ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

Vocal comfort in simulated room acoustic environments – experimental set-up and protocol development

Greta ÖHLUND WISTBACKA(1,*), Franz HEUCHEL(1), Viveka LYBERG ÅHLANDER(2,3), Johan MÅRTENSSON(3), Birgitta SAHLÉN(3), Jonas BRUNSKOG(1)

(1)Acoustic Technology Group, Technical University of Denmark, Kongens Lyngby DK-2800, Denmark
(2)Faculty of Art, Psychology and Theology, Åbo Akademi University, FI-20100 Åbo, Finland
(3)Department of Logopedics, Phoniatrics and Audiology, Lund University, SE-221 85 Lund, Sweden
(*)Corresponding author. Email: gmawi@elektro.dtu.dk

ABSTRACT

Voice problems are common among workers in occupations with high demands on oral communication. For sustaining vocal health, it is important to find efficient ways for how to reduce strain on the voice at work, and one way to do so is optimizing the room acoustics for communication. However, to date we do not know enough about what kind of room acoustic conditions that support voice and speech most efficiently. The purpose of this project is to investigate how different acoustic environments affect voice and speech in talkers. To do so, an experimental set-up has been developed for acoustic as well as visual simulation of different room conditions, using real-time auralization of a speaker's own voice in a 64-loudspeaker array in an anechoic room combined with virtual reality. Room conditions vary in room size, reverberation time and the speaker-oriented room acoustic parameters voice support and room gain. These parameters quantify the own-voice amplification provided by room reflections to the talker's own ears. In this paper, the experimental set-up and study protocol development will be presented and discussed.

Keywords: Room acoustics, Voice support, Virtual reality

1 INTRODUCTION

People working within vocally demanding occupations are at higher risk of developing vocal problems compared to workers within less vocally demanding fields [1, 2]. In order to minimize the risk for occupational voice disorders, it is important that the work environment supports oral communication efficiently. Room acoustics has been identified as one factor that has a direct influence on vocal behavior [3, 4]. This is due to its impact on the reflective part of the auditory feedback of one's own voice, the sidetone, hence the airborne sound reflections between the talker's mouth and own ears [5]. These sound reflections contain the information about the room acoustic conditions in which the person is speaking.

During the last decade, there has been a rising interest within the scientific community for investigating the impact of room acoustics on vocal behavior. The acoustic properties of real rooms have been measured and related to for instance vocal loudness adjustments [4, 6, 7, 8] and *speakers' comfort*, defined as the subjective experience of being heard by the listener efficiently and without a sensation of vocal effort, while speaking in a room [4, 6, 7]. As a way to investigate the individual role of separate acoustic parameters, laboratory studies have been conducted using simulated acoustic environments. In these studies, the reverberant field of a room model is simulated, applied to the participant's own voice and played back to the participant in real-time using either headphones [9, 3, 10, 11] or a loudspeaker array [3]. Results so far have shown on a negative linear relationship between vocal loudness and the room acoustic parameter voice support (ST_V, see Section 2.7) [9, 3], stating that when voice support decreases, the vocal loudness level tend to increase. However, the influence of room acoustics on vocal loudness adjustments seems to be strongly dependent on speech task[3], and to some extent also the speakers' gender [9]. Despite the recent progress in this field, much is still unknown regarding the complex relationship between room acoustics and the talker's vocal and speech behavior.

The laboratory environments enable for simulating specific acoustic conditions while keeping the visual environment constant. This brings a well-controlled acoustic environment, but with a rather unrealistic visual environment. Studies conducted in real rooms have a better ecological validity, but are challenging since they also bring less control over the experimental environment, and individual acoustic parameters, such as reverberation time and voice support, cannot be easily manipulated. Immersive 3D virtual reality (VR) using head-mounted displays (HMD), could enable for a more realistic visual environment while keeping the participant in a well-controlled acoustic laboratory setting. The purpose of this paper is to describe the development of an experimental set-up combining a simple VR environment with room acoustic simulations (Section 2), and to report some results from a pilot experiment using this combined set-up to investigate voice and speech behavior in different room environments (Section 3).

2 EXPERIMENTAL SETUP

The room acoustic simulation setup and study protocol is based on previous work by Pelegrin-Garcia [12] and Pelegrín-García and Brunskog [3]. Their original setup has been reimplemented with some changes for DTU's Audio-Visual Immersion Lab (AVIL), an anechoic room with a spherical loudspeaker array for sound field reproduction. The virtual acoustic environment is produced via real-time auralization of the reverberant sound field of the participant's voice, which is picked up by a head-mounted microphone. A corresponding visual virtual environment is presented via a virtual reality headset. The real-time auralization set-up consists of four steps: 1) room acoustic simulation, 2) computing convolution filters, 3) calibration of filters, and 4) real-time reproduction and data-acquisition.

