Audio Engineering Society

Conference Paper 12

Presented at the AES 5th International Conference on Audio for Virtual and Augmented Reality
2024 August 19–21, Redmond, WA, USA

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Expected Levels of Reproduced Speech

Nils Meyer-Kahlen[1], Sergio de las Heras[1, 2], and Tapio Lokki[1]

[1]*Acoustics Lab, Dpt. of Information and Communications Engineering, Aalto University, Finland*
[2]*Institute of Technical Acoustics, Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen University, Germany*

Correspondence should be addressed to Nils Meyer-Kahlen (nils.meyer-kahlen@aalto.fi)

ABSTRACT

In future Augmented Reality applications, the level of virtual sound sources should likely be aligned to the level of an equivalent real source. With the goal of working towards understanding the importance of such alignment, this paper studies the accuracy of listeners' expectations regarding the level of speech in a room. The presented experiment asks listeners to adjust the level of reproduced speech to match their expectations of how loud the corresponding real speaker would be. For the experiment, a dataset of speech recordings was used, where different speakers articulated a sentence at five distinct voice levels. The test was conducted in two different room acoustic conditions. In addition to adjusting the playback level, participants were asked to rate how well they knew the speaker. Results show that participants tended to underestimate the level, especially when the voice level was high. Moreover, the errors and their variability were smaller when participants knew the speaker well. There was no difference between the two room acoustic conditions, nor were we able to show strong adaptation to the room acoustics over time. This suggests that listeners are able to form level expectations instantaneously, incorporating the memory of source characteristics and the acoustic conditions of the room.

1 Introduction

The level of conversational human speech in a room is typically between 55 and 66 dB [1]. It varies depending on the distance between the source and receiver, the room's acoustic properties, and obviously, the sound pressure level of the produced speech, which we call voice level in this work. The voice level, in turn, influences acoustic features of the speech, such as its fundamental frequency, center frequency of the first formant [2], and spectral envelope [3]. Fig. 1 shows smoothed magnitude spectra of the same sentence uttered by the same person at five different voice levels.

Given these features, it is likely that human listeners can form expectations of the correct absolute level when listening to a speech recording. As a consequence, level mismatches could potentially be a determining factor of plausible [4] or transfer-plausible [5] virtual acoustics for augmented reality (AR) telepresence rendering; if virtual renderings of speech are reproduced too loud or too quiet, they might be less likely to be perceived as real.

In this paper, we ask how accurately human participants can adjust the level of reproduced speech based on their expectation of the real, absolute level. We study if familiarity between the subject and the speaker influences the level accuracy. Furthermore, we performed the test in two different room acoustic conditions, thereby examining to what extent listeners adapt

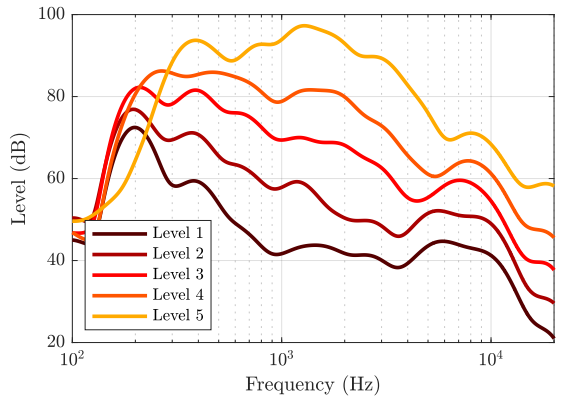


Fig. 1: Speech spectra of five samples of the same human speaker uttered at different voice levels.

