forum acusticum 2026
8-12 September · Graz, Austria

SPEECH RECOGNITION AND LISTENING EFFORT WITH REAL AND VIRTUAL SPEECH IN VARIABLE ACOUSTICS

Sara Ghorbani[1], Nils Meyer-Kahlen[1, 2], Johannes M. Arend[1]

[1] *Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland*
[2] *Dyson School of Design Engineering, Imperial College London, London, UK*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Augmented reality (AR) telepresence systems often aim to enhance immersion by adding room-adapted reverberation to remote speech, potentially at the cost of reduced intelligibility and increased listening effort. In this study, we investigated speech recognition and subjective listening effort in a speech-in-noise task across real and virtual acoustic environments with varying degrees of reverberation. Twenty participants performed a sentence recognition task in babble noise under seven conditions, including dry, reverberant, and highly reverberant scenarios, as well as mixed conditions in which the virtualized target remained dry rather than being matched to the acoustic environment in which the noise was placed. Results confirmed that increased reverberation reduces word recognition and increases listening effort. Notable discrepancies were observed between real and virtual presentation, particularly in dry conditions. Importantly, omitting room-adapted reverberation for the virtualized target speech provided only limited benefits for intelligibility and effort. These findings highlight trade-offs in AR audio design and suggest that the acoustics in which the background noise is placed may be more critical than that applied to the target speech regarding listening effort.

Keywords: *Augmented Reality, Telepresence, Speech-In-Noise, Listening Effort, Room Acoustics*

1. Introduction

The goal of augmented reality (AR) telepresence systems is to virtually place a remote interlocutor in a user's space, both visually and acoustically. Visually, this is achieved by displaying a world-locked three-dimensional "hologram" using semi-transparent AR glasses or intransparent head-mounted displays with camera-based pass-through. Acoustically, this is accomplished using head-tracked binaural rendering that renders room acoustics adapted to the user's acoustic environment.

In essence, room-adapted rendering adds matching reverberation to the remote-end speech signal with the aim of enhancing the user's holographic calling experience and creating a stronger sense of presence. Interestingly, the opposite objective—removing reverberation from remote-end speech—has long been an important part of many traditional telecommunication systems, aiming to avoid degraded speech intelligibility [1] and increased listening effort. In this context, listening effort refers to the mental exertion and cognitive resources required to understand speech, particularly under acoustically challenging conditions [2], [3].

Various studies have shown that listening to speech in noisy environments is more challenging if it takes place in a reverberant space. For example, Kuusinen *et al.* reported that speech reception thresholds and pupil-dilation-based listening effort measures were significantly higher in reverberant conditions than in dry conditions, indicating higher perceptual and cognitive demands [4]. Similarly, Lavandier and Culling showed that increasing reverberation reduced spatial unmasking benefits and generally worsened speech reception thresholds in noise [5]. Holube *et al.* showed that, when approximately matched by speech transmission index, hard reverberation-only conditions could elicit even higher subjective listening-effort ratings than hard noise-only conditions [6]. Assuming that such adverse effects of reverberation on intelligibility and listening effort exist, AR telepresence systems face an important trade-off: while adding room-adapted reverberation may improve the experience and enhance the sense of social presence, it may negatively affect speech intelligibility and increase listening effort.

Here, we present an experiment to assess how severe such problems might be when adding room-adapted acoustic rendering to a binaural presentation of speech.

***Corresponding author**: sara.ghorbani@alumni.aalto.fi.*

Since we did not expect high listening effort in a quiet, single-talker environment, regardless of the room acoustic conditions, we placed the target speech in surrounding background babble noise. This setup mimicked the situation of having an AR call in a noisy space. We then performed a classical speech-in-noise test, in which participants were instructed to repeat the sentence uttered by the target speaker. In addition, they rated subjective listening effort. Eye-movement data, recorded using a Meta Quest Pro headset worn by participants during the experiment, were also collected but are not described in detail here. Our main aim was to investigate the effect of adding matched reverberation to the target speech, as it would potentially be done in AR, compared to adding the acoustics of a drier space than what room adaptation would demand. The test was performed in a variable-acoustics environment, where different acoustic conditions could be created physically and virtually. However, since the room did not permit extreme reverberation conditions, we also tested cases in which both the target and the background noise were rendered virtually. In summary, we addressed the following research questions:

RQ1 Do more reverberant conditions yield lower intelligibility and higher self-reported listening effort than dry conditions?

