Effects of public speaking virtual reality on prosodic and gestural features

Io Valls-Ratés¹, Oliver Niebuhr², Pilar Prieto³,⁴

¹Universitat Pompeu Fabra
²Centre for Industrial Electronics, University of Southern Denmark, Sonderborg/DK
³ICREA

io.valls@upf.edu, olni@sdu.dk, pilar.prieto@upf.edu

Abstract

Previous studies have revealed that virtual reality (VR) is able to suggest to the human brain a real environment, as it was stated in 1989 when the term 'virtual reality' was coined. In the field of public speaking, studies have mainly focused on how VR environments can help reduce public speaking anxiety. However, there is no focus on VR training in educational settings and, more specifically, on how VR training can help improve subsequent public-speaking performances in front of a real audience. The present study aims at analyzing the prosodic and gestural effects of VR settings on speakers. A total of 31 secondary school students participated either in a VR oral-presentation training facing a VR audience, or in a control condition in which oral presentations were practiced alone in a room. Prosodic and gestural measures of speaking style were analyzed and compared between the two groups, each of which performed three rounds of practicing in three consecutive weeks. Training with VR resulted in longer speaking time and more pausing, an increase in f0 values and more gesturing, all of which suggests a stronger audience orientation. By contrast, the non-VR speakers developed in the opposite direction.

Index Terms: virtual reality, public speech, gesture, prosody.

1. Introduction

Boosting public speaking abilities at secondary-school level contributes to strengthening students' effectiveness with academic work, as well as their social skills, thus affording them more satisfactory interpersonal relationships [1,2] and preventing them from abandoning their studies prematurely [3,4]. In order to achieve that, schools need to realize the importance that oral abilities have in the development of students' self-confidence. Schools obviously need to take action and consider the challenge of involving students in more oracy settings to enhance their communication, empathy and, thus, to encourage them to take an active part of the community they live in [1].

As any other skill, public speaking needs practice. [5] outline that speech practices are the single greatest predictor of students’ oral presentation success and key for reducing public speaking anxiety (PSA).

The earlier speakers start preparing for their presentations, the more fluent (fewer total disfluencies) their speeches will become [6]. However, practice needs time, and it also needs to be oral. [7] reported that there is a high percentage of students spending most of their time writing the speech rather than rehearsing it orally. Also, [7] found that students spend only up to five minutes preparing for their presentations. [8] found that students in a public speaking course who rehearsed in front of an audience, compared to students rehearsing alone, obtained higher scores on their final classroom speech assessments, thus lending support to the claim that audience-based speech practice is a valid method to increase speech assessments.

As a way to enhance the oral practice of presentations when delivering speeches before an audience, VR simulations can be of great help. Practicing in virtual environments gives rise to the illusion of being in the environment depicted and makes the user participate in it, experiencing the sense of presence [9,10]. Studies with university students show [11,12,13] how practicing with VR audiences results in diminishing communication apprehension and enhancing self-perceived communication competence. The study by [14] concludes that employing VR for speech rehearsals, not only decreases the anxiety levels of students. In addition, students consider it as an innovative way to think about oral rehearsing and are willing to develop their skills. [15,16] also refer to the students’ excitement of participating in VR experiments as a different and motivating way to entice them to rehearse their speeches. However, [16] compared three conditions for practicing a speech (in front of a mirror; using video recording, and using a VR setting). They found that those using the VR setting did not show higher scores on their final classroom speech assessment their fellow students in the other two conditions. [16] also stress that participants in the VR settings reported that it was more demanding than other modes of practice, which underlines that digital, virtual audiences can elicit distress – compared to conditions without an audience.

Besides the fact that VR can provide a credible set of scenarios that allow for an immersive learning situation, when used for public speaking training, VR environments also seem to be conducive to a more listener-oriented speaking style. [17] showed that participants rehearsing within a VR environment performed their speech in a more conversation-like speaking style than participants in the control group, who practiced their speech alone in a classroom. They furthermore concluded that the delivery of those participants trained in the VR condition was more charismatic and more audience-oriented, showing reduced signs of erosion due to repeated rehearsing than the delivery after practicing alone in a classroom. In a very recent study by [18], VR training also proved to be effective in prompting vocal characteristics very similar to the ones used in classroom. The study was conducted with elementary school teachers who gave the same lesson in their classrooms and later in front of a VR audience. Results showed how performing in front of real and virtual audiences (compared to free speech in a control condition) significantly increased the participants’ f0 values, their f0 variation, and their voice intensity (in line with [17]), but there was no effect on the number of pauses nor on pause duration when speaking in VR.

