Evaluating Multimodal Behavioral Features for Public Speaking Assessment in Virtual Reality

Marion Ristorcelli
Aix-Marseille University, CNRS, LIS
Marseille, France
marion.ristorcelli@lis-lab.fr

Elodie Etienne
HEC Management School, QuantOM
Liège, Belgium
elodie.etienne@uliege.be

Michaël Schyns
HEC Management School, QuantOM
Liège, Belgium
M.Schyns@uliege.be

Rémy Casanova
Aix-Marseille University, CNRS, ISM
Marseille, France
remy.casanova@univ-amu.fr

Magalie Ochs
Aix-Marseille University, CNRS, LIS
Marseille, France
magalie.ochs@lis-lab.fr

ABSTRACT

Public speaking (PS) self-assessment remains challenging due to multiple behavioral dimensions and a lack of objective evaluation methods. Virtual reality (VR) offers immersive training environments with automatic performance analysis capabilities. However, current evaluation systems use ad hoc metrics lacking transparency and reproducibility. No comprehensive set of multimodal, context-independent behavioral cues exists for interpretable user feedback. We propose verbal and nonverbal features meeting three criteria: automatic measurement capability, context independence, and user interpretability. Using a multimodal corpus of VR presentations by 60 participants, we extracted 47 behavioral features via the Meta Quest Pro headset. Expert assessment used 7-point Likert scales. Correlation analysis and machine learning models demonstrate this feature set provides a relevant basis for automated PS assessment in VR.

CCS CONCEPTS

• **Human-centered computing → Virtual reality**.

KEYWORDS

Public speaking, Virtual Reality, Multimodal Behavior, Behavioral Features, Automatic Performance Assessment

ACM Reference Format:
Marion Ristorcelli, Elodie Etienne, Michaël Schyns, Rémy Casanova, and Magalie Ochs. 2025. Evaluating Multimodal Behavioral Features for Public Speaking Assessment in Virtual Reality. In *ACM International Conference on Intelligent Virtual Agents (IVA '25), September 16–19, 2025, Berlin, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3717511.3749301

1 INTRODUCTION

Speaking in front of an audience is not an easy task for everyone. Because of the diversity of skills required to make a successful presentation, this activity requires both training and practice to progress. To this end, virtual reality (VR) has proven effective in public speaking (PS) training, particularly due to its capacity to simulate immersive speaking environments and to automatically extract behavioral cues reflecting PS performance [2, 8, 14, 16, 24, 29]. A growing number of VR-based tools rely on multimodal behavioral cues such as gaze direction, speech rate, or body posture to provide feedback to users on their performance. However, several limitations persist in current systems. First, many cues are highly context-dependent: for example, body orientation is sometimes computed relative to a predefined target in the virtual room [23], making it difficult to generalize across scenarios. Similarly, in some systems, a fixed gaze duration (e.g. four seconds) towards a virtual agent is used to trigger specific audience behaviors [33], illustrating the arbitrary and context-specific nature of such cues. Although appropriate PS behavior may depend on the context, several studies have shown that a set of generic multimodal cues that characterize good PS performance can be defined, regardless of the context [25, 35]. For instance, high loudness [12] and eye contact with the audience [19] play critical roles, contributing to the speaker's overall effectiveness. Second, the feedback provided to users is often not interpretable. Many systems present raw data (e.g., number of words per minute) [23], or display aggregate performance scores via stars or colored gauges [9, 21, 24, 33], without explaining which specific behaviors influenced the result. Consequently, users are often unable to identify which aspects of their performance to improve. Finally, the relevance of the measured cues is rarely empirically validated [13, 33, 35]. In many cases, systems rely on a limited set of behavioral cues, without demonstrating their predictive power for actual PS performance [9, 13, 33]. While external sensors might provide more comprehensive tracking, our approach focuses on VR headset-only features to develop an autonomous, accessible tool usable by anyone without additional equipment.