RQ2 Can we virtualize a listening effort experiment, i.e., does a speech-in-noise test using real loudspeakers, or a virtual rendering thereof, provide the equivalent results?

RQ3 Does omitting room-adapted target rendering in favor of dry target speech rendering improve intelligibility and reduce effort?

With the results of the experiment, we aim to inform the design of AR audio systems that add reverberation to virtual speech.

2. Methods

2.1. Setup

The hardware setup consisted of a ring of loudspeakers placed around the listener, as illustrated in Fig. 1. One loudspeaker (a Genelec 8331A) was placed at a distance of 2.16 m directly in front of the listener (0° azimuth), at a height of 1.20 m. During the experiment, it was used to reproduce the target speech. In addition, eight Genelec 8030A loudspeakers were arranged at azimuth angles of ±35°, ±70°, ±105°, and ±145°, each approximately 2 m away from the listener, to reproduce spatially distributed babble noise during the experiment. Playback was controlled from a laptop running Pure Data and an RME MADIface XT audio interface, which in turn was connected to an RME M-16 AD digital-to-analog converter.

2.2. Virtual Rendering

To create the virtual listening conditions, Binaural Room Impulse Responses (BRIRs) and Headphone Transfer Functions (HpTFs) were measured with the same loudpeaker setup. The measurements were conducted using 4-second long logarithmic sine sweeps, generated in MATLAB at a sampling rate of 48 kHz. A GRAS KEMAR head and torso simulator (HATS) equipped with calibrated microphones was positioned at the listener location. MushRoom headphones [7] were used, which are designed to be especially transparent to real-world sounds. They were worn by the HATS during the BRIR measurements, so that the effect of the headphones on real sounds is included in the rendering as well. However, during the experiment, participants also wore a Quest Pro headset for eye gaze analysis, which is not presented here.

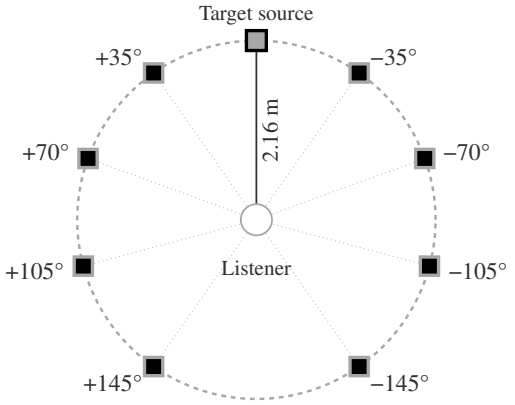
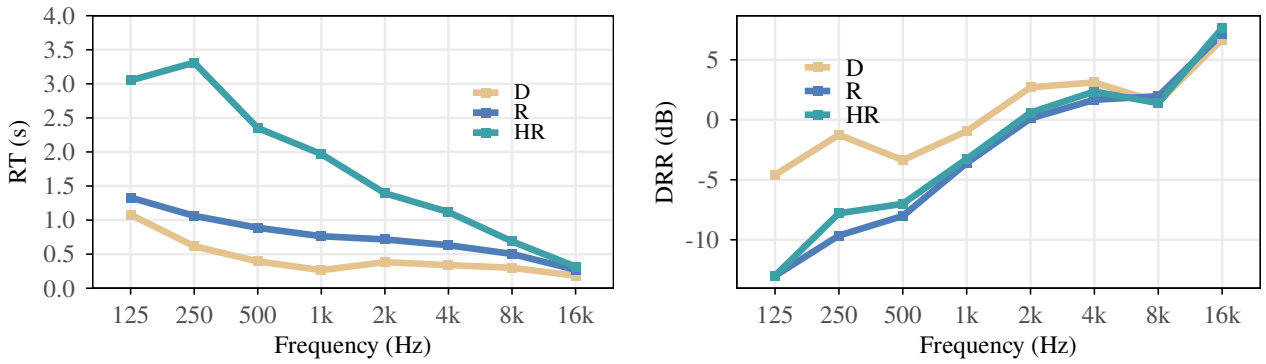


Figure 1: Schematic of the loudspeaker array used in the real-room measurements and rendering conditions. The listener was located at the center of the array.

