DOI: 10.61782/fa.2025.0765

SENSORY-DRIVEN VOICE MODULATION: EFFECTS OF AURALIZED ACOUSTICS AND VISUAL VR ON VOICE OUTCOMES

Charles J. Nudelman[1*] **Asritha Tunuguntla**[1] **Naomi Ha**[1]
Isabella Rogala[1] **Kaliyah House-Henry**[1] **Bella Lopez**[1]
Ricardo Perez Guerrero[1] **Pasquale Bottalico**[1]

[1] Department of Speech and Hearing Science, University of Illinois Urbana-Champaign, United States

ABSTRACT

The influences of auralized room acoustics and visual scenes in virtual reality (VR) on voice and the feasibility of VR-based voice therapy were investigated in this study.

Two groups of participants were involved – (Group one): 41 young adults (18-27 years) and (Group two): 10 pre-service teachers (18-19 years). Group one performed speech tasks in varying conditions: (i) auralized, (ii) visual-only, and (iii) audiovisual. Group two performed speech tasks in a VR intervention. With Group two, clinician-mediated feedback was provided in the VR. Voice parameters were analyzed.

Audiovisual VR environments significantly ($p <.05$) influenced voice outcomes. Larger, noisy, and densely occupied VR spaces contributed to more pronounced effects. The VR intervention resulted in significantly lower time dose ($p <.05$) compared to the control condition. Real-time clinician feedback within VR resulted in reduced sound pressure level ($p <.05$), fundamental frequency ($p <.05$), and time dose ($p <.05$) compared to instances without clinician feedback.

The results demonstrate that VR environments can significantly alter voice and may enhance voice therapy.

[*]***Corresponding author***: nudelma2@illinois.edu.

Keywords: *sensory input, virtual reality, voice production, auralized acoustics*

1. INTRODUCTION

Historically, the effects of room acoustic conditions on speech have been studied from the perspective of listeners. This is likely due to the overt detrimental effects that high noise levels and reverberation time have on speech intelligibility. The common challenge of understanding speech in noisy and reverberant environments has driven extensive research (e.g., [1-3]). Emerging evidence has shifted to examining the speaker's perspective, revealing significant detrimental associations between room acoustics and voice production, such that room acoustics can negatively influence a speaker to the point that they report symptoms of voice disorders (e.g., [4,5]).

High-quality recordings are pivotal for assessing voice disorders, as they inform clinical interventions. However, when clinical or research settings fail to replicate the sensory complexities of real-world environments, there is a risk of producing interventions that lack ecological validity. This limitation, referred to as the ecological validity crisis [6], arises from the mismatch between controlled laboratory conditions and natural communication scenarios. One potential solution lies in auralization technology, which simulates a binaural listening experience at a given position in a modeled space [7].

Visual simulations in virtual reality (VR) may complement auditory auralizations. VR simulations provide control over an individual's visual environment, enabling researchers to investigate how visual sensory input influences voice production. Immersive VR environments simulate real-world settings with high levels of presence and immersion. Key visual factors such as room size and occupancy have been shown to significantly modulate vocal behaviors, including sound pressure level (SPL) and fundamental frequency (F0). For instance, larger and densely occupied virtual spaces elicit higher SPL and F0 [8,9].

To better understand how voice production adapts to auralized room acoustics and simulated visual input, the present study involved two experiments. Two research questions guided the experiments:

(1) How do different levels of simulated sensory information (auditory, visual, and audiovisual) affect voice production?

(2) How do instances of clinician-mediated feedback in a VR intervention affect voice production?

2. METHODOLOGY

Two experiments were undertaken to address the research questions.