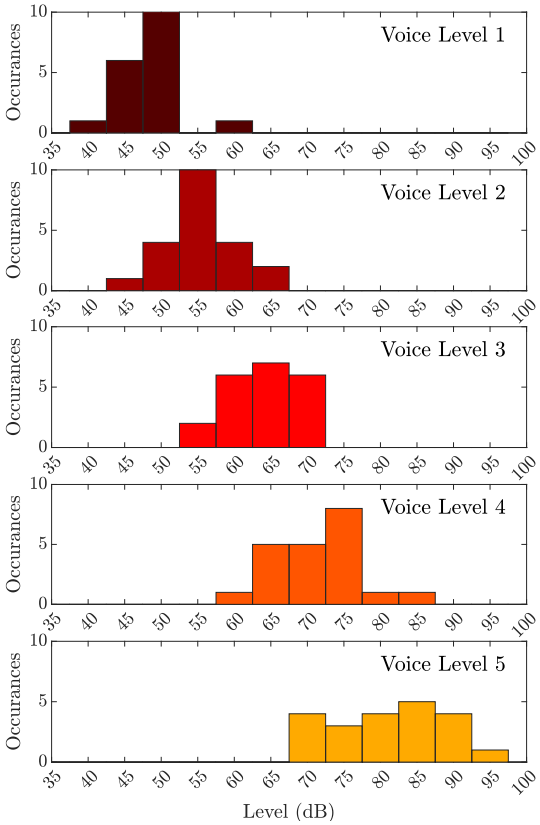


Fig. 2: All participants' A-weighted sound pressure levels at 1.5 m within the five voice level classes.

their level expectations to new room acoustics. A very similar experiment was already conducted by Wilkens and Bartel in 1977 [6]. However, their study consisted on a detection task rather than level adjustment. Also, different room acoustic conditions were not compared directly, and the variation between tested voice levels is not described. Nevertheless, we can compare some of our results to those obtained therein.

2 Methods

2.1 Signals

Anechoic recordings of 21 speakers (9 female, 13 male) were used as speech samples. These were part of a larger dataset in which participants recorded two sets of Harvard sentences and three scripted conversations each. The sessions took place in the anechoic chamber "Lampio" at the Aalto Acoustics Laboratory. A 1" low noise, G.R.A.S. Type 40HF measuring system was used for capturing. The distance between the speaker and the microphone was approximately 1.5 m. The microphone was placed roughly at the same height as the speaker's mouth. A concept picture showing the room and the microphone is presented in Fig. 3a. Also, a calibration tone from a B&K Type 4231 Sound Level Calibrator was recorded. More details regarding the dataset are presented in [7].

For the samples used in the experiment, speakers were asked to repeat the sentence, "On the island the sea breeze is soft and mild," at five different levels, between "barely not whispering" and "screaming as loud as you can," in five equally large level steps. In total, 105 samples were recorded. Fig. 2 shows the distribution of A-weighted sound pressure levels of the samples within each of these five "voice level classes". The mean levels range from $L_{\mathrm{Aeq}} = 49$ dB in the lowest voice level class to 81 dB in the highest. The variability between speakers increases with higher levels. The standard deviation for the samples in class 1 is 3.8 dB, while it is 7.7 dB for the samples in class 5. Already during the recording, it became clear that some volunteers could and were willing to shout much louder than others. The distributions are very similar in absolute levels and spread to those found in the anechoic recordings of different vocal levels performed in [8]. Also there, speakers were instructed to speak in five different voice levels. For categories 2-5, the mean levels found here and the mean levels found for female speech there are equivalent to 1 dB.



(a) Speech samples were recorded in the anechoic chamber. High voice level (Picture not from actual recording).

(b) Reproduction level calibration in the anechoic chamber.

(c) Listening experiment in the variable acoustics room (curtain made transparent for picture).

Fig. 3: An overview of the complete method: First, anechoic speech was recorded with different voice levels. Then, the level of a loudspeaker was calibrated to match the true level of the recorded speech on-axis. For the experiment, the loudspeaker was placed in the variable acoustics room, where participants were asked to adjust the level to match their expectations.

2.2 Reproduction

For the experiment, the collected speech samples were reproduced using a Genelec 8331A coaxial loudspeaker. The reproduction level was calibrated using the same G.R.A.S. Type 40HF measuring system. Therefore, the loudspeaker was placed in the center of the same anechoic chamber used for recordings along with the microphone at 1.5 m distance from the loudspeaker, on-axis with the center of the loudspeaker, see Fig. 3b.

Ten random samples were selected from the sample pool and recorded in the anechoic space. The microphone was calibrated using the same calibrator as used when doing the recordings. Then, the reproduction level was compared to that of the speech corpus recordings in order to obtain the required gain correction, ensuring that the on-axis reproduction level of the loudspeaker was equal to the level obtained during the speech recordings. For verification of the calibration procedure, five examples of recorded and reproduced speech spectra are shown in Fig. 4; they closely align for frequencies of up to 10 kHz.