The purpose of this study is two-fold. We propose a new set of features based on multimodal behavior in VR to assess speaking performance, and we want to investigate the relevance of these features for performance assessment. In order to define a generic and useful set of cues, we consider a multimodal behavioral cue as relevant if: (1) it influences the perceived quality of the presentation, (2) it can be measured automatically, (3) it is context-independent (e.g., the use or non-use of a visual aid, or the presence of words with emotional connotations, positive or negative, in the speech), (4) it can be explained and interpreted in a simple way by users (e.g.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
IVA '25, September 16–19, 2025, Berlin, Germany
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1508-2/2025/09
https://doi.org/10.1145/3717511.3749301

'speak slower' rather than 'we noted an average of 160 words per minute').

In this study, we rely on a public speaking corpus in which participants give 3-minute oral presentations. Their verbal and nonverbal behaviors were recorded, and a set of multimodal behavioral features was identified and extracted, based on the scientific literature, to satisfy the last three criteria mentioned above. The performances were then assessed by an expert using a 7-point Likert scale. Finally, we applied machine learning (ML) techniques to test whether these features can predict perceived presentation quality, thus verifying their relevance for performance assessment, in line with the first criterion.

2 VERBAL FEATURES

Verbal cues have a significant impact on the speaker's ability to engage, convince, and influence the audience [18, 34]. In particular, **acoustic features** play a key role in assessing how speakers use their voice in PS [6, 12, 17, 37]. Based on these research works, using the GeMAPS software package [15] from the OpenSMILE toolbox and Praat software for the automatic extraction, we have considered the following acoustic cues. The *fundamental frequency (F0)* refers to the pitch of the voice, measured in Hertz. We calculated two features to characterize pitch (i.e. how high or low a voice is perceived) based on the mean and standard deviation of F0. The *loudness* refers to vocal intensity during speech. We compute the loudness mean and standard deviation as well as the number of significant variations in vocal intensity per second corresponding to a sudden peak in the signal. The *silent pauses* refers to the number and duration of medium pauses (200-1000 ms) and long pauses (>1000 ms), as well as the total duration of all pauses. The *speech rate* refers to the flow of speech and is characterized by the number of words per unit of time.

Textual features, such as content-related measures, play an important role in assessing oral communication skills beyond fluency or pronunciation [38]. From Linguistic Inquiry and Word Count (LIWC) [5], based on the literature, we extracted the following relevant features. *Big words* refers to the percentage of words longer than seven letters, often used as a linguistic marker of lexical sophistication. *Fillers*, such as "like", "you know", or "I mean" are markers with no semantic value, typical of spontaneous spoken language [26]. As a first step, and to ensure the context-independence, we focused only on two verbal features.

3 NONVERBAL FEATURES

Nonverbal behavioral cues, such as gaze, postures, gestures and facial expressions, also have a major impact on the perception of oral performance [19, 30, 31]. Using the Meta Quest Pro VR headset, we extracted automatically the following nonverbal features. **The gaze direction** of the speaker during an oral presentation plays a crucial role in assessing the quality of the performance [7, 10, 19, 23, 27, 31, 32]. Based on the existing literature, we extracted a set of features related to gaze behavior using the eye-tracking technology from the VR headset. The *score of characters viewed* corresponds to the number of audience members viewed during the presentation. The *gaze fixation time* corresponds to the duration of uninterrupted gaze on each virtual agent. For this cue, we calculated the median and interquartile range of this duration. The *gaze duration* corresponds to the proportion of time spent focusing on the audience compared to other elements of the scene. The *hesitation time* corresponds to the proportion of time spent thinking during a presentation, looking at the ceiling or the floor when searching for words or ideas [4]. The *entropy of gaze direction* reflects the way in which the speaker's attention is distributed among the virtual agents in the audience.

