Virtual reality head-mounted displays affect sidetone perception

Wistbacka, Greta Öhlund; Shen, Weihan; Brunskog, Jonas

Published in:
Jasa Express Letters

Link to article, DOI:
10.1121/10.0014605

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Wistbacka, G. Ö., Shen, W., & Brunskog, J. (2022). Virtual reality head-mounted displays affect sidetone perception. Jasa Express Letters, 2(10). https://doi.org/10.1121/10.0014605
Virtual reality head-mounted displays affect sidetone perception

Greta Öhlund Wistbacka,a) Weihan Shen, and Jonas Brunskog
Acoustic Technology, Department of Electrical and Photonics Engineering, Technical University of Denmark, Kongens Lyngby DK-2800, Denmark
gmawi@elektro.dtu.dk, s212644@dtu.dk, jbr@elektro.dtu.dk

Abstract: The purpose of this study was to investigate whether head-mounted displays (HMDs) change the sidetone to an auditory perceivable extent. Impulse responses (IRs) were recorded using a dummy head wearing a HMD (IRtest) and compared to IRs measured without HMD (IRref). Ten naive listeners were tested on their ability to discriminate between the IRtest and IRref using convolved speech signals. The spectral analysis showed that the HMDs decreased the spectral energy of the sidetone around 2000–4500 Hz. Most listeners were able to discriminate between the IRs. It is concluded that HMDs change the sidetone to a small but perceivable extent.

1. Introduction

Immersive, three-dimensional (3D) virtual reality (VR) environments are becoming more and more common as tools within medical rehabilitation, psychological treatment, and education. Immersive virtual environments can be displayed through head-mounted-displays (HMDs), providing the user with the experience of actually "being" in the simulated environment, including a sense of disconnection from the real world.

HMDs add acoustic reflective surfaces around the head and face of the user that are not typically present. These surfaces create acoustic distortions that previous research has shown to affect the perception of spatial and timbre quality and accuracy of sound localization. Gupta et al. investigated alterations to the head-related transfer functions (HRTFs) between a sound source at 1 m distance to ear canal microphones of a dummy head wearing three different types of HMDs. Spectral differences were analyzed between different azimuth positions of the dummy head (0°–180°) for the ipsilateral and contralateral HRTF impulse responses, as well as the different HMDs. The results showed that the HMDs changed the spectral content of the HRTFs, with more prominent changes in the contralateral than the ipsilateral ear. Listening tests confirmed the objective results, showing perceptually noticeable timbre and spatial differences for most HMDs and azimuth positions, including 0°.

Previous research has focused on acoustic perturbations induced by the HMDs with the sound source located in the far field. However, there are several applications for immersive VR where the sound source will be located at the user’s mouth, i.e., when speaking or singing. For example, immersive VR has been implemented for public speaking training for university students, people with public speaking anxiety, and adults who stutter. Immersive VR systems have also been developed for group singing activities, making it possible to virtually take part in choir singing from home.

A talker/singer receives sensorimotor and auditory feedback on his own voice and speech production in real time. The auditory feedback, or sidetone, consists of a bone-conducted and an airborne part. The latter can be further divided into the direct pathway of the sound from the mouth to the talker’s own ears and the sound reflected by surrounding surfaces before reaching the talker’s eardrums. Changes to the sidetone have been shown to induce compensatory changes in vocal behavior, for instance, the Lombard effect stating that a talker will raise his vocal loudness when background noise increases and, on the contrary, decrease vocal loudness when sidetone loudness increases. Compensatory vocal behaviors have also been found for sidetone pitch alterations and formant shifts. Therefore, the sidetone has a direct impact on voice and speech regulation.

Since HMDs have been found to change HRTFs measured with an external sound source, they are likely to also affect HRTFs when the sound source is the user’s own voice. If so, the use of HMDs could risk inducing compensatory vocal behaviors of the user as a consequence of sidetone changes. The purpose of this study was therefore to investigate
how sound reflections from a commonly used HMD change the frequency spectrum of the sidetone and to investigate whether or not these spectral differences are perceivable to naive listeners.