2.3 Reproduction Room

For the experiment, the loudspeaker was placed in the variable acoustics room “Arni” at the Aalto Acoustics Laboratory. The room is equipped with 55 panels that can be controlled individually to create different room acoustic conditions [9]. Two very different conditions were selected for the task: one very dry, and one very reverberant. Fig. 5 shows their reverberation time (T_{20}) and direct-to-reverberant energy ratio (DRR) . The mid-frequency reverberation time (average between 500 Hz and 1 kHz bands, as in [10]) was 0.63 s for the dry and 0.99 s for the reverberant condition. The mid-frequency DRR was -0.56 dB vs. -5.75 dB, respectively.

2.4 Test Design

As seen in Fig. 3c, the reproduction loudspeaker was placed behind an acoustically transparent curtain (also made visually transparent for the figure). The intention behind this was that listeners would more easily be able to imagine a real, human speaker when forming their level expectations. Participants then sat down at a small desk with the computer showing the graphical user interface (GUI) for playback control.

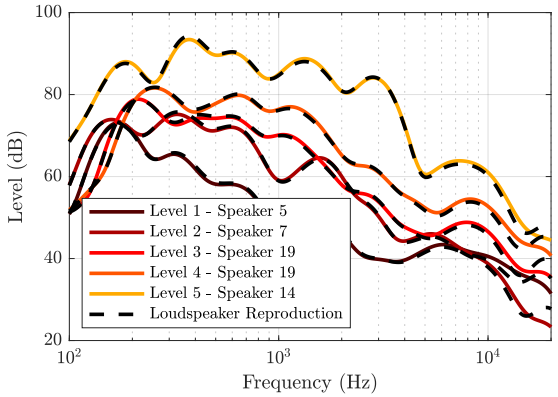


Fig. 4: Speech spectra of five randomly selected samples from different speakers comparing recorded and calibrated, reproduced speech. Except for very high frequencies above 10 kHz, recorded and reproduced sound overlap closely.

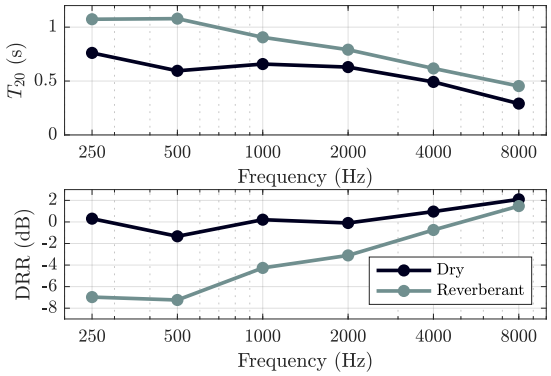


Fig. 5: Reverberation time and direct-to-reverberant energy ratio for the two room acoustic conditions.

The GUI is shown in Fig. 6. In each trial, a speech recording was presented at an initial level that was randomly chosen between −5 dB and 5 dB relative to the correct, on-axis reproduction level. Then, the task was to adjust the level to what seemed natural to the listener, i.e., to the level they would expect from a real speaker behind the curtain. To do so, four buttons were provided, two for each direction of adjustment, where the "++" / "– –" button changed the reproduction level by ±3 dB and the "+" / "–" by ±0.1 dB.

Participants were instructed outside the room and wore felt shoe covers to minimize the stepping sound when walking to the table. This was done to provide minimal information about the room acoustics before the test so that any possible room adaptation would start with the first sample of the test.

In total, each participant evaluated 50 samples. Half of them were played in the dry condition and half of them in the reverberant condition. Half of the participants listened to the dry condition first, and the other half started with the reverberant condition. There was a small break between the two parts to change the room acoustic configuration. Five samples from each of the five voice-level classes were presented in each of these room conditions. Each participant received a different selection of stimuli from the full pool of samples. At the end of the test, demographic data, experience with acoustics and audio technology, and general comments regarding used strategies were collected with a short questionnaire.

