Speech intelligibility in noise with varying spatial acoustics under Ambisonics-based sound reproduction system

Eugena Au[a], Shirley Xiao[a], C.T. Justine Hui[a], Yusuke Hioka[a,*], Hinako Masuda[b], Catherine I. Watson[c]

[a]*Acoustics Research Centre, Department of Mechanical Engineering, University of Auckland, Auckland 1142 New Zealand*
[b]*Faculty of Science and Technology, Seikei University, Tokyo, 180-8633 Japan*
[c]*Department of Electrical, Computer, and Software Engineering, University of Auckland, Auckland 1142 New Zealand*

Abstract

The present study investigates the effect of noise on speech intelligibility under an Ambisonics-based sound reproduction system, which is a commonly used technique for realising audio virtual reality. The sound reproduction system was built in an anechoic chamber using a 16 channel spherical loudspeaker array. A commercially available first-order Ambisonics microphone array was utilised for collecting room impulse responses of rooms with different acoustic characteristics, and a decoder for generating signals rendered from the loudspeaker array. A subjective listening test was conducted in the sound reproduction system examining how intelligibility of sentences is compromised by noise under virtually reproduced acoustics environments with various acoustical conditions. Particular focus was given to investigating the effect of the amount of reverberation in rooms as well as the angular separation between the target speech and noise sources, both of which are known to have a significant effect on speech intelligibility in real (i.e. non-virtual) acoustic environment.

The experimental results suggest that the intelligibility of speech masked by pink noise was significantly reduced under reverberant acoustical environments reproduced by the Ambisonics-based sound reproduction system compared to the baseline anechoic environment. However, only marginal differences were observed between the acoustical environments with different reverberation time. The effect of spatial release from masking was also observed under the reproduced acoustic environments and showed some correlation with the accuracy of source localisation realised by the sound reproduction system.

Keywords: speech intelligibility, spatial release from masking, virtual reality, Ambisonics, spatial acoustics, sound source localisation

1. Introduction

Speech is an essential medium for human communication, but its intelligibility is often challenged by various environmental factors. Speech intelligibility in adverse environments has been actively studied in various disciplines related to acoustics including psychology of hearing [1], building acoustics [2], and linguistics [3, 4]. The most commonly studied adverse condition is noise; the detrimental effect of noise covering the target speech resulting in poor speech intelligibility is known as speech masking [5]. Speech masking is primarily caused by the energy of the noise exceeding the energy of the target speech triggering auditory masking (energetic masking) [6]. This effect is used for engineering applications, such as speech masking [7, 8, 9, 10].

The acoustics of the environment also has a significant impact on speech intelligibility. For instance, reverberation produces an echo, which reaches the listeners' ears with a delay, compared to the direct sound. This causes another type of auditory masking called temporal masking, which results in degraded speech intelligibility [11, 12]. In addition, the existence of noise in reverberant spaces creates further significant decrease of speech intelligibility. In contrast, acoustics of the environment can also help improve the intelligibility of speech in noise by taking advantage of binaural hearing [13]. Studies have shown that our hearing will be "released" from masking when the positions of the target and noise sources are spatially separated. This phenomenon is known as *spatial release from masking*

*Corresponding author
Email address: yusuke.hioka@ieee.org (Yusuke Hioka)

[14, 15, 16]. Extensive research has been conducted exploring the effect of spatial release from masking and its effect in various applied scenarios in the real world, such as classrooms [17] and open-plan offices [18]. In addition, previous studies suggest that the effect of spatial release from masking is reduced when the amount of reverberation is increased [19].

These days speech communication over virtual reality (VR) [20] is attracting huge interests from various industries because of the wide variety of potential applications. Audio VR technologies that provide listeners with an experience as if they were in a different acoustic environment (also known as auralisation) [21, 22] have been emerging in the market. As the audio VR technologies advance, enabling the replication of realistic acoustic environments, the aforementioned problem regarding speech intelligibility under adverse condition also occurs, however, few studies have looked into the problem under audio VR [23, 24, 25].

Several approaches to realise audio VR have been investigated. Of those, binaural sound reproduction (BSR) using headphones would be the most commonly used approach because of its affordability from the users' perspective. A key requirement for BSR is the head-related transfer function (HRTF) [26], which encapsulates the spatial acoustical cues of the environment derived by both ears. Since measuring the HRTF of individuals is time consuming, it is often replaced by the HRTF measured by a dummy head. However, it is known that the deviation from the actual HRTF causes errors in the perception of reproduced sound. In addition, the fact that listeners have to wear headphones and the sound source image does not track the listener's head motion could limit its applications.

To address these limitations, other approaches utilise sound rendered from multiple channels of loudspeakers, i.e. a loudspeaker array, which unlike BSR, requires no HRTF measurement nor wearing headphones. Several rendering methods are studied for processing signals projected from the loudspeaker array. Of those, level difference/time delay-based panning methods generate perception of virtual sound sources by panning sound across loudspeakers. Similar to BSR, the technique also leverages the perceptual cues such as the inter-aural time difference (ITD) and inter-aural level difference (ILD). Sound field synthesis on the other hand aims to physically recreate the sound field that would have been generated had the sources been placed within the same space, rather than relying on perceptual cues.

Two main sound field synthesis techniques have been developed, namely wave field synthesis (WFS) [27] and Ambisonics [28], which use different representations of the sound field. WFS is based on the Kirchhoff-Helmholtz integral, thus, the sound is considered as being produced by either the original primary source or secondary sources spread across the wave front [29]. Ambisonics, on the other hand, is based on the theory of spherical harmonic decomposition of sound, which is able to consistently reproduce the sound field at the centre of the loudspeaker array, i.e. the sweet spot.

A disadvantage of sound field synthesis is the number of microphones and loudspeakers required to achieve accurate measurement and reproduction of the original environment. In general, implementations with a lower number of microphones and/or loudspeakers result in less accurate sound reproduction. For Ambisonics systems, this is caused by the limited order of spherical harmonic decomposition available, whereas WFS is affected by aliasing and the truncation effect. However, there are countermeasures available for Ambisonics. With Ambisonics, a spherical microphone array is utilised to measure the acoustical properties of the original environment. The recordings will be mapped to the signals projected by the loudspeaker array using an Ambisonics decoder [30]. Some decoders are designed in such a way that the reproduced sound field will achieve the accuracy with the spherical harmonic decomposition of higher order [31, 32, 33]. The availability of such techniques allows users to achieve high quality Ambisonics-based audio VR using a relatively small number of microphones and loudspeakers, allowing the technique to be more affordable and popular option in recent years.