2. Materials and methods

2.1 Impulse response measurements with and without head-mounted display

To investigate changes in sidetone signal content due to HMDs, impulse responses (IRs) were measured between the mouth loudspeaker and ear canal microphones of a head-and-torso simulator (HATS) (model 4128; Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark) while it was wearing a HMD (model HTC VIVE; HTC Corporation, New Taipei City, Taiwan) and compared to reference IRs measured in the same way without the HMD (Fig. 1).

All measurements took place in an anechoic chamber using 10.9 s long sinusoidal sweeps excited by the DIRAC software (version 7841, Brüel & Kjær Sound & Vibration Measurement A/S). The sampling frequency was 48 kHz. Six experimental IRs (IRtest) were measured with the HATS wearing the HMD. The HMD was removed from and put back on the dummy head between each measurement to mimic typical usage. Ten reference IRs (IRref) were measured without the HMD. Data analyses were performed using MATLAB (MathWorks, Natick, MA). After inspection of the energy-time curve, all IRs were truncated in time domain at 0.25 s as to reduce the influence of noise possibly contaminating the spectrum. The energy and frequency content of the truncated IRref and IRtest were further analyzed and compared in time and frequency domain, respectively, and used for convolving speech signals for a listening experiment.

2.2 Listening experiment

2.2.1 Equipment and test procedure

A listening experiment was performed to investigate whether naive listeners could discriminate between the IRref and IRtest. Two speech samples from the TIMIT speech corpora were used. The samples consisted of a male and a female voice saying the phrase “she had your dark suit in greasy wash water all year” in New England American English. Both speech samples were approximately 3 s long and had a sampling frequency of 16 kHz. Two yes/no auditory perception tasks were carried out, one with the female voice stimuli and one with the male voice stimuli, using a custom-made MATLAB application. The participants were instructed to listen to two consecutive signals and determine if they differed from each other (YES-option) or if they were identical (NO-option). For each signal pair, a reference signal consisting of the speech sample convolved with a randomly chosen IRref was played first, followed by the second signal, which was either identical to the first one (NO-stimuli) or consisted of the speech sample convolved with a randomly chosen IRtest (YES-stimuli). The signals were separated by 1.5 s, and each test consisted of 20 pairs of stimuli. The experiment was carried out in a double-walled acoustically treated booth, and the sound stimuli were presented using a Fireface UCX soundcard (RME, Haimhausen, Germany) and HD650 open headphones (Sennheiser, Wedemark, Germany). The signals were presented to the participants at an equivalent sound level of 57 dBA, adjusted by connecting the headphones to an artificial ear while measuring the sound pressure level using a Norsonic NOR139 sound level meter (Norsonic AS, Tranby, Norway).
2.2.2 Participants

Ten volunteers were recruited for the listening experiment, seven males and three females. The participants were university students and staff members in acoustics and/or hearing sciences, and all except two reported previous experience of participating in listening tests. The median age was 32 years (range 27–52 years). All participants reported normal hearing, and five reported to have had musical training for at least a few years. The participants were naive to the purpose of the study. All participants provided informed consent before participation, and ethical approval was obtained from the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

2.2.3 Data analysis

Data from the listening experiment were analyzed using signal detection theory (SDT). SDT separates sensitivity of signal discrimination ability from response bias (for an overview of the measures, please see, e.g., Stanislaw and Todorov33 and See et al.34). The SDT calculations are based on the HIT-rate and false alarm rate. The HIT-rate is defined as the ratio of the number of correctly identified IRref and IRtest stimulus pairs over the total number of IRref and IRtest stimulus pairs (i.e., the ratio of the number of correct YES-replies over the total number of YES-stimuli). The false alarm rate is defined as the ratio of the number of incorrectly identified IRref and IRtest stimulus pairs over the total number of IRref only stimulus pairs (i.e., the ratio of the number of incorrect YES-replies over the total number of NO-stimuli).

The sensitivity of signal discrimination was calculated using the non-parametric discriminability index, $A'$, first described by Pollack and Norman35 and calculated by33,36

$$A' = 0.5 + \left[ \frac{\text{sign}(H - F)}{4(H/F)} \left( \frac{H - F}{4(H/F)} + \frac{|H + F|}{4(H/F)} \right) \right],$$

(1)

where $H =$ HIT-rate and $F =$ false alarm rate.

