Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments

Ahrens, Axel; Marschall, Marton; Dau, Torsten

Published in:
Hearing Research

Link to article, DOI:
10.1016/j.heares.2019.02.003

Publication date:
2019

Document Version
Version created as part of publication process; publisher's layout; not normally made publicly available

Link back to DTU Orbit

Citation (APA):
Research Paper

Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments

Axel Ahrens*, Marton Marschall, Torsten Dau

Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Bygning 352, Ørsteds Plads, 2800, Kgs. Lyngby, Denmark

ARTICLE INFO

Article history:
Received 30 November 2018
Received in revised form 6 February 2019
Accepted 12 February 2019
Available online xxx

Keywords:
Speech intelligibility
Virtual audio
Modeling

ABSTRACT

Loudspeaker-based virtual sound environments provide a valuable tool for studying speech perception in realistic, but controllable and reproducible acoustic environments. The evaluation of different loudspeaker reproduction methods with respect to perceptual measures has been rather limited. This study focused on comparing speech intelligibility as measured in a reverberant reference room with virtual room reproductions using microphone array recordings and mixed-order ambisonics. Three factors that affect speech intelligibility were varied: reverberation, the spatial configuration and the type of the interferers (speech or noise). Two interferers were placed either colocated with the target, or were symmetrically or asymmetrically separated. The results showed differences between the reference room and the simulation-based reproductions when the target and the interferers were spatially separated but not when they were colocated. The reproduction utilizing the microphone array was most similar to the reference room in terms of measured speech intelligibility. Differences in speech intelligibility could be accounted for using a binaural speech intelligibility model which considers better-ear signal-to-noise ratio differences and binaural unmasking effects. Thus, auditory modeling might be a fast and efficient way to evaluate virtual sound environments.

© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

One of the challenges in hearing research is to understand the mechanisms involved in speech perception in complex acoustic scenarios, such as in a restaurant or at a social gathering, commonly referred to as a “cocktail-party” scenario (Bronkhorst, 2000; Cherry, 1953). To study the factors influencing speech perception in a given acoustic environment in a controllable and reproducible manner, virtual sound environments (VSEs) provide a valuable tool. For example, loudspeaker-based VSEs can reproduce acoustic scenes in a laboratory to investigate how the auditory system functions in realistic listening scenarios.

Using such a system, Koski et al. (2013) compared speech reception thresholds (SRTs) in a multi-talker scenario measured in a reference room, with corresponding SRTs measured in virtual room reproductions using microphone array recordings and directional audio coding (Pulkki, 2007). An increase of the SRT of up to 2 dB (i.e. decreased speech intelligibility) was found in some of the virtual conditions, but no significant differences appeared in the highest fidelity reproduction setup, which used up to nine loudspeakers and an anechoic reproduction room. Instead of microphone array recordings, Cubick and Dau (2016) used room acoustic simulations and a combination of higher-order ambisonics (HOA; Gerzon, 1973) and an approach to map early sound reflections to the nearest loudspeakers (NLM; Favrot and Buchholz, 2010). The setup included a target talker in the front direction and three speech-shaped noise interferers behind the listener. Speech intelligibility measurements revealed a 2 dB higher SRT in the virtual room, relative to the reference room, when using the NLM approach, and a 4 dB higher SRT when the reproduction was based on HOA. In contrast to the speech intelligibility results, classical room acoustic measures, i.e. reverberation time, clarity and interaural cross-correlation, were found to be similar in the virtual room and in the reference room, showing that these parameters are not sensitive enough to reveal differences in certain conditions. Whereas the studies of Koski et al. (2013) and Cubick and Dau (2016) used noise as interfering signals, Oreinos and...
Buchholz (2016) employed seven conversational pairs of talkers in a reverberant reference room. The reference room was reproduced either using a simulation-based NLM approach, as in Cubick and Dau (2016), or a HOA microphone array recording and reproduction technique. High correlations between the SRTs measured in the real and the virtual rooms were obtained. However, the simulation-based NLM approach led to lower SRTs and the microphone array-based HOA approach to higher SRTs than those obtained in the reference room.

Overall, the studies of Koski et al. (2013), Cubick and Dau (2016) and Oreinos and Buchholz (2016) demonstrated that, while speech intelligibility measures in VSEs provide a reasonable correlation with corresponding measurements in the real environment, deviations remained which have not yet been resolved. The goal of the current study was to further analyze these discrepancies between real and virtual environments, as well as the differences observed across the different reproduction methods, in relation to several main factors influencing speech intelligibility. Specifically, the effects of (i) masking of different types of interferers, (ii) their spatial positions relative to the target speech signal as well as (iii) the amount of reverberation in the environment on the intelligibility of a target speech were investigated. This was done by measuring SRTs in multiple conditions. In terms of the effects of speech masking, both speech interferers with a high similarity to the target speech and speech-modulated, spectrally-matched noise interferers were considered. While the speech masker was assumed to produce some amount of informational masking (IM; e.g. Brungart et al., 2001; Watson, 2005), the noise masker was considered to produce mainly energetic masking and only little IM.

