Evaluating a Loudspeaker-Based Virtual Sound Environment using Speech-on-Speech Masking

Axel Ahrens[1], Marton Marschall[1], Torsten Dau[1]

[1] *Hearing Systems group, Department of Electrical Engineering, Technical University of Denmark, DK-2800, Kgs. Lyngby, Denmark, E-Mail:{aahr, mm, tdau}@elektro.dtu.dk*

Introduction

The ability of the human auditory system to understand speech in a mixture of multiple talkers and background noise, the so-called cocktail-party problem, is a widely investigated area of hearing research since its definition by Cherry [1] in the 1950s. Since then, multiple aspects of the cocktail-party phenomenon have been studied with varying degrees of realism (see [2] for a review). To investigate the highly realistic cocktail-party situations, we need to be capable of simulating the acoustic environment as precisely as possible. In recent years, multiple sound reproduction methods have been developed that may be suitable for hearing research. The most well-known methods are wave-field synthesis (WFS, [3]), vector base amplitude panning (VBAP, [4]), directional audio coding (DirAC, [5]), and higher- and mixed order ambisonics (HOA/MOA, [6, 7]). The advantage of HOA over the other reproduction strategies is its aim to reproduce a physically accurate sound field in 3D. Furthermore, MOA allows a higher resolution representation in the horizontal plane, where most sound sources of interest are located, by increasing the number of transducers in or near the horizontal plane. Such a MOA loudspeaker array was recently installed at the Technical University of Denmark.

Several previous studies have investigated the accuracy and the precision of spatial audio reproduction techniques. These evaluations can consider both physical measures, like the sound pressure [8] and room acoustic parameters [8, 9], and perceptual measures, like quality attributes [10], localization accuracy and speech intelligibility [8, 9]. It is clear that a physically accurate reproduction of the sound field over the entire audible bandwidth in a head-sized area is currently not possible with practically realizable setups [11]. Therefore, if such systems are to be used for hearing research, the impact of the various simulation and reproduction methods on the outcome measures of interest must be evaluated.

In order to assess the applicability of virtual environments for audiological testing, a speech intelligibility test was performed in a cocktail-party-like environment. The test was carried out in an IEC-listening room [12] using loudspeakers to imitate talkers. The listening room setup served as a reference condition and was reproduced in a loudspeaker-based virtual sound environment. The reproduction was realized using (i) room acoustic simulations, and (ii) impulse response recordings from a spherical microphone array. Finally, speech intelligibility results in the reference and reproduced conditions were compared.

Methods

Stimuli and Spatial Setup

The material for target and interfering speech was taken from the multi-talker version of the Dantale II [13]. The Dantale II sentences consist of five word sentences (Name, Verb, Numeral, Adjective, Noun) with low context information and ten words per category. The name was presented as a call-sign and subjects were asked to identify the remaining four words on a user-interface displayed on a tablet computer. The answers were scored on a word basis and speech reception thresholds (SRT) were measured with an adaptive procedure at 50% correct intelligibility. The presentation level of the maskers was kept constant at 55 dB SPL, while the level of the target speech was adjusted adaptively, starting at 65 dB SPL. The speech material contains five female talkers with similar voice pitch.

SRTs were measured in two spatial conditions: a co-located condition with target and two interfering talkers presented from the front, and a separated condition with the target from the front but the interferers at ±30°. All conditions were repeated three times. For each SRT measurement a call-sign (name) was chosen randomly and kept for all sentences while the three talkers for target and interfering speech were chosen randomly for each sentence.

Reference Room

The speech intelligibility test was performed in a reference room, an IEC listening room (IEC standard 268-13, [12]) with a volume of 100 m^3 (7.52 m * 4.75 m * 2.8 m) and an average reverberation time of 0.35s. The listening position was centered along the longest dimension of the room and 1.35 m from the back wall (see Figure 1). The talkers were imitated using Dynaudio BM6P loudspeakers. The loudspeakers were located 2.4 m distant from the listener at 0° and at ±30° ahead of the listener. The loudspeakers were placed approximately at ear level (h = 1.17 m).

