SPEAKER DISCRIMINATION IN MULTISOURCE ENVIRONMENTS AURALIZED IN REAL ROOMS

Paper No.306/2020

[a)]Kristian Jambrošić, [b)]Marko Horvat, [c)]Dominik Kisić, [d)]Tin Oberman

[a-c)]*University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia, kristian.jambrosic@fer.hr, marko.horvat@fer.hr, dominik.kisic@fer.hr*
[d)] *University of Zagreb Faculty of Architecture, Zagreb, Croatia, tin.oberman@arhitekt.hr*

Abstract: *With the recent development of audio in modern VR/AR systems and the increasing capability of synthesizing natural sound fields over headphones with head tracking, the question of the ability of our hearing system to discriminate multiple concurrent sound sources has become important again. We must understand how psychoacoustical and psychophysical limitations of the hearing system cope with novel technologies of virtual acoustics that can simulate an almost unlimited number of sound sources. Previous research has shown that the capacity of human hearing to discriminate a reference sound source is limited when there is background noise, a reverberant surrounding, or when other, disturbing sound sources simultaneously mask the reference source. A set of listening tests based on the cocktail-party effect was designed to determine the intelligibility of speech emitted by a reference sound source, with one to six disturbing sound sources simultaneously emitting speech from different directions around the listener. The tests were repeated in three test rooms with different acoustical properties, and two test signals were used: logatomes and regular spoken sentences with specific keywords. The results have revealed the changes in speech intelligibility scores in relation to the number of disturbing sources, their positions, and acoustical properties of test rooms.*

Keywords: *speech intelligibility, multiple sound sources, room acoustics, auralization, cocktail party effect*

DOI: *1036336/akustika20203719*

1. INTRODUCTION

Audio has finally evolved in recent years. For more than half a century, the stereo system has dominated in virtually all applications of recorded sound such as Hi-Fi music, TV broadcast, computer games, car audio, or listening in motion via headphones. In recent years, the rise of virtual/augmented/mixed reality systems (VR/AR/MR) has been giving a strong incentive to rethink the paradigm of audio production. A large number of smartphones is present in the majority of households in the world. Most of these devices have built-in Inertial Measurement Units (IMU) sensors capable of detecting their orientation in space. Therefore, the technological prerequisites for the implementation and use of virtual reality audio are already there, and there is no need for purchasing new equipment.

All audio systems are designed based on how people perceive sound. Even if they are emitting sound simultaneously, human hearing can distinguish several sound sources around the listener, localize them and discriminate any one of them from the others. Moreover, higher brain functions allow us to understand speech even when not all words and syllables are heard clearly due to other distracting sounds or unfavorable environment. There has been a lot of research in the field of psychoacoustics in the last decades that tackles these hearing parameters [1, 2]. The findings of these research groups have become even more important today because of the development of the auralization technologies. These technologies enable us to synthesize sound environments, in which any number of sound sources can be placed anywhere around the listener, and this is the key point of research interest of this paper. The means for creating 3D audio with an arbitrary number of sound sources around the listener have been shown by many researchers [3, 4, 5].

The hearing mechanism in its most basic form is a two-channel microphone pair with its unique directivity, and it uses dynamic positioning of the ears/head to solve the source localization uncertainties. Therefore, to (re)create a perfect and natural listening experience by synthesizing the audio signals for any specific moment and for the current orientation of the listener, two independent audio channels and a way to determine the orientation of the head should be provided. This is already achievable using a smartphone equipped with a pair of headphones.

The usual sound reproduction systems, especially multichannel systems that utilize several spaced loudspeakers, use strongly correlated audio signals to drive individual channels. This is the key principle of stereophonic systems that utilize the amplitude panning law as a way of positioning the phantom sources around the listener [6, 7]. For example, when a stereo audio signal is reproduced over headphones instead of loudspeakers, the direction of the sound sources is found only by interaural level differences (ILDs), while interaural time differences (ITDs) and head-related spectral filtering (HRTF) are not taken into account. On the other hand, real sound sources are mostly uncorrelated and we are used to discriminate them

from one another. It can be argued that fewer sound artefacts are perceived if the sound reproduction system delivers sound to the ears in the manner similar to natural listening in real life, i.e. with proper ILD, ITD and HRTF cues. Sound reproductions systems that can deliver these cues properly are the Ambisonics system, and binaural systems with head position tracking [8, 9, 10].