2.1 Experiment 1: Methodology

Forty-seven participants (18 – 27 years; mean (SD) 21 (2) years) were enrolled and recruited through sequential convenience sampling. Inclusion criteria were being over the age of 18 years old, passing a voice and hearing screen, and reporting no history of speech, language, or hearing disorders. Voice screenings were performed by a certified speech-language pathologist (author, C.J.N.), and included (1) video stroboscopic imaging of the larynx, and/or (2) completion of the Voice Handicap Index-10 (VHI-10; [10]), (3) completion of the Voice-Related Quality of Life (V-RQOL; [11]), and (4) completion of the Vocal Fatigue Index (VFI; [12]). Hearing screenings included pure-tone audiometry at octave frequencies from 250 to 8000 Hz. A total of six participants were excluded from the study. One participant was excluded from the study due to the presence of hearing loss; four participants were excluded due to drop-out/incomplete participation, and one participant was excluded due to failing the voice screen. Thus, 41 participants (18 – 27 years; mean (SD) 21 (2) years) met the inclusion criteria. Thirty-three of these 41 participants reported their gender as female and eight as male. With protocol approval from the University of Illinois Urbana-Champaign Institutional Review Board (IRB #23336), speech samples of the participants were recorded in six different auditory simulations, six different visual VR simulations, and six audiovisual simulations. The recordings were performed in a sound attenuating double-walled Whisper Room. The effects of the simulations on acoustic voice parameters were assessed.

2.2 Experiment 2: Methodology

Ten participants (18–19 years; mean (SD) 18.5 (0.7) years) were enrolled and recruited through sequential convenience sampling. Of note, all ten participants completed the research study, however, due to a software error, only nine participants' voice samples were recorded. All participants were pursuing a bachelor's degree from the College of Education at the University of Illinois Urbana-Champaign and reported a desired career of being a teacher. All participants were female due to the vast majority of public school teachers in the United States being female. With protocol approval from the University of Illinois Urbana-Champaign Institutional Review Board (IRB #23336), speech samples of the participants were recorded in a VR intervention condition. To this end, the participants were alone in the Whisper Room wearing a Meta Quest 2 VR headset and open-backed headphones. Meanwhile, the clinician (C.J.N.) was seated outside wearing his own Meta Quest 2. Both the participant and clinician were virtually present in the same environment, which was facilitated by Ovation VR software. The intervention incorporated Conversation Training Therapy [13] using the Rehabilitation Treatment Specification System – a comprehensive framework that aims to guide a rehabilitative treatment [14]. To this end, prior to the VR intervention, the clinician performed stimulability testing and a pre-brief. During the intervention, the clinician provided feedback based on the participants' vocal intensity. Participants' acoustic voice parameters were assessed within the intervention condition according to each provision of clinician-mediated feedback.

2.3 Processing and Statistical Analyses: Experiment 1

For experiment 1, all participant recordings were processed to extract the SPL and F0. The recordings were processed with MATLAB R2023b (Mathworks, Natick, MA) and Praat 6.0.13 (Netherlands). Specifically, a custom MATLAB script was applied to estimate the SPL values of the voiced speech signals. The F0 of the speech signals was estimated with Praat using the autocorrelation method.

For all acoustic voice parameters, summary statistics were calculated to evaluate the uncertainty of their mean error. Prior to calculating the summary statistics, the interquartile technique was employed to remove outliers. Statistical analyses were conducted using R version 4.2.0. Linear mixed-effects (LME) models were fitted by restricted maximum likelihood. Models were selected based on the Akaike and Bayesian information criterion, with the model with the lowest value being preferred. 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.

2.4 Processing and Statistical Analyses: Experiment 2

For experiment 2, All participant recordings were processed to extract SPL, F0, and time dose (Dt%). The recordings were processed in the same manner as experiment 1. Dt% was estimated with Praat using the command "To Pitch (ac)," which creates a pitch object from every selected sound object within the window.

3. RESULTS

Across the two experiments, a total of six LMEs were conducted – three for experiment 1 (SPL, F0 (female), and F0 (male)) and three for experiment 2 (SPL, F0, Dt%).