2.3. Rendering Conditions

The experiment was conducted in the variable acoustics room "Arni" at Aalto University, which allows modification of the room acoustic characteristics by opening or closing wall panels, mounted in front of boxes with absorbing material.
During the experiment, the room was used in two conditions: an open-panel configuration providing dry acoustics, and a closed-panel configuration yielding a more reverberant sound. These conditions are referred to as ("R:D", "R:R") in Tab. 1, where the first "R" stands for "real". To assess RQ1, the dry and the reverberant states are compared. The same acoustic states were measured using the described procedure to create virtual rendering conditions ("V:D", "V:R"). In all virtual conditions (starting with "V"), the physical room was kept in the "dry" state during the test.
The reverberation time (RT) difference between the two room states is clearly noticeable, with 0.33 s at mid-frequencies in the dry state and 0.83 s in the reverberant state; see Fig. 2 for frequency-dependent RTs and direct-to-reverberant energy ratios (DRRs). As other studies on listening effort tested much larger RT differences, we decided to include a third condition. Given that all panels were already closed for the reverberant state, it could not be physically created in Arni. Also, it was impossible to move the complete setup between rooms during the experiment. Instead,



(a) **(b)**

Figure 2: Octave-band a) reverberation time, and b) DRR for the three measured setups used for virtualization: dry (D), reverberant (R), and highly reverberant (HR).

the configuration was moved to a more reverberant room only for a BRIR measurement. The room was a hallway with a mid-frequency RT of approximately 2.16 s (see Fig. 2). We reasoned that if we see no differences between the real and the virtual conditions for the two states in Arni (RQ2), we could assume that the virtual rendering approach is valid, and including the highly reverberant hallway only virtually ("V:HR") would be acceptable.

As the main objective was to gain insights into AR scenarios, where the question is whether room-adapted reverberation should be added, two mixed conditions were included in the test. For "V:D/R", a dry target was used, while the background noise was in a reverberant room. This condition mimics an AR case in which a designer has chosen to apply drier rendering to the target than room adaptation would demand. Comparing the results to "V:R", where both noise and target are in a reverberant state, will show whether there are potential benefits for intelligibility in forfeiting room-adaptation (RQ3). The same idea is behind V:D/HR, where the highly reverberant space was used for the background noise, while a dry rendering was used for the target. Note that in these two mixed conditions, the noise is virtual, due to the inability to move between Arni and the hallway during the test. In practical AR scenarios, the background noise would be real. In total, seven listening conditions were tested: two real and five virtual renderings.

Table 1: Experimental conditions and abbreviations used in the listening experiment. R = real room, V = virtual room, D = dry, R = reverberant, HR = highly reverberant.

Cond.	Target / Noise	Abbrev.
Real Room (Loudspeakers)		
1	Dry	R:D
2	Reverberant	R:R
Virtual Room (Headphones): Target / Noise		
3	Dry / Dry	V:D
4	Reverberant / Reverberant	V:R
5	Highly reverb. / Highly reverb.	V:HR
6	Dry / Reverberant	V:D/R
7	Dry / Highly reverb.	V:D/HR