The **body movements and postures** play also an important role in the perception of the speaker during the speech. Experts agree that certain postural behaviors should be avoided during oral presentations in order to achieve an effective presentation [19, 31, 39]. Based on the data collected with the headset and based on previous works, we have extracted the following features. The *amplitude of the horizontal displacement* refers to the speaker's use of space during speech. The *distance covered by the speaker* corresponds to the use of the stage and potential agitation characterized by the distance covered per unit of time. The *entropy of body movements* reflects the variability of the speaker's gestures, indicating whether the speaker moves frequently or remains rather static.

Speakers' **gestures** provide valuable indications of their state of mind and their message. Moreover, some behaviors have a positive impact on listeners [11] while others may be misinterpreted and then have a negative impact [3, 20]. In order to observe this impact, we decided to calculate the following features. The *proportion of open posture* corresponds to the proportion of presentation time during which the speaker adopts an open posture, defined by hands that are not clasped, crossed nor overlapped. The *duration of closed and open posture* refers to the normalized cumulative time spent in this configuration during the presentation. The *proportion of hand detection* corresponds to the proportion of time that hands are detected by the VR headset's hand tracking system, depending on their relative position to the headset. The *frequency of hand openness* corresponds to the frequency of hand positioning (neutral, palm up, palm down) and opening (neutral, open, closed) during the presentation. The *duration of palm position* corresponds to the median and interquartile range of durations of uninterrupted palm positions. The *entropy of hand openness and direction* reflects the dynamics of hand opening and palm position.

The **facial expressions** are also crucial in PS, as shown by [7, 36]. Using the headset's face-tracking technology, we calculated the following features based on the muscular activation of facial muscles characterized by blendshapes. The *frequency of positive or negative facial expressions* corresponds to activation frequency for ten blendshapes (Meta's Face Tracking API) deviating by more than 2 standard deviations from baseline. The *entropy of blendshapes* corresponds to the dynamics of the speaker's facial expressions, measuring the variability of expression across all blendshapes.

4 MULTIMODAL CORPUS FOR ASSESSING PUBLIC SPEAKING PERFORMANCE

We collected a multimodal corpus containing VR-recorded oral presentations to extract the behavioral features described above. Fifty-eight participants (M_{age} = 31.1, SD_{age} = 12.9; 29 men, 26 women, 3 non-binary) delivered six 3-minute presentations on predefined topics (e.g., self-introduction, personal passion, future project) to four virtual agents. The virtual environment, based on the system

described in [22], allows adapting the audience's gender and attitude to vary the level of difficulty of the oral presentations [28]. For each presentation, the set of features presented above was extracted from the raw data. Audio was recorded from the Meta Quest Pro VR headset and used to compute the acoustic and textual features. The VR headset was used to capture the nonverbal behavior behind the evoked nonverbal features. In addition, an expert with several years of experience in PS coaching assessed the overall performance using video recordings of each presentation via a 7-point Likert scale, ranging from 1 ("very poor") to 7 ("excellent").

5 METHODOLOGY

Once features are extracted, we examine the relevance of these features for assessing speaking performance using correlation analysis and ML methods. If our feature set correlates with the performance score, and if ML models can predict this score, we will be able to confirm the relevance of the proposed set of features for assessing public speaking performance. To predict the performance score using the feature set, we conducted two binary classification tasks using a **Multilayer Perceptron (MLP)** and a **Random Forest (RF)** model. These tasks used a set of 47 behavioral features presented in section 2 and section 3 and extracted from each presentation. The target variable $y \in \{0, 1\}$ represents the expert evaluation. For the *2-class classification*, scores from 1 to 4 were labeled as *Bad* (0) and scores from 5 to 7 as *Good* (1). For the *extreme classes classification*, only scores of 1 and 2 were considered *Bad* and scores of 6 and 7 as *Good*.