2.1 Laboratory facility

AVIL consists of 64 loudspeakers (KEF LS50, KEF Audio, Maidstone, UK) placed in a spherical array in an anechoic room (7 m × 8 m × 6 m), see Figure 1. The loudspeakers are positioned in seven circles around the center of the room, where also a chair for the participant is placed. The circle angles are at ±80°, ±56°, ±28° and 0° in relation to the participants head, with 2, 6, 12 and 24 loudspeakers uniformly placed on each respective circle. The radius of the spherical array is 2.4 m. The participants' voice signal is picked up and recorded using a head-mounted microphone (DPA-4088, DPA Microphones A/S, Denmark) fed into a PC over an audio interface (Tesira SERVER-IO, Biamp Systems, Beaverton, OR), processed with MAX 8 (Cycling '74, Walnut, CA) and played back over the loudspeakers via a set of amplifiers (Sonible d:24, sonible GmbH, Graz, Austria). This laboratory has been used in previous research on, e.g., sound localization [13] and speech intelligibility in virtual sound environments [14].

2.2 Real-time convolution system

The voice signal is convolved in real-time via zero-latency convolution [15] with a set of 64 FIR filters (see secs. 2.4 and 2.5) using the Hisstools toolbox [16]. The signal is then played back to the participant through the reproduction system, creating the sensation of talking in the simulated room. The total round-trip time of the reproduction system is around 22 ms, which is made up by the propagation time through air, the audio-interface's I/O buffers, Max 8's I/O buffers, and a look-ahead loudspeaker-protection limiter in the audio interface. The sampling rate is 48 kHz. To allow for interruption free audio processing, the 64 convolution processes aree running in parallel on four CPU cores via the 'mc.poly~' object.

2.3 Room acoustic simulation

The room acoustic models are created in accordance with Pelegrin-Garcia [12]. First, a computer-based geometrical room model is built using a CAD program, defining the room size and shape. Then, the model is imported into a geometrical acoustic simulation software, in our case Odeon [Odeon A/S; Kongens Lyngby, Denmark] in which the surface properties are set to achieve a diverse range of reverberation times. A sound source with a directivity pattern similar to average human speech (In Odeon: BB93NormalNaturalSO8) is placed at the

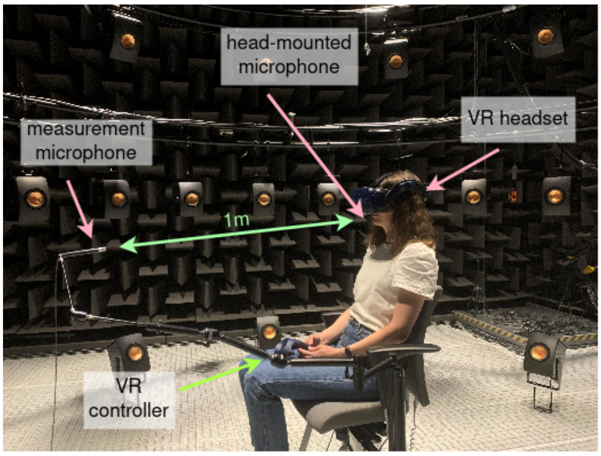
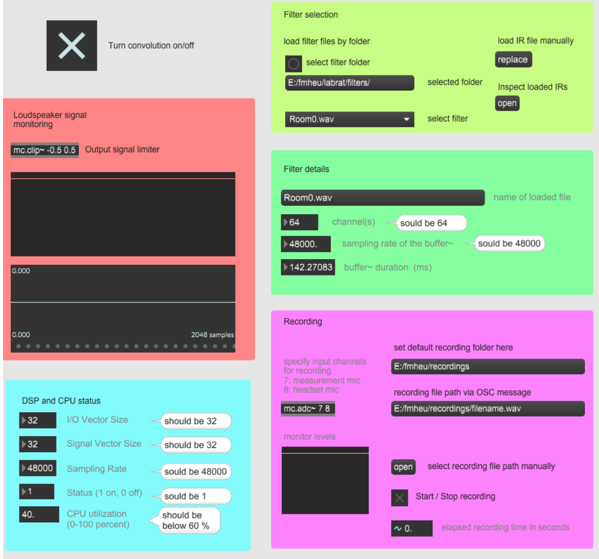


Figure 1. Experimental setup: voice is picked up by a head-mounted microphone while the measurement microphone at 1 m distance is used during calibration for estimating $H_{\text{direct,repr}}$.



Turn convolution on/off

Loudspeaker signal monitoring

mc.clip~ -0.5 0.5 Output signal limiter

0.000
0.000 2048 samples

Filter selection

load filter files by folder
select filter folder
E:/fmheu/labrat/filters/ selected folder
Room0.wav select filter
load IR file manually
replace
Inspect loaded IRs
open

Filter details

Room0.wav name of loaded file
64 channel(s) sould be 64
48000. sampling rate of the buffer~ sould be 48000
142.27083 buffer~ duration (ms)

Recording

specify input channels for recording
7: measurement mic
8: headset mic
mc.adc~ 7 8
set default recording folder here
E:/fmheu/recordings
recording file path via OSC message
E:/fmheu/recordings/filename.wav
monitor levels
open select recording file path manually
Start / Stop recording
0. elapsed recording time in seconds

DSP and CPU status

32 I/O Vector Size should be 32
32 Signal Vector Size should be 32
48000 Sampling Rate sould be 48000
1 Status (1 on, 0 off) sould be 1
40. CPU utilization (0-100 percent) should be below 60 %

Figure 2. Graphical interface in MAX 8 for monitoring and controlling real-time convolution, VR environment, calibration and voice recording.

talker's position. This position is located toward the audience and far enough from the surrounding surfaces to allow for a satisfactory reproduction when taking the system latency into consideration. A receiver is placed 1 m in front of the source, which should give a reflection pattern reasonably similar to the reflection pattern experienced at the ears, and allowing for calibration (see sec. 2.5) [12, 3]. The simulation parameters used were 5000 late rays, a maximum reflection order of 2000, a transition order of 3 for the early reflections and an impulse response length of 10000 ms.

