

APRIL 03 2024

Sex differences in vocal behavior in virtual rooms compared to real rooms

Georgios Papadimitriou; Jonas Brunskog ; Franz M. Heuchel; Viveka Lyberg Åhlander; Greta Öhlund Wistbacka



JASA Express Lett. 4, 045201 (2024)

<https://doi.org/10.1121/10.0025523>



View
Online



Export
Citation



ASA

Advance your science and career as a member of the
Acoustical Society of America

[LEARN MORE](#)

Sex differences in vocal behavior in virtual rooms compared to real rooms

Georgios Papadimitriou,¹ Jonas Brunskog,¹  Franz M. Heuchel,¹ Viveka Lyberg Åhlander,^{2,3} and Greta Öhlund Wistbacka^{2,4,a)}

¹Acoustic Technology Group, DTU Electro, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

²Department of Clinical Sciences, Logopedics, Phoniatrics and Audiology, Lund University, 22100 Lund, Sweden

³Faculty of Arts, Psychology, and Theology, Åbo Akademi University, 20500 Turku, Finland

⁴Speech Language Pathology, Department of Public Health and Caring Sciences, Uppsala University, 75237 Uppsala, Sweden

geopapadim94@gmail.com, jbru@dtu.dk, franz.heuchel@pm.me, viveka.lybergahlander@abo.fi, greta.ohlundwistbacka@abo.fi

Abstract: This study investigates speech production under various room acoustic conditions in virtual environments, by comparing vocal behavior and the subjective experience of speaking in four real rooms and their audio-visual virtual replicas. Sex differences were explored. Males and females ($N = 13$) adjusted their voice levels similarly to room acoustic changes in the real rooms, but only males did so in the virtual rooms. Females, however, rated the visual virtual environment as more realistic compared to males. This suggests a discrepancy between sexes regarding the experience of realism in a virtual environment and changes in objective behavioral measures such as voice level. © 2024 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

[Editor: Douglas D. O'Shaughnessy]

<https://doi.org/10.1121/10.0025523>

Received: 14 December 2023 Accepted: 21 March 2024 Published Online: 3 April 2024

1. Introduction

Voice and speech behavior are affected by several factors, one of them being the room acoustic conditions. Previous studies have pointed out poor room acoustic conditions as a possible risk factor for voice problems for people working within vocally demanding occupations.^{1–3} The link between room acoustics and vocal behavior can be explained through side-tone, which encompasses how a speaker hears his/her own voice while speaking.⁴ Studies aiming to disentangle the relationship between specific room acoustic parameters and vocal behavior, referred to as the *room effect*, show diverse results. The room effect seems to be sensitive to several confounding factors, such as the investigated speech task,⁵ the talker-listener distance,⁶ the vocal health of the participants,⁵ and background noise level.⁷ Two room acoustic parameters used for investigating the room effect are voice support (ST_v) and room gain (G_{RG}).^{5,8} Voice support measures the strength of the speaker's voice reflection as related to the strength of the direct sound between the mouth and the ears, and room gain measures the level of amplification that the room provides of the speaker's voice at his own ears.⁸

Studying voice production in varied room acoustic conditions can be challenging because, in real-world settings, it is difficult to isolate specific factors of interest while controlling for confounding variables. Virtual acoustic environments can aid this research by creating space- and time-efficient experiments in a laboratory setting, using real-time auralization provided either by headphones or a loudspeaker array. This has been done by, e.g., Pelegrín García and Brunskog⁵ and Rapp *et al.*,⁹ providing promising results on how to efficiently investigate room effects. A disadvantage of the laboratory setting is the laboratory environment in itself since the visual information also plays a fundamental part in regulating speech and voice behavior.^{6,8} Virtual reality (VR) using a head-mounted display (HMD) could aid this problem, by adding a simulated visual environment. In this way, both visual and auditory cues important for speech regulation can be controlled and changed accordingly in line with the research question of interest.

A laboratory setup for studying speech behavior in virtual environments, through acoustic and visual simulations, has recently been developed by our research team.^{10,11} Preliminary results investigating the room effect, defined as changes in vocal sound pressure level (SPL) with respect to ST_v , suggest a discrepancy in the room effect between males and females.¹¹ The results showed that males increased their vocal SPL when ST_v increased, to a magnitude of almost 0.70 dB vocal SPL dB/dB ST_v . For females, the increase was only 0.13 dB vocal SPL dB/dB ST_v . These results contradict previous research in two main aspects. First, in previous studies, vocal SPL has mainly decreased with increasing ST_v .^{5,8,9} Second, Rapp *et al.*,⁹ investigated sex differences in vocal SPL dB/dB ST_v , and found no differences between males and

^{a)} Author to whom correspondence should be addressed.

females. A major difference between our set-up and previous studies^{5,8,9} is the use of VR for visual simulation. The purpose of this study was therefore to explore the use of audio-visual virtual rooms as a method for investigating vocal SPL and fundamental frequency (f_0) as well as subjective perception of speaking under different room acoustic conditions. Another purpose was to investigate possible sex differences in room effects and talker experience. This was done by comparing the vocal output and the subjective perception of female and male talkers speaking under different acoustic conditions in real rooms and in their simulated acoustic and visual counterparts.