The previous studies investigating speech intelligibility under audio VR [23, 24, 25] are all limited to BSR. To date, no report has been made that investigated speech intelligibility under loudspeaker array-based audio VR. A state-of-the-art study by Dagan *et al.* [25] reported that only first order spherical harmonic decomposition was sufficient in order to benefit from spatial release from masking under an anechoic environment using BSR. However, they also suggest further research to investigate whether first order spherical harmonics is sufficient under other conditions. Being inspired by this suggestion, the present research investigates the effect of noise with various spatial acoustics on speech intelligibility using a first order Ambisonics-based 16 channel loudspeaker array sound reproduction system (SRS) built in an anechoic chamber. In terms of the varying spatial acoustics, a particular focus is given to the amount of reverberation as well as the angular separation between the target speech and noise, neither of which has been investigated in the previous study. The reasons for paying special attention to the first order

Ambisonics is not only because of the suggestion by the previous study but also because various first order Ambisonics microphone arrays are available in the current market. Understanding its performance would be of particular interest to practical engineers using audio VR. Hence, all the hardware and software used in this study to build the SRS are available in the current market in order to allow readers to access the same system.

The rest of this paper is organised as follows. The detailed specifications of the Ambisonics-based SRS used in this study are presented in Section 2. Section 3 reports the result of a preliminary experiment that measured the accuracy of sound source localisation realised by the SRS. Section 4 describes the detailed design of the subjective listening test, followed by its results and discussion in Section 5. The paper is finally concluded and future research in this area is identified in Section 6.

2. Ambisonics-based sound reproduction system

The overall process of the Ambisonics-based SRS used in this study, specified in Figure 1, consists of three steps: *Measurement*, *Integration* and *Playback*. In the Measurement step, the effect of room acoustics is collected by measuring the room impulse responses of actual rooms using an Ambisonics microphone array. In the Integration step, sound files of A-format Ambisonics recordings are generated by integrating the measured room impulse responses with a dry sound source. Finally, in the Playback step, the recordings passed from the Integration step is converted to files played from a spherical loudspeaker array using an Ambisonics decoder.

2.1. Measurement

Room impulse responses were measured to capture the acoustical characteristics of physical rooms. In this study, three rooms were selected with varying degrees of reverberation so as to model a variety of listening environments. The three rooms were living room, lecture theatre and church, the acoustical parameters of which are detailed in Table 1. The key factors of interest with regard to the acoustics of the rooms were reverberation time (RT60) and speech clarity (C50) [34]. The T20 measure was used for calculating the RT60 values in Table 1. As summarised in Table 1, the living room had the shortest RT60 while the church had the longest. C50 is an objective quantification of speech intelligibility, which takes into account the energy ratio between direct and reverberant component of a measured room impulse response. In general, a higher C50 suggests better speech intelligibility. Due to C50's dependency on multiple factors including RT60 as well as the distance between the source and receiver of the measurement, all rooms were measured at the same distance of 2.5 m, resulting in the living room measuring the highest C50 and the church the lowest C50.

Measurement specifications. The room impulse responses were measured by playing a swept sine signal of 131072 samples at sampling frequency of 96 kHz. The swept sine signal was played over a loudspeaker (Genelec 8020D) connected to an audio interface (Roland Octa-Capture), which was USB-connected to a Windows 10 laptop computer. A first order Ambisonics microphone (Rode NT-SF1) was utilised for recording the swept sine signal. The microphone contains four uni-directional microphone capsules organised into a tetrahedral arrangement, each of which was connected to the input channels of the audio interface. The sensitivity of each capsule and the amplifier gain of the input channels were calibrated to be equal to one another. The loudspeaker was placed at a variety of azimuth angles in each room in order to capture the 3D spatial sound effect of a single speaker facing the listener at different points in space. Both the cone of the loudspeaker and the centre of the Ambisonics microphone array were set at a height of 1.51 m, which was maintained across the rooms measured.

2.2. Integration

Once room impulse responses were measured, they were integrated with the signal of a sound source to generate the A-format Ambisonics recording. This process includes applying a high-pass filter to the sound source to remove low frequency noise below 60 Hz and normalising them to ensure the energy of the audio files are equal, before convolving it with the room impulse responses. Convolution took place at the sampling frequency of the sound source by downsampling the measured room impulse response. After convolution, the recording was sent to the Playback step. Matlab® was used for the processes in the Integration step.

2.3. Playback

The Ambisonics recording file generated in the Integration step was decoded and played from a 16-channel spherical loudspeaker array through an audio interface (MOTU 16A) at the sampling frequency of 44.1 kHz. To acquire the best performance of the SRS, listeners sat inside the array on a chair which placed their ears at the approximate centre of the sphere.

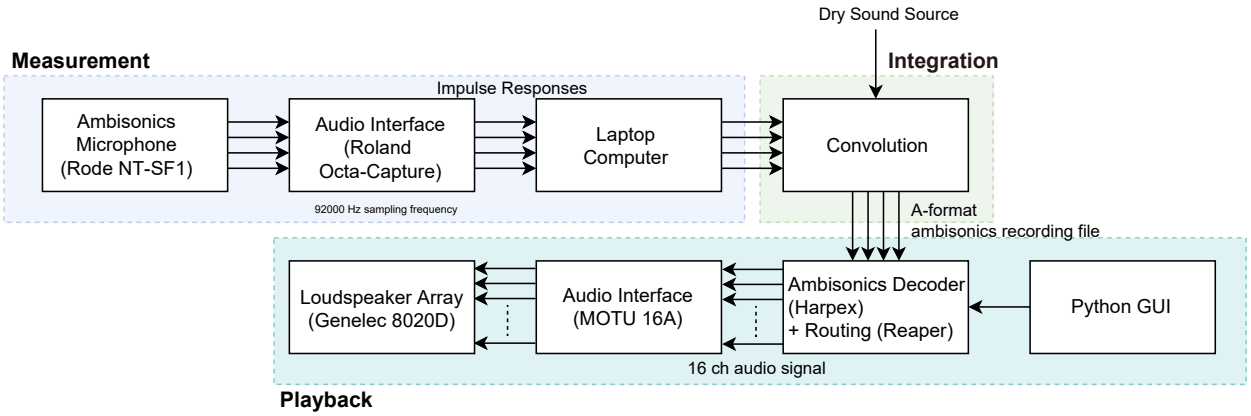


Figure 1: Flowchart of signal processing implemented in the developed sound reproduction system.

Table 1: Properties of measured rooms (RT60 and C50 values are averaged over the frequency range of 250 Hz to 4 kHz)

Room	Approx. volume (m^3)	RT60 (s)	C50 (dB)	Description
Living Room	80	0.40	10.33	A sound listening room at the University of Auckland. Simulates a typical domestic environment (living room). Small in volume. Carpeted, plain painted walls, curtains and soft wall panels, plush furnishings.
Lecture Theatre	900	0.86	7.96	Lecture theatre MLT1 at the University of Auckland. Used for lectures. Medium volume. Carpeted, plain painted walls/whiteboards (reflective surfaces at front), flat plastic desks and plush chairs.
Church	2500	2.03	4.96	Saint Lukes Church in Remuera, Auckland. Used for sermons, music performances, etc. Large volume. Carpet floor for half, wooden floor for other half, large stained glass windows, stone walls, high vaulted ceiling, wooden pews.