The influence of the spatial separation of the interferers was examined by considering three spatial conditions: a “colocated” condition, where the target and two interferers were presented from the frontal direction; a condition with “symmetrically separated interferers”, where the target was in the front and the interferers at ±30° azimuth; and an “asymmetric interferer condition”, where the two interferers were presented from the same location at 30° azimuth. Finally, the effect of reverberation was investigated by considering SRTs in an anechoic control condition, a reverberant reference room and virtual versions of the reference room. Three reproduction methods were considered in the present study. The reference room was either reproduced based on room acoustic simulations and rendered using NLM or HOA, similar to Cubick and Dau (2016), or based on impulse response measurements obtained with a HOA microphone array, as in Oreinos and Buchholz (2016). The stimuli were played back using a spherical loudspeaker array installed in an anechoic chamber.

To characterize the virtual rooms objectively, classical room acoustic measures were employed, such as the reverberation time. Furthermore, the computational speech intelligibility model of Jølfs et al. (2011) was considered to compare the predicted SRTs in the different conditions and to analyze the differences between the cues underlying speech intelligibility in the framework of the model.

2. Methods

2.1. Reference room

A standard listening room (IEC 268-13, 1985), reflecting the acoustics of a living room, with a volume of 100 m³ (7.52 m × 4.75 m × 2.8 m) and an average reverberation time of 0.4 s, was chosen as the reference environment for this study. The wooden floor of the room is covered with a carpet, the plastered walls are partially covered with different acoustic panels and diffusers and the ceiling is fully covered with acoustic panels. The acoustical properties of the room are unknown and were estimated (find the room model estimates in the accompanying dataset). The listening position was centered along the longest dimension of the room and 1.35 m from the back wall (see Fig. 1). The talkers were imitated using Dynaudio BM60 (Dynaudio A/S, Skanderborg, Denmark) loudspeakers, driven by custom-made amplifiers and a RME FIREFACE 800 (Audio AG, Haimhausen, Germany) sound card. The loudspeakers were located at 2.4 m distance from the listener at 0° and at ±30° azimuth and were placed approximately at ear level (h = 1.17 m, from the floor to the center of the woofer of the loudspeaker).

2.2. Acoustic scene generation and recording

The reference room was reproduced using two alternative approaches. Room acoustics were either simulated using a commercially available acoustic simulation software, or captured by recording impulse responses using a spherical microphone array.

To simulate the acoustics of the listening room, a geometrical model of the room was constructed in the room acoustics software Odeon version 13.04 (Odeon A/S, Kgs. Lyngby, Denmark), including the same source and receiver/listener positions as in the reference room. The directivity and frequency response of the loudspeakers were incorporated in the model as in Cubick and Dau (2016). The absorption coefficients of the room surfaces were optimized from initial estimates of the surface materials, using the Odeon genetic material optimizer (Christensen et al., 2014). The optimization was performed by employing measured reverberation times (T20, T30), as well as early decay time and clarity (C7, C50, C80) parameters as calculated (ITA-toolbox; Berzborn et al., 2017) from impulse responses measured in the reference room. The details of the impulse response measurement procedure are described below. The optimized absorption coefficients did improve the room acoustics model with respect to the measurements, however the error remained larger than previously reported by Christensen et al. (2014). The reason for the larger error is likely due to the size of the room used in the current study, which is small in comparison to rooms generally modeled using Odeon. From the optimized room acoustics model, direct sound, early reflections and energy decay curves were exported in eight octave bands from 63 Hz to 8 kHz and processed using the Loudspeaker-Based Room Auralization toolbox (LoRA; Favrot and Buchholz, 2010) to obtain impulse responses for each loudspeaker in the VSE. The optimized room acoustics model can be found in a dataset (Ahrens, 2018). Two processing strategies implemented in LoRA were applied: a nearest-loudspeaker mapping (NLM) and a mixed-order ambisonics (MOA) coding strategy. The NLM approach maps the direct sound and each of the early reflections to the geometrically closest loudspeaker. Late reflections were reproduced with energy envelopes represented in 1st order ambisonics and multiplied with uncorrelated noise for each loudspeaker (Favrot and Buchholz, 2010). For MOA, the same strategy was used for the late reflections as for the NLM. The direct sound and the early reflections were encoded using 7th order horizontal and 5th order periphanic
ambisonics. The loudspeaker signals were obtained from the MOA signals using a dual-band mode matching/max-rE decoder (Daniel, 2001), with a crossover frequency of 4 kHz.

Measurements in the reference room were undertaken with a 52-channel spherical microphone array with a radius of 5 cm (Marschall et al., 2012). Impulse responses (IRs) were recorded between the three source positions with the Dynaudio BM6P loudspeakers and the microphone array placed at the listening position. The IRs were measured using eight 16 s long logarithmic sweep signals (Müller and Massarani, 2001). The same MOA orders were used for encoding the array signals as for the simulations (7th order horizontal, 5th order periphronic). From the ambisonics components the loudspeaker signals were obtained using a dual-band mode matching/max-rE decoder (Daniel, 2001; Marschall, 2014) as for the simulation-based reproduction and a regularization parameter of $\lambda = 0.01$ (Marschall et al., 2012).

The two room acoustic simulation-based reproduction strategies are termed “simulated NLM” and “simulated MOA” and the microphone array recording-based reproduction is termed “recorded MOA” throughout the article.