2. Methods

2.1 Participants

A total of 13 engineering students, seven males and six females, were recruited as participants through convenience sampling. The age spread was 22 to 32 years. All reported normal hearing, normal or corrected normal vision, and no occurrence of any vocal disorder. Three male and two female participants reported extensive video game usage, and the remaining seven reported playing video games occasionally, mostly in their teens. None had notable VR experience.

2.2 Experimental conditions

The experimental conditions consisted of four real rooms (an anechoic room, a lecture room, a reverberation chamber, and a corridor, henceforth *real scenario*) and their virtual replicas, henceforth *virtual scenario*, giving a total of eight conditions. The virtual scenario consisted of corresponding audio-visual simulations of the rooms presented in the anechoic room using VR for visual, and a 64-loudspeaker array for auditory simulations (see Secs. 2.4 and 2.6). For the anechoic room, the acoustic condition was the same for both the real and virtual scenarios, the only difference being the visual condition provided by the HMD. The order of the two scenarios, real and virtual, was randomized for each participant. The order of the four rooms was also randomized for each participant, although consistent across the two scenarios.

In the real scenario, the participants sat in an office chair wearing a DPA 4088 head-mounted microphone (DPA Microphones A/S, Kokkedal, Denmark). In the virtual scenario, the participants sat in a similar chair at the center of the loudspeaker array wearing the same microphone and an HMD (see Sec. 2.6). The test-leader or a human avatar, henceforth referred to as the audience, stood 4 meters in front of the participant in the real and virtual environments, respectively. The virtual distance has been rated to a median of 4 meters ($IQR = 2$ m) by 80 participants in a previous yet unpublished data collection (Öhlund Wistbacka). In both scenarios, the participants' speech were recorded by a measurement microphone (B&K 4192, Brüel & Kjær Sound & Vibration Measurement A/S, Virum, Denmark) positioned one meter away.

2.3 Experimental task

In each experiment, the participants were asked to speak freely about a topic of their own choice, for instance, hobbies, travels, or studies. The participants had different native languages, and in order to obtain as fluent speech as possible,¹² they were instructed to speak in their own mother tongue: English ($n = 1$; male), Greek ($n = 5$; two males, three females), French ($n = 1$; male), Chinese ($n = 2$; one male, one female), Danish ($n = 2$; one male, one female), Thai ($n = 1$; male), or German ($n = 1$; female). The instruction was to "speak in a way so that it feels like the audience would be able to hear you." No feedback was given to the participant during the speech task. After speaking for 3 min, the participants were asked to rate statements about their subjective experience of speaking in the present condition, on a scale from 1 (*do not agree at all*), and 10 (*completely agree*). The statements, given in English, were: (1) *The room helps me to speak comfortably*; (2) *It is easy for me to make myself heard in the room*; (3) *My voice reverberates in the room as I speak*; (4) *The room feels acoustically dry or damped*; (5) *I succeeded well with the task I was asked to perform*; (6) *I needed to make an effort to perform as I did*; (7) *The task made me feel insecure, irritated or stressed*; (8) *The task was fun*; (9) *The sound environment feels realistic*; (10) *The visual environment feels realistic*. Statements 9 and 10 were asked solely in the virtual rooms as they only refer to the acoustic and visual simulations. In addition, the participants were asked to estimate their distance to the audience in meters for each room in both scenarios.

2.4 Room acoustic simulations

For the acoustic simulations, the room impulse responses (RIR) were first measured in the real rooms (classroom, reverberation chamber and corridor), in accordance with ISO 3382-2,¹³ using the Dirac software (Dirac Research AB, Uppsala, Sweden), a B&K 4292 loudspeaker, and a B&K 4192 microphone (Brüel & Kjær Sound & Vibration Measurement A/S, Virum, Denmark). Room dimensions were measured with a laser meter.