Ambisonics decoder specifications. For the decoding process, a Harpex Ambisonics decoder [32] was utilised, and Reaper, a free digital audio workstation, was used for routing the decoded signals to the loudspeakers by setting two tracks to cover the 16 channels. Each hardware send in Reaper was connected to a single output channel of the audio interface, which in turn was physically connected to each loudspeaker.

Loudspeaker array specifications. A spherical loudspeaker array system was built in the anechoic chamber of the Acoustics Laboratory at the University of Auckland as shown in Figure 2. The array consisted of 16 loudspeakers (Genelec 8020D) that were supported by aluminium frames shaped to form a sphere (diameter of 4.25 m) and positioned at regular intervals along the sphere. As can be seen in Figure 3, eight loudspeakers were placed along the plane at the level of the centre of the sphere (middle ring). Four were placed along the top ring, angled inwards at 45 degrees towards the centre of the sphere, and another four were placed along the bottom ring in a similar manner, again pointing at the centre of the sphere. To support each loudspeaker, a 3D-printed mounting bracket was made using PLA filament.

Individual loudspeakers were calibrated to ensure that the same sound level was achieved at each loudspeaker when presented with the same 1 kHz tone.

3. Localisation performance of SRS

To quantify the performance of the system in terms of reproducing spatial acoustics of real environments, a pilot test was carried out. The test examined how listeners perceive the intended direction of the sound source reproduced by the SRS, given that one of the key interests of this study is the effect of spatial release from masking. Participants were seated at the centre of the SRS,

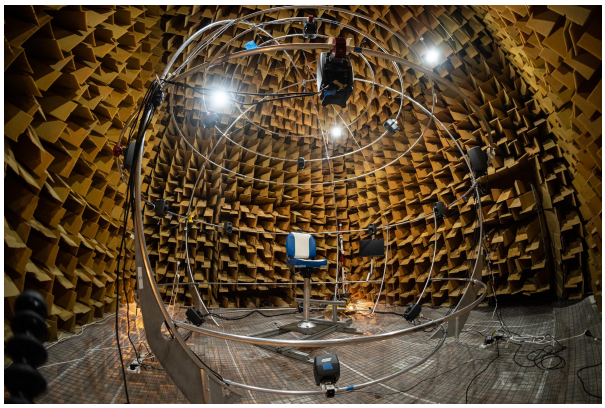


Figure 2: 16 channel spherical loudspeaker array built in the anechoic chamber of the University of Auckland.

and they were told to indicate on a circular dial the direction of where the target stimulus came from for each trial, similar to that used in [35]. The minimum angular resolution of the responses was one degree.

There were two types of target sounds: speech spoken by a female and pink noise. Each target sound was around 2 to 3 seconds long. These target sounds were convolved with the room impulse response measured at one of the eight directions (every 45 degrees) in one of the rooms as specified in Table 1 as per the procedure described in Section 2. As the baseline measurement was under an anechoic environment, the localisation performance was also measured by projecting the stimulus directly from one of the eight loudspeakers in the middle ring of the SRS.

The participants were given a practice run of 10 trials, where they familiarised with the system, using the same stimuli as those used in the main test. Altogether, each participant heard 64 stimuli, i.e. 8 angles $\times$ 2 target sounds $\times$ 4 room acoustics (3 room types + 1 baseline), per block. We collected data from 8 participants, with 5 of the participants taking part in 2 blocks, and the rest 1 block, giving us a total of 13 data points per condition. The reproducibility of the test was examined by the test-retest reliability via correlation of the two sets of responses given by the five participants who repeated the test (using the R function *testretest* in the package *psych* [36]). There was a 0.80 correlation of the responses over the two repetitions suggesting the test is reproducible.

Figure 4 shows the confusion matrices of the responses for the localisation test. As the responses were in the unit of one degree, they were clustered to their nearest 45 degrees angle before the analysis. The columns consist of the responses in terms of the four types of room acoustics and the rows are separated by the target type. The horizontal axis for each matrix indicates the target angle, where the stimulus was intended to sound from, and vertical axis indicates the response angle, where the participants responded the sound to have come from. The number in the tiles denotes the number of responses for the particular combination of target and response angles. The darker the tiles are, the more responses there are for that particular combination. A blank tile indicates zero responses. This means that if the participants responded correctly for all angles (e.g. answering 90 degrees for a target stimulus coming from 90 degrees), then the anti-diagonal tiles of the upper half of the matrices would be filled. This can be seen in Figure 4 for the anechoic case, where listeners were mostly able to localise the correct angle of the sound source image. Confusions for the anechoic case mostly occurred at 45 degrees on either side of the target angle. In contrast, more confusions were observed from the acoustics reproduced by the SRS. As seen in the right three columns of Figure 4, in general, participants were able to indicate somewhat correctly where the source was located, suggesting that the SRS was operating as expected. Participants were able to locate the target sound within a 90 - 135 degrees range, with some front-back confusion [37] made for the 0 and 180 degrees target angle.

To further analyse the localisation performance, Figure 5 shows the circular mean and mean resultant vector length [38], which are concerned with the accuracy and precision of the responses, respectively. Each plot shows the results of pink noise for each combination of room and angle as well as that for speech in each room at 0 degree only. These results will further be discussed in Section 5.

4. Experimental design

A subjective listening test was conducted in order to investigate the effect of spatial acoustics on speech intelligibility under the Ambisonics-based SRS summarised in Section 2.

4.1. Stimuli

The stimuli used in the test consisted of target speech and noise, which were projected simultaneously to evaluate the effect of the noise on the intelligibility of the speech. There were four key parameters involved in generating the stimuli: speech sentence, noise type, room acoustics, and spatial (angular) separation between the speech and noise sources.

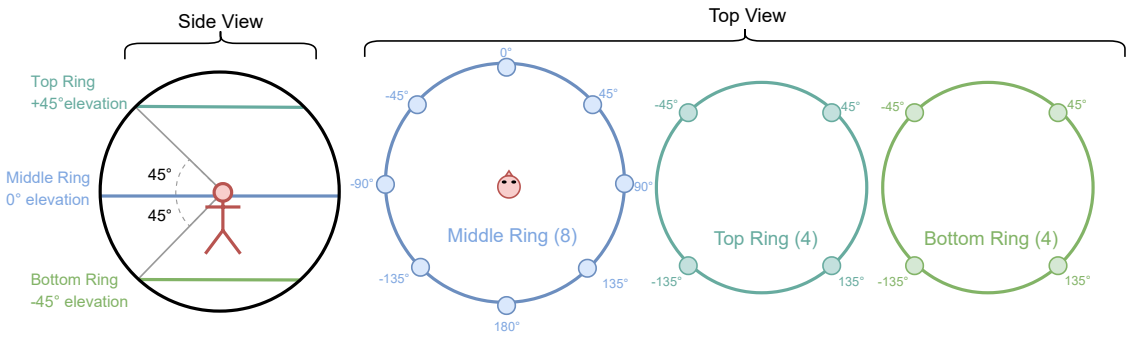


Figure 3: Side and top views of the loudspeaker rig. Shaded circles denote loudspeaker placement.