Against the outlined research background, we see that still too little is known about the relationship between practicing oral speeches with VR and its effects on voice parameters and gesture rate. Therefore, we investigate, in a between-subjects training experiment, whether training with VR makes a differ-
ence in how the voice sounds and if and how the speaker performs gestures. To our knowledge this is the first experiment that investigates the effects of prosody and gesture through a VR experiment with high-school students.

Three hypotheses are tested: (1) Compared to a baseline condition of non-VR speech training, VR-based speech training induce prosodic changes in speakers, reflecting a more audience-oriented way of speaking. (2) The more audience-oriented prosody triggered by VR-based speech training will be accompanied by a higher amount of gestures, unlike for the baseline condition speakers. Note that, with reference to the above reviewed studies, we expect the effects of VR-based training to differ from those of non-VR training in a quantitative way, e.g., in that they are strongly pronounced for VR.

2. Method

2.1. Participants

A total of 31 secondary-school students (mean age=16.95) participated in the study: the VR group (N=17) and the Non-VR group (N=14). Participation was voluntary, and all participants completed an informed consent process prior to enrolling in the study.

2.2. Materials

The topic selected for the baseline speaking task was “Do you think that adolescents should spend more time in nature?”. Before giving their speech, participants received a sheet of instructions in which they were asked to prepare and then deliver a two-minute speech such as to persuade three representatives of the Catalan Department of Education to increase funding for secondary school field trips to the countryside. To facilitate participants’ preparations, five possible lines of argumentation were offered. The instructions also noted that the participants would have two minutes to prepare their speech. Though they could take notes for that purpose if they wanted to, they were not allowed to use them when they delivered their speech. The speech topics for each of the three trainings were the following: “What would the house of my dreams be like?”, “Is graffiti a form of art?”, and “Can happiness be bought?”.

2.3. Training sessions

One week after performing the baseline speaking task, the same participants undertook three training sessions (one per week). Participants were allotted two minutes to prepare their speech and did so alone in an empty classroom. After the two minutes of preparation were over, they went to the adjacent classroom. For the VR participants, the experimenter fitted them with a Clip Sonic® VR headset, which was attached to a smartphone. It ran the the BeyondVR virtual reality interface application. Based on that, the VR headset created the 3D illusion that the participant was standing in front of an audience, see Figure 1. Note that a timer was visible in the view provided by the VR headset. It helped participants to not exceed the two-minute presentation-time limit. The Non-VR group of participants gave their speeches in an empty classroom, and they too could use a timer in front of them to control their speaking time. Participants of both groups were videotaped in all training sessions.

2.4. Procedure

All participants carried out the same baseline condition, which consisted of giving a brief speech in front of a live audience.

The basic experimental structure of this between-subjects study was a baseline speech followed by a training period of three sessions. The VR group performed the three training sessions delivering their speeches in front of a virtual audience, whereas the Non-VR group gave the speeches being alone in a classroom.

Figure 1: Screenshot of the VR scenario with a virtual audience generated by BeyondVR.

2.5. Data analysis: Acoustics and statistics

For each participant, the total durations of the recorded speeches were similar in the baseline and training speeches (M = 1:34 minutes; min–max = 1:15–2 minutes). A total of 93 speeches were analyzed, three for each of the 31 participants: baseline, Training 1 (TR01), and Training 3 (TR03). Each analysis included 16 different parameters. The acoustic-phonetic analysis was per-formed using the ProsodyPro script of [19] and the supplementary analysis script of [20], using the (gender-specific) default settings of PRAAT [21]. In the pitch or fundamental-frequency (f0) domain, we measured the minimum and maximum f0, the f0 variability (in terms of the standard deviation), the mean f0, and the f0 range. For all five f0 parameters, one value was determined per prosodic phrase. Measured values were checked manually for plausibility. Outliers or missing values were corrected by manual measurements. Moreover, all f0 values were recalculated from Hz to semitones (st) relative to a base value of 100 Hz.