3 Results

The average age of the 15 participants who completed the test was 30.1 years (standard deviation: 7.1 years). None of the participants reported hearing loss. All except three participants had at least one year of experience doing academic research in acoustics. The median number of years of experience was 3.

The main results of the test are shown in Fig. 7. The figure shows the signed errors of the adjusted levels, i.e., 0 dB indicates that the adjusted level was equal to the true level, and a negative error states that the adjusted level was lower than the true value. As the normality and homogeneity of variance assumptions for conducting a repeated measures ANOVA were not met in the sample, non-parametric statistical tests were conducted, separately for each of the factors.

3.1 Overall Accuracy and Voice Level

The overall median error of all conditions together was −3.6 dB. The interquartile range was 6.6 dB. Consequently, participants adjusted the gain to an on-axis level that was lower than the true level that would be obtained by a real speaker. Fig. 7a illustrates that this mismatch differed between the voice level classes. Specifically, participants underestimated the level of loud samples more severely than that of quiet samples.

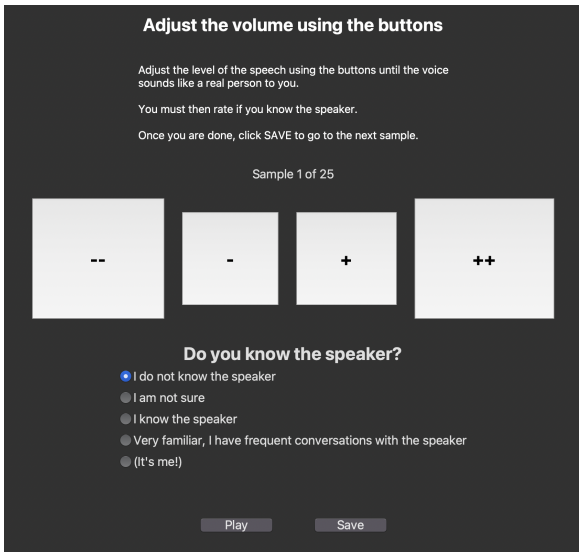


Fig. 6: User interface

While the error was only -1.7 dB for samples from voice level class 1, it increased to -6.2 dB for voice level class 5. A Kruskall-Wallis test confirms the existence of a difference between at least one pair of level classes, $\chi^2 = 71.1, p < 0.001$.

3.2 Familiarity

As different participants recognized different speakers, the number of data points collected within a given familiarity rating is not equal. These numbers together with the results are shown in Fig. 7b. Here, it becomes apparent that the error was lower for more familiar speakers. The median error for unfamiliar speakers was -5.4 dB, whereas it was between -3.3 dB and -2.8 dB for the speakers that were at least somewhat familiar. The existence of a difference in the medians is corroborated by a Kruskall-Wallis test, $\chi^2 = 29.04, p < 0.001$. The "It's me" answers were excluded from this analysis, as there were only 10 trials in which participants recognized themselves as the speaker. Including them, however, would not change the conclusion.

Besides the reduction of the median error, also the variability between answers decreased with increased familiarity. The interquartile range was 9.4 dB for completely unfamiliar and only 3.7 dB for the most familiar speakers.

3.3 Room Dependence and Adaptation

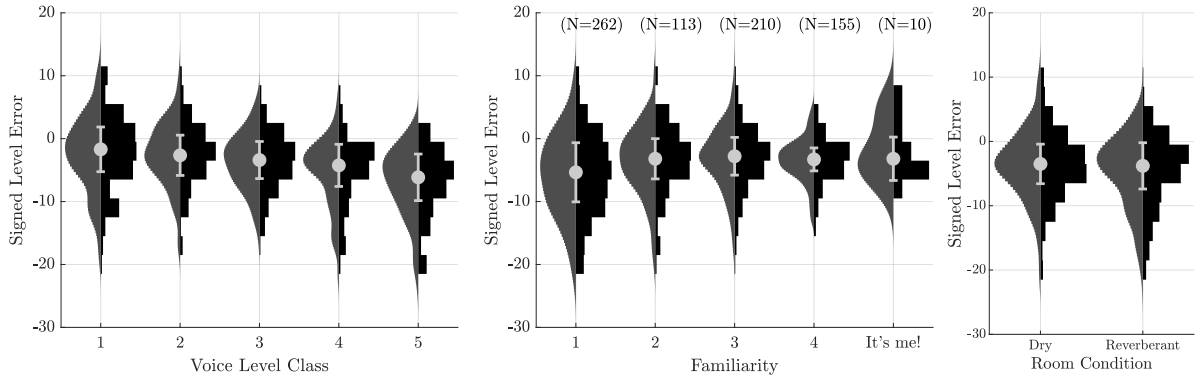
There was only a small difference in signed error between the room conditions. The median was -3.8 dB for the dry room setting and -3.5 dB for the reverberant room setting. A Mann-Whitney-U test, $Z = 1.71$, $p = 0.08$, does not show this difference to be significant.