3.1 SPL Results: Experiment 1

The auralization condition and the multisensory condition resulted in significant effects on SPL. Specifically, SPL values were approximately 2 and 3 dB higher (p <.001 in both cases), in the auralization and multisensory conditions, respectively compared to the visual VR conditions. Post-hoc comparisons confirmed that the increases in SPL across these conditions were statistically significant (p <.001, across both post-hoc comparisons), and revealed that the multisensory conditions resulted in increases in SPL by approximately 1 dB, which were significantly higher (p <.001) than the SPL values demonstrated during the auralization conditions. There was a significant effect of room on SPL. SPL values were approximately 0.8 dB higher (p <.001) in the medium room compared to the small room, and SPL values were approximately 0.5 dB higher (p <.001) in the large room compared to the small room. Post-hoc comparisons confirmed these effects with statistical significance (p <.001 in both cases). These results are displayed in Table 1 and Figure 1.

Table 1. LME model output and post-hoc comparison model output with SPL (experiment 1) as the response variable and condition and room as fixed factors.

Fixed factors	*Estimate (dB)*	*Std. Error(dB)*	*df*	*t*	*p*
Sound Pressure Level (Experiment 1)					
Small Room, Visual VR Condition (Intercept)	58.1	0.6	42	96.7	<0.001
Auralization	1.8	0.3	40	5.4	<0.001
Multisensory VR	3.0	0.3	40	10.8	<0.001
Room Medium	0.8	0.1	1350	6.3	<0.001
Room Large	0.5	0.1	1356	4.1	<0.001
Post-Hoc Comparisons: Sound Pressure Level (Condition)					
Fixed Factors	*Estimate (dB)*	*Std. Error(dB)*	*z*	*p*	
Auralization – Visual VR	1.8	0.3	5.4	<0.001	
Multisensory VR – Visual VR	3.0	0.3	10.8	<0.001	
Multisensory VR – Auralization	1.2	0.2	5.9	<0.001	

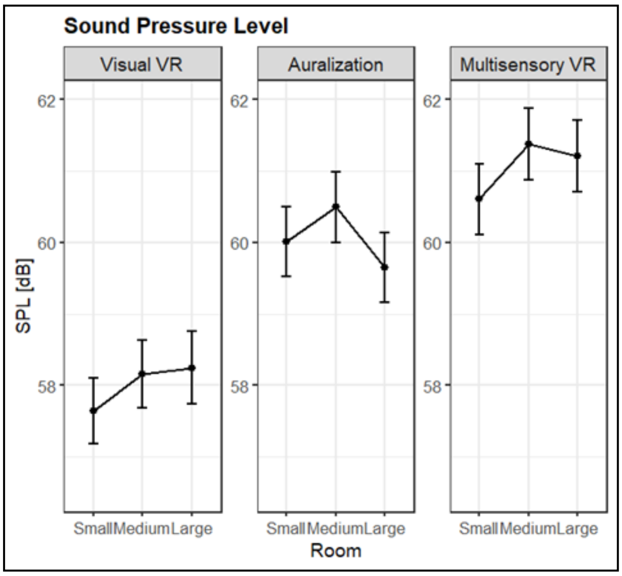

Figure 1. Mean and standard error of SPL for all participants in each room and occupancy across experiment 1 (Visual VR, Auralization, and Multisensory VR).

3.2 F0 Results: Experiment 1

Considering the female participants, the auralization condition and the multisensory condition resulted in significant effects on F0. Specifically, F0 values were approximately 3 and 5 Hz higher in the auralization and multisensory conditions, respectively compared to the visual VR conditions (p =.022 and p <.001, respectively). Post-hoc comparisons confirmed that the increases in F0 across these conditions were statistically significant (p <.001 and p =.039) and revealed that the multisensory conditions resulted in increases in F0 by approximately 2.3 Hz, which were significantly higher (p <.001) than the F0 values demonstrated during the auralization conditions. There was also a significant effect of room on F0, with F0 approximately 1 Hz higher in the medium and large room compared to the small room (p =.003 and .043, respectively) and this effect was confirmed only for the medium versus small room by post-hoc comparisons (p =.010). The results are displayed in Table 2 and Figure 2.