2.4. Source Signals

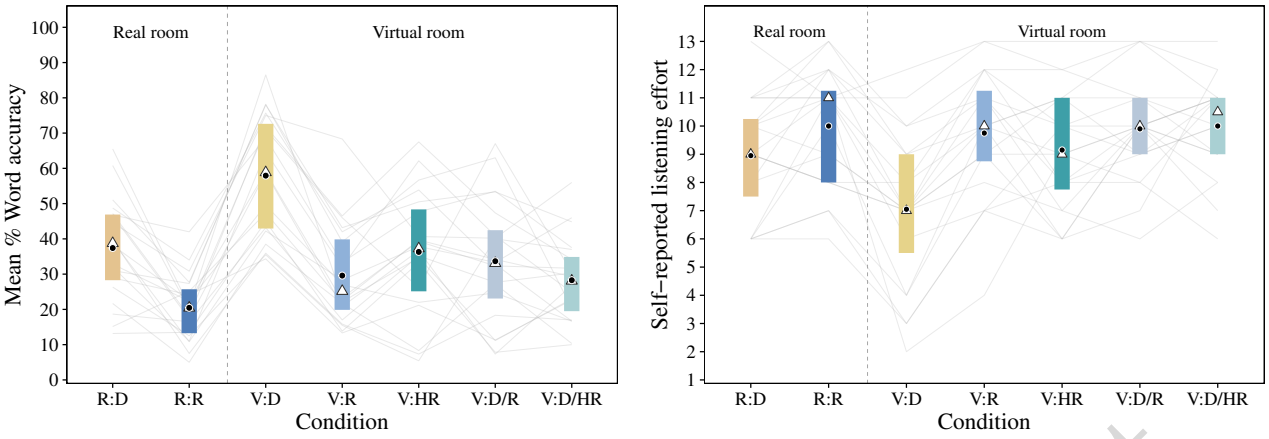
The target speech material was recorded in the large anechoic chamber "Lampio" at the Aalto Acoustics Lab, using a calibrated GRAS 40HF low-noise microphone. The speaker was a male native British English talker from our lab. Speech material was derived from the CSTR VCTK Corpus developed at the University of Edinburgh [8], which contains approximately 400 sentences. From this corpus, 170 short sentences were selected, and those with durations under two seconds were used in the listening test for consistency between trials.

A multi-talker babble noise masker was generated from a speech corpus in which 21 speakers were recorded in the same anechoic room as the target speech using the same calibrated microphone [9]. To create the babble noise, each speech file was segmented by randomly extracting 1 s snippets. To reduce abrupt onsets and offsets, each snippet was windowed using a Tukey window with a 20% cosine taper, corresponding to 100 ms fade-in and 100 ms fade-out per snippet. These snippets were then placed at random starting times within a 10 s noise buffer, and overlapping snippets were summed. This procedure was repeated independently for each of the eight noise loudspeaker channels, resulting in a multi-channel babble noise masker.

Playback levels were calibrated at the listener position, with the babble noise presented at approximately 65.2 dB LAeq and the target speech at 58.2 dB LAeq. This corresponds to a signal-to-noise ratio (SNR) of −7 dB in all conditions.

2.5. Experimental Procedure

All participants provided informed consent prior to testing. The condition order was randomized to minimize the influence of order effects on the results. Each participant heard seventy distinct, non-repeated sentences (ten trials per condition). After each trial, participants repeated the sentence aloud, and their



(a) Word accuracy **(b)** Self-reported listening effort

Figure 3: Results for (a) percentage of correctly recognized words and (b) subjective listening effort rating for each listening condition. Thin gray lines connect individual participants across conditions, thick colored bars indicate the interquartile range (IQR), white triangles mark the median, and black circles mark the mean.

verbal responses were recorded using a Rode NT1-A microphone for later speech recognition analysis. Participants rated their perceived listening effort on a 13-point categorical scale ranging from 1 ("no effort") to 13 ("extreme effort") after each condition. The scale included seven labeled categories and six unlabeled intermediate categories, following earlier listening-effort studies [10], [11]. A short questionnaire was also presented at the end of the session.

2.6. Participants

21 students (14 males, 7 females) participated in the experiment. One participant was excluded due to technical issues, resulting in a final sample of 20 participants. The participants were between 21 and 39 years old ($M = 27.55$ years). Their experience with acoustics ranged from none to eight years, with 13 participants (65%) reporting more than two years of experience and five (25%) reporting no prior experience. All participants reported normal hearing, except for one who reported mild bilateral tinnitus. The majority (85%) of the participants were non-native English speakers. Only three identified as native English speakers, and three others classified their English proficiency as intermediate or lower. 56% of the participants reported being familiar with the speaker.