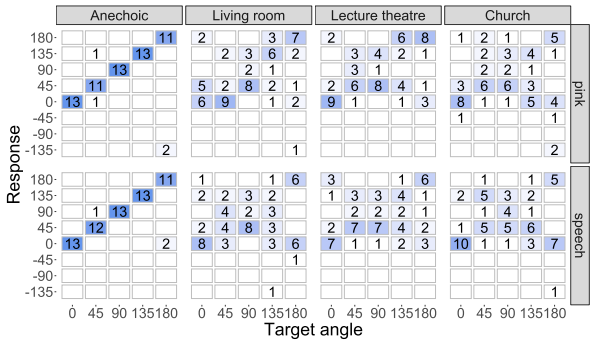


Figure 4: Confusion matrix of localisation test.

4.1.1. *Speech sentences*

Nonsense sentences were used as the target speech in order to avoid the possibility of inference arising from the presence of lexical patterns and contextual cues. Sentences were selected from the previous study [39] using the speech corpus reposited in the Online Speech/Corpora Archive and Analysis Resource [40]. The sentences are grammatically correct and consist of monosyllabic words randomly generated from the 2000 most commonly used words in English [41]. Each sentence consists of four keywords with an even spread of consonants and vowels. With all sentences abiding by the same structure of:

The *(adjective)* *(noun)* *(verb, past tense)* the *(noun)*,

it allows participants to focus on phonemes without considerations about sentence length and meaning. All sentences were spoken by a female speaker with North American English accent at a rate of 130 words per minute, which is slightly below a comfortable talking speed [42], and were sampled at 22.05 kHz.

4.1.2. *Noise type*

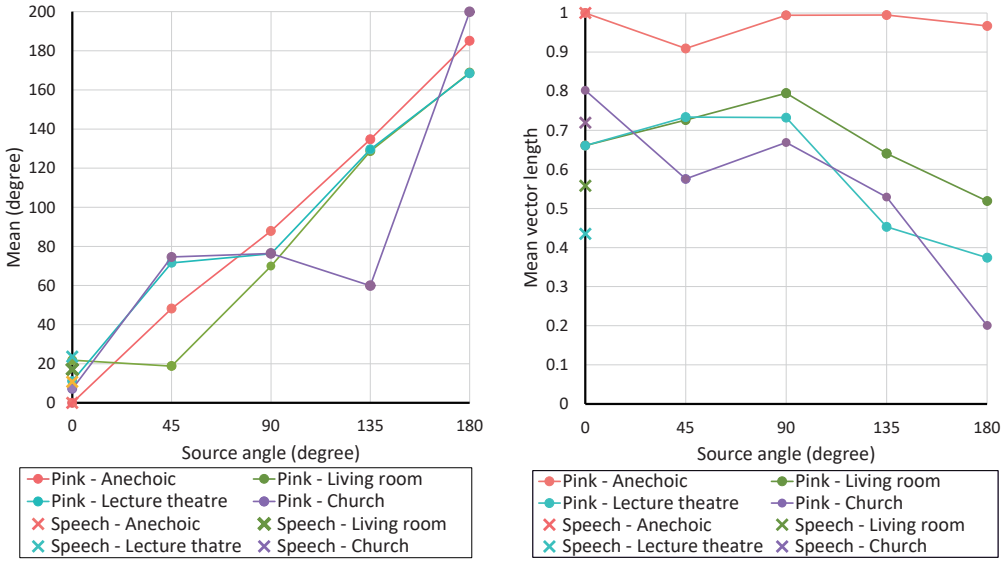
Pink noise, which is a commonly used noise in previous studies investigating the effect of speech masking, was used as the noise source in this study. It represents typical stationary noises seen in real world such as heating, ventilation and air conditioning noise. The noise was generated by Adobe Audition at the sampling rate of 22.05 kHz.

Figure 6 shows the average power spectral density of the target speech (calculated from average of five sentences used in the experiment) and pink noise used for generating the stimuli. Energy of both signals are dominant in the lower frequency band wherein accurate sound reproduction can be realised by first order Ambisonics [43].

Noise level and Calibration. Due to time constraints on the listening test, and also to follow the condition of previous studies [7, 8, 9, 10, 44], only a single target-masker-ratio of -3 dB was used in this study. This value showed neither a ceiling nor flooring effect possibility in a pilot test we ran beforehand. Therefore, the full system, utilising 16 loudspeakers, was calibrated to ensure that the target sound was played at 53 dBA (± 1 dBA), with the masking noise played at 56 dBA to achieve a target-masker-ratio of -3 dBA at the participant's seat. Calibration was performed through the use of a G.R.A.S 46AE free field microphone positioned at the centre of the loudspeaker array and pointed directly upwards.

4.1.3. *Room acoustics*

To investigate the effect of acoustic properties on speech intelligibility, three rooms with different acoustical properties presented in Section 2.1 were also used in the experiment. Alongside the three simulated room acoustics, stimuli were directly played from loudspeakers without the use of the SRS, i.e the speech was played from the 0 degree azimuth loudspeaker and noise was



(a) Circular mean

(b) Mean resultant vector length

Figure 5: Statistics of localisation test result.

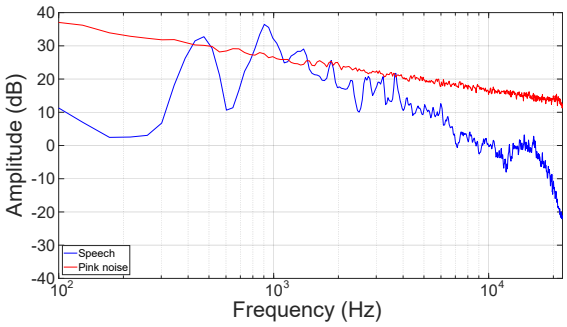


Figure 6: Average spectrum of the speech sentences and pink noise used for generating the stimuli.

played from either the same or a different loudspeaker located at the angles from 45 to 180 degrees on the middle ring. It was expected that results from this case would be a projection of the baseline performance of spatial release from masking.

4.1.4. Spatial separation

The effect of spatial release from masking was investigated by varying the position of the noise source from five angles on the right, namely 0, 45, 90, 135 and 180 degrees while the angle of the target speech was always fixed at 0 degree. Due to time constraints only the angles on the right side were tested. Colocation, or the placement of the noise at 0 degrees, defined the case where target speech and noise were projected from the same position. The effect of elevation changes were out of the scope of this study.

4.2. Testing environment

Testing was performed in the anechoic chamber at the University of Auckland where the SRS was built, with the chamber measuring a negligible reverberation time of 0.04 s. A monitor was installed below the 0 degree azimuth loudspeaker and controlled by a wireless keyboard for participants to enter their answer through a graphical user interface (GUI) as described in Section 4.4.3.