Table 2. LME model output and post-hoc comparison model output with F0 (female (experiment 1)) as the response variable and condition and room as fixed factors.

Fixed factors	*Estimate (Hz)*	*Std. Error(Hz)*	*df*	*t*	*p*
Fundamental Frequency Female (Experiment 1)					
Small Room, Visual VR Condition (Intercept)	152.8	3.3	33	47.0	<0.001
Auralization	2.5	1.0	32	2.4	0.022
Multisensory VR	4.8	0.8	32	6.0	<0.001
Room Medium	1.3	0.5	1086	2.9	0.003
Room Large	0.9	0.5	1094	2.0	0.043
Post-Hoc Comparisons: Fundamental Frequency (Condition)					
Fixed Factors	*Estimate (Hz)*	*Std. Error(Hz)*	*z*	*p*	
Auralization – Visual VR	2.4	1.0	2.4	<0.001	
Multisensory VR – Visual VR	4.8	0.8	6.0	0.039	
Multisensory VR – Auralization	2.3	0.6	3.7	<0.001	

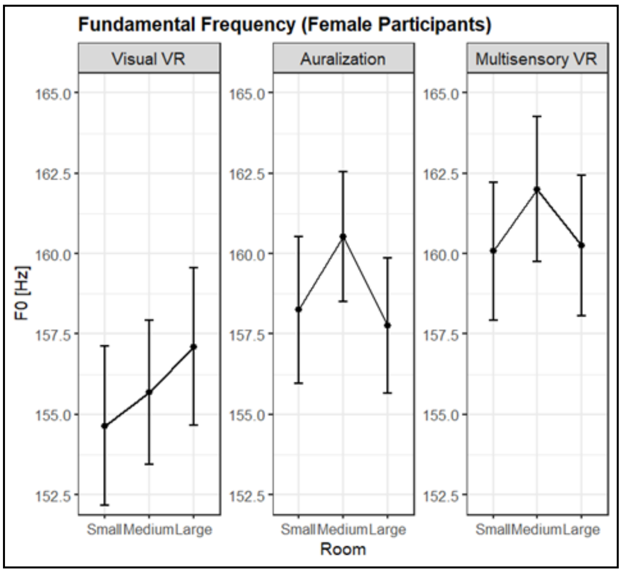

Figure 2. Mean and standard error of F0 for female participants in each room and occupancy across experiment 1 (Visual VR, Auralization, and Multisensory VR).

Considering the male participants, the auralization and multisensory conditions were significantly different for measures of F0. Specifically, F0 values were approximately 2 Hz higher in the multisensory conditions compared to the auralization conditions (*p* =.006). There was a significant effect of room on F0, with F0 approximately 2 Hz higher in the medium room compared to the small room (*p* =.003) and this effect was confirmed by post-hoc comparisons (*p* =.009). These results are displayed in Table 3 and figure 3.

Table 3. LME model output and post-hoc comparison model output with F0 (male (experiment 1)) as the response variable and condition and room as fixed factors.

Fixed factors	*Estimate (Hz)*	*Std. Error(Hz)*	*df*	*t*	*p*
Fundamental Frequency Male (Experiment 1)					
Small Room, Visual VR Condition (Intercept)	101.8	4.0	7	25.8	<0.001
Auralization	1.9	1.8	7	1.0	0.333
Multisensory VR	4.0	2.0	7	2.1	0.078
Room Medium	1.8	0.6	261	2.9	0.003
Room Large	1.1	0.6	261	1.9	0.062
Post-Hoc Comparisons: Fundamental Frequency (Condition)					
Fixed Factors	*Estimate (Hz)*	*Std. Error(Hz)*	*z*	*p*	
Auralization – Visual VR	1.9	1.8	1.0	0.528	
Multisensory VR – Visual VR	4.0	2.0	2.1	0.087	
Multisensory VR – Auralization	2.2	0.7	3.0	0.006	