2.4 Computing reproduction filters

For the playback via the reproduction system, the room-acoustic simulations between the source and receiver are exported via a list of early reflections and directional energy curves for the late reverberation in the 63 Hz to 8 kHz bands. The direct sound component is removed, as only the reflections are to be reproduced. To compensate for the system latency of 22 ms, each early reflection in the first 22 ms (from surfaces closer than 3.8 m) is moved by 22 ms back in time. The data is converted into 64 finite impulse-response filters via the LoRA toolbox [17, 18], which renders the early reflections using the nearest-loudspeaker method (this is a modification as compared to [17] where higher-order ambisonics were used) and the late reflections using first-order ambisonics; the latter is constructed by the envelopes of the room impulse responses multiplied with uncorrelated noise.

2.5 Calibration of reproduction filters

The FIR filters so obtained need to be calibrated in order to match the relationship of direct-to-reverberant sound field in the simulated and reproduced environments and to compensate for the location of the head-worn microphone relative to the talker's mouth. Again, the procedure is inspired by the work of Pelegrin-Garcia [12] and Pelegrín-García and Brunskog [3]. Let's separate the total transfer-function between source and receiver into direct and reflected sound components and denote them by $H_{\text{direct,sim}}$ and $H_{\text{refl,sim}}$. In the reproduction environment, let $H_{\text{direct,repr}}$ be the transfer-function between the head mounted microphone and a measurement microphone at 1 m distance and let $H_{\text{refl,repr}} = \mathbf{A}^T \mathbf{W} C$ be the transfer function between the head-mounted microphone and the talker's head position via the reproduction system, where $\mathbf{A}$ is the 64 element vector of transfer functions between the convolution engine and the participant's head through each loudspeaker,

$\mathbf{W}$ is a 64 element vector containing the frequency responses of the reproduction filters and C is the calibration filter. Assuming invertibility of the direct sound transfer-functions, the equivalence of reflected-to-direct sound is expressed as

$$\frac{H_{\text{refl,sim}}}{H_{\text{direct,sim}}} = \frac{H_{\text{refl,repr}}}{H_{\text{direct,repr}}}. \tag{1}$$

Using a regularized inverse of $\mathbf{A}^T\mathbf{W}$ with regularization parameter ε, the calibration filter is then given by

$$C = \frac{H_{\text{refl,sim}}}{H_{\text{direct,sim}}} H_{\text{direct,repr}} \frac{\left(\mathbf{A}^T\mathbf{W}\right)^*}{|\mathbf{A}^T\mathbf{W}|^2 + \varepsilon}.$$

The regularization parameter was chosen manually to be around the noise variance in the measurement of the impulse responses $\mathbf{A}$.

In general, C will not be causal and as such not realizable. Instead, we smooth the magnitude response of C in octave bands and approximate it by a minimum-phase FIR filter (with DFT C_{min}) with $N = 511$ taps and a simple (negative) delay τ such that

$$C(\omega) = C_{min}(\omega)\exp(-\mathrm{j}\omega\tau).$$

The delay $\tau = 22\,\text{ms}$ was estimated via cross-correlation.

When combining the calibration and reproduction filters, the negative delay is implemented by removing the first 22 ms of samples (which are approximately zero due to shifting of the early reflections) at the beginning of the reproduction filters. The calibration thus accounts for both gain and time differences between the simulated and reproduced environments in contrast to the method of Pelegrín-García and Brunskog [3], which only calibrated the magnitude response.

The direct sound $H_{\text{direct,repr}}$ is estimated from 1 minute of speech for each participant. First, cross- and autospectral densities are estimated with Welch's method [19] and a window of M = 1024 samples. Then, a FIR filter is designed via a causally constrained, regularized Wiener filter [20].

2.6 Virtual reality simulation

The visual environments are simulated using a HTC VIVE Pro Virtual Reality head-mounted display (HTC Corporation, New Taipei City, Taiwan). The environments are built with Unity (Unity Technologies, San Francisco, CA) and projected to the VR-headset using the SteamVR plugin (Valve Corporation, Bellevue, WA). Unity and MAX 8 are running on separate PCs to avoid buffer under-runs in the audio loop.

In the virtual environment (see Figure 3), the participant was placed in the middle of a room of similar shape and size as the room models in the acoustic simulation. An audience consisting of human avatars (packages "realpeople female" and "realpeople male", 3drt.com) are seated on simple chairs in front of the participant position. The position of the chairs/audience in relation to the participant are the same in each room, regardless of room size. The avatars moves slightly, but are not responsive to the participant. Besides the modelled audience, the room is empty.

