DOI: 10.61782/fa.2023.0215

forum acusticum 2023

VOICE PRODUCTION IN VIRTUAL REALITY: EFFECTS OF VISUAL INPUT

Charles J. Nudelman[1*] **Pasquale Bottalico**[1]
[1] Department of Speech and Hearing Science, University of Illinois Urbana-Champaign, United States

ABSTRACT

Purpose To examine the relationships between visual input and voice production in virtual reality with healthy participants.
Methods Voice samples from 30 participants were recorded in six virtual conditions. After each condition, the participants rated their vocal status. The voice recordings were processed to calculate acoustic parameters. The effects of the virtual reality conditions on these voice acoustic parameters and the vocal status ratings were analyzed.
Results The full virtual reality rooms resulted in significantly worse vocal fatigue and vocal discomfort ratings. The virtual reality room size had statistically significant effects on mean sound pressure level and mean pitch strength.
Conclusions This study demonstrated that different types of visual input have distinct effects on voice production and self-reported vocal status. Visual size affected voice acoustic outcomes, while visual fullness affected self-reported outcomes.

Keywords: *virtual reality, voice production*

1. INTRODUCTION

The relationship between room acoustics and voice production holds precautionary value and may assist in reducing the incidence of voice disorders in professional voice users (i.e., individuals whose occupations are associated with increased voice demands such as singing, teaching, performing, lecturing, shouting, etc.) [1]. One aspect of a speaking environment that influences voice production is room size, which is believed to affect how a speaker produces their voice in that setting [2].

1.1 Virtual Reality as a Methodological Tool

Visual input is known to contribute to the majority of motor control processes, learning, and language construction [3]. However, the role of visual input is not described in the currently-accepted voice and speech production models (e.g., The Directions Into Velocities of Articulators (DIVA) model of speech production [4]). One reason this gap in the literature may exist are the methodological difficulties associated with investigating the individual role of visual input in real-life environments. Virtual reality (VR) has emerged be a tool to simulate a variety of environments, which could allow for the examination of the influence of visual sensory input on voice production. The current study was designed to answer the following question: How are objective voice parameters and self-reported vocal status in healthy speakers associated with the visual perception of the size and fullness of virtual rooms? Addressing this question will provide an initial understanding of the mechanistic effects of visual input on objective and subjective voice parameters.

2. METHODS

Thirty participants (19–44 years; mean (SD) 25 (5) years) were enrolled in the study. Twenty of the participants reported their sex as female and 10 as male. Inclusion criteria for the present study was being over the age of 18 years old, passing a voice and hearing screen, and reporting no history of speech, language, or hearing disorders.

*__Corresponding author__: nudelma2@illinois.edu.

Speech samples of the participants were recorded in six different virtual conditions. The recordings were performed in a sound attenuating double-walled Whisper Room. The effects of the virtual reality conditions on 1) the self-reported vocal status on visual analog scales (VASs), 2) sound pressure level (SPL) values, and 3) the mean pitch strength (PS) were evaluated.

2.1 Instructions and Conditions

In each of the six virtual conditions, participants responded to open-ended questions [5] for a minimum of three minutes and read aloud the "The Rainbow Passage," a standardized text [6]. Following each condition, the participants rated their vocal effort, fatigue, and discomfort, on separate VASs for each construct.
The six VR conditions were:

- 1) A sparsely occupied small room (virtual meeting room) with two audience members present in the room
- 2) The same small room densely occupied with seven audience members present in the room
- 3) A sparsely occupied medium room (virtual lecture hall) with 16 audience members
- 4) The same medium room densely occupied with 65 audience members
- 5) A large room (virtual theater) sparsely occupied with 108 audience members
- 6) The same large room densely occupied with 1,343 audience members

The number of audience members reflected approximately 45% and 75% of the capacity of each room for the sparse and dense occupancies, respectively. Of note, this occupancy is approximated due to a distancing metric which was applied to the sparse conditions, so that audience members were spaced equally. This distancing metric is included within the virtual reality software. The locations of audience members were fixed across all conditions to include at least one audience member occupying the farthest possible room position from the participant. All conditions were presented without external noise added and were randomized.

2.2 Equipment

The speech material was recorded by an M2211 microphone (NTi Audio, Tigard, OR, United States). The VR equipment involved an HTC Vive Pro 2 (HTC Corporation, Taiwan). Vizard (Santa Barbara, CA) software (Ovation VR, www.ovationvr.com) was used to display the virtual rooms based on head position and orientation.