The dataset was split into a training (80%) and test set (20%) with stratification based on the target label. Hyperparameter optimization was conducted using the Optuna framework [1]. We performed 10 different trials, and for each trial, a different combination of hyperparameters was randomly sampled and evaluated by 10-fold cross-validation on the training set. The best model was selected using the average weighted F1-score and evaluated on the test set.

6 RESULTS AND DISCUSSION

To study the multimodal features' relevance for assessing public speaking in VR, we first analyzed the Pearson correlations between objective behavioral features and the expert's overall score. The strongest positive correlation was found for the distance covered by the hands ($\rho = 0.29$, $p < .0001$) while the strongest negative correlation was obtained for the standard deviation of loudness ($\rho = -0.21$, $p < .0001$). This analysis also reveals that 12 features are not correlated with the performance score. These include features related to fillers, which contain around 40% null values, as well as several features related to facial expressions. This result is not surprising, as the participants' facial expressions were partially obscured by the VR headset during the experiment. Consequently, the expert's assessment probably did not take these cues into account when evaluating performance. Although correlation coefficients are relatively low, most features are statistically significant. These low values can be explained by the fact that the quality of a performance is not based on a single cue, but results from a combination of behaviors. Thus, even weak correlations reflect the contribution of these features to the public speaking assessment.

Classification results are reported in Table 1. Interestingly, using all features set, the RF model consistently outperforms the MLP in both classification tasks. The RF model is known to be more robust to noisy features and less sensitive to irrelevant or uninformative features. In contrast, MLPs can be more affected by these features, particularly when the dataset is limited in size and no feature selection is applied. These observations suggest that tree-based models may be particularly well suited to this type of behavioral analysis, at least in the context of current data. The results also show that the RF and MLP models perform significantly better when classifying only extreme cases, suggesting that the most polarized performances are easier to distinguish based on available features. This highlights the potential difficulty of reliably assessing average performance. Although not very high for 2-class classification, the result indicates that relevant predictive information is present in certain features, justifying further exploration and refinement of the model. In the future, it may be interesting to experiment with other ML models and use feature selection methods to extract the most important features for evaluating speaker performance in VR.

Model	F1-Score	Precision	Recall
RF 2 classes	61.30%	61.88%	62.04%
MLP 2 classes	57.48%	57.71%	57.41%
RF Ext. classes	93.23%	94.49%	94.12%
MLP Ext. classes	81.40%	85.37%	77.78%

Table 1: Binary classification results of Random Forest (RF) and Multilayer Perceptron (MLP) models for performance prediction, using either a standard 2-class task or a task restricted to extreme classes (Ext. classes).

7 CONCLUSION

Our research aims to assess the suitability of a set of multimodal behavioral cues for the automatic evaluation of speaking performance in VR, with the final objective of providing users with constructive feedback to improve their speaking skills. The results of the correlation and ML analyses tend to prove the relevancy of the proposed features set for the evaluation of public speaking in virtual reality. Further research can be carried out to improve the performance prediction model, but also to identify more precisely the most important features among those presented in this article.

ACKNOWLEDGMENTS

This research was funded under the ANR REVITALISE grant ANR-21-CE33-0016-02.

REFERENCES

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2623–2631.

[2] Ligia Batrinca, Giota Stratou, Ari Shapiro, Louis-Philippe Morency, and Stefan Scherer. 2013. Cicero-towards a multimodal virtual audience platform for public speaking training. In *Intelligent Virtual Agents: 13th International Conference, IVA 2013, Edinburgh, UK, August 29-31, 2013. Proceedings 13*. Springer, 116–128.

[3] Tobias Baur, Ionut Damian, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. 2013. Nova: Automated analysis of nonverbal signals in social interactions. In *Human Behavior Understanding: 4th International Workshop, HBU 2013, Barcelona, Spain, October 22, 2013. Proceedings 4*. Springer, 160–171.

[4] Geoffrey W Beattie. 1979. Planning units in spontaneous speech: Some evidence from hesitation in speech and speaker gaze direction in conversation. (1979).