Pictures modelled after the boardgame Story Cubes [21] was implemented in the virtual environment, with the purpose of providing the participant with speech stimuli. Nine pictures appeared randomly when the participant pressed a button on the VR-controller, and the pictures were visible for five seconds (see Figure 3)

2.7 Definitions and measurements of the room acoustic parameters

So far, the effective room acoustic conditions simulated for the participant are unknown, and need to be measured directly. This is done by measuring the impulse response between the loudspeaker mouth and microphone earcanals of a head-and-torso simulator (HATS; Brüel & Kjær Sound & Vibration Measurement A/S; Nærum, Denmark, model 4128). As VR is to be used in the experiment, and VR HMDs have been shown to affect the head-related-transfer function [22] as well as sound localization [13], the HATS wears the HMD during these measurements. The HATS is placed in the participant position and personalized reproduction filters are created in the same way as for a human participant, but by using white noise instead of speech for estimating $H_{\text{direct,repr}}$. The filters for each room acoustic model are then activated one at a time, and for each one a

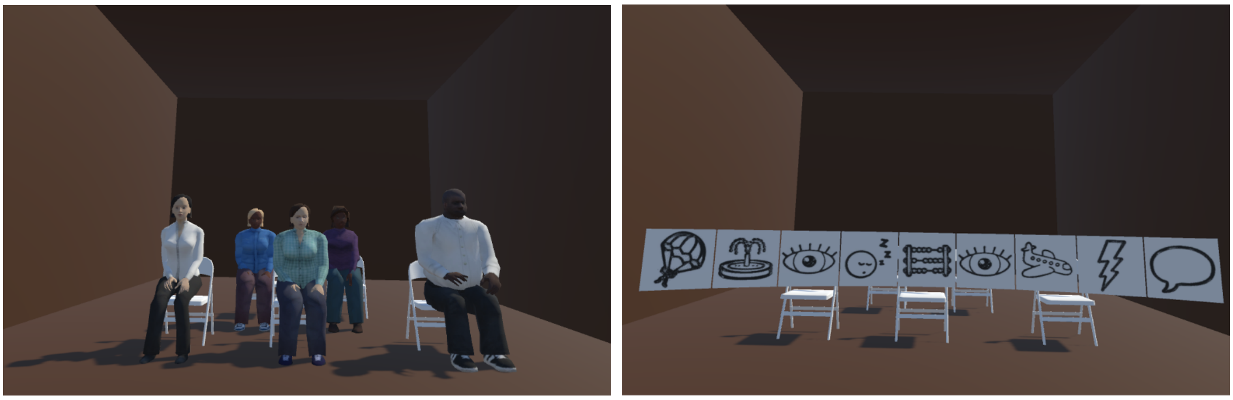


Figure 3. View of the virtual environment with and without avatar audience. The right hand figure show an example of the pictures used as speech stimuli, implemented from the board game Story Cubes [21]

number of impulse responses are measured (we used n = 40 sinusoidal sweeps á 240 s for each filter). Impulse responses are also measured without any active filter, as to obtain reference values for the physical room (AVIL) and the room gain (see below). The measures for the room acoustic parameters are derived from the impulse responses, hence the effective measures represent the actual acoustic environment at the ears of the participant. The measurements are not averaged over different room positions. As the talker and listener in this case are located in the same position (since they are the same person), this is not a problem. The parameters chosen for the purpose of this setup were reverberation time T_{30} (average over octave bands 500-1000 Hz), room gain G_{RG}, voice support ST_V and decay time DT_{40}.

Room gain and voice support are room acoustic parameters relating to the degree of amplification that the room provides to the talker's own voice while talking [4]. The room gain is defined as:

$$G_{RG} = 10\log\frac{E_D + E_R}{E_D}\,[\mathrm{dB}], \tag{2}$$

where E_D is the airborne direct sound energy and E_R the reflected sound energy. Room gain can also be defined as:

$$G_{RG} = L_E - L_{E,d}, \tag{3}$$

where L_E is the total energy level, containing both the direct and the reflective parts between the mouth and the ears, and $L_{E,d}$ is the energy level of only the direct path between the mouth and the ears. The magnitude varies between 0 and 2 dB in normal rooms.

Voice support is defined as:

$$ST_V = 10\log\frac{E_R}{E_D}\,[\mathrm{dB}] \tag{4}$$

or

$$ST_V = L_{E,r} - L_{E,d}, \tag{5}$$

where $L_{E,r}$ is the reflected sound. Both the direct and reflected sound are derived from a single impulse response between the mouth and the ears. The voice support values were speech weighted, in accordance with Pelegrín-García, Brunskog, Lyberg-Åhlander, and Löfqvist [23].