An $A'$ value close to 1 indicates good discriminability between the signals, whereas an $A'$ value of 0.5 indicates chance performance. An $A'$ value below 0.5 indicates sampling error or response confusion.33 The response bias, hence, the participant’s underlying criterion before replying “Yes” was analyzed using the non-parametric $B'_{00}$,37 which in turn is calculated by34

$$B'_{00} = \frac{(1 - H)(1 - F) - HF}{(1 - H)(1 - F) + HF}.$$  

(2)

The $B'_{00}$ value ranges between −1 and 1, where a value of 0 indicates no bias, a positive value indicates a tendency to reply “No” (conservative bias; in this case, “the signals are identical”), and a negative value indicates a tendency to reply “Yes” (liberal bias; in this case, “the signals are different”).37 All SDT calculations were performed using MATLAB and plotted using R.38 Possible differences in rating performance due to musical background as well as due to speech stimulus gender were analyzed using Wilcoxon signed rank tests in R.

3. Results

3.1 Spectral differences with and without head-mounted display

The differences between the IRtest and IRref can be seen in the time and frequency domain, respectively, in Fig. 2. The magnitude of the IRtest was up to approximately 2 dB lower compared to the IRref in the frequency bands ca.
180–1500 Hz, up to approximately 13 dB lower in the bands ca. 2100–4700 Hz, and up to approximately 10 dB lower in the frequency bands ca. 6000–7100 Hz. The IRtest exceeded the IRref with approximately 1 dB around 2000 Hz and with approximately 10 dB around the frequency bands 7600–7700 Hz.

3.2 Listening experiment

Medians and data distribution results from the SDT analysis are found in Fig. 3. The HIT-rate varied between Min = 0.00 and Max = 1.00 for both the female and male voice stimuli, with a Med = 0.65 for the female voice stimuli, and Med = 0.62 for the male voice stimuli. The corresponding false alarm rates were Med = 0.12, Min = 0.00, and Max = 0.50 for the female voice stimuli and Med = 0.19, Min = 0.00, and Max = 0.58 for the male voice stimuli. The A' discriminability index varied between Min = 0.68 and Max = 0.96 with a Med = 0.82 for the female voice stimuli. The A' index could not be calculated for two participants since their HIT-rate and false alarm rate equaled 0, giving an n = 8 for the A' index results for the female voice stimuli. Corresponding A' index results for the male voice stimuli were Med = 0.89, Min = 0.79, and Max = 0.98 after excluding two outliers with A' < 0.50, indicating too low sampling or response confusion. One participant presented HIT- and false alarm rates of 0, and therefore the A' index could not be calculated, giving an n = 9 for the A' index results for the male voice stimuli. The bias measure B'00 varied between Min = −1.00 and Max = 1.00 for both the female and male voice stimuli, with a Med = 0.34 for the female voice stimuli and Med = 0.28 for the male voice stimuli. The analyses of differences in rating performance due to musical background and speech stimulus gender both presented non-significant results.

4. Discussion

The purpose of this study was to investigate how a HMD affects the sidetone and whether these possible alterations are auditory perceivable. The results showed a decrease in spectral energy over several frequency bands when the HMD was worn (IRtest) compared to the no-HMD condition (IRref). The results of the listening test showed that naive listeners were mostly able to discriminate between the IRs when they were presented as convolved speech signals. Previous research has shown that HMDs affect the spectral energy levels of the HRTFs as well as the accuracy of sound localization when using a sound source in the far field. When investigating sidetone, the sound source (i.e., the mouth) is located at equal distance from the receivers (i.e., the ear canals) at all times, most equivalent to 0° azimuth in the far field. Our results are in line with those reported by Gupta et al., who found perceptually detectable differences in timbre quality at 0° azimuth when using a similar HMD (HTC Vive pro) as in the current study. They, however, found the most prominent spectral differences (4 dB lower with HMD compared to without HMD) in the high frequency region (5–16 kHz and at 0° azimuth), whereas we found the largest decreases in magnitude in the frequency band 2100–4700 Hz.