The novelty of the audio in virtual reality and related systems is that the sound field it is created in different ways than traditional recordings or traditional audio production. Most of audio production for VR systems is done using audio objects, e.g. sound sources with a certain position or movement trajectory in the virtual/augmented space around the listener. As smartphone and computer hardware are not a limiting factor anymore, an arbitrary number of such sound sources can be introduced into the virtual audio environment. Therefore, a new research question arises, or, to be more precise, an old one has again gained importance: how many sound sources placed around can a listener still distinguish from one another, in terms of their position and the content they emit?

There are many more applications of novel audio systems. Many autonomous robots are equipped with microphones, often emulating binaural hearing with two sensors, and it is essential to examine the performance of such systems with regard to sound source localization and separation in real-life, non-anechoic environments [11, 12]. Such systems should be able to localize sources with overlapping frequency spectra, even for moving sources, based on ITDs and ILDs. In some applications, these systems can even outperform the human hearing apparatus. Reverberant and noisy environments are especially demanding and challenging, since the localization performance suffers due to strong reflections and background noise sources added to the direct sound of the observed sound source [13]. To improve performance, mobile robots often use more than two microphones configured as an array that can be optimized for different applications, sound source types or environments [14].

Another hot topic in multiple sound source processing is the extraction of the content of a particular sound source from a recording that contains a certain amount of directional information, but is not binaural by nature. A good example is the Ambisonics system, in which the sound field recordings are made with various Ambisonics microphones ranging from first to fourth order. For this system, different algorithms and methods have been developed to improve the speech intelligibility / music quality / channel separation for a sound source positioned in the chosen direction [15].

The evaluation of any sound localization experiment is traditionally done by listening tests with a group of test persons. It is crucial to design a proper test that can reliably examine any parameter under test, and to do it on a sufficiently large sample of test subjects to get statistically significant results. This is because there is always variability, sometimes quite a lot, on how each person perceives sound. Many researchers have proposed grounding principles for perceptual testing that is routinely done today, since they allow the tests to be optimized for various types of tested sound source settings [16, 17]. Many decisions must be made before starting the test design. One must choose between in-situ tests in a natural environment, and tests in laboratories where the sound field is recreated using a sound reproduction system. These systems are again differentiated according to the use of loudspeakers and their setup, or headphones, with or without head tracking. If the tests are done in-situ, there are always more interfering conditions that can influence the test results, such as background noise, atmospheric conditions, temperature, etc. On the other hand, laboratory spaces are sometimes too reverberant, and not all that similar to a free field environment that is often found in in-situ experiments, which can also have an influence on the results of the tests.

This paper presents the results of several tests used to recreate a typical cocktail-party scenario where there is always one speaker and one or more simultaneous, interfering speakers. The tests were done in three acoustically different rooms to examine the influence of reverberant environments on the speech intelligibility results as well. The main goal was to determine the capacity of the hearing system to suppress unwanted signals, which is of uttermost importance when designing scenarios for VR/AR/MR systems, designing hearing aid processors, etc. Other authors have done similar research, but not always in a natural, reverberant environment, in which real-life conditions can be simulated [18, 19]. Moreover, these experiments were rarely conducted using many simultaneous natural sound sources that simulate a typical cocktail party setup. The final goal of VR and related systems is to have a seemingly transparent audio-visual system that would allow the listeners to experience the sound field in the same way as they would in a natural environment. These systems regularly use head tracking IMU sensors. Fig. 1 shows an overview of typical binaural systems with head tracking and lists how we evaluate their quality [20].

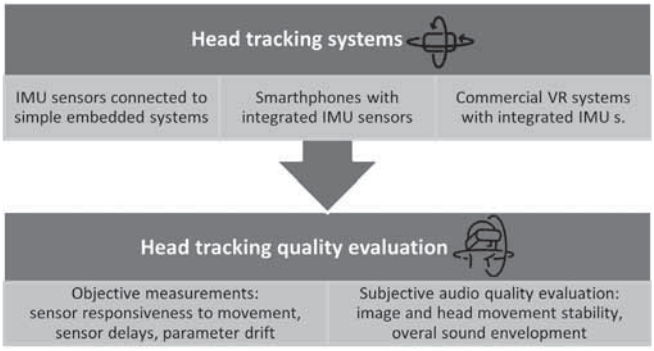

Fig. 1: Use of binaural reproduction with head-tracking for simulating natural sound sources in VR/AR/MR scenarios