Moreover, Fig. 8 shows the results as a function of the trial number within the two parts of the test. There is only a slight tendency for reduced errors over time. Fitting a regression line to the first set, and evaluating it at the first and the last trial index reveals that the error decreases from -5.33 dB to -4.16 dB, indicating improvement during the first set. However, this can only be considered a trend, as the regression analysis was insignificant, $R^2 = 0.0039$, $p = 0.226$. The errors at the start of the second set are -3.99 dB and slightly decrease further to -3.25 dB. Yet, again, the regression analysis was not significant yielding $R^2 = 0.0017$ and $p = 0.43$.

If adaptation to the room were a dominant factor, we would expect to see a more evident difference between the beginning and end of each of the sets, and a drop in performance after changing the room for the second set.

4 Discussion

The obtained results allow for discussion in several regards. First, it is interesting to note the overall tendency to underestimate the level, i.e., participants tend to set the playback level too low. Before searching for perceptual explanations, purely acoustical explanations should be ruled out. Using the described procedure, the on-axis level was calibrated. That is, the loudspeaker produced an on-axis level at 1.5 m that was identical to that obtained during the speech recording (see Fig. 4). One possible factor contributing to the observed underestimation of -3.6dB could be the difference between the directivity of speech and that of the loudspeaker. If the loudspeaker had a lower directivity index (DI) than real speech, this could explain the underestimation. In that case, the loudspeaker would radiate more sound in other directions besides the front compared to a human speaker. This would imply that for the same on-axis level, there would be more overall energy in the room, ultimately leading participants to set the playback to a lower level to compensate for this mismatch.



(a) Voice Level Classes **(b)** Familiarity **(c)** Room Conditions

Fig. 7: Results represented as signed error in relation to the correct level. Violinplots on the left and histograms on the right. Gray dots indicate medians and whiskers show quartiles. The level distribution in the vocal effort classes is shown in Fig. 2, the room conditions are detailed in Fig. 5 and the specific descriptions for the familiarity classes can be seen in the GUI (Fig. 6).

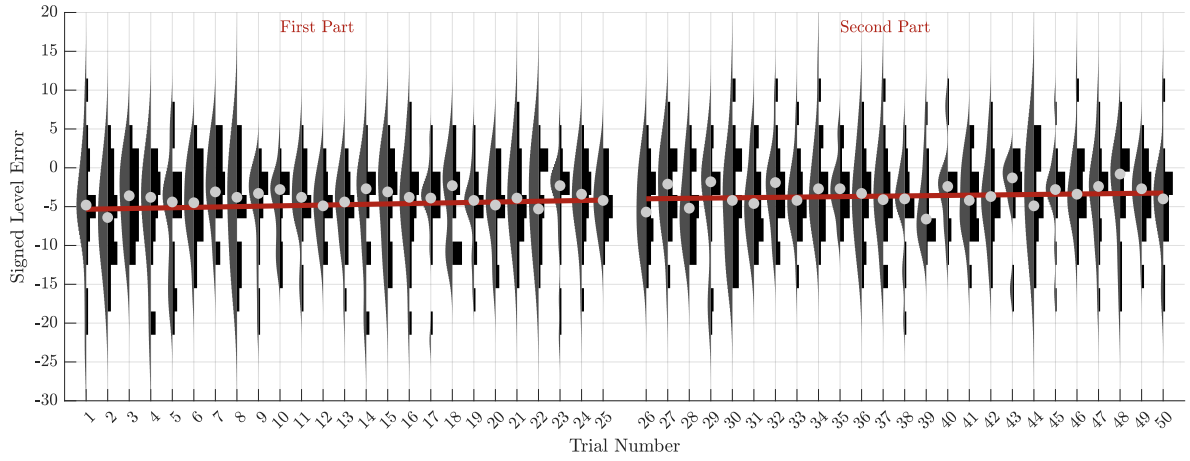


Fig. 8: Signed level error by trial number throughout the complete test. The first 25 trials correspond to the first room condition, and the following 25 to the second. There is a slight tendency for improved performance over time, albeit the correlation between performance and trial index is not significant in either of the parts.