Decay time DT_{40} is the time it takes for the reverse integrated energy curve of an impulse response derived between the mouth and the ears to decay 60 dB after the arrival of the direct sound, but calculated from the initial decay at 40 dB. It is defined as:

$$DT_{40} = \frac{60}{40}(t_{-40\,\mathrm{dB}} - t_{0\,\mathrm{dB}})[\mathrm{dB}] \tag{6}$$

where $t_{-40\,\mathrm{dB}}$ is the time when the reverse integrated energy curve is 40 dB lower than before the arrival of the direct sound, which happens at t_0. The reverberation time T_{30} is defined as usual as

$$T_{30} = \frac{60}{30}(t_{-35\,\mathrm{dB}} - t_{-5\,\mathrm{dB}})[\mathrm{dB}]. \tag{7}$$

3 PILOT STUDY

A small pilot study was conducted as to test the set-up with combined VR and room acoustic simulations. The main purpose of the pilot was to gather information needed to finalize the study protocol for the upcoming main study. In addition to testing the room simulations, we also wanted to investigate at which distance the virtual audience was perceived to be located and if the presence of the audience affected the participants speech behavior as measured by speech rate and voice level. In this paper, only results from the distance perception as well as speech behavior differences with and without audience will be presented. All participants provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). Participants were given the opportunity to obtain an optional compensation fee for their time.

3.1 Participants

Ten volunteers (8 men, 2 women) were recruited as participants. They consisted of a convenience sample of university students and employees. The mean age was 33.4 years and age spread 22-60 years. All participants were fluent in English, but none were native English speakers. All participants were screened for hearing loss with pure-tone audiometry and all except one had a minimum of 20 dB better ear hearing level (HL) in octaves from 125 to 4000 Hz. One participant had a better ear HL of 40 dB at 4000 Hz, and below 20 dB at frequencies 2000 Hz and below. Three participants reported voice problems to a small extent based on the question *Does your voice tire, strain or get hoarse when you talk? Disregard symptoms that depend on current cold or upper-airway infection. The voice symptoms may vary but try to estimate an average.* This question has been used in a previous cohort study investigating the prevalence of voice problems in a general population [2]. Seven participants reported no voice problems. All recruited participants were included in the pilot.

3.2 Task

The participant was placed on the chair in the center of the loudspeaker array, wearing the head-mounted microphone and the VR HMD, see Figure 1. He/she was asked to speak in English for 3 minutes in each room, to a virtual audience of five avatars sitting on chairs or to an imaginary audience sitting on empty chairs. The participants were informed beforehand that they were going to be asked to speak freely, and were encouraged to think of topics to talk about in advance. The test leader suggested topics such as travels or hobbies. If the participants ran out of things to talk about during a test, they had the possibility to look at pictures for inspiration in the VR environment. The pictures were implemented from a game called Story Cubes[21], see also Figure 3. When the participant pressed a button on the VR-controller, nine pictures appeared for five seconds. The total picture bank consisted of 54 pictures, as in the original game. The participants were free to use the picture bank as much or as little as they preferred. After speaking for three minutes the participant was asked to rate the distance to the audience/empty chairs in meters.

3.3 Conditions

A total of 13 room simulations were tested in the pilot. Twelve were combinations of acoustic and visual simulations, and one room included only the visual simulation combined with the real acoustic properties of the anechoic chamber in which the participant was physically sitting. In the VR environment, half of the rooms had audience, and half of the rooms contained only chairs with no audience present. Background noise levels varied between the rooms; one third of the rooms contained no extra background noise (only the default background noise in the anechoic chamber which was below 20 dBA at the talker's position). One third of the rooms had a background noise level of 33 dBA pink noise at the talker's position, which was barely audible. The last third of the rooms had a background noise level of 43 dBA pink noise. The background noise was generated via MAX 8 and played back over the loudspeakers. Room order, noise levels in each room and presence of audience in each room was randomized for each participant, using the rand function in Matlab. Due to time limitations, the experiments were carried out during a maximum total time of 2 hours for each participant. This resulted in a different number of rooms being completed for each participant. The simulated rooms varied in size and acoustic conditions. See Table 1 for room specifics.

Table 1. Mean and standard deviations of the acoustic conditions of the simulated room environments. All values are based on N = 5 measures of each room filter. The room "AVIL" is the anechoic chamber without convolution. Room volumes are based on the predefined room acoustic models. The room volumes in the VR environment were set to equivalent sizes as the acoustic models, based on the default scale in the VR software.

Room	Volume [m3]	$T_{30,500-1000Hz}$[s]	ST_V[dB]	G_{RG}[dB]	DT_{40}[s]
AVIL	336	0.05 (0.003)	-18.24 (0.29)	0.21 (0.027)	0.08 (0.01)
1	315	0.34 (0.028)	-16.54 (0.41)	0.29 (0.032)	0.69 (0.03)
2	315	0.64 (0.037)	-16.01 (0.46)	0.30 (0.032)	0.76 (0.02)
3	810	0.49 (0.047)	-17.47 (0.33)	0.24 (0.028)	0.60 (0.03)
4	810	0.75 (0.081)	-17.38 (0.34)	0.24 (0.028)	1.10 (0.08)
5	550	0.14 (0.026)	-17.66 (0.30)	0.24 (0.027)	0.41 (0.15)
6	550	0.46 (0.043)	-16.92 (0.38)	0.26 (0.028)	0.69 (0.12)
7	1500	0.48 (0.079)	-17.58 (0.31)	0.24 (0.026)	0.81 (0.04)
8	1500	0.76 (0.123)	-17.65 (0.32)	0.23 (0.026)	1.23 (0.18)
9	1800	0.76 (0.121)	-17.73 (0.32)	0.23 (0.027)	1.33 (0.03)
10	1800	0.11 (0.038)	-18.10 (0.32)	0.22 (0.027)	0.33 (0.03)
11	1120	0.50 (0.090)	-17.66 (0.30)	0.23 (0.026)	0.94 (0.30)
12	1120	0.72 (0.099)	-17.51 (0.31)	0.24 (0.027)	1.02 (0.06)