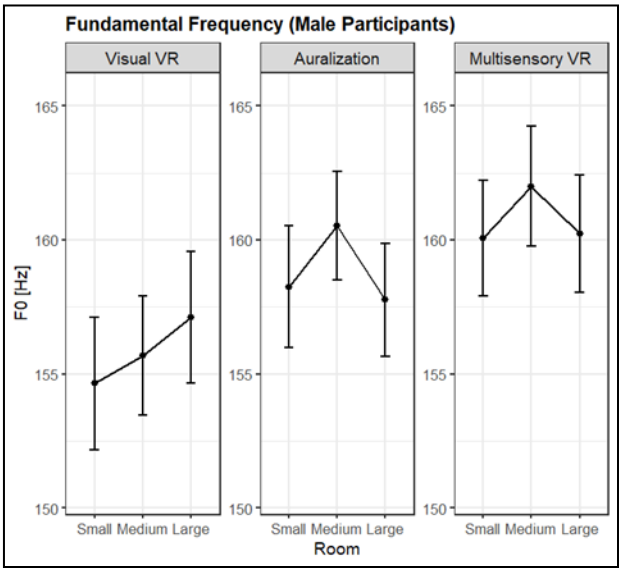

Figure 3. Mean and standard error of F0 for male participants in each room and occupancy across experiment 1 (Visual VR, Auralization, and Multisensory VR).

3.3 SPL Results: Experiment 2

To assess how instances of clinician-mediated feedback in the VR intervention condition (experiment 2) influenced SPL, an LME was fitted with a single predictor, Feedback. The reference level was the absence of feedback. The random effects term was random intercepts for participant ID. Within the VR intervention condition, there was a significant effect of the presence of clinician-mediated feedback on SPL, with values approximately 1.5 dB lower after the provision of feedback compared to instances of no clinician-mediated feedback (*p* <.001). These results are displayed in Table 4 and Figure 4.

Table 4. LME model output with SPL (experiment 2) as the response variable and presence of feedback as the fixed factor.

Fixed factors	*Estimate (dB)*	*Std. Error(dB)*	*df*	*t*	*p*
SPL Within VR Intervention (Experiment 2)					

Absence of feedback (Intercept)	62.5	1.2	8	50.7	<0.001***
Presence of Feedback	-1.5	0.3	115	-5.0	<0.001***

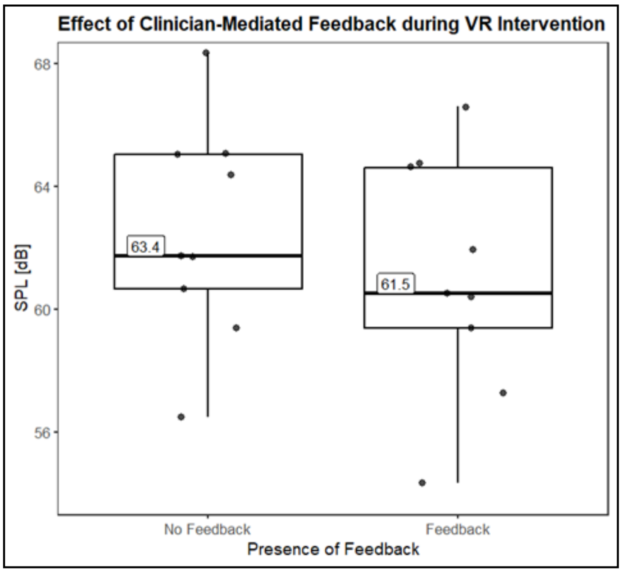

Figure 4. Boxplots indicating the median SPL values during the VR intervention condition in reference to the absence/presence of clinician-mediated feedback (No Feedback median = 63.4 dB, Feedback median = 61.5 dB). The boxplots represent the distribution of SPL measurements across their respective feedback instances, with the boxes displaying the interquartile range and the whiskers indicating variability outside the upper and lower quartiles.