2. LISTENING TESTS SETUP

The goal of this research was to determine how the number and position of concurrent, disturbing sources of speech influences the intelligibility of the main speaker signal, thus finding the limitation of the hearing system to resolve the "cocktail-party effect" problem. With this in mind, a test setup was designed to simulate multiple speakers positioned at different directions who speak at the same time, thus reducing the ability of the listeners to single out the speaker they want to listen to and understand. This setup was repeated in three acoustically different rooms in order to examine the influence

of acoustical properties of closed spaces, and to perform the experiment in a realistic environment. It should be noted that all sources were placed at the same height, in the horizontal plane at ear level, which is the usual location of speakers in realistic cocktail-party scenarios. Sources placed in positions outside the horizontal plane would certainly have an additional influence on the outcome of the experiment, but such a setup would be more relevant for other types of disturbing sound sources (music, ventilation, air-conditioning, etc.) than for speech sources.

2.1. Test configuration

The listeners were placed in the center of a circle with the radius of 2 m. The loudspeakers were placed along the circle at fixed angular positions (…, -60°, -30°, 0°, 30°, 60°, …), and put at the ear-level height of a seated person (1.3 m above the floor). The loudspeakers were 2-way full-range active studio monitors with a frequency response within ± 3 dB in the range of 70 - 20000 Hz. The average sound pressure level of the speech signals from a single source at the listener position was calibrated to 70 dBA. Since up to 7 sources with speech signals could be active at the same time depending on the test case, the total sound pressure level at listener position was between 70.0 and 76.5 dBA. All loudspeakers were connected to a multichannel sound card, and a DAW software was used to feed the loudspeakers with synchronized audio signals. All audio signals were sampled at 44.1 kHz, and had 16 bit dynamic resolution.

The loudspeakers were spaced at regular angular intervals of 30°, and the reference loudspeaker was always at the azimuth of 0°, i.e. right in front of the listener, thus mimicking the usual position of the person one wants to speak with. All possible positions of the loudspeakers are shown in Fig. 2.

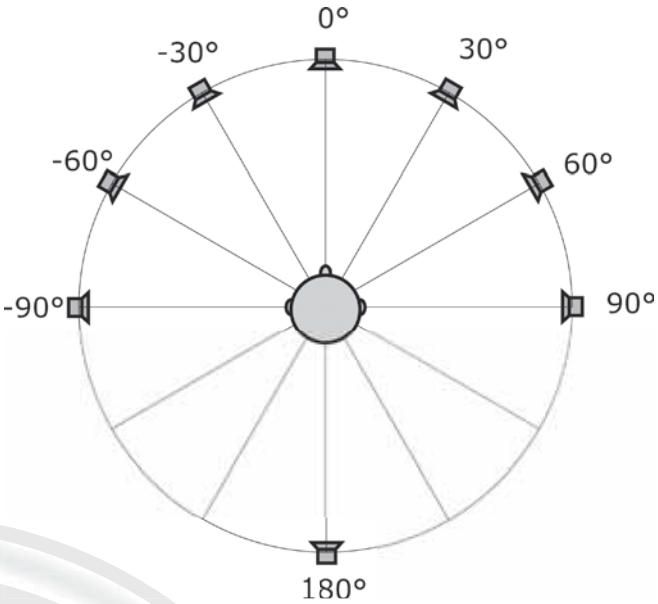

Fig. 2: Plan view of the position of sound sources in regard to the listener

All the tested cases with different number and position of disturbing sound sources are described below. As indicated in in fig. 2, there are 12 possible positions for loudspeakers set along the full circle at regular 30° intervals. The frontal position at azimuth 0° was reserved for the reference loudspeaker, which leaves 11 positions for the disturbing sources. Out of those, 6 positions in the front part of the horizontal plane were used, along with the seventh position at azimuth 180°, i.e. right behind the listener.

2.2. Test rooms

The tests were performed in three different rooms. The loudspeaker configuration was always the same, as shown in Fig. 2. The rooms were rectangular, shoebox-shaped university classrooms. The width and the height of all three rooms was the same, and only the length differed from one room to the next. Acoustical properties of these rooms are different, which reflects in the values of reverberation time. Room 1 is acoustically treated and serves as a listening room. The plan views of all test rooms are shown in Fig. 3. The basic dimensions of the rooms, their corresponding volume, measured reverberation time averaged at middle frequencies, calculated hall radii, and measured background noise levels are shown in Tab. 1. It can be seen that room 1 and room 3 are very similar in shape and volume, although the reverberation time of room 3 is about 2.5 times longer compared to room 1, thus making it a much less favorable room in terms of acoustic comfort. Room 2 is smaller than room 1, but still with twice as long reverberation time, again being less comfortable in terms of reverberation.