3.4 Processing of voice recordings

The voice levels were adjusted to filter out the pink noise (33 dBA or 43 dBA) as well as to calibrate the voice levels. The transfer-function between the head-mounted microphone and the measurement microphone 1 m away ($H_{\text{refl,repr}}$) was used to normalize the sound pressure level to 1 m distance. An average sound pressure level of the 3 minute recordings in each room was measured for each participant. Speech rates were extracted automatically using a Praat [24] script based on syllable nuclei identification [25].

3.5 Analyses

Statistical analyses were conducted in R version 4.2.0 [26]. Distance perception for each participant and as function of room volume were analysed using descriptive statistics and linear regression, respectively. Voice level and speech rate differences between conditions with and without audience was analysed on group level by t-tests.

3.6 Results

Perception of audience distance for each participant and room volume can be seen in Figures 4 and 5, respectively. The average perceived distance was M = 5.6 m (SD = 2.6 m). Simple linear regression was used to determine if room volume was a predictor for the perceived distance. The results showed that it was not ($R^2 = 0.00, F = 0.048, p = 0.82$).

Voice levels and speech rates with and without audience can be seen in Figures 6 and 7, respectively. Re-

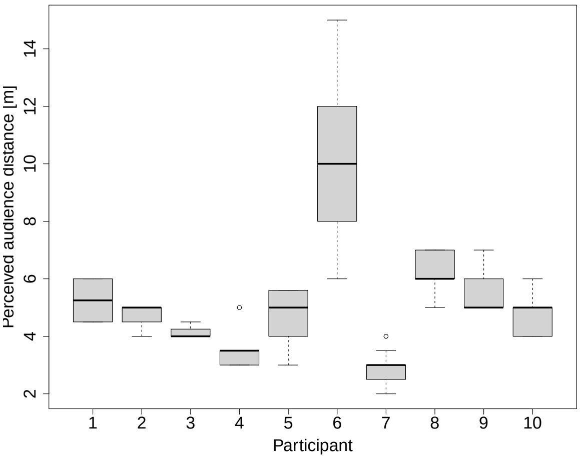


Figure 4. Boxplots of perceived audience distance for each participant

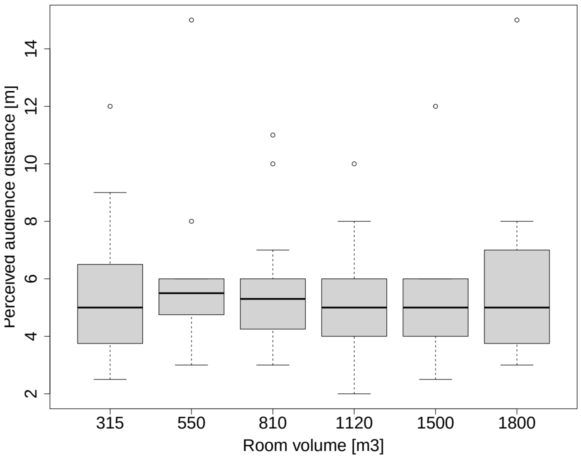


Figure 5. Boxplot of perceived audience distance in different room volumes.

sults from paired two-sided t-tests showed no statistically significant differences between voice levels ($t = -1.1639, df = 50, p = 0.25$) nor speech rates ($t = -1.6267, df = 50, p = 0.1101$) with and without audience.

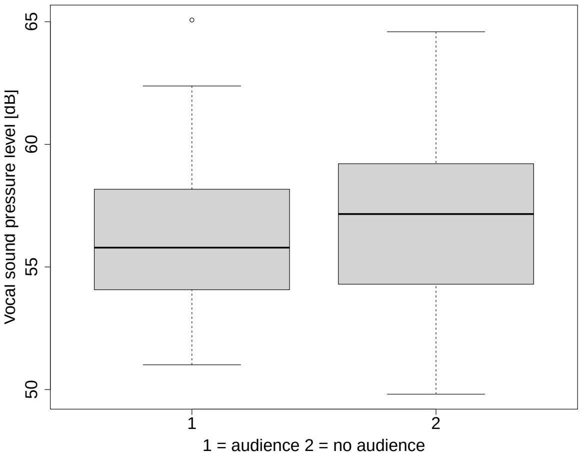


Figure 6. Boxplot of voice levels with and without audience in the virtual reality environment

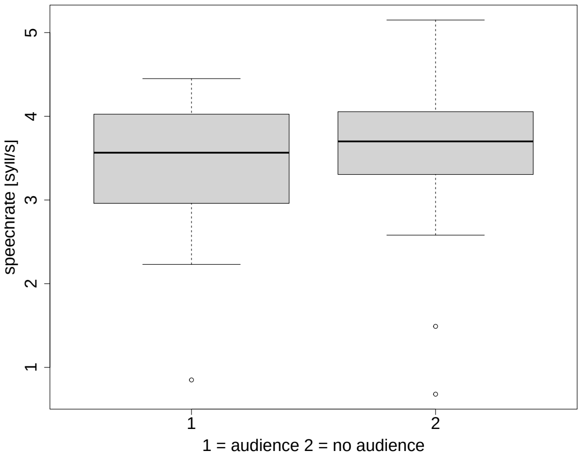


Figure 7. Boxplot of speech rates with and without audience in the virtual reality environment

3.7 Pilot conclusions

Based on the pilot tests, it can be concluded that distance perception in the VR environment varied between participants but not with room volume. The presence of audience did not seem to affect speech behavior based on speech rates and voice levels, however, most participants noted to the test leader afterwards that it felt more natural to speak with the audience present.