3.4 F0 Results: Experiment 2

To assess how instances of clinician-mediated feedback in the VR intervention condition influenced F0, an LME was fitted with a single predictor, Feedback. The reference level was the absence of feedback. The random effects term was random intercepts for participant ID. Within the VR intervention condition, there was a significant effect of the presence of clinician-mediated feedback on F0, with values approximately 6 Hz lower after the provision of feedback compared to instances of no clinician-mediated feedback (p =.011). These results are displayed in Table 5 and Figure 5.

Table 5. LME model output with F0 (experiment 2) as the response variable and presence of feedback as the fixed factor.

Fixed factors	*Estimate (Hz)*	*Std. Error(Hz)*	*df*	*t*	*p*
F0 Within VR Intervention (Experiment 2)					
Absence of feedback (Intercept)	183.3	3.9	9	47.3	<0.001
Presence of Feedback	-5.6	2.2	115	-2.6	0.011

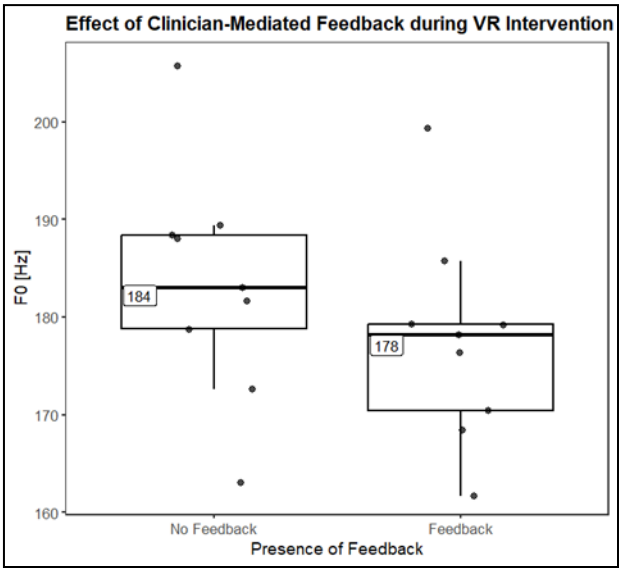

Figure 5. Boxplots indicating the median F0 values during the VR intervention condition in reference to the absence/presence of clinician-mediated feedback (No Feedback median = 184 Hz, Feedback median = 178 Hz). The boxplots represent the distribution of F0 measurements across their respective feedback instances, with the boxes displaying the interquartile range and the whiskers indicating variability outside the upper and lower quartiles.

3.5 Dt% Results: Experiment 2

To assess how instances of clinician-mediated feedback in the VR intervention condition influenced Dt(%), an

LME was fitted with a single predictor, Feedback. The reference level was the absence of feedback. The random effects term was random intercepts for participant ID. Within the VR intervention condition, there was a significant effect of the presence of clinician-mediated feedback on Dt(%), with values approximately 5% lower after the provision of feedback compared to instances of no clinician-mediated feedback (p =.001). These results are displayed in Table 6 and Figure 6.

Table 6. LME model output with Dt% (experiment 2) as the response variable and presence of feedback as the fixed factor.

Fixed factors	*Estimate (%)*	*Std. Error(%)*	*df*	*t*	*p*
Dt% Within VR Intervention (Experiment 2)					
Absence of feedback (Intercept)	59.9	2.3	10	25.7	<0.001
Presence of Feedback	-4.7	1.4	115	-3.3	0.001