2.3 Analysis

All participant recordings were processed to calculate mean SPL values and mean PS with MATLAB R2022b (Mathworks, Natick, 284 MA, USA) and Praat 5.4/5.4.17 (Netherlands). Statistical analyses were conducted using R version 4.2.0 (R Development Core Team, 2022). Linear Mixed-Effects (LME) models were fitted by restricted maximum likelihood (REML). Tukey's post-hoc pairwise comparisons were performed to examine the differences between all levels of the fixed factors of interest. These are pairwise z tests, where the z statistic represents the difference between an observed statistic and its hypothesized population parameter in units of the standard deviation. The LME output included the estimates of the fixed effects coefficients, the standard error associated with the estimate, the degrees of freedom (df), the test statistic (t), and the p-value. The Satterthwaite method was used to approximate degrees of freedom and calculate p-values.

3. RESULTS

3.1 Results: Sound Pressure Level

3.1.1 Sound Pressure Level: Reading Task

For the reading task, there was a significant relationship between condition and SPL, in which SPL increased by approximately 1 dB (SPL) ($p < 0.001$) when speaking in the large room compared to the small room and approximately 0.5 dB (SPL) ($p < 0.001$) when speaking in the medium room compared to the small room. Post-hoc comparisons confirmed that the increases in SPL comparing the large room to the small room, the medium room to the small room, and the large room to the medium room are statistically significant (SE = 0.17, z = 6.41, $p < 0.001$, SE = 0.17, z = 3.10, p = 0.006, and estimate = 0.57 dB (SPL), SE = 0.17, z = 3.31, p = 0.003, respectively).

3.1.2 Sound Pressure Level: Spontaneous Speech Task

As with the reading task, there was a significant relationship between condition and SPL during the spontaneous speech task, in which SPL increased by approximately 0.7

dB (SPL) ($p < 0.001$) when speaking in the large room compared to the small room. Post-hoc comparisons confirmed that the increases in SPL comparing the large room to the small room and the large room to the medium room are statistically significant (SE = 0.16, z = 4.33, $p < 0.001$ and SE = 0.16, z = 3.00, $p < 0.008$). There were no significant relationships between fullness and SPL for either task. These relationships are displayed in Table 1.

3.2 Results: Pitch Strength

3.2.1 Pitch Strength: Reading Task

For the reading task, there was a significant relationship between condition and mean pitch strength in which mean pitch strength increased by 0.01 (p = 0.004) when speaking in the large room compared to the small room and by 0.01 when speaking in the medium room compared to the small room (p = 0.041). Post-hoc comparisons confirmed that the increases in mean pitch strength comparing the large room to the small room are statistically significant (SE = 0.003, z = 2.90, p = 0.010).

3.2.2 Pitch Strength: Spontaneous Speech Task

As with the reading task, there was a significant relationship between condition and mean pitch strength during the spontaneous speech task, in which mean pitch strength increased by 0.01 ($p < 0.001$) when speaking in the large room compared to the small room. Post-hoc comparisons confirmed that the increases in mean pitch strength comparing the large room to the small room are statistically significant (SE = 0.002, z = 3.42, p = 0.002). There were no significant relationships between fullness and mean pitch strength for either task. These relationships are displayed in Table 2.

3.3 Results: Vocal Status Ratings

The vocal fatigue ratings had a statistically significant relationship with the fullness of the virtual rooms in which the densely occupied conditions were rated approximately 5 points higher on the vocal fatigue VAS (p = 0.012) compared to the sparsely occupied conditions.
The vocal discomfort ratings had a statistically significant relationship with the fullness of the virtual rooms in which the densely occupied conditions were rated approximately 6 points higher on the vocal discomfort VAS (p = 0.048) compared to the sparsely occupied conditions.
There was no statistically significant association between vocal effort ratings on the VAS and condition.
The relationships among the three vocal status ratings and virtual room conditions are displayed in Table 3.

Table 1: LME models output run with SPL as the response variable and the large room conditions, medium room conditions, and densely occupied conditions as fixed factors.

Fixed factors	*Estimate (-)*	*Std. Error(-)*	*df*	*t*	*p*
SPL (reading task)					
(Intercept: Small Room)	66.18	0.50	30	131.48	<0.001***
Large room	1.10	0.17	137	6.41	<0.001***
Medium room	0.54	0.17	137	3.10	0.002**
Densely Occupied Conditions	>0.00	0.14	137	0.02	0.985
SPL (spontaneous speech task)					
(Intercept: Small Room)	66.01	0.51	30	129.42	<0.001***
Large room	0.71	0.16	137	4.33	<0.001***
Medium room	0.22	0.16	137	1.35	0.181
Densely Occupied Conditions	>0.00	0.13	137	0.04	0.966
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1					

Table 2: LME models output run with mean pitch strength as the response variable and the large room conditions, medium room conditions, and densely occupied conditions as fixed factors.