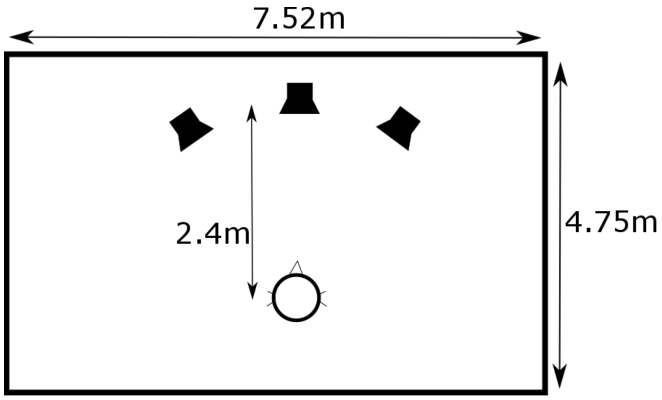

Figure 1: Experimental setup in the reference room.

Acoustic Scene Generation and Recording

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

To simulate the room acoustics of the IEC listening room, a geometrical model of the room was constructed in the room acoustics software ODEON [14] including the same source and receiver/listener positions as in the reference condition. The absorption coefficients of the room surfaces were optimized from initial estimates with the ODEON genetic material optimizer [15], using measured reverberation times (T20, T30), early decay time and clarity (C7, C50, C80) parameters as calculated (with the ITA-toolbox [16]) from impulse responses measured in the reference room. From the optimized room acoustics model, direct sound, early reflections, and energy decay curves in eight octave bands from 63 Hz to 8 kHz were exported and processed using the LoRA-toolbox [17]. The LoRA-toolbox uses the ODEON time, amplitude and spatial information to compute impulse responses for loudspeaker-based auralization. Two processing strategies are implemented in LoRA: 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 [17]. For MOA, the same strategy was used for the late reflections as for the NLM. However, the direct sound and the early reflections were encoded using 7^{th} order horizontal and and 5^{th} order periphonic ambisonics. The loudspeaker signals were obtained from the MOA signals using a two-band mode matching / max-rE decoder, with a crossover frequency of 4000 Hz [17].

The microphone array recordings were made with a 52-channel spherical array [18]. Impulse responses (IRs) were recorded in the reference room between the three source positions and the listening position using eight 16s long exponential sweeps [19]. The same MOA orders were used for encoding the array signals as for the simulations (7^{th} order horizontal, 5^{th} order periphonic). From the ambisonics components the loudspeaker signals were obtained using a mode-matching decoder [20] in this case.

Virtual Sound Environment

The acoustic scenes were reproduced in a loudspeaker-based virtual sound environment (VSE). The VSE consists of a 64 channel spherical loudspeaker (KEF LS50) array housed in an anechoic chamber (6 m*7 m*8 m). The empty anechoic chamber is specified to be anechoic above 100 Hz. The loudspeaker array is arranged on seven rings with 2, 6, 12, 24, 12, 6, 2 loudspeakers from the top to the bottom ring with an equiangular distribution of the loudspeakers on each ring. The rings are elevated by -80°, -56°, -28°, 0°, 28°, 56°and 80° relative to the listener. The distance from the loudspeakers to the listening position in the center of the sphere is 2.4 m. This setup allows mixed-order ambisonics reproduction up to the 11^{th} order in the horizontal plane and up to the periphonic order of five. In this study the horizontal order was limited to seven due to the limitation of the order of the spherical microphone array.

Experimental Procedure

The listening experiment was performed with five young normal hearing listeners. All conditions were repeated three times. In addition to the previously presented reproduction conditions a control condition was tested where the two spatial conditions (co-located, separated) were reproduced without reverberation. This leads to a total of 2 x 5 conditions: (1) reference room, (2) simulation-based NLM, (3) simulation-based MOA, (4) recording-based MOA, (5) anechoic control.