### 2.3. Virtual sound environment (VSE)

The virtual sound environment consists of a spherical array of 64 loudspeakers located in an anechoic chamber (7 m*8 m*6 m), with the listener’s head positioned in the center of the sphere of 2.4 m radius. A depiction of the loudspeaker array can be seen in Fig. 2. The empty anechoic chamber is considered anechoic above 2.4 m radius. A depiction of the loudspeaker array can be seen in Fig. 2. The empty anechoic chamber is considered anechoic above 100 Hz according to ISO 26101 (ISO26101, 2012). The loudspeakers are mounted on seven rings elevated by ±80°, ±56°, ±28° and 0° with respect to the head position, with 2, 6, 12 and 24 loudspeakers uniformly distributed on the respective rings.

The loudspeakers are of type KEF LS50 (KEF Audio, Maidstone, UK) and driven by three sonible d:24 amplifiers (sonible GmbH, Graz, Austria) and controlled via two biamp TESIRA Server digital signal processing (DSP) units and sixteen TESIRA SOC-4 digital-to-analog converters (biamp Systems Inc., Beaverton, USA). Level, time, and frequency response corrections were applied using the DSP units, based on IR measurements at the midpoint of the loudspeaker array.

### 2.4. Room acoustic measures

Three objective room acoustic parameters were investigated and compared between the reference room and its virtual versions created with the three reproduction techniques (simulated NLM, simulated MOA and recorded MOA). Three energy parameters, reverberation time (T30), early decay time (EDT) and speech clarity (C50) were calculated from the impulse responses (IRs) measured between the three source positions and the listener position (as shown in Fig. 1). These parameters have been shown to correlate with speech intelligibility (Bradley, 1986). An omni-directional ¼” pressure-field microphone (Type 4192, Brüel & Kjær, Nærum, Denmark) was used to acquire the room impulse responses (RIRs) to calculate the energy parameters. In the reference room, the RIRs were directly measured at the listening position using the three loudspeakers corresponding to the three source positions. In the VSE, IRs were measured from each of the 64 loudspeakers to the omni-directional microphone positioned at the center of the array, pointing upwards. Subsequently, these 64 IRs were convolved with the impulse responses generated for each loudspeaker by one of the three reproduction methods, and summed to obtain the reproduced RIRs. All IRs were truncated to 0.7 s. T30, EDT and C50 were calculated from the RIRs using the ITA-toolbox (Berzborn et al., 2017).

### 2.5. Speech intelligibility experiment

The speech material for the experiment was taken from the multi-talker version of the Dantale II matrix sentence test (Behrens et al., 2007; Wagener et al., 2003). The sentences have a five-word structure (Name, Verb, Numerical, Adjective, Noun) with low context information and ten words per word-category. The word-category “name” was presented as a call-sign and subjects were asked to identify the remaining four words on a user-interface displayed on an iPad Air 2 screen (Apple Inc., Cupertino, USA). The responses were scored on a word basis and speech reception thresholds (SRTs) were measured with an adaptive procedure at 70% correct intelligibility (Brand and Kollmeier, 2002). The presentation level of each of the maskers was kept constant at a sound pressure level (SPL) of 60 dB, while the level of the target speech was adjusted adaptively, starting at 70 dB SPL. The multi-talker version of the Dantale II contains five female talkers with similar voice pitch. Three of the five talkers with the closest average root-mean-square levels were selected to reduce level differences in the test (talkers 1, 4 and 5).

SRTs were measured in three spatial conditions as shown in Fig. 3: a co-located condition with target and two interferences presented from the front, a symmetrically separated condition with the target from the front but the interferers at ±30°, and an asymmetrically separated condition with the two interferers at −30°. The difference between the colocated and the given non-colocated spatial sound source configuration is commonly considered to reflect a spatial benefit (SB). In the present study, the difference between the colocated and the asymmetrical interferer configuration was defined as the SB. The difference between the symmetric and asymmetric interferer locations was considered to reflect the effect of long-term better-ear listening. The long-term better-ear listening advantage is in the current paper termed “asymmetry benefit” (AB) to clearly distinguish from short-term better-ear listening effects. Benefits for both symmetric and

![Fig. 2. Depiction of the virtual sound environment consisting of 64 loudspeakers. The gray surface represents the wire-mesh floor and the black sphere the listening position with the facing direction indicated by the line.](image)

![Fig. 3. The three spatial configurations with two interfering sources colocated (I), symmetrically separated (II) and asymmetrically separated (III) with respect to the target.](image)
asymmetrical interferers as compared to the colocated case are referred to as spatial release from masking in this study.

Two kinds of interfering signals were used: speech interferers using sentences spoken by different talkers from the Dantale II database, and noise interferers. To create the noise interferers, for each sentence, the broadband Hilbert envelope was extracted and low-pass filtered at 40 Hz as in Best et al. (2013) and Westermann and Buchholz (2015). Subsequently, the envelope was multiplied with a speech-shaped noise having the same long-term magnitude spectrum as the particular sentence. The speech interferer is contextually similar to the target and can be expected to produce a high amount of informational masking (IM), while the noise interferer is expected to produce less IM but has similar envelope statistics and spectral content as the speech masker (Agus et al., 2009; Best et al., 2013; Ewert et al., 2017; Westermann and Buchholz, 2015). For each SRT measurement, the call-sign (name) for the target sentence was chosen randomly and kept for the following sentences, while the three target and interfering talkers were randomly permuted for each sentence. The call-sign was shown on the user interface to the listener before the start, and continuously throughout each condition. The interfering sentences did not contain the same words as the target.