Fixed factors	*Estimate (-)*	*Std. Error(-)*	*df*	*t*	*p*
Mean Pitch Strength (reading task)					
(Intercept: Small Room)	0.34	0.01	32	46.39	<0.001***
Large room	0.01	>0.00	137	2.91	0.004**
Medium room	0.01	>0.00	137	2.07	0.041*
Densely Occupied Conditions	>0.00	>0.00	137	0.58	0.562
Mean Pitch Strength (spontaneous speech task)					
(Intercept: Small Room)	0.36	>0.00	29	41.47	<0.001***
Large room	0.01	>0.00	137	3.42	<0.001***
Medium room	>0.00	>0.00	137	1.66	0.099.
Densely Occupied Conditions	>0.00	>0.00	137	0.14	0.889
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1					

Table 3: LME models output run with vocal fatigue, vocal discomfort, and vocal effort response variables and the large room conditions, medium room conditions, and densely occupied conditions as fixed factors.

Fixed factors	*Estimate (-)*	*Std. Error(-)*	*df*	*t*	*p*
Vocal Fatigue Ratings					
(Intercept: Small Room)	23.31	3.56	46	6.55	<0.001***
Large room	3.39	2.48	137	1.37	0.173
Medium room	2.18	2.48	137	0.88	0.381
Densely Occupied Conditions	5.13	2.02	137	2.54	0.012*
Vocal Discomfort Ratings					
(Intercept: Small Room)	20.00	4.08	42	4.91	<0.001***
Large room	2.39	2.86	135	0.84	0.404
Medium room	-2.25	2.86	135	-0.79	0.433
Densely Occupied Conditions	5.71	2.86	135	2.00	0.048*
Vocal Effort Ratings					
(Intercept: Small Room)	32.79	4.46	36	7.35	<0.001***
Large room	5.07	2.60	135	1.95	0.053.
Medium room	3.64	2.60	135	1.40	0.163
Densely Occupied Conditions	3.11	2.60	135	1.20	0.234
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1					

4. DISCUSSION

The primary aim of the present study was to examine the effects of visual input on two specific outcome measures: 1) objective voice parameters and 2) subjective vocal status ratings. The larger virtual rooms resulted in significantly higher SPL and PS compared to smaller virtual rooms. Also, densely occupied rooms resulted in significantly higher ratings of vocal fatigue and vocal discomfort compared to sparsely occupied rooms. These results imply that room size has different mechanistic effects on voice production compared to room fullness in VR rooms.

4.1 Visual Input: Influence on Performance and Perceived Fatigue

Room size and fullness in VR may be related to performance and perceived fatigue, respectively. Performance fatigue is indicated through physiologic outcome measures [7], while perceived fatigue is generally self-reported. In the present study, room size input was related to the voice acoustic outcome measures (reflecting performance fatigue), while the room fullness input was related to self-reported vocal status ratings (reflecting perceived fatigue).

5. CONCLUSION

The present study found that VR room size and fullness have different mechanistic effects on voice production. Larger VR rooms resulted in significantly higher voice acoustic parameters compared to smaller virtual rooms. More densely occupied rooms resulted in higher ratings of vocal fatigue and vocal discomfort compared to sparsely occupied rooms. We propose that these visual size and fullness input may be linked to performance fatigue and perceived fatigue, respectively.

6. REFERENCES

[1] L. Allen and A. Hu, "Voice disorders in the workplace: A scoping review," *Journal of Voice*, 2022.

[2] D. Pelegrín-García, B. Smits, J. Brunskog, and C.-H. Jeong, "Vocal effort with changing talker-to-listener distance in different acoustic environments," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 1981–1990, 2011.

[3] P. Kuhl and M. Rivera-Gaxiola, "Neural substrates of language acquisition," *Annu. Rev. Neurosci.*, vol. 31, pp. 511–534, 2008.

[4] J. A. Tourville and F. H. Guenther, "The diva model: A neural theory of speech acquisition and production," *Language and cognitive processes*, vol. 26, no. 7, pp. 952–981, 2011.

[5] A. Bradlow, L. Ackerman, L. Burchfield, L. Hesterberg, J. Luque, and K. Mok, "Allsstar: Archive of l1 and l2 scripted and spontaneous transcripts and recordings," in *Proc. of the International Congress on Phonetic Sciences*, pp. 356–359, 2010.

[6] G. Fairbanks, "The rainbow passage," *Voice and articulation drillbook*, vol. 2, pp. 127–127, 1960.

[7] E. J. Hunter, L. C. Cantor-Cutiva, E. van Leer, M. Van Mersbergen, C. D. Nanjundeswaran, P. Bottalico, M. J. Sandage, and S. Whitling, "Toward a consensus description of vocal effort, vocal load, vocal loading, and vocal fatigue," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 2, pp. 509–532, 2020.