4.3. Participants

Twenty participants aged between 18-49 years old, who speak English as the first language, were recruited from the student and staff of the University of Auckland. For the definition of first language, for this study, all participants were either born in an English speaking country or arrived there before the age of seven [45] and most of them were speakers of New Zealand English. Participants were provided with a questionnaire prior to the commencement of the test and only those without self reports of hearing issues proceeded.

4.4. Test Procedure

Testing was performed by following the procedures summarised below, which was approved by the University of Auckland Human Participants Ethics Committee.

Table 2: Parameters tested in the experiment

Room acoustics	Anechoic Living room Lecture theatre Church
Speech-noise separation	0°, 45°, 90°, 135°, 180°

4.4.1. Pre-test preparation

Upon arrival, participants were presented with an information sheet outlining the testing procedure, including the chosen stimuli and the importance of head orientation and maintaining it within the sweet spot. Participants were then seated to a chair placed in the middle of the loudspeaker array, with adjustments made to the height of the chair ensuring that their ears were level with the middle ring loudspeakers and parallel to the ±90 degrees loudspeakers.

4.4.2. Test Format

Participants were required to transcribe speech sentences that were played simultaneously with noise. The sentences were randomised to prevent participants adapting to the stimuli but the same order was kept consistent between all participants.

A practice test of five sentences was used to ensure participants were confident with testing procedures, with the initial practice sentence being played without noise to allow participants to become familiar with the pitch of the speaker and sentence syntax. The latter sentences each replicated different room acoustics, with the first two from the anechoic (baseline) case, and the last two sentences resembling the easiest and hardest cases of the stimuli using the SRS. Participants proceeded to the main test once they were confident with the testing protocols.

The test involved the stimuli discussed in Section 4.1 where each combination of the parameters included four repetitions, thus, participants were required to transcribe a total of 80 masked sentences (4 room acoustics $\times$ 5 spatial separation $\times$ 4 repetitions).

4.4.3. Graphical user interface

A GUI was developed in Python using Reaper Application Programming Interface (API) action commands and was used to control the playback of sentences as well as recording participants' responses. Participants were given 30 seconds to enter their answers before the GUI automatically moved on to the following sentence. The option of bypassing the timer and pressing 'Submit' to submit answers before the 30 seconds elapsed was also provided. Figure 7 shows an example view of the GUI.

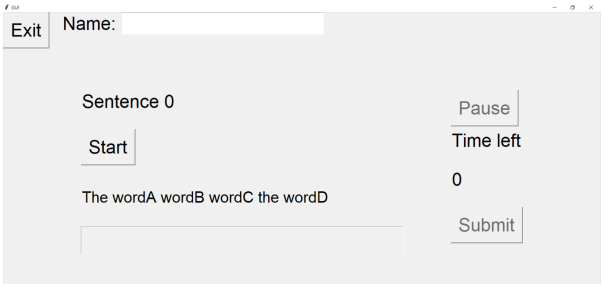


Figure 7: Example view of the GUI used in the test.

4.5. Marking rubric

To quantify speech intelligibility, marking was performed manually according to the common errors discussed by Nye *et al.* [39]. Marks were presented as a score out of a possible four marks (one mark for each non-"the" word in a sentence) before being converted to a percentage for ease of interpretation. Inaccuracies were generally classified into two categories, "phonetic" and "word". The former encompasses substitution (e.g. neck for net), insertion (e.g. find for fine) and deletion (e.g. fend for friend) of phonemes. These were rewarded half marks as it demonstrated that the participant was able to identify a similar word. The word category highlights the removal of words (omission), or the incorrect positioning of correctly identified words (transposition). While cases for omission were penalised by the number of missing words (e.g. three marks out of four if one word was missing), cases for transposition were still awarded full marks as these could be ascribable to bad memory. Further considerations, given that many of the participants were speakers of New Zealand English, homophones and words with the phonemes /iə/ and /eə/ (these vowels have merged for New Zealand English speakers [46], creating homonyms such as ear/air, hear/hair, spear/spare), were also left unpenalised due to the inability to differentiate between phonetically similar words without context.

4.6. Statistical analysis

The speech intelligibility results in terms of the percentage correct were analysed using a linear mixed effect model (LME) with the R [47] package *lme4* [48] and model fitting was carried out using the step function from *lmerTest* [49]. Interactions between two and more factors were included when it improved the fitness of the model. Significance in fixed effects was determined

using a likelihood ratio test by comparing between a model with the effect in question with a model without the effect. Post-hoc pairwise comparisons of the models were carried out using the *emmeans* package [50] with p-values adjusted using the Tukey method.

5. Results and discussion

5.1. Results

Figure 8 shows the mean and the 95% confidence interval of the percentage correct results from the intelligibility test in terms of room type (anechoic, living room, lecture theatre, church). In each plot, the x-axis represents the speech-noise separation (the angular separation between the target speech and noise from 0 to 180 degrees), and the y-axis shows the percentage correct from the speech intelligibility test. The raw data shows that in general, speech intelligibility scores were the highest for the anechoic case compared to the three rooms with reverberation reproduced by the SRS. Higher accuracy rates were observed for 45 to 180 degrees compared to 0 degree, where the noise position overlapped the target speech.

To analyse the data, we constructed a model with a two-way interaction between the speech-noise separation and the room as fixed effects. For the random effect, the participant ID was included. We found a significant two-way interaction using the model ($\chi^2(12) = 139.04$, $p < 0.0001$) from a likelihood ratio comparison. Figure 9 shows the predicted probabilities of percentage correct for the three reverberant rooms and anechoic case from the LME model. The effect of the room can be observed in the speech intelligibility score for the different speech-noise separation, confirming the two-way interaction from the linear mixed model analysis.

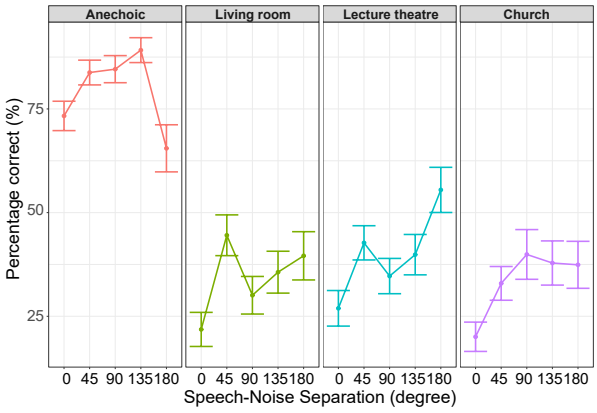


Figure 8: Raw percentage correct results for the intelligibility test.