[5] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin* 10 (2022).

[6] Caterina Breitenstein, Diana Van Lancker, and Irene Daum. 2001. The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition & Emotion* 15, 1 (2001), 57–79.

[7] Mathieu Chollet, Pranav Ghate, Catherine Neubauer, and Stefan Scherer. 2018. Influence of individual differences when training public speaking with virtual audiences. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 1–7.

[8] Mathieu Chollet, Kalin Stefanov, Helmut Prendinger, and Stefan Scherer. 2015. Public speaking training with a multimodal interactive virtual audience framework. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 367–368.

[9] Mathieu Chollet, Torsten Wörtwein, Louis-Philippe Morency, and Stefan Scherer. 2016. A multimodal corpus for the assessment of public speaking ability and anxiety. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 488–495.

[10] Mathieu Chollet, Torsten Wörtwein, Louis-Philippe Morency, Ari Shapiro, and Stefan Scherer. 2015. 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*. 1143–1154.

[11] Bianca Malaika Ciuffani. 2017. *Non-verbal Communication and Leadership: the impact of hand gestures used by leaders on follower job satisfaction*. B.S. thesis. University of Twente.

[12] Kariane Pereira Dos Santos, Vanessa Veis Ribeiro, Larissa Thais Donalonso Siqueira, Larissa Cruz Brugnara, Inaiê Caroline Brugnolo Rosa, and Ana Paula Dassie-Leite. 2022. Does shyness influence the self-perception of vocal symptoms, public speaking, and daily communication? *Journal of voice* 36, 1 (2022), 54–58.

[13] M El-Yamri, A Romero-Hernandez, M Gonzalez-Riojo, and B Manero. 2019. Comunicarte: A virtual reality game to improve public speaking skills. In *EDULEARN19 Proceedings*. IATED, 9061–9066.

[14] Elodie Etienne, Anne-Lise Leclercq, Angélique Remacle, Laurence Dessart, and Michaël Schyns. 2023. Perception of avatars nonverbal behaviors in virtual reality. *Psychology & Marketing* 40, 11 (2023), 2464–2481.

[15] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[16] Yann Glémarec, Jean-Luc Lugrin, Anne-Gwenn Bosser, Cédric Buche, and Marc Erich Latoschik. 2022. Controlling the stage: a high-level control system for virtual audiences in Virtual Reality. *Frontiers in Virtual Reality* 3 (2022), 876433.

[17] Alexander M Goberman, Stephanie Hughes, and Todd Haydock. 2011. Acoustic characteristics of public speaking: Anxiety and practice effects. *Speech communication* 53, 6 (2011), 867–876.

[18] Joshua J Guyer, Leandre R Fabrigar, and Thomas I Vaughan-Johnston. 2019. Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion. *Personality and Social Psychology Bulletin* 45, 3 (2019), 389–405.

[19] Fasih Haider, Loredana Cerrato, Nick Campbell, and Saturnino Luz. 2016. Presentation quality assessment using acoustic information and hand movements. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2812–2816.

[20] David McNeill. 2000. *Language and gesture*. Vol. 2. Cambridge University Press.

[21] Anh-Tuan Nguyen, Wei Chen, and Matthias Rauterberg. 2015. Intelligent presentation skills trainer analyses body movement. In *International Work-Conference on Artificial Neural Networks*. Springer, 320–332.

[22] Magalie Ochs, Marion Ristorcelli, Alexandre D'Ambra, Rémy Casanova, and Jean-Marie Pergandi. 2024. REVITALISE: viRtual bEhaVioral skIlls TrAining for pubLIc SpEaking. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. 1–3.

[23] Fabrizio Palmas, Jakub Cichor, David A Plecher, and Gudrun Klinker. 2019. Acceptance and effectiveness of a virtual reality public speaking training. In *2019 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 363–371.