Computer-based room models were created using SketchUp,¹⁴ and imported to the ODEON Room Acoustics Software.¹⁵ The simulations were calibrated using the measured RIRs and the Genetic Material Optimizer.¹⁵ Speaker and listener positions were inserted according to the real room positions. The sound source approximated human voice directivity (ODEON's 'BB93NormalNaturalSO8'). The receiver was an omnidirectional microphone placed 1 m away from the source. Early reflections and directional energy curves were extracted from the simulated room acoustic model, using 5000 late rays, a maximum reflection order of 2000, early reflections' transition order 3, and an impulse response length of

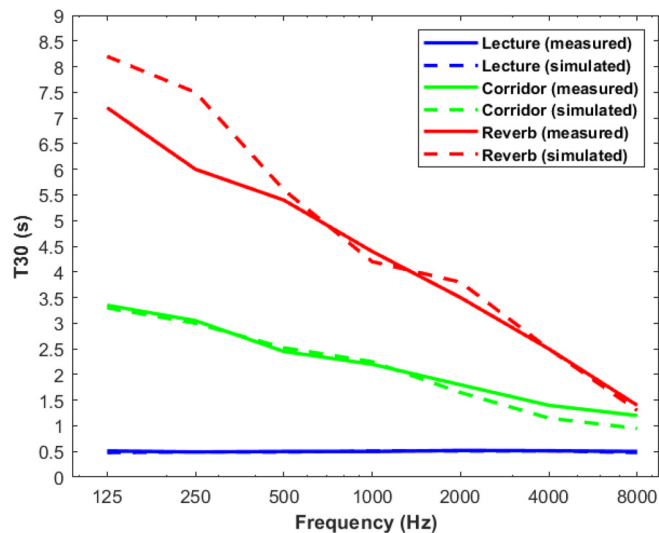


Fig. 1. T_{30} measured and tuned simulations in Odeon (before including the audience materials). Note that the JND is 5%, see Sec. 2.5, being fulfilled in all cases except for the lowest frequencies for the reverberation chamber.

15 000 ms. Figure 1 shows a comparison between the tuned simulated and the measured reverberation times (T_{30}) for the rooms. We notice a discrepancy between the real room and the simulated room at low frequencies in the reverberation chamber.

The early reflections and directional energy curves were processed by the LoRA toolbox¹⁶ to generate 64 FIR filters for reproducing the simulated acoustic environment at the center of a spherical 64-channel loudspeaker array in the anechoic room. Details regarding the laboratory facility can, e.g., be found in two studies by Ahrens *et al.*^{17,18} The input to the reproduction system was the participant’s voice, picked up the head-mounted microphone. Details regarding the real-time auralization including a calibration process for each participant is described in a previous paper.¹⁰

2.5 Effective room acoustic parameters

The effective room acoustic parameters for each room and scenario are found in Table 1. In both the real and virtual scenarios, the acoustic conditions as quantified by room acoustic parameters and perceived at the location of the participants were calculated from RIRs between the loudspeaker-mouth and microphone-ears of a head-and-torso simulator (HATS 4128; HBK Sound & Vibration Measurements A/S; Virum, Denmark). The HATS was positioned at the participant’s location and the measurements were conducted under the same conditions as during the experiment. The RIRs were estimated from averages over 30 exponential sweeps of 240 s, using a frequency range of 1 Hz–20 kHz at a sample rate of 48 kHz, as to achieve a good signal-to-noise ratio. For the measurements in the virtual scenarios, HATS was wearing the HMD, as the HMD can influence the head-related transfer function,¹⁹ sound localization,¹⁸ and sidetone.²⁰ The personalized reproduction filters for HATS were created in a similar manner as for a human participant.¹⁰

The room acoustic parameters of interest were reverberation time T_{30} , room gain G_{RG} , voice support ST_V , and the mouth-to-ear decay time DT_{40} . For the computation of T_{30} , the first 5 ms were removed from the impulse responses to avoid the influence of the direct sound, in accordance with Pelegrin-Garcia and Brunskog.⁵ The just noticeable difference (JND) of T_{30} is 5%, as established by ISO3382-1:2009.²¹ The differences between the real and

Table 1. Room acoustic parameters measured with the ear-canal microphones of a head-and-torso simulator in the real and virtual scenarios. The acoustic conditions are identical in the anechoic room for both scenarios. T_{30} , reverberation time; G_{RG} , room gain; ST_V , voice support.

Rooms	Volume [m ³]	T_{30} [s]	G_{RG} [dB]	ST_V [dB]
Anechoic Room	336	0.20	0.08	−16.89
Corridor (Real)	450	2.00	1.10	−5.54
Corridor (Virtual)		1.90	0.80	−7.10
Lecture Room (Real)	186	0.60	0.20	−13.06
Lecture Room (Virtual)		0.58	0.17	−13.95
Reverb. Chamber (Real)	240	2.78	1.60	−3.50
Reverb. Chamber (Virtual)		2.60	1.00	−5.93