4 DISCUSSION AND CONCLUDING REMARKS

The purpose of this paper was to describe a combined set-up for acoustic and visual room simulations, to be used in research concerning voice and speech adaptations to different room acoustic environments. A previously developed set-up for real-time auralization of the talker's own voice [12, 3] has been reimplemented with some improvements at DTU's auditory-visual immersion lab, now also including virtual reality as visual simulation of room sizes as well as an audience for the participant to speak to.

While Pelegrin-Garcia [12] and Pelegrín-García and Brunskog [3] implemented their setup in a semi-anechoic chamber, AVIL is fully anechoic. Another difference is the way the system was calibrated. Pelegrin-Garcia *et al.* realized the ratio of reflected-to-direct sound (1) only in magnitude by passing the speech signal through a 1/3 octave-band equalizer. By applying the calibration to the reproduction filters instead of the speech signal, we can align the reflections also in time. Moving the early reflections by the system latency of 22 ms will have little impact on the room gain or voice support, as the energy of the first reflections (in the first 50 ms) is not affected. This also means that the room acoustic measures used will not be affected by this modification of the impulse responses.

In previous laboratory studies using real-time auralization for investigating voice use in different room acoustics, the visual environment for the participant has been that of the laboratory. The advantages of implementing VR to the set-up is the possibility to place the participant in a more ecologically valid setting. The use of VR within voice research is so far scarce, but previous studies investigating immersive VR as a possible tool for public-speaking training for people who stutter have shown that a virtual audience seems to evoke the same type of behavioral and cognitive experience as a real audience [27]. Comments from the participants in the pilot study suggests that the virtual rooms and audience make the experience more realistic compared to a pure laboratory environment. Although speech rate and vocal SPL were not affected by the presence of audience in the VR environment in our pilot study, most participants reported afterwards that they preferred the rooms with the audience in them, as it felt more natural to speak to someone than to just an empty room with chairs. The perception of distance to the audience/chairs varied between participants. Most participants perceived the distance to be around 4 to 7 meters, however one perceived it to be around 3 meters and one perceived large distance variations of 6 to 15 meters in the different rooms. The distance did however not change between conditions. These differences in distance perception needs to be taken into account when analyzing speech and voice behavior, as people tend to rise their voice levels with increased distance to the listener [28]. The question regarding distance perception is therefore important, as a way to control for the distance factor when analyzing voice level changes in various room acoustic conditions.

Based on the findings from the pilot study, we decided to keep the question regarding distance perception as well as to implement the audience into all rooms. Although not formally analyzed in this paper, we found the speech stimuli [21] implemented in the VR environment to work well and we will continue using it in our upcoming work. Other adjustments made to the study protocol after piloting includes broadening the range in the acoustic parameters in the room acoustic simulations as well as including more background questions to the participants. The rationale behind these adjustments will be reported elsewhere.

In this paper, we have described the laboratory set-up now in use at DTU for investigating vocal adjustments to different room acoustic conditions in combination with virtual reality. The set-up is already in use for data collection, and results from our upcoming work will be presented at a later time.

ACKNOWLEDGEMENTS

This work was funded by the Swedish Research Council FORTE, as part of the project *Communication, a challenge for the older employee: Need for a communication supporting workplace*, project ID 2019-01329. The authors would like to thank David Pelegrín García for consultation during the reimplementation of the set-up.

REFERENCES

[1] Vilkman, E. "Occupational risk factors and voice disorders". In: *Logopedics Phoniatrics Vocology* 21.3-4 (Jan. 1996), pp. 137–141. DOI: 10.3109/14015439609098881.

[2] Lyberg-Åhlander, V, Rydell, R, Fredlund, P, Magnusson, C, and Wilén, S. "Prevalence of Voice Disorders in the General Population, Based on the Stockholm Public Health Cohort". eng. In: *Journal of Voice: Official Journal of the Voice Foundation* 33.6 (Nov. 2019), pp. 900–905. DOI: 10.1016/j.jvoice.2018.07.007.

[3] Pelegrín-García, D and Brunskog, J. "Speakers' comfort and voice level variation in classrooms: Laboratory research". en. In: *The Journal of the Acoustical Society of America* 132.1 (July 2012), p. 249. DOI: 10.1121/1.4728212.

[4] Brunskog, J, Gade, AC, Bellester, GP, and Calbo, LR. "Increase in voice level and speaker comfort in lecture rooms". en. In: *The Journal of the Acoustical Society of America* 125.4 (Apr. 2009), p. 2072. DOI: 10.1121/1.3081396.