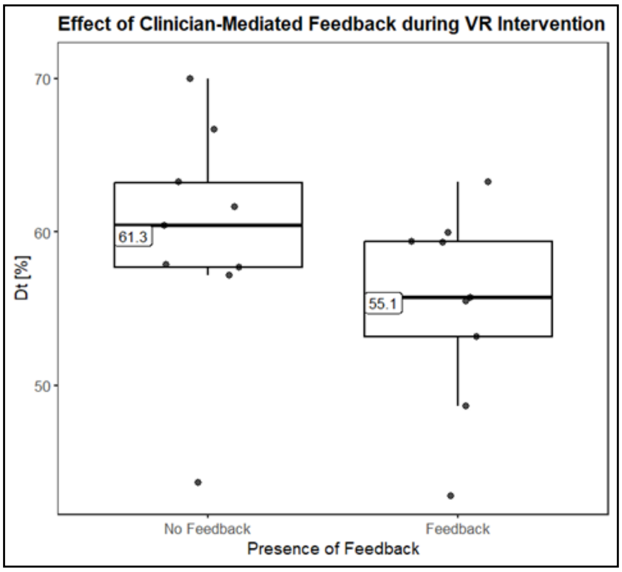

Figure 6. Boxplots indicating the median Dt(%) during the VR intervention condition in reference to the absence/presence of clinician-mediated feedback (No Feedback median = 61.3%, Feedback median = 55.1%). The boxplots represent the distribution of Dt(%) measurements across feedback instances, displaying the interquartile range and variability outside the upper and lower quartiles.

4. DISCUSSION

This study examined the influence of varying sensory input (auditory, visual, and audiovisual) on voice production in VR environments and assessed the feasibility of VR-based voice therapy. The results demonstrate that voice outcomes were significantly altered by multisensory VR simulations. Furthermore, within a VR intervention, real-time clinician-mediated feedback was associated with significant reductions in SPL, F0, and time dose (Dt%). These results highlight the the potential of VR to serve as a platform for research and therapeutic applications in the area of voice disorders. Future research will benefit from examining the distinct sensorimotor profiles (e.g., [16]) of speakers with regard to their candidacy for VR-based voice therapy. Related considerations are the potential influences of biological sex (e.g., [17]) on sensorimotor integration, the infleunce of language (e.g., [18]), and the influence of cultural factors on self-pereceived voice symptoms (e.g., [19]). In addition to implementing VR within clinical environments, future work will benefit from the exploration of wearable voice dosimeters (e.g., [20]) and auditory feedback devices (e.g., [21]) in daily life.

4.1 Limitations

The current study has several limitations. First, the auralizations incorporated a head-and-torso simulator with fixed anatomical features. However, auralizations were provided to participants with unique anatomical features. Second, the visual VR conditions did not reflect real-life environments with fidelity. This hinders the ability to generalize the results. Third, there was sensory incongruence associated with the auralizations and visual VR conditions in experiment 1which diminishes generalizability.

5. CONCLUSION

This study highlights the distinct influences of auditory, visual, and multisensory VR simulations on voice production and the feasibility of a VR-based intervention. In experiment 1, multisensory VR conditions elicited the greatest changes in acoustic voice parameters compared to unimodal conditions. In experiment 2, real-time clinician-mediated feedback in VR resulted in significantly reduced voice acoustic parameters, indicating improved vocal efficiency. These findings demonstrate evidence for VR technology as a tool for both voice-related research and therapy.

6. ACKNOWLEDGMENTS

The authors acknowledge Dr. Edward Hashimoto for his generous VR gift. This study was supported by a 2024 Ph.D. scholarship from The Council of Academic Programs in Communication Sciences and Disorders and a 2024 Raymond H. Stetson Scholarship in Phonetics and Speech Science from the Acoustical Society of America (both awarded to the first author, C.J.N.).

7. REFERENCES

[1] Bottalico, P., Murgia, S., Puglisi, G. E., Astolfi, A., & Ishikawa, K. (2021). Intelligibility of dysphonic speech in auralized classrooms. *Journal of the Acoustical Society of America*, *150*(4), 2912–2920.

[2] Nudelman, C. J., D'Orazio, D., & Bottalico, P. (2023). Reducing noise in restaurants: Feasibility and recommendations. *The Hearing Journal*, *76*(02), 6–7.

[3] Ishikawa, K., Nudelman, C., Park, S., & Ketring, C. (2021). Perception and acoustic studies of vowel intelligibility in dysphonic speech. *Journal of Voice*, *35*(4), 659-e11.

[4] Cutiva, L. C. C., & Burdorf, A. (2015a). Effects of noise and acoustics in schools on vocal health in teachers. *Noise & Health*, *17*(74), 17.