virtual measurements for the lecture room and the corridor were within the 5% range. In the reverberation chamber, the T_{30} in the virtual room measurement was 2.60 s and in the real measurement 2.78 s, giving a difference between measurements exceeding the 5% JND with 0.04 s. The room gain is defined by $G_{RG} = 10 \log [(E_D + E_R)/E_D]$ dB, where E_D is the airborne direct sound energy and E_R is the reflected sound energy.⁸ Voice support ST_V is the energy ratio in dB between the reflected and direct airborne sound defined by $ST_V = 10 \log E_R/E_D$ dB. The direct and reflected sounds were extracted from the averaged energy responses between the loudspeaker-mouth and the two microphone-ear canals.^{8,22} Furthermore, a speech weighting in the frequency octave bands 125–4000 Hz for ST_V was applied, in accordance with Pelegrín García *et al.*,²³ in order to more accurately adapt the parameter to auditory perception. The mouth-to-ear decay time DT_{40} is the 60 dB reverberation time extrapolated from the energy decay over the first 40 dB drop. Despite previously being useful,⁵ it turned out strongly correlated with T_{30} in the current measurements and was therefore dropped from further analysis.

2.6 VR simulation

VR versions of all four real rooms were built using Unity (Unity Technologies, San Francisco, CA) to simulate the visual environments. The corridor and reverberation chamber models were empty, windowless rooms, as they were in the real versions. The anechoic chamber was also modeled as empty, hence the model did not show the 64-loudspeaker array present in the real room. The classroom model was furnished in the same way as the real room including windows. The display was presented *via* HTC VIVE Pro Virtual Reality head mounted display (HTC Corporation, New Taipei City, Taiwan) with the use of the SteamVR plugin (Valve Corporation, Bellevue, WA). The test participant and audience were positioned similarly to the talker and listener in the real room experiment. The audience avatar was chosen from a package called “realpeople.male” (3drt.com). To avoid drop-outs in the audio loop, Unity and the real-time signal convolution were run on separate computers.¹⁰

2.7 Processing of voice recordings

Voice SPL and fundamental frequency (f_o) were extracted from the voice recordings by averaging the 3-min recordings measured in each room for each participant. The transfer function between the head-mounted microphone and the measurement microphone 1 m away was used to normalize the SPL to a distance of 1 m. The signals were further processed to exclude speech pauses. For each signal, a short-time Fourier transform with 43 ms ($N = 2048$) Hann-windowed blocks with 25% overlap was computed, and each block was classified as voiced or unvoiced according to a spectral flatness criterion of its power spectrum S : voiced blocks have a spectral flatness,

$$\frac{\exp\left(\frac{1}{N} \sum_{n=0}^{N-1} \ln \max\left(S_n^2, \frac{1}{N} \sigma^2\right)\right)}{\frac{1}{N} \sum_{n=0}^{N-1} \max\left(S_n^2, \frac{1}{N} \sigma^2\right)} < 0.05. \quad (1)$$

Other blocks were considered unvoiced and excluded when computing the equivalent voice level of the speech signal. The flatness measure was robustified by lower bounding the power spectral density with a background noise variance estimate σ^2 . The mean f_o was extracted automatically using the software Praat.²⁴

2.8 Statistical analysis

Data were analyzed using linear mixed-effect models. The analyses were performed in R²⁵ version 4.3.0 with the *lme4*²⁶ and *MuMin* packages.²⁷ The general mixed linear models are described by the expression $y = X\beta + Z\alpha + \epsilon$, where y is a vector of observations of the outcome variable, X is a matrix of the predictor variables (fixed effects), β is a vector of unknown regression coefficients of the fixed effects, Z is the design matrix for the random effects, α is a vector of random effects, and ϵ is a column vector of errors.²⁸ The outcome variables, y , were the objective measures of SPL and f_o , and the subjective responses to each of the ten statements. For each of the 12 y , three models were tested, one for each room acoustic parameter as a fixed effect (ST_V , G_{RG} , or T_{30} ; continuous variables). For outcome variables SPL and f_o , the additional fixed effects X were *Distance* (the subjective estimation of distance to the audience in meters, continuous variable), *Sex* (male/female, categorical variable), and *Scenario* (real/virtual, categorical variable). For statements 1–8, the additional fixed effects were *Sex* and *Scenario*. For statements 9–10, which were solely answered in the virtual scenario, the only additional fixed effect was *Sex*. All possible one- and two-way interactions of the fixed effects were included in each model. *Participants* were always included as the random effect Z . A top-down approach was implemented when building the models, starting with the full models and removing the non-significant predictors one at a time. This was done by evaluating the analysis of variance (ANOVA) (significance level $p < 0.1$) and comparing the Akaike information criterion (AIC). The lower the AIC score, the better the fit of the model.²⁹

3. Results

3.1 Perceived distance to the audience

Males rated the perceived distance to the audience to be between 2 and 5.5 meters in the real scenario rooms, with a median matching the real distance of 4 meters. Females rated the distance to be between 2 and 5 meters in the real rooms, also with a median of 4 meters. In the virtual scenario, the perceived distance for males was between 1 and 4 meters, with a median of 2.5 meters. For females, the perceived distance varied between 1.5 and 3 meters, with a median of 2.8 meters.