2.7. Data Analysis

Audio responses were transcribed with the Whisper-Large automatic speech recognition model [12]. Speech intelligibility was then quantified for each trial as the word accuracy (WA), computed from the word error rate (WER) based on the word-level Levenshtein distance (LD) between the lowercased, tokenized transcription and reference sentence [13]. The LD corresponds to the number of word insertions, deletions, and substitutions required to transform the transcription into the reference. Specifically,

$$\mathrm{WA} = (1 - \mathrm{WER}), \quad \text{with} \quad \mathrm{WER} = \frac{\mathrm{LD}}{N_{\mathrm{ref}}},$$

where N_{ref} is the number of words in the reference sentence. To avoid negative scores when $LD > N_{\mathrm{ref}}$, WA values were bounded below at 0%.

To answer our research questions, we conducted Bayesian data analysis on WA and the subjective ratings. For WA, we calculated a difference score between two given conditions for each participant. We then fit a Bayesian intercept-only model with Gaussian likelihood using *brms* [14], where the intercept represents the population mean difference score. A weakly informative Student-t prior $t(3, 0, 15)$ was placed on the intercept. The standard deviation had a half-student-t prior with the same parameters. Furthermore, we assume that the ordinal 13-point self-report scale can be approximated as an interval scale, so that we used the same procedure, with a student-t prior $t(3, 0, 2)$ on the intercept and the corresponding half-student-t prior on the standard deviation. For both models, four chains with 2000 samples each were used, where the first 1000 samples were discarded as warm-up.

The posterior distribution of the inferred intercept was used to estimate the mean paired contrast. For analyzing the difference between dry and reverberant conditions (RQ1), contrasts along with the credibility intervals (CrIs) and the probability of direction (pd) are presented. For testing the equivalence between real presentation and virtual renderings (RQ2), we perform Region Of Practical Equivalence (ROPE) tests, defining WAs between -5% and 5% and subjective scores with ± 0.5 points as negligibly small effects in the context of this study. ROPE was also computed for comparing cases with fully reverberant rendering to those where the target was kept drier (RQ3).

3. Results

The results for WA and self-reported effort are shown in Figs. 3a and 3b, respectively. Mean WAs between approximately 20% and 58% indicate that the task was not trivial, but not impossible either. The mean listening effort across conditions was about 9.26 on the

13-point scale. Some differences between conditions can be observed, especially in the WA results, which are now analyzed in more detail.

3.1. Reverberation (RQ1)

When comparing the WA between the two real acoustic conditions, one dry and one reverberant ("R:D" vs "R:R"), a clear difference was found. The mean posterior difference was 16.29% with a CrI of $[9.78\%, 22.75\%]$, which can be regarded as a credible result as the CrI clearly excludes 0%; pd was exactly 1. In terms of self-reported effort, the difference was also evident. The estimated posterior mean difference was 0.99 rating points with a CrI of $[0.12, 1.87]$. The estimated pd was 0.97.

3.2. Real/Virtual Comparison (RQ2)

When comparing real and virtual presentations (i.e., "R:D" vs. "V:D" shown in yellow and "R:R" vs. "V:R" shown in blue in Fig. 3), a difference in WA is observed as well. For the dry condition, the difference was the strongest. The estimated mean difference between virtual dry and real dry was 19.9% with a CrI of $[13.62\%, 25.80\%]$ and a pd of 1. None of the posterior samples were in the ROPE.

A clear difference was also found for self-reported effort, with a posterior mean of 1.8 (CrI: $[0.79, 2.82]$, pd ≈ 1); no samples were within ROPE, either.

For the comparison between real and virtual in the reverberant condition, the difference was smaller, but still present. Mean WA differed by 8.82% (CrI: $[3.31, 14.57]$, pd $\approx$ 1) with 5% of samples within ROPE. For the effort ratings, no credible difference was found; the mean was 0.23 (CrI: $[-0.87, 1.38]$, pd: 0.65). 62.63% of samples were within the ROPE.