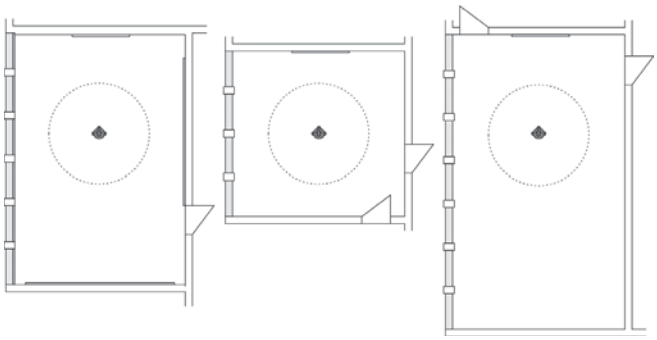

Fig. 3: Plan view of test rooms from left to right: room 1, room 2, and room 3. Sound absorbing materials are shown as grey stripes on walls, listener and orientation is indicated with the listener's head icon, and the loudspeaker rig is shown with the dotted circle

Room	Length (m)	Width (m)	Height (m)	Volume (m^3)	Reverberation time RT (s)	Hall radius (m)	Noise level L_{EQ} (dBA)
Room 1	10.20	7.05	3.20	230	0.57	1.14	31.7
Room 2	6.83	7.05	3.20	154	1.05	0.69	32.9
Room 3	11.95	7.05	3.20	270	1.39	0.79	41.8

Tab. 1: The basic dimensions, volume, reverberation time, hall radius, and background noise levels for the three test rooms

2.3. Test signals

Since the goal of this research is to check the limitations of the hearing system when processing and understanding one speaker among multiple other speakers as distractors, the obvious choice of the test signals for this experiment are speech signals.

Two different types of speech signals were used in the tests as the stimuli. One type of stimuli are logatomes as three-letter syllables without meaning. For the purpose of this test, only the logatomes were used that consist of speech sounds in the configuration: consonant/vocal/consonant. The logatomes were adapted to Croatian language, as all the listeners were native speakers of Croatian. The advantage of using logatomes is obvious: there are no higher cognitive brain functions that can help fill in the missing parts of these logatomes if they were unintelligible for any reason, since these syllables have no meaning and are not pre-memorized. Both the reference speaker and the distracting speakers spoke out the logatomes in synchronicity. With the increasing number of simultaneous speakers, this task was getting increasingly harder to fulfill, as shown in the results. The list of used logatomes for this experiment is shown in Tab. 2. All the logatomes were read by the same male narrator. The task for the listeners was to write down the logatomes as they heard them. An example of a test case in shown in Fig. 4 left.

beć	der	ječ	jer	jos	ker	kis	loz	mar	mis
mos	muk	nan	nat	nog	nok	per	pij	roz	sat
sis	šer	tan	tar	tij	tuk	van	vat	vok	žeć

Tab. 2: Logatomes used in the experiment as the first speech test signal

The second and a more common test signal was a set of pre-recorded spoken sentences. Altogether, 184 sentences were taken from Croatian literature and read by a male narrator to serve as reference speech sounds. The length of the chosen sentences was quite similar, and none of them exceeded 5.5 s in length when read out loud. In each sentence, four keywords were chosen, and the test examined if the listeners have heard these words correctly. The overall intelligibility score in these tests was determined as the percentage of the correctly written keywords out of the total number of keywords in a given test sample. The average RMS level of each recorded sentence was equalized to -24 dBFS, thus avoiding the problems with gain adjustments in the audio chain. The distracting signals were the same kind of sentences read by other male narrators, but they were taken from regular radio programs, typically from news or other spoken programs. The distracting signals were also free from any audible background noise, and their length was set to six seconds. Fade-in and fade-out was applied to the signals in the first and last 250 ms of the recording. The distracting sentences were also equalized to an RMS value of -24 dBFS.