3.2 Voice SPL and fundamental frequency (f_o) for males and females in the real and virtual scenario

The SPL and f_o for male and female participants for each scenario and room are shown in Fig. 2. In the real rooms, the median SPLs differed by up to 3 dB between sexes. The relative changes in SPL between the real rooms were similar for males and females. When comparing SPL in the real rooms to the virtual ones, male participants presented 1–2 dB higher SPL median values in all virtual rooms except for the anechoic one in which the median SPL was 1 dB lower compared to the real room. The relative changes in SPL between the four different rooms followed the same pattern in the virtual scenario as in the real scenario for males. Females, on the other hand, lowered their median SPL in the virtual scenario and did not change their SPL as much between the four virtual scenario rooms, compared to the four real scenario rooms.

For males, a slight increase in f_o was found in the virtual anechoic room compared to the real counterpart, and a drop in f_o in the virtual corridor compared to the real one. In the lecture room and reverberation chamber, the f_o was kept rather steady regardless of room or scenario. For the female participants, the f_o was kept steady in the anechoic room regardless of the scenario. In the other rooms, the f_o dropped 5–10 Hz in the virtual scenario compared to the real one, however, the relative F0 changes between the three rooms were similar between scenarios.

Based on the AIC, the best-fitted statistical model for SPL as an outcome variable included the significant ($p < 0.05$) fixed effects ST_V , Sex, and Scenario as well as the two-way interactions $Distance \times Scenario$ and $Sex \times Scenario$. The fixed effect explaining most of the variance of SPL was the ST_V , giving a room effect vocal SPL dB/dB ST_V of 0.27 dB/dB. The effect of perceived distance on SPL was +1.4 dB SPL higher per +1 meter distance in the virtual scenario compared to the real one ($Distance \times Scenario$). Both males and females decreased their SPL in the virtual scenario compared to the real one, but males did so to a lesser extent ($Sex \times Scenario$). The best fitted model for f_o included the fixed effects T_{30} , Sex, Scenario, and the two-way interaction $T_{30} \times Sex$. The results showed a decrease in f_o for females when T_{30} increased (-6.9 Hz/s), and a slight increase in f_o for males when T_{30} increased ($-6.9 + 7.3 = 0.4$ Hz/s). f_o decreased by approximately 9 Hz in the virtual scenario compared to the real one for both sexes. The full results from these models are presented in Table 2. All six models tested are presented in the supplementary material.

3.3 Subjective results

All results from the mixed models of the ten statements are found in the supplementary material. For statements 1–4 and 6–8, the best fitted models all included T_{30} , which also was the fixed effect explaining most of the variance. Additional statistically significant fixed effects were Scenario (statements 1, 3, 4, and 7), Sex (statements 2 and 10), and the two-way interaction $T_{30} \times Scenario$ (statements 1–4, 6 and 8). No fixed effects were statistically significant for statements 5 (*I succeeded well with the task I was asked to perform*) and 9 (*The sound environment feels realistic*), in either model.

In statement 1 (*The room helps me to speak comfortably*), the room was rated as more comfortable to speak in when T_{30} decreased, to a magnitude of $-14.0/T_{30}$ in the real scenario and $-5.5/T_{30}$ in the virtual scenario. A similar effect was found for statement 2, (*It is easy for me to make myself heard in the room*), to a magnitude of $-7.2/T_{30}$ in the real scenario and $-2.4/T_{30}$ in the virtual scenario. In this statement, an effect of Sex was also found, showing that males rated this statement on average 1 point lower than females. Statement 3, (*My voice reverberates in the room as I speak*), was

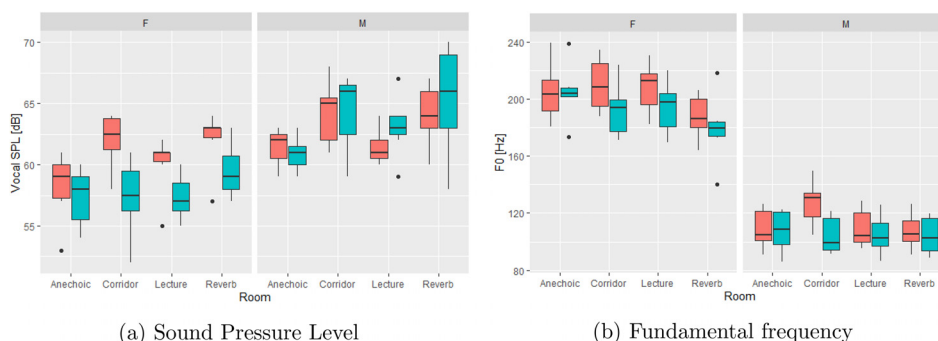


Fig. 2. Medians and interquartile ranges of voice SPL (left) and fundamental frequency (f_o , right) in the different rooms in the real (in red) and virtual (in blue) scenarios for female (F) and male (M) participants.