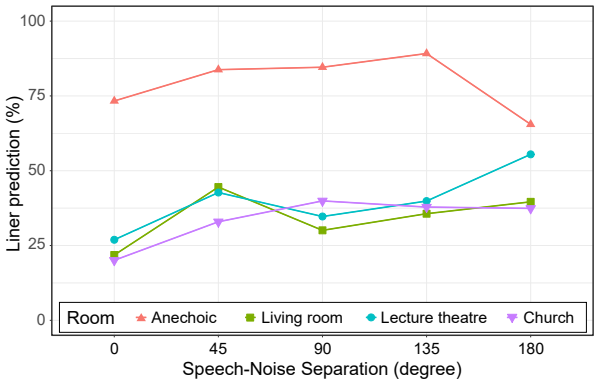


Figure 9: Predicted probabilities of percentage correct from the linear mixed effect model.

Table 3 shows the post-hoc results of pairwise comparisons between the speech-noise separation for the anechoic case. We found significant differences between 0 and 45/90/135 degrees, as well as between 180 and 45/90/135 degrees. Table 4 shows the results of post-hoc comparisons between the speech-noise separation for the three room types. For the living room, participants scored the highest at 45 degrees of separation, and lowest at 0 degree compared to 45, 135 and 180 degrees, but there was no significant difference between 0 and 90 degrees. For the lecture theatre, there were significant differences between 0 degree and all angles except for 90 degrees, where the performance for 0 degree was significantly lower than 45/135/180 degrees. The percentage score at 180 degrees was significantly higher than 0, 45, 90 and 135 degrees. For the church, there were only significant differences between 0 degree and all the other angles, with 0 degree having a lower percentage score than the other angles. The other angles did not differ in how listeners perceived the speech in terms of intelligibility.

Table 5 shows the post-hoc results of pairwise comparisons between the anechoic case with the three reverberant rooms. The performance in the anechoic case is significantly higher than all three rooms at all speech-noise separation angles. Table 6 shows the results of post-hoc comparisons between the simulated room acoustics for the speech-noise separation angles. In terms of the room contrasts, the living room had a significantly lower score for 180 degrees compared to the lecture theatre and for 90 degrees compared to the church. The lecture theatre had a higher score for 45 and 180 degrees compared to the church.

Table 3: Pairwise comparisons of contrasts between speech-noise separation for the anechoic case

Contrast	Estimate(SE)	t.ratio	p.value
0 - 45	-10.46(3.14) 1	-3.34	**0.008**
0 - 90	-11.27(3.14)	-3.60	**0.003**
0 - 135	-15.86(3.14)	-5.06	**<0.0001**
0 - 180	7.81(3.14)	2.49	0.093
45 - 90	-0.81(3.14)	-0.26	1
45 - 135	-5.40(3.14)	-1.72	0.42
45 - 180	18.28(3.14)	5.83	**<0.0001**
90 - 135	-4.59(3.14)	-1.46	0.59
90 - 180	19.09(3.14)	6.09	**<0.0001**
135 - 180	23.68(3.14)	7.55	**<0.0001**

5.2. Discussion

Based on the previous study by Nábělek *et al.* [11], we first hypothesised that reverberation reproduced by the SRS would cause an overall reduction in speech intelligibility. We expected the intelligibility scores of the reverberant rooms to be lower than the baseline (anechoic) and that increased amount of reverberation would cause further reductions in speech intelligibility. We also hypothesised listeners to benefit from spatial release from masking even with the SRS using the first order Ambisonics microphone array [25]. We expected the trend to reflect the result exhibited in the study conducted by Bronkhorst *et al.* [14], suggesting spatial release from masking being least beneficial when the noise source is located at either 0 degree or 180 degrees; for 180 degrees case possibly due to the front-back confusion making the noise source being colocated with the speech. Moreover, we predicted that reverberation would lessen the benefit of spatial release from masking where the extent of reduction would be dependent on the amount of reverberation [19]. Thus, we expected to see the intelligibility curve flattening with increased reverberation.

Pairwise comparisons of speech-noise separation in the baseline case, as found in Table 3, agrees with the results presented in [14]. As in Table 5, it is also evident that the speech intelligibility attained under the SRS is significantly lower than the baseline, thus deeming the hypothesis regarding reduced speech intelligibility with the reproduced reverberation as correct. Despite this, in Table 6, we find that with the exception of a few cases, there was no evidence of significant differences between the different rooms. Thus, we are unable to confirm that intelligibility reduces with increased amount of reverberation realised by the SRS. Lastly, from the results shown in Tables 3 and 4, it is evident that in most cases there are significant differences between 0 degree and 45/90/135 degrees. This suggests that spatial release from masking was maintained under the acoustical environments rendered by the SRS, which align with the results presented by Dagan *et al.* [25]. The contrasts between 0 degree and 90 degrees for the living room and lecture theatre being not significantly different will be discussed later in this section. No clear evidence was found to support the hypothesis of a flattening intelligibility curve by the increase of reverberation. Only a trend was observed when comparing the church to the baseline.

In addition to the above results, several other findings from Tables 3, 4 and 6 are of interest. Firstly, we found that 0 degree and 180 degrees remained significantly different for all environments under the SRS as opposed to the baseline. A plausible explanation for this occurrence can be inferred from the localisation test described in Section 3, where we found the mean resultant vector length for the pink noise was significantly lower at 180 degrees, a trend consistent between all environments as shown in Figure 5(b). While the circular mean values shown in Figure 5(a) are close to 180 degrees, the low mean resultant vector length suggests the stimuli sounded fairly spread out, hindering the ability of the listener localising the noise source with high precision. As a result, the lowered speech intelligibility at 180 degrees observed in the previous study [14] no longer occurred as the listeners failed to recognise the stimuli as being directed from behind, thus resulting in an increase in intelligibility. Secondly, as pointed out earlier, difference of environment did not significantly affect the intelligibility when the speech-noise separation was between 45 and 135 degrees, but with some exceptions. In fact, these exceptions occurred when the mean resultant vector length values at the angles were reasonably different, again suggesting the localisation performance of the SRS affected the perception. It is also interesting to note the unexpectedly lower intelligibility scores at 90 degrees in the living room and lecture theatre, where the 0-90 degrees pairs were no longer significantly different. A possible explanation can be found in the confusion matrix generated from the localisation test, as seen in Figure 4. We found that in the said cases, the localisation of 90 degrees appears to be fairly bimodal with the majority of responses being at 45 degrees and 135 degrees, thus suggesting the importance of high localisation precision for maintaining the effect of spatial release from masking. From these observations, we conclude the effect of spatial release from masking under audio VR environments would have a correlation with the localisation performance of the used SRS. Further

Table 4: Pairwise comparisons of contrasts between speech-noise separation across living room, lecture theatre and church acoustics