Fig. 3. Median and data distribution results from the SDT analysis. (a) shows the HIT-rate (i.e., the ratio of the number of correct YES-replies over the total number of YES-stimuli, N = 10); (b) shows the false alarm rate (the ratio of the number of incorrect YES-replies over the total number of NO-stimuli, N = 10); (c) shows the A' discriminability index (n = 8 for the female and n = 9 for the male voice sample, respectively); and (d) shows the B'00 bias measure (N = 10).
The results of the SDT analyses of the listening test showed that the ability to distinguish between the IRtest and the IRref varied between the participants. The $A'$-value for some participants approached 1, indicating almost perfect performance. These participants tended to comment afterward that they had noticed a small shift in timbre for some of the stimuli but still experienced the task as difficult and the signals overall to be quite similar. A few participants demonstrated no ability to distinguish between the IRs at all, which confirms that the task was indeed difficult and the differences between the IRs minor. The bias measure $B_{D}$ varied between the extremes $-1$ and $1$, indicating a big variation in the participants’ underlying criterion when deciding upon whether they perceived a difference or not.

However, as some participants demonstrated a clear ability to discriminate between the IRs, we find that the result supports our hypothesis that the HMD changes the sidetone to a perceivable extent. This, in turn, raises the question of whether the sidetone alteration is big enough to affect voice or speech production. There is substantial evidence that shifts in sidetone loudness or pitch induce compensatory behaviors in vocal loudness and pitch regulation, respectively. The same is true for sidetone alterations of formants, as several studies have demonstrated compensatory behaviors in vowel articulation when the first and/or second formant frequency have been shifted. We have demonstrated small changes in spectral energy levels over frequency bands overlapping typical first and second formant frequencies and more prominent changes in frequency bands overlapping typical third and fourth formant frequencies (for an overview of formant frequencies in American English, see Kent and Vorperian). This suggests that the HMD could induce formant shifts that, in turn, could provoke compensatory behaviors.

Real-time perceptual differences and speech characteristic changes due to wearing a HMD need to be addressed in further studies to take possible compensatory strategies into account if HMDs are to be used within fields in which small changes in vocal behavior could affect the outcome of interest. Such areas are, for instance, medical voice therapy or singing training. Although immersive VR already has been implemented for different speaking and singing activities, such as public speaking training and choir singing, the focus of this research has been on aspects other than the individual speaker’s/singer’s voice use. Compensatory vocal and speech behaviors could be investigated through real-time auralization, in which the participant’s speech signal would be convolved with the IRtest and IRref, respectively, and played back to the participant during continuous speech. This would also allow for subjective analysis of sidetone perception, using the speaker’s own voice instead of pre-recorded speech stimuli.

There are different types of HMDs on the market, and a limitation of the current study is that only one type of HMD was investigated. Therefore, the results are not transferable to models other than the one tested. Gupta et al. analyzed three different HMDs from different manufacturers and found the biggest spectral differences as well as perceptually detectable timbre changes when using the HTC Vive pro compared to the other two tested (Oculus Rift and HoloLens). The authors suggest that this could be due to the slightly larger size of the HTC Vive pro compared to the other two HMDs. The HMD used in the current study, HTC Vive, is similar in size and shape to the one used by Gupta et al., and it is possible that other types of HMDs would alter the sidetone differently and to a lesser extent. It would therefore be important to investigate sidetone changes using HMDs from different manufacturers.

5. Conclusions
The HTC Vive head-mounted VR display changes the spectral energy levels of the sidetone, particularly in the frequency band 2100–4700 Hz. These spectral differences are perceivable to naive listeners. There is a need to investigate whether these changes affect vocal behavior during speaking or singing with a HMD.

Acknowledgments
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.” The authors wish to thank master student in acoustics Zhengmeng Li for assistance with data analysis and all participants for their time.

References and links
1. R. Gupta, R. Ranjan, J. 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).
2. A. 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).
3. K. Kryston, H. Goble, and A. Eden, “Incorporating virtual reality training in an introductory public speaking course,” J. Commun. Pedagogy 4, 133–151 (2021).
4. F. Vanni, C. Conversano, A. Del Debbio, P. Landi, M. Carlini, C. Fanciullacci, M. Bergamasco, A. Di Fiorino, and L. Dell’Osso, “A survey on virtual environment applications to fear of public speaking,” Eur. Rev. Med. Pharmacol. Sci. 17(12), 1561–1568 (2013).
5. M. H. Lim, V. Aryadoust, and G. Esposito, “A meta-analysis of the effect of virtual reality on reducing public speaking anxiety,” Curr. Psychol. (published online 2022).
6. B. Brundage and A. B. Hancock, “Real enough: Using virtual public speaking environments to evoke feelings and behaviors targeted in stuttering assessment and treatment,” Am. J. Speech Lang. Pathol. 24(2), 139–149 (2015).
M. Rapp, D. Cabrera, and M. Yadav, “Effect of voice support level and spectrum on conversational speech,” J. Acoust. Soc. Am. 128(3), 2288–2300 (2010).