The speech intelligibility experiment was performed with ten young, normal-hearing listeners with an average age of 24.7 years (σ = 4.5y) and pure-tone audiogram thresholds below 20 dB HL at the octave band frequencies between 250 Hz and 8 kHz. In addition to the previously presented reproduction conditions, a control condition was also included where the three spatial conditions (co-located, symmetrically and asymmetrically separated) were reproduced in the loudspeaker environment without reverberation (i.e. anechoic presentation using single loudspeakers). The interferers were either speech or noise. Thus, two interferer types, three spatial conditions, and five reproduction methods were tested, leading to a total of 30 conditions, with the 5 reproduction methods being: (1) reference room, (2) simulation-based NLM, (3) simulation-based MOA, (4) recording-based MOA, (5) anechoic control. The conditions were presented in random order. Five of the ten subjects started the experiments in the reference room whereas the other five started in the VSE. Each of the conditions was repeated three times in the reference room and once in the VSE. In total, the experiments lasted about 4 h for each listener. All listeners were financially compensated on an hourly basis and provided informed consent. The experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

2.6. Speech intelligibility modeling

The binaural speech intelligibility model of Jelfs et al. (2011) was used to predict the speech intelligibility data in the conditions considered in the present study. The model uses binaural room impulse responses (BRIRs) measured between the target and interferer locations and the listening position as input signals and computes the target-to-interferer ratio. The target-to-interferer ratio comprises a binaural masking level difference or binaural unmasking (BU) component and a long-term better-ear signal-to-noise ratio (BE-SNR) component. The implementation of the model was taken from the auditory modeling toolbox (Soendergaard and Majdak, 2013). The BRIRs were obtained as described above, but using a head and torso simulator (HATS, Type 4100, Bruel & Kjaer, Nærum, Denmark) instead of an omnidirectional microphone as for the room acoustic measures. The BRIRs were presented to the model at 0 dB SNR, i.e. with the BRIR at the target location having the same energy as the BRIRs at the interferer locations.

3. Results

3.1. Room acoustic measures

The obtained objective room acoustic measures for the reference room and the three reproduction methods are shown in Fig. 4 for octave frequency bands. Panels A–C show the energy parameters T30, EDT, and C50, respectively. The results represent averages over the three source positions. The gray shaded area represents just-noticeable differences (JNDs) for the results obtained in the reference room. The reported JNDs for T30 and EDT are 5% (Vorländer, 1995) and 1.1 dB for C50 (Bradley et al., 1999).

The reverberation times in the simulation- and recording-based reproductions were found to match the reference room well. The results were within or close to the JNDs at most frequencies. However, at 125 Hz, the reverberation time was slightly overestimated with the two simulation-based methods whereas the recording-based reproduction led to a slight underestimation. The EDT and C50 were reproduced accurately with the recorded-MOA method, whereas differences beyond the corresponding JNDs were found with the simulation-based reproductions.

3.2. Speech intelligibility

Fig. 5 shows speech reception thresholds (SRTs, SNR at 70% correct words) in dB target-to-masker ratio (TMR). The results with the speech interferers are shown in panel A. The results obtained with the noise interferers are shown in panel B. The white, light blue and dark blue boxes represent the spatial locations of the two interfering signals: colocated, symmetrically separated and asymmetrically separated from the target, respectively. The various reproduction methods, i.e. the reference room, the three virtual rooms and the anechoic condition, are indicated on the abscissa.
To analyze the outcomes of the speech intelligibility experiment, a linear mixed effects model was fitted to the SRTs and analyzed employing an analysis of variance, using the statistics software R and the step function included in the lme4Test package (Kuznetsova et al., 2014). The factors interferer location, interferer type, repetitions and reproduction method were treated as fixed effects. The factor listener was treated as a random effect, including its interactions with the fixed effects. The factor repetitions was not found to have a significant effect on the SRT \( [F(2,382) = 2.38, p = 0.09] \) and was removed from the final model. The factors interferer location \( [F(2,14.81) = 59.4, p < 0.0001] \), interferer type \( [F(1,9,02) = 115.23, p < 0.0001] \) and reproduction method \( [F(4,384) = 91.97, p < 0.0001] \), as well as the interactions between interferer location and interferer type \( [F(2,384) = 146.53, p < 0.0001] \), and between interferer location and reproduction method \( [F(8,384) = 16.09, p < 0.0001] \), were found to be significant. The interaction of interferer type and reproduction method \( [F(4,378) = 1.93, p = 0.11] \) and the 3-way interaction \( [F(8,370) = 1.56, p = 0.13] \) were not found to be significant, but were nevertheless kept in the model because interactions on a level basis were suspected. To analyze differences between levels, a post-hoc multiple comparison analysis was performed. The post-hoc analysis was performed by contrasting least-square means using the “lsmeans” library (Lenth, 2016). Resulting p-values were corrected for multiple comparisons using the Tukey method.

### 3.2.1. Training effect and test-retest variability
The conditions measured in the reference room were repeated three times to investigate a possible training effect and the test-retest variability of the Dantale II-based speech test. A training effect over the three repetitions could not be found \( [F(2,382) = 2.38, p = 0.09] \). The test-retest variability was estimated as the standard deviation of the repetitions and averaged over conditions and subjects. It was found to be 1.5 dB and comparable to other speech intelligibility tests (Plomp and Mimpen, 1979).