The conditions were presented in a random order. Two subjects started the experiments in the reference room and three started in the VSE. Each of the conditions was repeated three times leading to a total duration of about 3-4 h per listener. All participants provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark.

Results

Repetitions and Test-Retest Variability

In the pre-analysis, the effect of repetitions on speech intelligibility was investigated. It was suspected that a learning effect might influence the outcomes, in particular in the spatially separated condition. Therefore, a linear model was fitted with the factors spatial condition and repetitions on speech reception threshold (SRT). A 2-way ANOVA revealed no significant effect of repetitions ($F(1,146) = 1.16$, $p = 0.28$) nor interaction between repetitions and spatial condition ($F(1,146) = 0.08$, $p = 0.78$). Thus, no systematic effect of repetitions on SRT was found.

The test-retest variability was estimated from the results to acquire a range where the reproduction techniques can be defined as accurate. It was calculated as the mean over subjects of the standard deviation over the three repetitions. The analysis was done for both spatial conditions but in the reference room only. The test-retest variability was estimated to be 1.3 dB.

Effect of Reverberation

The effect of reverberation on speech intelligibility was investigated by comparing the anechoic control condition to the reference condition (IEC listening room). The results are shown in Figure 2. In the co-located condition, mean SRTs were 2.3 dB target-to-masker ratio (TMR) and 3.1 dB TMR for the control and the reference condition, respectively. The difference of 0.8 dB was found to be statistically non-significant using a Wald-test ($p=0.3$). In the separated condition, the mean SRT was increased by 5.9 dB due to the addition of reverberation (control: -12.9 dB TMR; reference: -6.9 dB TMR). The effect was found to be significant (Wald-test, $p < 0.0001$).

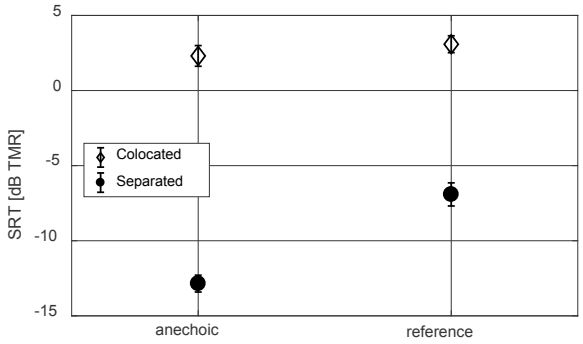

Figure 2: Speech reception thresholds (SRT) in dB TMR (target-to-masker ratio) in the anechoic control condition and in the reference room (IEC listening room).

Effect of Reproduction Technique

In the following, speech intelligibility results are compared in the reference condition (IEC listening room) and in the virtual sound environment with reproduction based on nearest loudspeaker mapping (simulated-NLM), mixed-order ambisonics coding (simulated-MOA) and microphone array recording (recorded-MOA). The mean SRTs for the reference condition and the three reproduction techniques are shown in Figure 3. For co-located target and maskers, no differences were found between the reference and the three reproduction techniques (simulated-NLM, simulated-MOA, and recorded-MOA) with respective p-values from the Wald-test of p=0.36, p=0.29, and p=0.22. For separated target and maskers, SRTs were found to be significantly lower for the conditions with simulated room acoustics. Simulated-NLM led to a 2.7 dB lower SRT and simulated-MOA to a 3.7 dB lower SRT than for the reference (Wald-test: $p<0.0001$ and $p<0.001$). No statistically significant difference was found between the two simulation-based methods (p=0.19). In contrast, the SRT obtained with the microphone array recording was 5.2 dB higher in comparison to the reference ($p < 0.0001$).