	Living room			Lecture theatre			Church		
Contrast	Estimate(SE)	t.ratio	p.value	Estimate(SE)	t.ratio	p.value	Estimate(SE)	t.ratio	p.value
0-45	-22.70(3.15)	-7.22	**<.0001**	-15.79(3.14)	-5.03	**0.0001**	-12.87(3.15)	-4.09	**0.0004**
0-90	-8.24(3.15)	-2.62	0.07	-7.78(3.14)	-2.48	0.096	-19.85(3.15)	-6.31	**<.0001**
0-135	-13.80(3.15)	-4.39	**0.0001**	-12.94(3.14)	-4.13	**0.0004**	-17.79(3.15)	-5.65	**<.0001**
0-180	-17.77(3.16)	-5.63	**<.0001**	-28.55(3.14)	-9.10	**<.0001**	-17.35(3.15)	-5.52	**<.0001**
45-90	14.46(3.14)	4.61	**<.0001**	8.01(3.14)	2.56	0.079	-6.98(3.14)	-2.22	0.17
45-135	8.90(3.14)	2.84	**0.037**	2.85(3.14)	0.91	0.89	-4.91(3.14)	-1.57	0.52
45-180	4.94(3.15)	1.57	0.52	-12.76(3.14)	-4.07	**0.0005**	-4.48(3.14)	-1.43	0.61
90-135	-5.56(3.14)	-1.77	0.39	-5.16(3.14)	-1.65	0.47	2.06(3.14)	0.66	0.96
90-180	-9.53(3.15)	-3.03	**0.02**	-20.78(3.14)	-6.63	**<.0001**	2.50(3.14)	0.80	0.93
135-180	-3.97(3.15)	-1.26	0.72	-15.61(3.14)	-4.98	**0.0001**	0.44(3.14)	0.14	1

investigations focusing on the correlation between the effect of spatial release from masking and the localisation accuracy would be an open problem for future studies.

6. Conclusions

The present study investigated speech intelligibility in noise under various acoustic environments virtually reproduced by an Ambisonics-based SRS. A particular focus was given to investigating the impact of varying the amount of reverberation in rooms as well as the angular separation between the target speech and noise sources. This study was an extension of [25], where they observed spatial release from masking under first order Ambisonics-based binaural sound reproduction. The experimental results of our study suggest that overall, speech intelligibility reduces immensely under reverberant environments reproduced by the SRS compared to the anechoic condition, however, there were no significant differences between environments despite the variation in their reverberation times. The data showed evidence that spatial release from masking continues to be observed under virtually reproduced reverberant environments, however, the effect of varying reverberation time on spatial release from masking could not be concluded. Furthermore, speech intelligibility under the SRS seems to correlate with the localisation accuracy of the SRS. This is supported by the fact that lower localisation accuracy was often observed in the cases where the trend of the speech intelligibility score did not agree with similar previous reports studied in real (non virtual) acoustic environments.

The scope of this study has been limited to testing speech intelligibility under pink noise, directed from the azimuthal angles on the right semicircle, and using a specific design of SRS with a first order Ambisonics microphone array. Testing could be performed by introducing alternative types of maskers, and extending their directions to both left and right angles. Further study using a higher order Ambisonics-based SRS is also recommended to investigate the relationship between speech intelligibility and the localisation performance of the SRS. Lastly, comparing the result with the speech intelligibility measured in the actual rooms would also provide interesting insights.

Acknowledgements

The authors would like to thank all anonymous participants of the subjective listening tests conducted for this study. Special thanks to Dr George Dodd of the Acoustics Research Centre, The University of Auckland, for arranging the authors' visit to measure the room impulse responses and the dimension of Saint Lukes Church. This work was supported by Faculty Research Development Fund at the University of Auckland.

References

[1] G. A. Miller, J. C. R. Licklider, The intelligibility of interrupted speech, The Journal of the Acoustical Society of America 22 (2) (1950) 167–173.

[2] T. Houtgast, H. J. M. Steeneken, A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria, The Journal of the Acoustical Society of America 77 (3) (1985) 1069–1077.

[3] L. H. Mayo, M. Florentine, S. Buus, Age of second-language acquisition and perception of speech in noise, Journal of Speech, Language, and Hearing Research 40 (3) (1997) 686–693.

[4] M. L. G. Lecumberri, M. Cooke, A. Cutler, Non-native speech perception in adverse conditions: A review, Speech Communication 52 (11-12) (2010) 864 – 886.

Table 5: Pairwise comparisons of contrasts between the anechoic case and the room types at different speech-masker separation

	Anechoic - Living room			Anechoic - Living room			Anechoic - Church		
Angle	Estimate(SE)	t.ratio	p.value	Estimate(SE)	t.ratio	p.value	Estimate(SE)	t.ratio	p.value
0	51.48(3.15)	-16.36	**<0.0001**	46.39(3.14)	14.79	**<0.0001**	53.25(3.15)	16.93	**<0.0001**
45	39.24(3.14)	12.51	**<0.0001**	41.06(3.14)	13.09	**<0.0001**	50.84(3.14)	16.21	**<0.0001**
90	-54.51(3.14)	17.38	**<0.0001**	49.89(3.14)	15.91	**<0.0001**	44.68(3.14)	14.25	**<0.0001**
135	53.54(3.14)	17.07	**<0.0001**	49.31(3.14)	15.72	**<0.0001**	51.32(3.14)	16.37	**<0.0001**
180	25.90(3.15)	8.23	**<0.0001**	10.02(3.14)	3.20	**<0.008**	28.09(3.14)	8.96	**<0.0001**

Table 6: Pairwise comparisons of contrasts between the room types at different speech-masker separation

	Living room - Lecture theatre			Living room - Church			Lecture theatre - Church		
Angle	Estimate(SE)	t.ratio	p.value	Estimate(SE)	t.ratio	p.value	Estimate(SE)	t.ratio	p.value
0	-5.09(3.15)	-1.62	0.37	1.77(3.16)	0.56	0.94	6.86(3.15)	2.18	0.13
45	1.82(3.14)	0.58	0.94	11.6(3.14)	3.7	**0.001**	9.8(3.14)	3.11	**0.01**
90	-4.62(3.14)	-1.48	0.45	-9.84(3.14)	-3.14	**0.009**	-5.21(3.14)	-1.66	0.34
135	-4.22(3.14)	-1.35	0.53	-2.21(3.14)	-0.71	0.89	2.01(3.14)	0.64	0.91
180	-15.87(3.15)	-5.05	**<.0001**	2.19(3.15)	-0.70	0.90	18.06(3.14)	5.76	**<.0001**

[5] D. S. Brungart, P. S. Chang, B. D. Simpson, D. Wang, Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation, The Journal of the Acoustical Society of America 120 (6) (2006) 4007–4018.

[6] D. D. Greenwood, Auditory masking and the critical band, The Journal of the Acoustical Society of America 33 (4) (1961) 484–502.

[7] A. Ito, A. Miki, Y. Shimizu, K. Ueno, H. Lee, S. Sakamoto, Oral information masking considering room environmental condition part1: Synthesis of maskers and examination on their masking efficiency part1 synthesis of maskers and examination of their masking efficiency, in: Proceedings of Inter-Noise 2007, 2007.

[8] B. Jing, A. Liebl, P. Leistner, J. Yang, Sound masking performance of time-reversed masker processed from the target speech, Acta Acustica United with Acustica 98 (4) (2012) 135–141.