### 3.2.2. Effect of reverberation on speech intelligibility
To investigate the effect of reverberation on speech intelligibility, differences between the reference room and the anechoic condition were investigated. The speech intelligibility results are shown in Fig. 5 and the significance values of the pairwise comparisons are shown in Table 1. In the colocated configuration (white boxes), no influence of reverberation was found for speech while a significant effect was found for the noise interferers. In the case of the symmetrically separated interferers, reverberation resulted in

### Table 1
Statistical overview of comparisons between reproduction methods for speech reception thresholds.

<table>
<thead>
<tr>
<th></th>
<th>Collocated</th>
<th>Separated Symmetric</th>
<th>Separated Asymmetric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference Room vs Simulated NLM</td>
<td>( t(372) = -0.67, p = 0.96 )</td>
<td>( t(372) = 3.82, p = 0.0015 )</td>
<td>( t(372) = 6.43, p &lt; 0.0001 )</td>
</tr>
<tr>
<td>Reference Room vs Simulated MOA</td>
<td>( t(372) = 0.68, p = 0.96 )</td>
<td>( t(372) = 3.8, p = 0.0016 )</td>
<td>( t(372) = 3.97, p = 0.0008 )</td>
</tr>
<tr>
<td>Reference Room vs Recorded MOA</td>
<td>( t(372) = -0.43, p = 0.99 )</td>
<td>( t(372) = -1.73, p = 0.42 )</td>
<td>( t(372) = -3.96, p = 0.0008 )</td>
</tr>
<tr>
<td>Reference Room vs Anechoic</td>
<td>( t(372) = -0.36, p = 1.0 )</td>
<td>( t(372) = -3.4, p = 0.0066 )</td>
<td>( t(372) = -7.59, p = 0.0001 )</td>
</tr>
<tr>
<td>Simulated NLM vs Simulated MOA</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Simulated NLM vs Recorded MOA</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Simulated MOA vs Recorded MOA</td>
<td>( t(372) = -4.53, p = 0.0001 )</td>
<td>( t(372) = 4.93, p = 0.0001 )</td>
<td>( t(372) = 4.93, p = 0.0001 )</td>
</tr>
</tbody>
</table>

Please cite this article as: Ahrens, A et al., Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments, Hearing Research, https://doi.org/10.1016/j.heares.2019.02.003
an average increase of SRT by 5.4 dB for speech, and by 4.9 dB for the noise interferers. For the asymmetric interferers, the effect of reverberation was 6.9 dB for speech and 8.4 dB for the noise interferers.

3.2.3. Effect of reproduction methods on speech intelligibility

In the colocated configuration, no difference was found between the reproduction methods and the reference condition, neither for speech nor for the noise interferers. The significance values of all pairwise comparisons are shown in Table 1.

For the symmetrically separated interferers, the simulation-based reproduction methods showed statistically significant differences to the reference condition. For the speech interferers, 2.7 dB lower SRTs (better speech intelligibility) were found for both the simulated NLM and the simulated MOA methods. For the noise interferers, the SRTs decreased by 3.6 dB for the simulated NLM, and by 2.5 dB for the simulated MOA method relative to the reference condition. The SRTs obtained with the recorded MOA method were not significantly different from the reference condition, neither for the speech, nor for the noise interferers.

When comparing the two simulation-based approaches using NLM and MOA, no significant effect was observed with symmetric interferers. However, when comparing the simulation-based to the recording-based approach, significantly higher SRTs were observed in the microphone array-recording condition. These differences were found to be 3.9 dB for the speech interferers, both in the case of the NLM and MOA reproduction. For the noise interferers, the corresponding SRT differences were 4.4 dB in the case of NLM reproduction and 3.2 dB in the case of MOA reproduction.

For the asymmetrical interferers, the simulation-based reproduction methods again showed significant differences from the reference condition. For the speech interferers, the SRTs decreased by 3.3 dB for the simulated NLM, and by 2.8 dB for the simulated MOA method, relative to the reference room. For the noise interferers, SRTs were 4.6 dB lower for the simulated NLM and 3 dB lower for the simulated MOA method than obtained in the reference room. The recording-based reproduction method did not show a significant difference to the reference with noise interferers, but with the speech interferers the SRTs increased by 2.8 dB in relation to the reference room.

The two simulation-based reproduction methods, using NLM and MOA, showed no significantly different SRTs with asymmetric interferers, with both the speech and the noise interferers. However, lower SRTs were found in the two simulation-based methods in relation to the recording-based method. The difference was about 6 dB for the simulated NLM method with both interferer types. Differences in SRT of 5.6 dB with speech and 4.3 dB with noise interferers were obtained between the simulation- and recording-based MOA methods.

3.2.4. Effect of spatial separation on speech intelligibility

Fig. 6 shows the SB (light blue), i.e. the difference between the colocated and the symmetrically separated interferer condition, and the AB values (dark blue), i.e. the difference between the symmetrically and asymmetrically separated interferer conditions. The significance values of the pairwise comparisons are shown in Table 2. For the speech interferers, a significant SB was found in all reproduction conditions. For the noise interferers, no significant SB was found in the reference room nor for the MOA reproductions. However, a significant SB of 2.9 dB was found for the NLM reproduction and in the anechoic condition (2.5 dB).