Table 2. Summary results of the mixed-model ANOVA for the best-fitted models for voice SPL and fundamental frequency (f_o) based on the lowest AIC. The effect size is presented as the conditional R^2 . ST_V , voice support [dB]; T_{30} , reverberation time [s]; M, male sex; V, virtual scenario; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; NS, not significant.

	Fixed effects	Estimate	F statistic	p
Effect of ST_V on SPL AIC = 456.6 conditional $R^2 = 0.76$	Intercept	65.5		
	ST_V	0.27	$F(1, 88.7) = 52.5$	<0.001***
	Distance	-0.61	$F(1, 94.3) = 0.04$	NS
	Sex (M)	2.21	$F(1, 10.9) = 11.4$	0.006**
	Scenario(V)	-6.56	$F(1, 87.9) = 9.5$	0.003**
	DistanceXScenario(V)	1.37	$F(1, 88.6) = 7.6$	0.007**
Effect of T_{30} on f_o AIC = 807.7 conditional $R^2 = 0.95$	Sex (M)Xscenario(V)	3.23	$F(1, 86.7) = 19.9$	<0.001***
	Intercept	211.59		
	T_{30}	-6.85	$F(1, 88) = 10.1$	0.002**
	Sex(M)	-98.21	$F(1, 13.7) = 133.3$	<0.001***
	Scenario(V)	-9.01	$F(1, 88) = 19.5$	<0.001***
	T_{30} XSex(M)	7.29	$F(1, 88) = 12.9$	<0.001***

rated higher with increasing T_{30} , to a magnitude of $18.9/T_{30}$ in the real scenario, and $8.5/T_{30}$ in the virtual scenario. The opposite effect was found in the ratings of statement 4, (*The room feels acoustically dry or damped*), showing a lower rating with an increasing T_{30} , to a magnitude of $-18.1/T_{30}$ in the real scenario, and $-8.9/T_{30}$ in the virtual scenario. For statement 6 (*I needed to make an effort to perform as I did*), the participants rated their performance as more effortful when T_{30} increased, to a magnitude of $4.4/T_{30}$ in the real scenario, and $1.5/T_{30}$ in the virtual scenario. For statement 7 (*The task made me feel insecure, irritated or stressed*), the participants rated their feelings more negatively affected when T_{30} increased ($1.3/T_{30}$). The participants rated their feelings as on average 2.6 points more negatively affected in the real scenario, compared to the virtual one. Statement 8, (*The task was fun*) was rated lower when the T_{30} increased, to a magnitude of $-3.5/T_{30}$ in the real scenario and $-0.8/T_{30}$ in the virtual scenario. In statement 10 (*The visual environment feels realistic*), females rated the visual environment in the virtual scenario as on average 2.6 points more realistic than males did.

4. Discussion

The purpose of this study was to investigate differences in vocal behavior (SPL and f_o) and the subjective experience of speaking in four real rooms with varying room acoustic conditions, compared to audio-visual virtual replicas of the same rooms. An additional purpose was to investigate possible sex differences. The results showed that males adjusted their SPL to changes in ST_V in the same manner in both the real and virtual scenarios. Females, on the other hand, adjusted their SPLs less to changes in ST_V in the virtual scenario, compared to the real one. Females tended to lower their f_o as an effect of increased T_{30} in both scenarios, whereas males slightly increased their f_o when T_{30} increased. The subjective ratings were mostly affected by T_{30} in interaction with *Scenario*, and to some extent by *Sex*.

The results from the SPL analyses are in line with some reported results,¹¹ but differ from others.^{5,6,9} The increase in SPL when the ST_V increases is contradictory to most previous research, in which a negative slope between the two parameters has been found.⁵⁻⁹ One explanation for the different results could be the chosen speech task, as that has been shown to affect the room effect in previous studies.⁶