The duration/time domain consisted of the following four measured parameters: total number of silent pauses (> 300 ms, which is above the perceived disfluency threshold in continuous speech [22]), total time of the presentation (incl. silences), total speaking time (excl. silences), and net syllable rate (or articulation rate) in syl/s. All temporal measurements were conducted based on the analyzed presentation as a whole.

The domain of voice-quality measurements included the most measured seven parameters of previous studies: harmonic-amplitude difference (h1-h2, f0 corrected), cepstral peak prominence (cpp), harmonics-to-noise ratio (HNR), spectral center of gravity (CoG), formant dispersion (F1-F5), jitter, and shimmer. Like for the f0 parameters, voice-quality measurements were conducted based on the prosodic phrase, i.e. one value per prosodic phrase was calculated. Also, all values were manually checked and corrected, if required.

All measurements were statistically analyzed by means of a mixed-model two-way multivariate ANOVA (MANOVA) based on the two-level fixed factors Training (VR vs. non-VR) and Presentation (baseline vs. TR01 vs. TR03). Training was a between-subject and Presentation a within-subject factor. Participants were included as a random factor. Additional pairwise-comparisons tests used the Sidak correction to the adjust the
alpha-error level for multiple testing. Note that we are testing interaction hypotheses in the present paper. That is, we expect differences between the three levels of Presentation, but, crucially, these differences are expected not to be the same for the VR and the non-VR groups of participants.

2.6. Assessment of Gesture

The amount of gesturing that was present in the speakers’ speeches of the two conditions VR vs. Non-VR was quantitatively assessed over the entire data set. All instances of co-speech gestures were identified and manually annotated by the first author. Nongestural movements (self-adaptors, e.g., scratching, touching one’s hair [23]) were excluded.

3. Results

The significant main effects of Training in the MANOVA concerned almost all analyzed 16 parameters. Many of these main effects do not coincide with a significant Training* Presentation interaction. That is, what the MANOVA shows in these cases are offset differences between the two compared groups, the VR group and the non-VR group. For example, for the participants in the VR group we generally measured higher f0 values (F[1,1932] = 7.86, p < 0.01, η² = 0.12), including larger f0 ranges (F[1,1932] = 23.06, p < 0.01, η² = 0.18), as well as higher h1-h2 values (F[1,1932] = 25.55, p < 0.001, η² = 0.13), lower CoG values (F[1,1932] = 4.44, p = 0.04, η² = 0.05) and fewer silent pauses (F[1,1932] = 6.29, p = 0.02, η² = 0.10) than for the non-VR participants, independently of Training.

Regarding the main effects of Presentation, we also found a few effects that occurred independently of VR or non-VR training, i.e. without an interaction effect. This includes that training in general significantly increased the speakers’ f0 variability (F[2,87] = 3.35, p = 0.04, η² = 0.07) and the number of pauses they made in their speeches (F[2,87] = 3.95, p = 0.02, η² = 0.08), see Figure 2. In addition, the speakers’ jitter level went up during training (F[1,1193] = 62.21, p < 0.001, η² = 0.05), whereas the CoG level developed in opposite direction and went down (F[1,1193] = 10.77, p = 0.001, η² = 0.01).

The interaction effects we were looking for primarily did emerge for a number of parameters. None of these parameters concerned f0, though. In the duration/time domain, there was a significant Training*Presentation interaction for the total time of the presentation (F[1,1932] = 5.23, p = 0.03, η² = 0.08) as well as for the total speaking time (F[1,1932] = 6.08, p = 0.02, η² = 0.09) and the net syllable rate (F[1,1932] = 5.89, p = 0.02, η² = 0.08). According to the multiple comparisons tests, these interactions are all caused by a significant change in the VR group that is absent in the non-VR group. That is, unlike for non-VR training, the VR training made speakers present longer (ø + 14.74 s, p = 0.026) and speak more (ø +67.71 s, VR = 71 s, p < 0.05) and faster (ø +0.14 syl/s p = 0.006) in their last training session (TR03). Note, although no significant interactions related to f0 showed up, there was one marginally significant trend for an interaction effect. It concerned the f0 range. The latter showed a decrease from TR01 to TR03 in the non-VR condition that was absent in the VR condition.