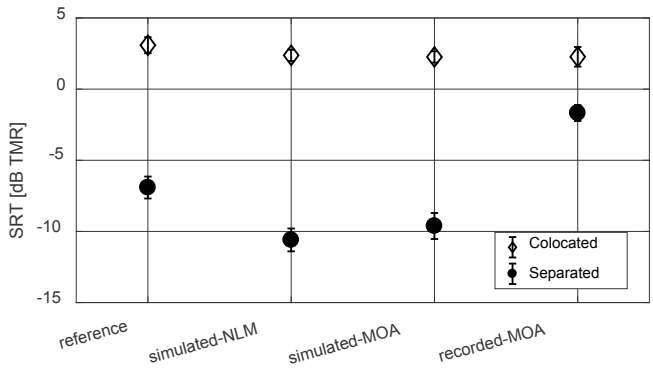

Figure 3: SRTs in dB TMR (target-to-masker ratio) in the reference room (IEC listening room) and in the 3 reproduced conditions. The simulated conditions were generated using room acoustics modelling and the recorded condition by measured impulse responses using a microphone array.

Discussion and Conclusions

The main goal of this study was to evaluate the accuracy of a loudspeaker-based virtual sound environment using speech intelligibility as an outcome measure. For co-located target and maskers, no differences between the reproduction techniques were found. However, no difference was found between the anechoic control condition and the reference room either, which suggests that reverberation had no effect on speech intelligibility in this condition. Since the SRTs were at around 3 dB target-to-masker ratio (TMR) subjects might have used a level cue in the absence of spatial cues, to discriminate the target speech from the interferers, which had similar structure and voice pitch. Hence, the obtained SRTs in the co-located condition can be ascribed mainly to the TMR being correctly reproduced, while any variation in spatial attributes and room reflections had no effect.

In contrast to the co-located condition, for separated target and maskers, reverberation decreased speech intelligibility by 5.9 dB, when comparing the anechoic control condition to the reverberant reference room. Therefore, the separated condition can be used to evaluate the reproduction of spatial cues, as well as of room reflections.

The simulation based reproduction methods resulted in somewhat lower SRTs than in the reference room (differences of 2.7 dB and 3.7 dB) for the separated condition. These differences might have occurred due to inaccuracies in the material properties of the room acoustic model. This highlights a particular disadvantage of room simulations, in that detailed a-priori knowledge about the room and its acoustic properties is necessary. In-situ recordings with microphone arrays on the other hand do not require this information. Therefore, errors due to geometrical simplifications and incorrect material definitions are omitted. However, microphone array recordings have a limited frequency range due to physical restrictions (see [18] for details). These limitations, in particular at low frequencies, might have led to the higher SRT as obtained with the microphone array recordings. In order to control the amplification of noise at low frequencies, a regularization factor needs to be applied which leads to a decrease in the effective order of the microphone array at low frequencies [18, 21]. We hypothesize that with the rather conservative regularization applied, interaural time differences that are important to localize sounds at low frequencies were not reproduced accurately and the separation of target and interfering speech became more difficult. When the ability to separate the sources in the spatial domain is decreased, the separated condition becomes more similar to the co-located condition as it can be seen for the recorded-MOA condition in Figure 3.

In contrast to the current study, a previous study [9] did find a difference between the NLM and HOA reproduction methods for simulated room acoustics. The different findings may be due to the different virtual sound environments that were used. The VSE used in the current study allowed for higher ambisonics orders, which may explain the improved performance of ambisonics in this case.

In conclusion, it was shown that speech intelligibility in the reference room could not be fully matched in the virtual sound environment for all conditions. Room acoustic simulations strongly depend on the accurate modelling of the acoustic properties, while microphone array recordings may lead to perceptual errors due to physical and signal processing restrictions. It may be possible to reduce the observed perceptual differences by further optimizing the room models or the signal processing parameters. Even though the characteristics of a specific room may be difficult

to reproduce in detail, the simulated environments led to plausible results in terms of the chosen outcome measure, SRTs, and are encouraging regarding the application of virtual environments for audiological tests.

Acknowledgments
This research was supported by the Oticon Centre of Excellence for Hearing and Speech Sciences (CHeSS). The multi-talker version of the Dantale II speech test was provided by Eriksholm Research Centre.

Furthermore, the authors would like to thank everyone who supported us in building the virtual sound environment, in particular Ralf Baumgartner, Jiho Chang, Jens Cubick, Johannes Kaesbach, Peter Sciri, and Johannes Zaar.