Fig. 6. The spatial release from masking (SRM) due to separating target and interfering talkers (left) and the benefit due to asymmetric versus symmetric interferers (right) for the different reproduction methods with speech and noise interferers. (The boxes represent the median and the 1st/3rd quartile. The whiskers include 1.5 times the interquartile range.)

Table 2
 Statistical overview of comparisons between reproduction methods for spatial benefit and asymmetry benefit.

<table>
<thead>
<tr>
<th>Reproduction Method</th>
<th>Speech Interferers</th>
<th>Noise Interferers</th>
<th>Spatial Benefit</th>
<th>Asymmetry Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference Room</td>
<td>t(28.18) = 8.83, p &lt; 0.0001</td>
<td>t(28.18) = 0.12, p = 0.99</td>
<td>t(165.96) = 4.62, p &lt; 0.0001</td>
<td>t(165.96) = 0.91, p = 0.63</td>
</tr>
<tr>
<td>Simulated NLM</td>
<td>t(104.76) = 9.34, p &lt; 0.0001</td>
<td>t(104.76) = 2.96, p = 0.0104</td>
<td>t(336.27) = 3.38, p = 0.0023</td>
<td>t(336.27) = 1.66, p = 0.22</td>
</tr>
<tr>
<td>Simulated MOA</td>
<td>t(104.76) = 8.35, p &lt; 0.0001</td>
<td>t(104.76) = 1.93, p = 0.13</td>
<td>t(336.27) = 2.87, p = 0.0122</td>
<td>t(336.27) = 1.13, p = 0.5</td>
</tr>
<tr>
<td>Recorded MOA</td>
<td>t(104.76) = 4.49, p &lt; 0.0001</td>
<td>t(104.76) = 0.59, p = 0.83</td>
<td>t(336.27) = 0.93, p = 0.62</td>
<td>t(336.27) = -0.1, p = 0.99</td>
</tr>
<tr>
<td>Anechoic</td>
<td>t(104.76) = 11.84, p &lt; 0.0001</td>
<td>t(104.76) = 2.56, p = 0.0316</td>
<td>t(336.27) = 4.42, p &lt; 0.0001</td>
<td>t(336.27) = 4.58, p &lt; 0.0001</td>
</tr>
</tbody>
</table>

Please cite this article as: Ahrens, A et al., Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments, Hearing Research, https://doi.org/10.1016/j.heares.2019.02.003
A significant AB of 2.4 dB was found in the reference room for the speech interferers, but not for the noise interferers. Similarly, the AB effect was significant for the speech but not the noise interferers in the case of the simulated NLM and the simulated MOA methods. For the recording-based reproduction, no AB was found for either the speech or the noise interferers. In the anechoic condition, the AB effect was significant for both speech and noise interferers.

3.3. Speech intelligibility modeling

Fig. 7A shows the results from the simulations obtained with the Jelfs et al. (2011) model in the conditions with the symmetrically (left panel) and asymmetrically (right panel) separated noise interferers. The colocated condition is omitted because the model takes only impulse responses into consideration, thus no model outcome is seen when all sources are presented from the same location. The model outcome (squares) is shown as the sum of the two contributors, the BU (circles) and the BE-SNR (triangles). Since the BE-SNR contribution can be below zero, the total model outcome can be lower than the BU contribution. In the configuration with the interferers placed symmetrically left and right, the BE-SNR is close to zero for all reproduction methods, as expected. The model predicts the highest BU in the anechoic condition. The contribution of BU is similar, about 1 dB, in the reference room and with the recorded MOA method. The simulated NLM and simulated MOA methods show a predicted BU contribution of about 1.7 dB. For the asymmetric interferer configuration, the contribution of BU to the model output is smaller than for the symmetric interferers, with values between 0.5 and 1 dB. Overall, the modeled BU is similar between the reproduction methods, except for the anechoic control condition where a contribution of 2.5 dB is predicted. The asymmetric interferer configuration was expected to result in a SNR advantage in one ear. However, the simulated BE-SNR shows values close to zero for both the reference and the recording-based MOA reproductions. The simulation-based NLM and MOA reproductions, on the other hand, show a 2 dB and 1.1 dB higher BE-SNR than the reference, respectively. The highest predicted BE-SNR of 5.5 dB was found in the anechoic condition. Fig. 7B shows the total model outcome together with the corresponding speech intelligibility data from the present study with noise interferers. The comparison was limited to the noise interferers over the speech interferers because the model is not able to incorporate IM. The model was fitted to the median SRT obtained in the reference room for each spatial configuration. The model captures the differences between the reproduction methods in relation to the reference room fairly well. Nevertheless, the symmetric interferer configuration (Fig. 7, left) is not captured as well as the asymmetric interferer configuration (Fig. 7, right).

4. Discussion

The present study investigated the discrepancies that appear between speech intelligibility tests in real and virtual environments, and the effect of various reproduction methods on these differences. Several common factors influencing speech intelligibility were varied: the spatial position and type of interferers, as well as the presence of reverberation.