To rule out that the underestimation could originate from this discrepancy, a comparison between the directivity index (DI) of averaged speech, as tabulated in [11], and the DI of the used loudspeaker, computed from data presented in [12], is shown in Fig. 9. The directivity index relates the level of sound radiated to the frontal direction to that radiated to all directions. It is defined as

$$\mathrm{DI}(f) = 10\log_{10}\frac{4\pi|p(f,0^\circ,0^\circ)|^2}{\int_0^{2\pi}\int_0^{\pi}|p(f,\phi,\vartheta)|^2\sin\vartheta\, d\vartheta\, d\phi}, \tag{1}$$

where p is the measured pressure, f denotes frequency, ϕ, ϑ are the azimuth and elevation angles, respectively. In practice, the integral is replaced by discrete measurements and an appropriate spherical quadrature rule.

For most frequencies, the DI of the loudspeaker turns out to be *higher* than that of speech. There is, however,

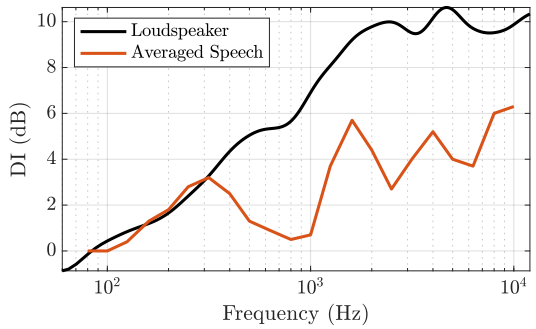


Fig. 9: Directivity index of the used Genelec loudspeaker (from [12]) and average directivity index of speech (from [11]).

a frequency region around 200-300 Hz, where the DI of the loudspeaker and the averaged speech are similar. Note that in another measurement [13], the DI for speech is generally higher than in the given measurement; however, it is not above that of the loudspeaker.

It may also be noted that the underestimation is strongest for high voice levels. This may be due to the participants unconsciously avoiding the unpleasantness of being shouted at at high levels. Also, it is unlikely that a person would shout at such high levels as presented here in some conditions when being aware of the relatively close distance to their conversation partner.

Interestingly, [6] also found slight, general underestimation. However, this underestimation was only found for unfamiliar speakers and the voice levels of the speech used for the evaluation are not indicated.

Apart from the overall level underestimation, it is worth looking into the results of the two room acoustics conditions. The DRR comparison in Fig. 5 shows that there was approximately 5 dB more direct sound energy than reverberant energy in the dry condition compared to the reverberant condition. Since the direct reproduced sound was the same for both, this implies that there was 5 dB more sound energy detectable at the receiver in the reverberant condition than in the dry condition. Nevertheless, participants adjusted the level similarly regardless of the room acoustic conditions, as shown in Fig. 7c.

Consequently, participants were able to adjust their level of expectations depending on the room's acoustic condition immediately. They seem to compensate for the fact that if the room is more reverberant, the same speech is louder, being able to separate source and space. Such loudness constancy is well known for distance experiments. The original study [14] showed that the source level was judged equally, independent of source-receiver distance, indicating that listeners were able to factor out distance when making their loudness judgment. Here, we show that participants are also able to factor out the room's acoustic properties to a large extent when forming level expectations.

Lastly, errors were lower for familiar speakers, as hypothesized. The same was found in [6] as well. In addition, the variability in the answers was lower for the most familiar speakers. This observation points to the hypothesis that absolute speech level judgment does not only depend on the internal consistency of the stimulus but also on long-term memory of non-verbal speech qualities.