[5] Nudelman, C. J., Bottalico, P., & Cantor-Cutiva, L. C. The Effects of Room Acoustics on Self-reported Vocal Fatigue: A Systematic Review. *Journal of Voice*, S0892-1997.

[6] Aanstoos, C. M. (1991). Experimental psychology and the challenge of real life. *American Psychologist*, *46*(2), 77–78.

[7] Kleiner, M., Dalenbäck, B. I., & Svensson, P. (1993). Auralization-an overview. *Journal of the Audio Engineering Society*, *41*(11), 861-875.

[8] Nudelman, C. J., & Bottalico, P. Investigating the Impact of Visual Input on Voice Production in Virtual Reality. *Journal of Voice*, S0892-1997.

[9] Daşdöğen, Ü., Awan, S. N., Bottalico, P., Iglesias, A., Getchell, N., & Abbott, K. V. The Influence of Multisensory Input On Voice Perception and Production Using Immersive Virtual Reality. *Journal of Voice,* S0892-1997.

[10] Rosen, C. A., Lee, A. S., Osborne, J., Zullo, T., & Murry, T. (2004). Development and validation of the voice handicap index-10. *The Laryngoscope*, *114*(9), 1549–1556.

[11] Hogikyan, N. D., & Sethuraman, G. (1999). Validation of an instrument to measure voice-related quality of life (V-RQOL). *Journal of Voice*, *13*(4), 557–569.

[12] Nanjundeswaran, C., Jacobson, B. H., Gartner-Schmidt, J., & Verdolini Abbott, K. (2015). Vocal Fatigue Index (VFI): Development and Validation. *Journal of Voice*, *29*(4), 433–440.

[13] Gartner-Schmidt, J., Gherson, S., Hapner, E. R., Muckala, J., Roth, D., Schneider, S., & Gillespie, A. I. (2016). The development of conversation training therapy: A concept paper. *Journal of Voice*, *30*(5), 563–573.

[14] Van Stan, J. H., Dijkers, M. P., Whyte, J., Hart, T., Turkstra, L. S., Zanca, J. M., & Chen, C. (2019). The Rehabilitation Treatment Specification System: Implications for improvements in research design, reporting, replication, and synthesis. *Archives of Physical Medicine and Rehabilitation*, *100*(1), 146–155.

[15] Bottalico, P., Codino, J., Cantor-Cutiva, L. C., Marks, K., Nudelman, C. J., Skeffington, J., ... & Rubin, A. D. (2020). Reproducibility of voice parameters: The effect of room acoustics and microphones. *Journal of Voice*, *34*(3), 320-334.

[16] Nudelman, C. J. (2024). Sensorimotor Integration in Patients with Voice Disorders: A Scoping Review of Behavioral Research. *Folia Phoniatrica et Logopaedica*, *76*(6), 501-520.

[17] Nudelman, C., Webster, J., & Bottalico, P. (2024). The effects of reading speed on acoustic voice parameters and self-reported vocal fatigue in students. *Journal of Voice*, *38*(1), 243-e1.

[18] Cantor-Cutiva, L. C., Bottalico, P., Webster, J., Nudelman, C., & Hunter, E. (2023). The effect of bilingualism on production and perception of vocal fry. *Journal of Voice*, *37*(6), 970-e1.

[19] Nudelman, C. J., Bottalico, P., van Mersbergen, M., & Nanjundeswaran, C. (2024). Toward enhanced voice-related self-reports: Translation, cross-cultural adaptation, and validity. *Journal of Voice*.

[20] Bottalico, P., & Nudelman, C. J. (2023). Do-it-yourself voice dosimeter device: A tutorial and performance results. *Journal of Speech, Language, and Hearing Research*, *66*(7), 2149-2163.

[21] Nudelman, C. J., Codino, J., Fry, A. C., Bottalico, P., & Rubin, A. D. (2022). Voice biofeedback via bone conduction headphones: effects on acoustic voice parameters and self-reported vocal effort in individuals with voice disorders. *Journal of Voice*.