K. G. Munhall, E. N. MacDonald, S. K. Byrne, and I. Johnsrude, “Talkers alter vowel production in response to real-time formant perturbation,” J. Acoust. Soc. Am. 132(4), 2288–2297 (2012).

M. Xu, R. O. Tachibana, H. Hagiwara, R-i Hashimoto, and F. Homae, “Unconscious and distinctive control of vocal pitch magnitudes,” J. Acoust. Soc. Am. 140(2), 105202 (2022).

R. Alemi, A. Lehmann, and M. L. D. Deroche, “Adaptation to pitch-altered feedback is independent of one’s own voice pitch sensitivity,” Sci. Rep. 10, 16860 (2020).

M. Arbeiter, S. Petermann, U. Hoppe, C. Bohr, M. Doellinger, and A. Ziethe, “Analysis of the auditory feedback and phonation in normal voices,” Ann. Otol. Rhinol. Laryngol. 127(2), 249–260 (2012).

C. R. Larson, K. W. Allman, H. Liu, and T. C. Hain, “Interactions between auditory and somatosensory feedback for voice F0 control,” Exp. Brain Res. 187(4), 613–621 (2008).

J. A. Laster-Smith, J. H. Kim, A. Hilger, C.-L. Chan, and C. R. Larson, “Auditory-motor control of fundamental frequency in vocal vibrato,” J. Voice (published online 2021).

M. Xu, R. O. Tachibana, K. Okanoya, H. Haginawa, R-i Hashimoto, and F. Homae, “Unconscious and distinctive control of vocal pitch and timbre during altered auditory feedback,” Front. Psychol. 11, 1224 (2020).

E. N. MacDonald, R. Goldberg, and K. G. Munhall, “Compensations in response to real-time formant perturbations of different magnitudes,” J. Acoust. Soc. Am. 127(2), 1059–1068 (2010).

E. N. MacDonald, D. W. Purcell, and K. G. Munhall, “Probing the independence of formant control using altered auditory feedback,” J. Acoust. Soc. Am. 129(2), 955–965 (2011).

T. Mitsuya, E. N. MacDonald, D. W. Purcell, and K. G. Munhall, “A cross-language study of compensation in response to real-time formant perturbation,” J. Acoust. Soc. Am. 130(5), 2978–2986 (2011).

T. Mitsuya, E. N. MacDonald, K. G. Munhall, and D. W. Purcell, “Formant compensation for auditory feedback with English vowels,” J. Acoust. Soc. Am. 138(1), 413–424 (2015).

V. M. Villacorta, J. S. Perkell, and F. H. Guenther, “Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception,” J. Acoust. Soc. Am. 122(4), 2306–2319 (2007).

D. W. Purcell and K. G. Munhall, “Compensation following real-time manipulation of formants in isolated vowels,” J. Acoust. Soc. Am. 119(4), 2288–2297 (2006).

K. G. Munhall, E. N. MacDonald, S. K. Byrne, and I. Johnsrude, “Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate,” J. Acoust. Soc. Am. 125(1), 384–390 (2009).

S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, Timit Acoustic Phonetic Continuous Speech Corpus Linguistic Data Consortium, Philadelphia, PA, 1993.

H. Stanislaw and N. Toddor, “Calculation of signal detection theory measures,” Behavior Res. Methods Instrum. Comput. 31(1), 137–149 (1999).

E. J. See, J. S. Warm, W. N. Dember, and S. R. Howe, “Vigilance and signal detection theory: An empirical evaluation of five measures of response bias,” Hum. Factors 39(1), 14–29 (1997).

D. Pollack and D. A. Norman, “A non-parametric analysis of recognition experiments,” Psychon. Sci. 1(1), 125–126 (1964).

J. G. Snodgrass and J. Corwin, “Pragmatics of measuring recognition memory: Applications to dementia and amnesia,” J. Exp. Psychol. 117(1), 34–50 (1988).