[9] Y. Hioka, J. Tang, J. Wan, Effect of adding artificial reverberation to speech-like masking sound, Applied Acoustics 114 (2016) 171–178.

[10] Y. Hioka, J. James, C. I. Watson, Masker design for real-time informational masking with mitigated annoyance, Applied Acoustics 159 (2020) 1070–1073.

[11] A. Nábělek, P. Robinson, Monaural and binaural speech perception in reverberation for listeners of various ages, The Journal of the Acoustical Society of America 71 (5) (1982) 1242–1248.

[12] A. Nábělek, T. Letowski, F. Tucker, Reverberant overlap- and self-masking in consonant identification, The Journal of the Acoustical Society of America 86 (4) (1989) 1259–1265.

[13] J. Blauert, Spatial hearing : the psychophysics of human sound localization, rev. ed.. Edition, MIT Press, Cambridge, Mass., 1997.

[14] A. Bronkhorst, R. Plomp, The effect of head-induced interaural time and level differences on speech intelligibility in noise, The Journal of the Acoustical Society of America 83 (4) (1988) 1508–1516.

[15] A. W. Bronkhorst, The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions, Acta Acustica united with Acustica 86 (1) (2000) 117–128.

[16] R. Y. Litovsky, Spatial release from masking, Acoustics today 8 (2) (2012) 18–25.

[17] R. Y. Litovsky, Speech intelligibility and spatial release from masking in young children, The Journal of the Acoustical Society of America 117 (5) (2005) 3091–3099.

[18] M. Yadav, J. Kim, D. Cabrera, R. de Dear, Auditory distraction in open-plan office environments: The effect of multi-talker acoustics, Applied Acoustics 126 (2017) 68 – 80.

[19] G. Kidd, C. R. Mason, A. Brughera, W. M. Hartmann, The role of reverberation in release from masking due to spatial separation of sources for speech identification, Acta acustica united with acustica 91 (3) (2005) 526–536.

[20] G. C. Burdea, P. Coiffet, Virtual reality technology, John Wiley & Sons, 2003.

[21] M. Kleiner, B.-I. Dalenbäck, P. Svensson, Auralization-an overview, J. Audio Eng. Soc 41 (11) (1993) 861–875.

[22] M. Vorländer, Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality, Springer Science & Business Media, 2007.

[23] N. R. Shabtai, I. Nehoran, M. Ben-Asher, B. Rafaely, Intelligibility of speech in noise under diotic and dichotic binaural listening, Applied Acoustics 125 (2017) 173–175.

[24] A. Rungta, N. Rewkowski, C. Schissler, P. Robinson, R. Mehra, D. Manocha, Effects of virtual acoustics on target-word identification performance in multi-talker environments, in: Proceedings of the 15th ACM Symposium on Applied Perception, SAP '18, 2018.

[25] G. Dagan, N. R. Shabtai, B. Rafaely, Spatial release from masking for binaural reproduction of speech in noise with varying spherical harmonics order, Applied Acoustics 156 (2019) 258 – 261.

[26] W. Zhang, P. Samarasinghe, H. Chen, T. Abhayapala, Surround by Sound: A Review of Spatial Audio Recording and Reproduction, Applied Sciences 7 (5) 532.

[27] E. Hänsler, G. Schmidt (Eds.), Wave Field Synthesis Techniques for Spatial Sound Reproduction, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 517–545.

[28] F. Zotter, M. Frank, Ambisonics - A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality-, 1st Edition, Springer International Publishing, 2019.

[29] J. Daniel, S. Moreau, R. Nicol, Further Investigations of High-

Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging, in: Audio Engineering Society Convention 114, 2003.
[30] M. A. Gerzon, Surround-sound psychoacoustics, Wireless World 80 (1468) (1974) 483–486.
[31] J. Vilkamo, T. Lokki, V. Pulkki, Directional audio coding: Virtual microphone-based synthesis and subjective evaluation, Journal of the Audio Engineering Society 57 (9) (2009) 709–724.
[32] S. Berge, N. Barrett, High angular resolution planewave expansion, in: Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics May, 2010, pp. 6–7.
[33] A. Politis, S. Tervo, V. Pulkki, Compass: Coding and multidirectional parameterization of ambisonic sound scenes, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 6802–6806.
[34] V. Pulkki, M. Karjalainen, Communication acoustics: an introduction to speech, audio and psychoacoustics, John Wiley & Sons, 2015.
[35] M. Schoeffler, S. Westphal, A. Adami, H. Bayerlein, J. Herre, Comparison of a 2D-and 3D-based graphical user interface for localization listening tests, in: Proc. of the EAA Joint Symposium on Auralization and Ambisonics, 2014, pp. 107–112.
[36] W. R. Revelle, psych: Procedures for personality and psychological research (2017).
[37] J. Middlebrooks, Chapter 6. Sound localization, Handbook of clinical neurology 129C (2015) 99–116.
[38] P. Berens, CircStat: a MATLAB toolbox for circular statistics, Journal of Statistical Software 31 (10) (2009) 1–21.
[39] P. Nye, J. Gaitenby, The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences, Haskins Laboratories status report on speech research 37 (38) (1974) 169–190.
[40] S. Brouwer, K. J. V. Engen, L. Calandruccio, A. R. Bradlow, Linguistic contributions to speech-on-speech masking for native and non-native listeners: language familiarity and semantic content., The Journal of the Acoustical Society of America 131 (2) (2012) 1449–64.
[41] E. L. Thorndike, I. Lorge, The teacher's word book of 30,000 words. (1968).
[42] E. Rodero, A comparative analysis of speech rate and perception in radio bulletins, Text & Talk 32 (3) (2012) 391–411.
[43] S. Bertet, J. Daniel, E. Parizet, O. Warusfel, Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources, Acta Acustica united with Acustica 99 (4) (2013) 642–657.
[44] H. Masuda, Y. Hioka, J. James, C. I. Watson, Protecting speech privacy from native/non-native listeners - effect of masker type, in: Proc. of the 19th International Congress of Phonetic Sciences, 2019, pp. 3070–3074.
[45] H. Nicholas, P. M. Lightbown, Defining child second language acquisition, defining roles for L2 instruction, Second language acquisition and the younger learner: Child's play (2008) 27–51.
[46] M. A. Maclagan, E. Gordon, Out of the AIR and into the EAR: Another view of the New Zealand diphthong merger, Language Variation and Change 8 (1) (1996) 125–147.
[47] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2019).
URL https://www.R-project.org/
[48] D. Bates, Linear mixed model implementation in lme4, Manuscript, University of Wisconsin 15 (2007).
[49] lmerTest Package: Tests in Linear Mixed Effects Models, Journal of Statistical Software 82 (13) (2017).
[50] R. Lenth, emmeans: Estimated Marginal Means, aka Least-Squares Means (2019).
URL https://cran.r-project.org/package=emmeans