The interaction effects obtained for voice-quality are more complex in that we mostly see in the multiple comparisons tests significant changes from baseline to TR03 for both the VR and the non-VR groups – but in opposite directions. For example, the VR groups showed a significant increase in cpp values, whereas those of the non-VR group decreased significantly from baseline to TR03 speeches (F[1,1932] = 5.71, p = 0.02, η² = 0.05). Similarly, the f0-corrected h1-h2 values decreased (and actually changed from positive into negative values) during VR training but increased (from negative towards positive values) during non-VR training (F[1,1193] = 14.25, p < 0.001, η² = 0.12), see Figure 3. For formant dispersion, we found a decrease for the non-VR speakers from TR01 to TR03 that was absent for the VR speakers (F[1,1193] = 14.25, p < 0.001, η² = 0.12). In fact, they rather showed a slight increase. Finally, shimmer increased and HNR decreased from TR01 to TR03, but solely for VR speakers, not for non-VR speakers; F[1,1193] = 8.68, p = 0.003, η² = 0.07; F[1,1193] = 14.58, p < 0.001, η² = 0.12).

![Figure 2: Effects of training on number of pauses.](image)

![Figure 3: Effects of training on f0-corrected h1-h2.](image)

The amount of gesture performed by each group differed from baseline to TR01 and from TR01 to TR03. Both groups performed similarly in terms of gestures in the baseline speech. But, from baseline to TR01 there was a significant decrease in the amount of gestures (p = 0.048) for the Non-VR group. Furthermore, from TR01 to TR03, the VR group showed a significant increase in the amount of gestures (p = 0.044), whereas the Non-VR group kept on with the same reduced gesture level used in TR01.

4. Discussion and Conclusions

The purpose of this experiment was to examine the impact of VR training, i.e. speaking in front of a virtual audience, on secondary-school students’ prosody and gesture patterns. To achieve that, we designed a between-subjects experiment with a baseline and three training sessions so that we could compare changes between VR training and a control condition of Non-VR training. The whole experiment extended over four weeks, with one training session per week.
Acoustic analyses were performed for both groups. Regarding the effects of VR on prosodic parameters, results showed significant changes across groups on f0 values from baseline to TR01, meaning that intonation patterns varied when speakers moved on from facing a real audience to the first training session (with VR or Non-VR); as was found by [18] with teachers who performed the same lesson in class and with a virtual audience, or by [17] whose participants got lectures on the importance of pitch in persuasive investor pitches prior to training. The prosodic changes found here during VR (and non-VR) training are largely consistent with both [18] and [17].

Our results showed that VR-based speech training made the participants’ final training-session speeches longer, more verbose, and faster in terms of the (net) tempo. For voice quality, we found that VR speakers increased their HNR levels and shimmer levels, both a characteristic acoustic change that occurs from neutral to happy speech [24] – whereas the exact opposite applied to the non-VR speaker sample. The speakers of the VR sample developed a clearer, stronger and less "shaky" voice. The longer speaking times and lower HNR levels argue additionally in favor of a stronger audience-oriented way of speaking of the VR group. By contrast, the non-VR speakers developed in the opposite direction.

Results from baseline to Training 1 (TR01) showed for the VR group a significant increase of f0 sd, f0 range and f0 mean whereas the speaking rate decreased significantly. Regarding voice quality, we found that VR speakers increased their CPP levels. The Non-VR group only showed a significant increase of f0 sd. By contrast, f0 range increased only marginally and non-significantly, whereas f0 mean, speaking rate and number of syllables all decreased significantly. In the domain of voice quality, CPP levels decreased significantly too. Regarding gesture, there is no significant difference across groups.