Few studies have systematically investigated differences between sexes in vocal behavior in relation to room acoustics. Rapp *et al.*⁹ did so but found no sex differences in voice SPL changes in relation to changes in ST_V . Their study was performed using headphones for real-time auralization in an anechoic room.⁹ In the present study, the room effect was similar between sexes in the real scenario but differed in the virtual one. This discrepancy between scenarios could perhaps be explained by possible sex differences regarding how the user perceives and accepts him/herself as actually “being there” in a VR-simulated situation, despite knowing it to be an illusion; i.e., the concept of “Presence.”³⁰ The sense of presence has been evaluated through self-report or by measuring physical responses such as heart rates,³⁰ and some sex differences in self-report measures have been previously reported.³¹ Vocal behavioral measures have, to the best of our knowledge, not been used as a physical response measure for investigating presence. The results from the present study show some potential for using voice measurements, particularly SPL, as physical presence markers, especially when considering the discrepancy between the SPL adjustments and the ratings of the subjective statements. Despite adjusting SPL less than males in the VR scenario, females still rated the visual realism of the virtual scenario higher than males. With the exception of this statement, and the statement regarding the sense of easiness to be heard, females and males rated the real and virtual scenarios in a similar manner. An effect of the scenario was found in the ratings of most statements, presented as higher rating scores in the real scenario compared to the virtual one. The direction of the slopes of the

regression lines were however the same regardless of scenario in all statements, suggesting similar but less extreme experiences in the virtual scenario compared to the real one.

In order to confirm these results, the study should be replicated with a larger sample size. The distance to the audience was rated as four meters in the real scenario but under three meters in the virtual one. The expected outcome of a shorter distance to the listener would be a decrease in SPL,⁶ which is seen for the female participants. It is thus possible that the discrepancy in distance ratings has impacted the female results. The seven different native languages used might have affected the results, as vocal behavior has been shown to vary between languages (see, for example, Van Bezooijen³²). However, as language is a within-subject factor, it is taken into account when including *Participant* as a random factor in the statistical model. Including participants with different native languages also makes the findings more general across languages.

5. Conclusions

This study suggests that using virtual audio-visual rooms, integrating real-time auralization and a VR HMD, is a valid setup for studying how males' vocal behavior relates to room acoustics. However, for females, the VR environment appears to decrease the accuracy of their vocal response to acoustic changes compared to the real environment. Subjective perceptions differed slightly between real and virtual scenarios for both sexes, mostly aligning with room acoustic changes.

Supplementary Material

See the supplementary material for complete mixed model results of the six models tested for vocal sound pressure level and fundamental frequency as outcome variables and for complete mixed model results of the subjective statements as outcome variables.

Acknowledgment

This research was supported by the Swedish Research Council for Health, Working Life and Welfare (FORTE) for the project "Communication, a challenge for the older employee: Need for a communication supporting workplace," project ID 2019-01329, for authors JBR, VLÅ and GÖW, and Gösta Branders research fund, Åbo Akademi Research Foundation, for author GÖW.

Author Declarations

Conflict of Interest

The authors have no conflicting interests to report.

Ethics Approval

Ethical approval was obtained from the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). All participants confirmed informed consent in writing before taking part in the study.

Data Availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to ethical restrictions.

References

- ¹V. L. Åhlander, R. Rydell, and A. Löfqvist, "Speaker's comfort in teaching environments: Voice problems in Swedish teaching staff," *J. Voice* **25**(4), 430–440 (2011).
- ²V. L. Åhlander, D. P. Garcia, S. Whiting, R. Rydell, and A. Löfqvist, "Teachers' voice use in teaching environments: A field study using ambulatory phonation monitor," *J. Voice* **28**(6), 841.E5–841.E15 (2014).
- ³C. J. Nudelman, P. Bottalico, and L. C. Cantor-Cutiva, "The effects of room acoustics on self-reported vocal fatigue: A systematic review," *J. Voice* (published online 2023).
- ⁴C. Pörschmann, "Influences of bone conduction and air conduction on the sound of one's own voice," *Acta Acust. united Ac.* **86**(6), 1038–1045 (2000).
- ⁵D. Pelegrín García and J. Brunskog, "Speakers' comfort and voice level variation in classrooms: Laboratory research," *J. Acoust. Soc. Am.* **132**(1), 249–260 (2012).
- ⁶D. Pelegrín García, B. Smits, J. Brunskog, and C.-H. Jeong, "Vocal effort with changing talker-to-listener distance in different acoustic environments," *J. Acoust. Soc. Am.* **129**(4), 1981–1990 (2011).
- ⁷M. Cipriano, A. Astolfi, and D. Pelegrín-García, "Combined effect of noise and room acoustics on vocal effort in simulated classrooms," *J. Acoust. Soc. Am.* **141**(1), EL51–EL56 (2017).
- ⁸J. Brunskog, A. C. Gade, G. P. Bellester, and C. L. Reig, "Increase in voice level and speaker comfort in lecture rooms," *J. Acoust. Soc. Am.* **125**(4), 2072–2082 (2009).
- ⁹M. Rapp, D. Cabrera, and M. Yadav, "Effect of voice support level and spectrum on conversational speech," *J. Acoust. Soc. Am.* **150**(4), 2635–2646 (2021).