References

[1] Cherry, E.C.: Some Experiments on the Recognition of Speech, with One and with Two Ears. The Journal of the Acoustical Society of America 25 (1953), 975-979

[2] Bronkhorst, A.W.: The cocktail-party problem revisited: early processing and selection of multi-talker speech. Attention, Perception, & Psychophysics 77 (2015), 1465–1487

[3] Berkhout, A.J.: A Holographic Approach to Acoustic Control. Journal of the Audio Engineering Society 36 (1988), 977-995

[4] Pulkki, V.: Virtual Sound Source Positioning Using Vector Base Amplitude Panning. Journal of the Audio Engineering Society 45 (1997), 456-466

[5] Pulkki, V.: Spatial Sound Reproduction with Directional Audio Coding. Journal of the Audio Engineering Society 55 (2007), 503-516

[6] Gerzon, M.A.: Periphony: With-Height Sound Reproduction. Journal of the Audio Engineering Society 21 (1973), 2-10

[7] Daniel, J.: Representation de champs acoustiques, application a la transmission et a la reproduction de scenes sonores complexes dans un context multimedia. Ph.D. thesis (2001), Universite Paris, pp. 1–319

[8] Oreinos, C. and Buchholz, J.M.: Evaluation of Loudspeaker-Based Virtual Sound Environments for Testing Directional Hearing Aids. Journal of the American Academy of Audiology 27 (2016), 541-556

[9] Cubick, J. and Dau, T.: Validation of a Virtual Sound Environment System for Testing Hearing Aids. Acta Acustica united with Acustica 102 (2016), 547-557

[10] Guastavino, C. and Katz, B.F.G.: Perceptual evaluation of multi-dimensional spatial audio reproduction. The Journal of the Acoustical Society of America 116 (2004), 1105-1115

[11] Spors, S., Wierstorf, H., Raake, A., Melchior, F., Frank, M., and Zotter, F.: Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State. Proceedings of the IEEE 101 (2013), 1920-1938

[12] IEC Recommendation 268-13: Sound system equipment, Part 13: Listening tests on loudspeakers. International Electrotechnical Commission, Geneva (1985)

[13] Behrens, T., Neher, T., and Johannesson, R.B.: Evaluation of a Danish speech corpus for assessment of spatial unmasking. 1st International Symposium on Auditory and Audiological Research (2007)

[14] Rindel, J.H. and Naylor, G.M.: Odeon - A Hybrid Computer Model for Room Acoustic Modelling. 4th Western Pacific Regional Acoustics Conference (1991), Brisbane

[15] Christensen, C.L., Koutsouris, G. and Rindel, J.H.: Estimating absorption of materials to match room model against existing room using a genetic algorithm. FORUM ACUSTICUM (2014)

[16] Dietrich, P., Guski, M., Klein, J., Muller-Trapet, M., Pollow, M., Scharrer, R. and Vorlander, M.: Measurements and Room Acoustic Analysis with the ITA-Toolbox for MATLAB. Fortschritte der Akustik (2010)

[17] Favrot, S. and Buchholz, J.M.: LoRA: a loudspeaker-based room auralization system. Acta Acust United Acust 96 (2010), 364–375

[18] Marschall, M., Favrot, S. and Buchholz, J.M.: Robustness of a mixed-order Ambisonics microphone array for sound field reproduction. In: 132nd Convention of the Audio Engineering Society (2012)

[19] Farina, A.: Simultaneous measurement of impulse response and distortion with a swept-sine technique. In: 108th Convention of the Audio Engineering Society (2000)

[20] Zotter, F., Pomberger, H., and Frank, M.: An Alternative Ambisonics Formulation: Modal Source Strength Matching and the Effect of Spatial Aliasing. In: 126th Convention of the Audio Engineering Society (2009)

[21] Moreau, S., Daniel, J., and Bertet, S.: 3D sound field recording with higher order ambisonics-Objective measurements and validation of spherical microphone. In: 120th Convention of the Audio Engineering Society (2006)