorithm's ability to rapidly process data without real-time observation or manual data entry, highlights its potential for significant time savings and scalability in the assessment process. The algorithm's score of 7 out of 10 points, compared to expert assessments, reinforces its validity and reliability in evaluating public speaking skills. The use of VR technology in this assessment method brings an innovative dimension to public speaking training and evaluation. VR offers a realistic environment for practice, while the algorithm provides an objective and efficient assessment mechanism. However, this study also has its limitations. The relatively small dataset of 20 recordings may limit the generalizability of the findings to broader populations, particularly in diverse cultural and linguistic contexts. Additionally, the algorithm's performance for participants with severe disabilities or highly atypical speech patterns was not comprehensively tested, which highlights an important avenue for future research. Future studies should aim to expand the dataset to include more diverse samples and evaluate the algorithm's adaptability to different

demographic and cultural groups. Further research could also focus on optimizing the algorithm's processing time, enhancing its ability to interpret more complex non-verbal cues, and testing its efficacy in real-world training scenarios. The integration of VR and automated analysis promises to revolutionize public speaking training, offering immersive training experiences combined with precise and rapid feedback. This approach could significantly impact educational practices and professional development, enhancing the art of public speaking in the digital age.

Acknowledgment

This work was supported by the National Center for Research and Development under the Things Are for People competition, contract number: "Things are for People"/0056/2020-00, the title of the project "E-ZAWODY – Development of technological solutions with the use of VR allowing people with disabilities to improve their professional competencies through the implementation of work in virtual space".

Abstract: This research develops and validates an algorithm for automated analysis of communication skills in public speaking within a Virtual Reality (VR) environment, addressing limitations in traditional training methods. The study introduces an algorithm to analyze comprehensive data, including speech patterns, non-verbal cues, and tailored feedback, particularly for users with disabilities. The method involves multiple steps: processing input data, analyzing gestures, spatial movements, voice tone and timbre, and speech rate. The algorithm was tested against expert evaluations using a dataset of 20 VR public speaking recordings. This VR-based approach provides an immersive, adaptive training environment, crucial for high-stakes fields such as emergency services, where effective communication can impact outcomes in life-threatening situations.

Streszczenie: Niniejsze badanie opracowuje i weryfikuje algorytm automatycznej analizy umiejętności komunikacyjnych w wystąpieniach publicznych w środowisku wirtualnej rzeczywistości (VR), rozwiązując ograniczenia tradycyjnych metod szkoleniowych. Badanie wprowadza algorytm do analizy kompleksowych danych, w tym wzorców mowy, wskazówek niewerbalnych i dostosowanych informacji zwrotnych, w szczególności dla użytkowników niepełnosprawnych. Metoda obejmuje wiele kroków: przetwarzanie danych wejściowych, analizę gestów, ruchów przestrzennych, tonu i barwy głosu oraz tempa mowy. Algorytm został przetestowany w oparciu o oceny ekspertów przy użyciu zestawu danych 20 nagrań wystąpień publicznych VR. To podejście oparte na VR zapewnia immersyjne, adaptacyjne środowisko szkoleniowe, kluczowe dla dziedzin o wysokiej stawce, takich jak służby ratunkowe, w których skuteczna komunikacja może mieć wpływ na wyniki w sytuacjach zagrażających życiu.

Keywords: Virtual Reality, public speaking, communication skills, automated analysis, speech analysis, emergency response training

Słowa kluczowe: wirtualna rzeczywistość, wystąpienia publiczne, umiejętności komunikacyjne, automatyczna analiza, analiza mowy, szkolenie z reagowania kryzysowego

References

- Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., Scherer, S. (2013). *Cicero-towards a multimodal virtual audience platform for public speaking training*. California: Intelligent Virtual Agents.
- Campbell, N., Kane, J., Layher, G., Neumann, H., Scherer, S. (2012). *An audiovisual political speech analysis incorporating eye-tracking and perception data*. Istanbul: ELRA.
- Chen, L., Feng, G., Joe, J., Leong, C.W., Kitchen, C., Lee, C.M. (2014). *Towards automated assessment of public speaking skills using multimodal cues*. In Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul.
- Giraud, T., Soury, M., Hua, J., et al. (2013). *Multimodal expressions of stress during a public speaking task: Collection, annotation, and global analyses*. In Affective Computing and Intelligent Interaction (ACII) Humaine Association Conference. Geneva: IEEE.
- Goto, M., Igarashi, T., Kurihara, K., Matsusaka, Y., Ogata, J. (2007). *Presentation sensei: a presentation training system using speech and image processing*. In Proceedings of the 9th international conference on Multimodal interfaces. Nagoya: ACM.
- Kleinsmith, A., Bianchi-Berthouze, N. (2013). *Affective body expression perception and recognition: A survey*. IEEE Transactions on Affective Computing.
- Nguyen, A.-T., Chen, W., Rauterberg, M. (2012). *Online feedback system for public speakers*. In IEEE Symp. e-Learning, e-Management, and e-Services. Citeseer.
- Schreiber, L.M., Paul, G.D., Shibley, L.R. (2012). The development and test of the public speaking competence rubric, *Communication Education*, 61(3), 205–233.