- ¹⁰G. Öhlund Wistbacka, F. Heuchel, V. Lyberg Åhlander, J. Mårtensson, B. Sahlén, and J. Brunskog, “Vocal comfort in simulated room acoustic environments - experiment set-up and protocol development,” in *Proceedings of the 24th International Congress on Acoustics*, Gyeongju, Korea (October 24–28, 2022), pp. 1–12.
- ¹¹G. Öhlund Wistbacka, F. Heuchel, V. Lyberg-Ahlander, B. Sahlén, R. Rydell, J. Mårtensson, and J. Brunskog, “Voice levels in simulated room acoustic environments. Sex and age differences,” in *Proceedings of the 10th Convention of the European Acoustics Association, Forum Acusticum*, Torino, Italy (September 11–15, 2023).
- ¹²E. Guz, “Establishing the fluency gap between native and non-native speech,” *Res. Lang.* **13**(3), 230–247 (2015).
- ¹³ISO3382-2: “Acoustics—Measurement of room acoustic parameters—Part 2: Reverberation time in ordinary rooms” (ISO, Geneva, Switzerland, 2008).
- ¹⁴T. Inc, *SketchUp Pro Desktop*, my.sketchup.com (Last viewed March 5, 2023).
- ¹⁵Odeon A/S, *ODEON Room Acoustics Software User’s Manual* (Odeon, Lygby, Denmark, 2021).
- ¹⁶S. E. Favrot and J. Buchholz, “Lora: A loudspeaker-based room auralization system,” *Acta Acust. united Ac.* **96**(2), 364–375 (2010).
- ¹⁷A. Ahrens, M. Marschall, and T. Dau, “Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments,” *Hear. Res.* **377**, 307–317 (2019).
- ¹⁸A. Ahrens, K. D. Lund, M. Marschall, and T. Dau, “Sound source localization with varying amount of visual information in virtual reality,” *PLoS One* **14**(3), e0214603 (2019).
- ¹⁹R. Gupta, R. Ranjan, i He, and G. Woon-Seng, “Investigation of effect of VR/AR headgear on head related transfer functions for natural listening,” in *Proceedings of the Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*, Redmond, WA (August 20–22, 2018).
- ²⁰G. Öhlund Wistbacka, W. Shen, and J. Brunskog, “Virtual reality head-mounted displays affect sidetone perception,” *JASA Express Lett.* **2**(10), 105202 (2022).
- ²¹ISO3382-1: “Acoustics—Measurement of room acoustic parameters—Part 1: Performance spaces” (ISO, Geneva, Switzerland, 2009).
- ²²D. Pellegrin-Garcia, “Comment on ‘Increase in voice level and speaker comfort in lecture rooms’ [J. Acoust. Soc. Am. **125**, 2072–2082 (2009)] (L),” *J. Acoust. Soc. Am.* **129**, 1161–1164 (2011).
- ²³D. Pelegrín García, J. Brunskog, V. Lyberg-Ahlander, and A. Lofqvist, “Measurement and prediction of voice support and room gain in school classrooms,” *J. Acoust. Soc. Am.* **131**(1), 194–204 (2012).
- ²⁴P. Boersma and W. David, “Praat: Doing phonetics by computer (version 6.2.01) [computer program],” <http://www.praat.org> (Last viewed June 27, 2023).
- ²⁵R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2022).
- ²⁶D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4,” *J. Stat. Softw.* **67**(1), 1–48 (2015).
- ²⁷K. Bartón, “MuMIn: multi-model inference. Rpackage version 1.47.5” (2023), <https://CRAN.R-project.org/package=MuMIn> (Last viewed September 28, 2023).
- ²⁸J. Jiang and T. Nguyen, *Linear and Generalized Linear Mixed Models and Their Applications* (Springer, New York, 2007).
- ²⁹H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Automat. Contr.* **19**(6), 716–723 (1974).
- ³⁰M. Slater, D. Banakou, A. Beacco, J. Gallego, F. Macia-Varela, and R. Oliva, “A separate reality: An update on place illusion and plausibility in virtual reality,” *Front. Virtual Real.* **3**, 914392 (2022).
- ³¹A. Felnhöfer, O. D. Kothgassner, L. Beutl, H. Hlavacs, and I. Kryspin-Exner, “Is virtual reality made for men only? Exploring gender differences in the sense of presence,” in *Proceedings of the International Society on Presence Research*, Philadelphia, PA (October 24–26, 2012), pp. 103–112.
- ³²R. Van Bezooijen, “Sociocultural aspects of pitch differences between Japanese and Dutch women,” *Lang. Speech* **38**(3), 253–265 (1995).