[24] Fabrizio Palmas, Ramona Reinelt, Jakub E Cichor, David A Plecher, and Gudrun Klinker. 2021. Virtual reality public speaking training: Experimental evaluation of direct feedback technology acceptance. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 463–472.

[25] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2016. Multimodal analysis and prediction of persuasiveness in online social multimedia. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6, 3 (2016), 1–25.

[26] Abdur Rahman and RT Subaraksha. [n. d.]. Computational Quantification of Oral Fluency in English. ([n. d.]).

[27] Vikram Ramanarayanan, Chee Wee Leong, Lei Chen, Gary Feng, and David Suendermann-Oeft. 2015. Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In *Proceedings of the 2015 acm on international conference on multimodal interaction*. 23–30.

[28] Marion Ristorcelli, Emma Gallego, Kévin Nguy, Jean-Marie Pergandi, Rémy Casanova, and Magalie Ochs. 2023. Investigating the impact of a virtual audience's gender and attitudes on a human speaker. In *Companion Publication of the 25th International Conference on Multimodal Interaction*. 363–367.

[29] Sarah Saufnay, Elodie Etienne, and Michael Schyns. 2024. Improvement of Public Speaking Skills using Virtual Reality: Development of a Training System. In *ACII 2024*. Institute of Electrical and Electronics Engineers, New-York, United States.

[30] Jan Schneider, Dirk Börner, Peter Van Rosmalen, and Marcus Specht. 2016. Enhancing public speaking skills-an evaluation of the Presentation Trainer in the wild. In *Adaptive and Adaptable Learning: 11th European Conference on Technology Enhanced Learning, EC-TEL 2016, Lyon, France, September 13-16, 2016, Proceedings 11*. Springer, 263–276.

[31] Jan Schneider, Dirk Börner, Peter Van Rosmalen, and Marcus Specht. 2017. Presentation Trainer: what experts and computers can tell about your nonverbal communication. *Journal of Computer Assisted Learning* 33, 2 (2017), 164–177.

[32] Ha Trinh, Reza Asadi, Darren Edge, and T Bickmore. 2017. Robocop: A robotic coach for oral presentations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–24.

[33] Bao Truong, Trung-Nghia Le, Khanh-Duy Le, Minh-Triet Tran, and Tam V Nguyen. 2022. Public speaking simulator with speech and audience feedback. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 855–858.

[34] Thomas I Vaughan-Johnston, Joshua J Guyer, Leandre R Fabrigar, Grigorios Lamprinakos, and Pablo Briñol. 2024. Falling vocal intonation signals speaker confidence and conditionally boosts persuasion. *Personality and Social Psychology Bulletin* (2024), 01461672241262180.

[35] Torsten Wörtwein, Mathieu Chollet, Boris Schauerte, Louis-Philippe Morency, Rainer Stiefelhagen, and Stefan Scherer. 2015. Multimodal public speaking performance assessment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 43–50.

[36] Yutaro Yagi, Shogo Okada, Shota Shiobara, and Sota Sugimura. 2022. Predicting multimodal presentation skills based on instance weighting domain adaptation. *Journal on Multimodal User Interfaces* 16, 1 (2022), 1–16.

[37] Zixiaofan Yang, Jessica Huynh, Riku Tabata, Nishmar Cestero, Tomer Aharoni, Julia Hirschberg, et al. 2020. What makes a speaker charismatic? Producing and perceiving charismatic speech. In *Proc. 10th International Conference on Speech Prosody*, Vol. 2020. 685–689.

[38] Klaus Zechner and Isaac Bejar. 2006. Towards automatic scoring of non-native spontaneous speech. In *Proceedings of the Human Language Technology Conference of the NAACL, main conference*. 216–223.

[39] Hangyu Zhou, Yuichiro Fujimoto, Masayuki Kanbara, and Hirokazu Kato. 2021. Virtual reality as a reflection technique for public speaking training. *Applied Sciences* 11, 9 (2021), 3988.