[5] Pörschmann, C. "Influences of Bone Conduction and Air Conduction on the Sound of One's Own Voice". In: *Acta Acustica united with Acustica* 86.6 (Nov. 2000), pp. 1038–1045.

[6] Bottalico, P, Graetzer, S, and Hunter, EJ. "Effects of speech style, room acoustics, and vocal fatigue on vocal effort". en. In: *The Journal of the Acoustical Society of America* 139.5 (May 2016), p. 2870. DOI: 10.1121/1.4950812.

[7] Bottalico, P and Astolfi, A. "Investigations into vocal doses and parameters pertaining to primary school teachers in classrooms". en. In: *The Journal of the Acoustical Society of America* 131.4 (Apr. 2012), p. 2817. DOI: 10.1121/1.3689549.

[8] Puglisi, GE, Astolfi, A, Cantor Cutiva, LC, and Carullo, A. "Four-day-follow-up study on the voice monitoring of primary school teachers: Relationships with conversational task and classroom acoustics". In: *The Journal of the Acoustical Society of America* 141.1 (2017), pp. 441–452.

[9] Rapp, M, Cabrera, D, and Yadav, M. "Effect of voice support level and spectrum on conversational speech". en. In: *The Journal of the Acoustical Society of America* 150.4 (Oct. 2021), p. 2635. DOI: 10.1121/10.0006570.

[10] Cipriano, M, Astolfi, A, and Pelegrín-García, D. "Combined effect of noise and room acoustics on vocal effort in simulated classrooms". en. In: *The Journal of the Acoustical Society of America* 141.1 (Jan. 2017), EL51. DOI: 10.1121/1.4973849.

[11] Cantor-Cutiva, LC, Bottalico, P, Ishi, CT, and Hunter, EJ. "Vocal Fry and Vowel Height in Simulated Room Acoustics". In: *Folia Phoniatrica et Logopaedica* 69.3 (2017), pp. 118–124. DOI: 10.1159/000481282.

[12] Pelegrin-Garcia, D. *The role of classroom acoustics on vocal intensity regulation and speakers' comfort: PhD thesis*. en. Kgs. Lyngby: DTU Electrical Engineering, 2011.

[13] Ahrens, A, Lund, KD, Marschall, M, and Dau, T. "Sound source localization with varying amount of visual information in virtual reality". In: *PloS one* 14.3 (2019), e0214603.

[14] Ahrens, A, Marschall, M, and Dau, T. "Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments". In: *Hearing research* 377 (2019), pp. 307–317.

[15] Gardner, WG. "Efficient convolution without latency". In: *Journal of the Audio Engineering Society* 43 (1993), p. 2.

[16] Harker, A and Tremblay, PA. "The HISSTools impulse response toolbox: Convolution for the masses". In: *Proceedings of the international computer music conference*. The International Computer Music Association. 2012, pp. 148–155.

[17] Favrot, S and Buchholz, JM. "LoRA: A loudspeaker-based room auralization system". In: *Acta acustica united with Acustica* 96.2 (2010), pp. 364–375.

[18] Marschall, M and Ahrens, A. *Lora Toolbox repository*. https://bitbucket.org/hea-dtu/lora.

[19] Welch, P. "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms". In: *IEEE Transactions on audio and electroacoustics* 15.2 (1967), pp. 70–73.

[20] Proakis, J and Manolakis, D. *Digital Signal Processing*. Prentice Hall international editions. Pearson Prentice Hall, 2007.

[21] Cubes, RS. *Story Cubes via the Internet Archive*. https://web.archive.org/web/20210615132612/https://www.storycubes.com/en/games/rorys-story-cubes-classic/. Accessed: 202-06-30. 2021.

[22] Gupta, R, Ranjan, R, He, J, and Woon-Seng, G. "Investigation of effect of VR/AR headgear on Head related transfer functions for natural listening". In: *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society. 2018.

[23] Pelegrín-García, D, Brunskog, J, Lyberg-Åhlander, V, and Löfqvist, A. "Measurement and prediction of voice support and room gain in school classrooms". en. In: *The Journal of the Acoustical Society of America* 131.1 (Jan. 2012), p. 194. DOI: 10.1121/1.3665987.

[24] Boersma, P and Weenink, D. *Praat: doing phonetics by computer (Version 6.2.01)*. 2021. URL: http://www.praat.org.

[25] Jong, NH de and Wempe, T. "Praat script to detect syllable nuclei and measure speech rate automatically". en. In: *Behavior Research Methods* 41.2 (May 2009), pp. 385–390. DOI: 10.3758/BRM.41.2.385.

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: https://www.R-project.org/.

[27] Brundage, SB and Hancock, AB. "Real enough: Using virtual public speaking environments to evoke feelings and behaviors targeted in stuttering assessment and treatment". In: *American Journal of Speech-Language Pathology* 24.2 (2015), pp. 139–149.

[28] Pelegrín-García, D, Smits, B, Brunskog, J, and Jeong, CH. "Vocal effort with changing talker-to-listener distance in different acoustic environments". en. In: *The Journal of the Acoustical Society of America* 129.4 (Apr. 2011), p. 1981. DOI: 10.1121/1.3552881.