3.3. Mixed Target/Noise Conditions (RQ3)

We now compare the completely reverberant virtual rendering with a case in which the dry room rendering was used for the target speech. For the reverberant room (i.e., "V:R" vs. "V:D/R"), shown as light blue boxes in Fig. 3a, the difference in WA was small. The mean was 3.85 (CrI: [2.24, 9.88], pd: 0.90). 65.71% of samples were within the ROPE, meaning that participants became only slightly better at recognizing the speech when the target was kept dry, but there is insufficient evidence for stating equivalence either.

In the case of the highly reverberant room (i.e., "V:HR" vs. "V:D/HR"), the difference in WA was small again, but this time in the opposite direction. Participants tended to make more errors when the target was kept dry: the estimated absolute mean difference was 7.28 (CrI: [-1.25, 15.35], pd: 0.96); 28.34% of samples were within the ROPE.

In the case of self-reported effort, there were no notable differences. For the reverberant room, the mean difference was 0.15 (CrI: [-1.25, 0.94], pd: 0.61, 64.87% in ROPE). For the highly reverberant room, the trend was in the opposite direction again, with forfeiting room matching making the task marginally more effortful; the absolute mean difference was 0.78 (CrI: [-0.19, 1.78], pd: 0.93, 26.26% in ROPE).

4. Discussion

Our study confirms that reverberation clearly affects word recognition and, to a lesser extent, listening effort when listening to speech in noise. This confirms the results of other studies.

Notably, the interpretation of the absolute WA values should consider participant-related factors. Absolute WA was likely lowered by the high proportion of non-native listeners (85%), compared to native listeners. Non-natives can require up to 6 dB SNRs than native listeners to achieve the same recognition performance for complex sentence material [15]. Conversely, 56% of participants were familiar with the speaker, which may have provided an advantage in recognition.Neither of these factors should impact the within-participant contrasts across the experimental conditions.

4.1. Explanations for Real/Virtual Differences

Surprisingly, we found a difference between real and virtual presentation of the speech in noise test; the difference was most pronounced for the dry condition. Since playback levels were calibrated across conditions, this real/virtual mismatch is unlikely to be due to overall level differences. Yet, with the steep psychometric functions in the context of speech recognition in noise, small acoustic differences are enough to distort the result. One possible explanation is the difference in directivity index between the ears of the HATS and those of most listeners. Also, even though the acoustic effect of the headset that people wore during the test is expected to be small, it may have worsened the SNR in real conditions, especially since headset-induced impairments are primarily seen for frontal directions [16]. Another explanation might be that the non-individual rendering of the frontal source was perceived as more internalized than the other sources and was therefore easier to identify. Further studies are required to understand the effect.

Moreover, one methodological choice of our study could be investigated in more detail. Even though the microphone used to record the participants' answers was placed close to them, the stronger physical reverberation in the "R:R" case might have worsened automatic transcription performance. If this were the case, the difference between "R:R" and "V:R" may be reduced. All other conditions are unaffected, as the physical state was always "dry" except for "R:R".

4.2. AR Implications of Mixed-Case Results

In the context of the result just discussed, the results for RQ3 must be interpreted with caution. Here, the noise was virtual, as the highly reverberant room that was not physically available during the test could be included. In a practical AR scenario, the noise would be real. As we saw differences between real and virtual presentations (RQ2), the results for using matched rendering and more dry rendering for the target (RQ3) are not fully representative of an AR use case. Nevertheless, they provide a first insight into

whether changing reverberation only for the target has an effect, and the answer is rather clear: There was an effect for one condition, but it was much smaller than the effect of placing both target and noise in different environments. For the highly reverberant condition, there was even a trend in the opposite direction. This highlights that our everyday experience of it being more effortful to listen to a target speaker in a reverberant environment may not depend primarily on target reverberation, but rather on the reverberant background noise, which adds overall noise energy to the acoustics field and reduces the ability to benefit from binaural unmasking, thereby limiting the perceptual separation of target speech and masker.