4.1. The role of spatial configuration

The three spatial configurations of the interferers provided different levels of separation between the target and the interfering signals in terms of spatial cues. In the colocated condition, no such differences were available to the listener. Previous studies suggested that in a situation with similar target and
interferer and no spatial separation, a positive TMR is needed for segregation, implying a level cue for selecting the target (Best et al., 2012; Brungart et al., 2001). Consequently, the reproduction method must mainly capture the sound levels of the sources to reflect speech intelligibility correctly when the interferers are colocated. Indeed, no differences between any of the reproduction methods for speech interferers were found, as the SNR was correctly reproduced. With noise interferers, reverberation does play a role, as reflected by the lower SRTs obtained in the anechoic condition compared to the reverberant reference room. However, no differences were observed between the reference and the reproduction methods for the colocated noise interferers either.

When the target and interferers are symmetrically separated, spatial location differentiates the source signals. However, due to the left-right symmetry of the interferer positions, and as long as no head movement occurs, there is no long-term SNR benefit at either ear, and the auditory system must rely on binaural cues, i.e. interaural-time differences, or short-term better-ear listening (Brungart and Iyer, 2012; Glyde et al., 2013). This is supported by the predictions obtained with the model by Jeffs et al. (2011) (see Fig. 7), showing a close to zero BE-SNR advantage and a main contribution of all reproduction methods. This model only considers a long-term better-ear advantage, and would not reflect any short-term advantage that may exist. Furthermore, the model does not take head motion into account, which might have let to intelligibility advantages during the experiments, where subjects were explicitly allowed to move their heads.

When the target and interferers are asymmetrically positioned, a long-term SNR benefit may be available at one ear. The asymmetric configuration resulted in the largest spatial release from masking overall. However, contrary to the expectation, no spatial release from masking was observed in the reference room for either of the separated spatial configurations with a noise masker, suggesting that a long-term better-ear advantage was not, in fact, available. This is in line with the model predictions (Fig. 7), which showed a BE-SNR advantage of about 0 dB for the reference room also for the asymmetric configuration. Thus, the low amount of reverberation was sufficient to negate the effect of asymmetric positioning in terms of long-term SNR at the ears, as it was found in the anechoic condition.

In the symmetrically and asymmetrically separated configurations, differences emerge between the reproduction methods. Results from the recording-based reproduction compared favourably to the reference and a significant difference only appeared for one condition, with asymmetric speech interferers. The role of the interferer type is discussed further below. In contrast, the simulation-based reproductions led to consistently lower SRTs, or, in other words, a larger amount of spatial release from masking than in the reference room. Oreinos and Buchholz (2016) investigated speech intelligibility in VSEs in aided hearing-impaired listeners using a similar setup, but with seven conversational sources as interferers. They also found lower SRTs in their simulation-based virtual room than in the reference environment. However, the differences in that study were small and comparable to the test-retest variability of the speech test. In the current study, these differences were found to be somewhat larger, in the order of 2–3 dB, compared with the estimated test-retest variability of 1.5 dB. Despite using the same simulation framework as in Oreinos and Buchholz (2016), they considered a spatially more distributed masker configuration with a larger number of talkers, as well as the longer reverberation time in a larger room, which might have contributed to reduced reproduction errors. The fact that lower SRTs were observed for both ambisonics and nearest-loudspeaker presentation in the present study suggests that the deviations likely originate from the room acoustics modeling rather than the playback method, as also indicated by the room acoustic measures.

4.2. The role of reverberation

Reverberation is known to reduce speech intelligibility (Duquesnoy and Plomp, 1980; Houtgast et al., 1980; Plomp, 1976), which was the case for all conditions when compared to the anechoic control, except for the condition with colocated speech interferers. Thus, the acoustics of the room had an effect on the resulting SRTs in all but one case. It follows that an accurate reproduction of the acoustics is necessary to obtain SRTs that match those measured in the reference room. The simulation-based reproduction methods resulted in lower SRTs compared to the reference and the recording-based method when the target and interferers were separated. This suggests that some aspects of the room’s acoustics were not correctly captured with these methods. Indeed, the deviations apparent for the two simulation-based methods in terms of clarity, and especially early decay time (see Fig. 4), which has been shown to be negatively correlated with speech intelligibility (Grimm et al., 2016), indicate that early reflections are not correctly reproduced by the room model. Early reflections have been shown to improve speech intelligibility (Arweiler and Buchholz, 2011; Bradley et al., 2003; Lochner and Burger, 1964; Soulodre et al., 1989), thus, it is not surprising that it is insufficient to just correctly simulate the overall reverberation time in a room. The early reflection pattern also needs to be correct in order to obtain SRTs that closely correspond to the reference room. A general challenge with the room modeling approach is that it may be difficult to obtain detailed enough information about the room (geometry, material properties, etc.) to enable such an accurate simulation. In contrast, the recording-based approach captured the detailed acoustic response of the room, at least for the measured source-receiver positions, leading to a closer match to the reference room both in terms of room acoustic parameters, as well as measured SRTs. Favrot and Buchholz (2010) showed that the changes of the room acoustic parameters due to the reproduction system itself are within the listeners’ perceptual difference limens. Thus, the differences observed in the current study most likely result from inaccuracies in the room acoustic simulation.

For both simulation-based reproduction methods, the same late reverberation is reproduced using 1st order ambisonics. This method aims to create perceptually reasonable, but not physically accurate late reverberation. It has been shown that room acoustics parameters (e.g. EDT, T30, C50) are only affected marginally by this method (Favrot and Buchholz, 2010). Thus, it is more likely that the inaccuracies of the early reflections have the largest effect on the speech intelligibility.