In future experiments, we plan to assess level expectations in a plausibility test similar to [4]. In this case, the task would be to decide whether a real person or a loudspeaker is placed behind the curtain. The question is whether mismatched reproduction levels would decrease the number of times participants believe that there is a real person speaking. Furthermore, this could be coupled with visual rendering using a head-mounted display, as it would be the case in practical AR/VR applications. Fortunately, the used dataset also includes 3D point cloud recordings.[1]

5 Summary

To summarize the results, we have found that participants have expectations of the level of reproduced speech. When asking to adjust the level to what they would expect from a real speaker, an overall median error of -3.6 dB was found. Thus, participants believed a lower reproduced level to be correct. Such underestimation was found to be larger for samples that have a higher voice level. Moreover, errors were greatest for unfamiliar speakers, which confirms the results obtained in [2]. With some degree of familiarity between participant and speaker, error and variability decreased. Lastly, we showed that listeners' level expectations were nearly constant with respect to different room acoustic conditions and that only subtle

[1] The data is available upon request

adaptation might take place. Thus, the expectation is formed quickly, taking speech characteristics and room acoustic features into account when making a level judgment.

References

[1] Olsen, W. O., "Average Speech Levels and Spectra in Various Speaking/Listening Conditions: A Summary of the Pearson, Bennett, & Fidell (1977) Report," *Am. J. Audiol*, 7(2), pp. 21–25, 1998, doi:10.1044/1059-0889(1998/012).

[2] Traunmüller, H. and Eriksson, A., "Acoustic effects of variation in vocal effort by men, women, and children," *J. Acoust. Soc. Am.*, 107(6), pp. 3438–3451, 2000, doi:10.1121/1.429414.

[3] Sundberg, J. and Nordenberg, M., "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech," *J. Acoust. Soc. Am.*, 120(1), pp. 453–457, 2006, doi:10.1121/1.2208451.

[4] Lindau, A. and Weinzierl, S., "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acust.*, 98(5), pp. 804–810, 2012.

[5] Wirler, S., Meyer-Kahlen, N., and Schlecht, S. J., "Towards transfer-plausibility for evaluating mixed reality audio in complex scenes," in *AES International Conference on Audio for Virtual and Augmented Reality*, 2020.

[6] Wilkens, H. and Bartel, H.-H., "Wiedererkennbarkeit der Originallautstärke eines Sprechers bei elektroakustischer Wiedergabe," *Acustica*, 37, pp. 45–49, 1977.

[7] Hofmann, A., Meyer-Kahlen, N., and Schlecht, S. J., "Audiovisual Congruence and Localization Performance in Virtual Reality: 3D Loudspeaker Model vs. Human Avatar," *J. Audio Eng. Soc.*, 2024, accepted.

[8] Pearsons, K., Bennett, R., and Fidell, S., "Speech Levels in Various Noise Environments." Technical Report EPA/600/1-77/025 (NTIS PB270053), U.S. Environmental Protection Agency, Washington, D.C., 1977.

[9] Prawda, K., Schlecht, S. J., and Välimäki, V., "Evaluation of Reverberation Time Models with Variable Acoustics," in *Proceedings of the 17th Sound and Music Computing Conference*, 2020, doi:10.5281/zenodo.3903573.

[10] International Organization for Standardization, "ISO 3382-1:2009 Acoustics. Measurement of Room Acoustic Parameters. Part 1: Performance Spaces," 2009.

[11] Leishman, T. W., Bellows, S. D., Pincock, C. M., and Whiting, J. K., "High-resolution spherical directivity of live speech from a multiple-capture transfer function method," *J. Acoust. Soc. Am.*, 149, pp. 1507–1523, 2021.

[12] Gallien, A., Prawda, K., and Schlecht, S. J., "Matching early reflections of simulated and measured RIRs by applying sound-source directivity filters," 2024.

[13] Pörschmann, C. and Arend, J. M., "Analyzing the Directivity Patterns of Human Speakers," in *DAGA*, Hannover, 2020.

[14] Zahorik, P. and Wightman, F. L., "Loudness constancy with varying sound source distance," *Nat. Neurosci.*, 4(1), pp. 78–83, 2001, doi:10.1038/82931.