When analyzing the performances from TR01 to TR03, i.e. from first to final training session, a clear difference between the two training groups emerged: VR participants remained at the same high performance level (compared to their baseline speech) in terms of f0 sd and f0 range, and there was even a significant increase in the number of syllables, in the number of pauses, as well as in the speaking rate and the total speaking time. Moreover, the amount of gestures increased significantly. Regarding voice quality, CPP levels and H1*-H2* decreased significantly. The non-VR participants, on the other hand, did not talk more, nor did they make more pauses, or talked faster, or for a longer time; and they did not use more gestures either. For voice quality measures, we saw a significant increase in H1*-H2*, meaning that their voices got breathier across the training. In summary, while training with VR resulted in a livelier, more strongly audience-oriented way of speaking, the non-VR speakers developed in the opposite direction.

Regarding gesture, the different behavior found from baseline to TR01, in which participants performed significantly fewer gestures could be the result of wearing the VR glasses for the first time, while giving a speech and not being able to see ones hands while speaking. This could have caused a feeling of insecurity and artificiality. However, perhaps due to a familiarization with the technology, VR participants again performed significantly more gestures (thus, returning to the same levels as when speaking in front of a real audience) in TR03 than in TR01. This did not apply to the Non-VR participants, whose gesture rate also decreased from baseline to TR01, but then remained at this lower level until TR03. A possible explanation for this effect could be that Non-VR speakers do not feel the sense of presence [9] that VR speakers feel and that makes them present as if the audience in front of them was real. The gain in confidence that the VR technology achieved in our secondary-school setting, can be key to encourage students to practice their speeches voluntarily on a regular basis, therefore, extend VR rehearsals in communication-skills courses.

The VR sample’s prosody and gesture is indicative of a more audience-oriented way of speaking that was triggered by the VR training.

The present study has some limitations. First of all, the size of the sample was relatively small; a larger sample would have yielded more robust results and thus a clearer picture of how VR training sessions affects 17-year-old public speaking abilities. Also, in order to obtain positive effects of VR on f0 parameters during training, i.e. to change or improve participants’ intonation, the study could have added explicit feedback strategies, such as those tested and implemented by [25].

These limitations aside, note that the majority of students enjoyed the VR sessions and found them valuable to face their upcoming oral projects, as other studies also remarked [15,16].

Future studies have to examine in detail what effects would encompass adding a session with a live audience in order to compare it to the baseline speech. Doing so, we would be able to study the sustainability of the VR (and non-VR) training effects, i.e. after having taken off the VR glasses and having to perform in front of a real audience again. Whether the prosody and gesture patterns found during our training sessions would persist under these conditions must be tested in future studies.

Moreover, it would be interesting to study in more detail what beneficial after-training changes in speech prosody VR training is able to trigger "automatically", i.e. merely by using this technology, and what changes require an extra effort, for example, in the form of explicit feedback strategies. That f0-related melodic changes can basically be learned through training such that they last until speeches held after training was demonstrated by [25]. Unlike in our VR condition, the training condition of [25] included an explicit visualization and color-coded real-time evaluation of speech melody, though.

In addition, future studies should measure the students’ public-speaking anxiety before performing the baseline speeches in order to test whether these individual anxiety levels get reduced thanks to VR technology, as previous studies suggested [e.g. 11,12,13]. Also, adding instructions that explicitly encourage speakers to gesture would be interesting in order to study whether the higher gesture rate simultaneously positively affects acoustic-prosodic features, see [26,27].

To sum up, practicing speeches as realistically as possible may contribute to an overall better performance [8,5]. Promoting realistic ways of rehearsing oral skills may enhance the whole experience of delivering a speech and motivate students to practice voluntarily and regularly in educational settings.

We believe that this study serves as a starting point to further develop the relationships between VR practice and its effects on the speakers' prosody and gesture.

5. Acknowledgements

This work benefited from funding awarded by the Spanish Ministry of Economy and Competitiveness [PGC2018-097007-B-100] and the Generalitat de Catalunya [2017 SGR_971]. We also acknowledge the support from the Recercaixa Project [RecercAixa 2017ACUP 00249] and the Department of Translation at Universitat Pompeu Fabra. The 2nd author is the founder and CEO of the speech-technology company AllGoodSpeakers ApS. Please visit the following link for a conflict-of-interest statement: https://www.allgoodspeakers.com/coi
6. References