4.3. The role of interferer type

Two interferer types, speech and noise, were applied to investigate any differences in the reproduction methods with respect to IM. As expected, lower SRTs were found for the noise interferers than for the speech interferers. The high SRTs with speech interferers were due to the high similarity (same sentence structure, same gender) of the speech interferers with the target speech, which leads to a high probability that target and interferers are confused. An SRM with speech interferers was found in both conditions with and without reverberation. With noise interferers, a spatial release from masking was found in the anechoic condition, but in the reverberant reference room, the spatial release from masking disappeared, both in the symmetric and the asymmetric configurations. Comparable results of a reduced or diminishing release from masking in reverberant conditions were found in
previous studies (Freyman et al., 1999; Westermann and Buchholz, 2015), arguing for a spatial release from IM, which only occurs when the amount of IM is high, as in the speech interferer condition of the present study. With the noise interferers, an SB was only found in the NLM condition (and in the anechoic condition), which further suggests that the early reflections but also the diffuseness of the late reverberation in these conditions are not correctly reproduced.

4.4. The role of ambisonics reproduction

One defining feature of sound sources reproduced using ambisonics is that they have a higher spatial energy spread, i.e. a higher number of loudspeakers playing simultaneously, than the single loudspeaker used in the reference room (Gerzon, 1992; Stitt et al., 2016; Zotter and Frank, 2012). It was hypothesized that the larger energy spread could lead to reduced interaural level differences, and thus a reduced spatial release from masking, especially for the asymmetric condition. A comparison between the two simulation-based methods, employing ambisonics versus the mapping to loudspeaker used in the reference room (Gerzon, 1992; Stitt et al., 2016) found in the NLM condition (and in the anechoic condition), which further suggests that the early reflections but also the diffuseness of the late reverberation in these conditions are not correctly reproduced.

4.6. Limitations and perspectives

One of the limitations of this study is that only a single room was considered. Since the room acoustic parameters of the simulated virtual rooms did not match those of the real room, conclusions regarding the applicability of room acoustic simulations for the reproduction of rooms need to be taken with care. Furthermore, the considered room was small in relation to the general room size, for which the room acoustic software has been developed. Thus, future work should include various rooms with different levels of early reflections and reverberation to provide a more complete picture of the advantages and disadvantages of the room acoustics simulation and the reproduction techniques.

Only normal-hearing listeners were tested in this study in an effort to focus on a comparison between the reproduction techniques, as speech intelligibility results from hearing-impaired listeners typically show a markedly higher variance than those measured with normal-hearing listeners. As a next step, hearing-aids or other communication devices should be considered as well, as these devices might behave differently than human listeners in the generated sound fields and the processing algorithms, such as beamformers, might interact in unexpected ways with the applied reproduction methods. However, in the most important frequency range for speech up to about 6 kHz (ANSI, 2017), in which these devices typically operate, the sound field is relatively well controlled by the applied reproduction techniques, and previous work showed only a slight reduction in the efficacy of, e.g., beamforming algorithms (Cubick and Dau, 2016; Oreinos and Buchholz, 2016). Nonetheless, since one of the main application areas of VSEs is the evaluation of such communication devices and their benefit to the user, the interaction between advanced processing algorithms, hearing impairment, and virtual sound environments needs to be explored further. Outcome measures other than speech intelligibility, such as listening effort, scene awareness and headmovements, as for example considered in Hendrikse et al. (2018), might also be explored, as they can be relevant for hearing-aid applications.

5. Conclusions

This study examined the accuracy of speech intelligibility measurements in a virtual sound environment (VSE) in comparison to a reference room in several conditions and with computational auditory modeling as an analysis tool. Three reproduction methods and specific factors that influence speech perception were considered: room reverberation, interferer type and spatial location of the interferers.

The reproduction based on impulse responses measured with a microphone array provided the closest match to the reverberant reference room in terms of speech reception thresholds (SRTs). The two methods based on room acoustic simulations showed significantly lower SRTs compared to the reference room, but only when target and interferers were separated, while no differences were found when target and interferer were colocated. Lower SRTs in the
simulation-based reproductions could be explained by errors in the simulated early reflections, despite a correctly reproduced total reverberation time. The measured SRTs in the real and virtual rooms could be predicted using the auditory model. Overall, it was demonstrated that room acoustic models, which are successful in capturing average properties of a room, may be limited in their ability to match the exact details of the response at a specific location, which in turn can lead to differences in measured speech intelligibility. This may only be a relevant shortcoming if capturing the response of a specific room at a specific location is crucial. If this is the case, measurement-based methods provide a clear advantage.

Acknowledgments

This work was supported by the Technical University of Denmark and the Oticon Center of Excellence for Hearing and Speech Sciences (CheSS). The multi-talker version of the Dantale II speech material was provided by Eriksholm Research Center. Furthermore, we would like to thank Jens Cubick for the work in setting up the virtual sound environment and Adam Westermann for an early version of the graphical user interface test interface.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heares.2019.02.003.

References


Please cite this article as: Ahrens, A et al., Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments, Hearing Research, https://doi.org/10.1016/j.heares.2019.02.003