4.3. Other collected listening effort measures

In addition to subjective listening effort, we also examined eye-movement metrics and physiological measures, specifically electrodermal activity and electrocardiography. While the different measures were sensitive to the temporal structure of the auditory events, no statistically significant differences were observed between the conditions. This demonstrates that further research is required for obtaining reliable eye movement-based and physiological measures of relatively small differences in listening effort.

5. Conclusion

This study examined speech intelligibility and listening effort in a speech-in-noise test in various real and virtual acoustic environments. The results clearly showed differences between acoustic conditions in both outcome measures. However, there were also differences between real and virtual presentations that warrant further investigation. When keeping the target dry, in a reverberant noise rendering, the benefit was small to non-existent. However, in the former case, it was inconclusive whether adding room-matched reverberation is entirely harmless in AR applications.

6. Acknowledgments

Thanks to Ruijie Wang for technical support and to all participants. We acknowledge the use of HUCE infrastructure of Aalto School of Electrical Engineering.

References

[1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation* (Signals and Communication Technology). London: Springer, 2010.

[2] D. J. Strauss and A. L. Francis, "Toward a taxonomic model of attention in effortful listening," en, *Cogn. Affect. Behav. Neurosci.*, vol. 17, no. 4, pp. 809–825, 2017.

[3] M. B. Winn and K. H. Teece, "Listening Effort Is Not the Same as Speech Intelligibility Score," en, *Trends Hear.*, vol. 25, 2021.

[4] A. Kuusinen, K. Kondraciuk, and T. Lokki, "Effects of Masker Type and Reverberation on Speech-in-Noise Recognition Thresholds and Listening Effort as Indexed by Pupil Dilation Responses," in *Proc. 154th Audio Eng. Soc. Conv.*, Espoo, Finland, 2023.

[5] M. Lavandier, S. Jelfs, J. F. Culling, A. J. Watkins, A. P. Raimond, and S. J. Makin, "Binaural Prediction of Speech Intelligibility in Reverberant Rooms with Multiple Noise Sources," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 218–231, 2012.

[6] I. Holube, K. Haeder, C. Imbery, and R. Weber, "Subjective listening effort and electrodermal activity in listening situations with reverberation and noise," *Trends Hear.*, vol. 20, 2016.

[7] A. Mülleder and F. Zotter, "Ultralight circumaural open headphones," in *Proc. 154th Audio Eng. Soc. Conv.*, Espoo, Finland, 2023.

[8] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," University of Edinburgh, Tech. Rep., 2019.

[9] A. Hofmann, N. Meyer-Kahlen, and S. J. Schlecht, "Audiovisual Congruence and Localization Performance in Virtual Reality: 3D Loudspeaker Model vs. Human Avatar," *J. Audio Eng. Soc.*, vol. 72, no. 10, pp. 679–690, 2024.

[10] J. Rennies, H. Schepker, I. Holube, and B. Kollmeier, "Listening effort and speech intelligibility in listening situations affected by noise and reverberation," *J. Acoust. Soc. Am.*, vol. 136, no. 5, pp. 2642–2653, 2014.

[11] H. Luts et al., "Multicenter evaluation of signal enhancement algorithms for hearing aids," *J. Acoust. Soc. Am.*, vol. 127, no. 3, pp. 1491–1505, 2010.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, pp. 28 492–28 518.

[13] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[14] P.-C. Bürkner, "**brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*," *J. Stat. Softw.*, vol. 80, no. 1, 2017.

[15] A. Warzybok, T. Brand, K. C. Wagener, and B. Kollmeier, "How much does language proficiency by non-native listeners influence speech audiometric tests in noise?" *Int. J. Audiol.*, vol. 54, no. sup2, pp. 88–99, 2015.

[16] P. Lladó, T. Mckenzie, N. Meyer-Kahlen, and S. J. Schlecht, "Predicting Perceptual Transparency of Head-Worn Devices," *J. Audio Eng. Soc.*, vol. 70, no. 7, pp. 585–600, 2022.