The average spectra of the narrators, both reference and distracting, was analyzed as well. There were no notable differences that would indicate that spectral content could be used as a cue for better processing the referent speech signal. The reference and distracting sentences were played simultaneously, but the distracting sentences started 250 ms before the reference sentence. The task for the listeners was to write down the sentences as they heard them. Before each test sample, a 500 ms sine signal with the frequency of 250 Hz was played back from the reference loudspeaker at azimuth 0° as a cue that the next test sample is about to begin. The distracting speech signals started 500 ms after the sine signal ended, and the reproduction of the reference sentence started another 500 ms later. After the test sample ended, there was a 23 s long period of silence, giving time to the test persons to write down what they have heard, and this completed one test sample. An example of a test sample in shown in Fig. 4 right.

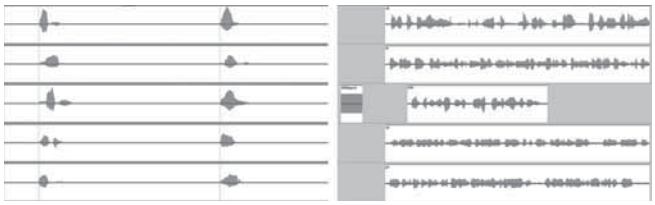

Fig. 4: Left: example of the DAW audio track arrangement in tests with logatomes with one referent and four distracting logatomes; right: example of the arrangement in tests with sentences for one reference speaker (middle row) and 4 distracting speakers

2.4. Listening experiments

Two listening experiments were conducted. The first one was performed using the described set of logatomes, and the second one utilized spoken sentences as the stimuli. Each experiment consisted of several listening tests that differed by the number of distracting sources and their position relative to the reference source, marked with the azimuthal angle of the disturbing source relative to the reference source. Each experiment was conducted in two acoustically different rooms, so that the influence of room acoustics could be examined as well. The first experiment with logatomes was conducted in rooms 1 and 2, and the second experiment with sentences was made in rooms 1 and 3.

The listeners who took part in the experiment were all students and university staff of both sexes. The average age of all the listeners was 26, and 17 listeners were available for each test case. None of the listeners reported any kind of hearing problems.

The first experiment consisted of 4 different test cases, as shown in tab. 3. Test cases 1, 3, and 4 examined the influence of 2, 3, and 4 real distracting sources, respectively. Test case 2 was devised as an additional case in order to examine whether the phantom sources, as perceived in all stereophonic systems, will yield different intelligibility of the logatomes in comparison with real sources. The phantom sources in this test case were created by having appropriate pairs of loudspeakers reproduce the same signals (100% correlation), instead of having real sources at the designated positions. Namely, the phantom reference source at 0° was created by two loudspeakers at -30° and 30°. The two disturbing phantom sources theoretically positioned at -60° and 60° were created by two pairs of loudspeakers positioned at -90° and -30°, and at 30° and 90°, respectively.

Test case	The number of distracting sources	The azimuth of the distracting sources	The azimuth of the reference source
1	2	-60°, 60°	0°
2	2 (4)	-60° (-90° and -30°), 60° (30° and 90°)	0° (-30° and 30°)
3	3	-90°, 90°, 180°	0°
4	4	-60°, -30°, 30°, 60°	0°

Tab. 3: Four test cases defined in experiment 1 with logatomes. The angles in parentheses in case 2 show the position of real sources used for the creation of phantom images at angles indicated in front of the parentheses

The second experiment involved 11 different test cases, as shown in tab. 4. The test cases were sorted in ascending order according to the number of distracting sources, which ranged from 1 to 6. The cases with the same number of distracting sources were additionally sorted according to the increasing angular distance between the reference source (always at 0°) and the disturbing sources.

Test case	The number of distracting sources	The azimuth of the distracting sources	The azimuth of the reference source
1	1	-30°	0°
2	1	-60°	0°
3	1	-90°	0°
4	1	-180°	0°
5	2	-30°, 30°	0°
6	2	-60°, 60°	0°
7	2	-90°, 90°	0°
8	3	-180°, -90°, 90°	0°
9	4	-60°, -30°, 30°, 60°	0°
10	4	-90°, -60°, 60°, 90°	0°
11	6	-90°, -60°, -30°, 30°, 60°, 90°	0°

Tab. 4: Eleven test cases defined in experiment 2 with spoken sentences

3. TEST RESULTS AND DISCUSSION

3.1. Experiment 1 - logatomes

The results of the tests conducted for four test cases in experiment 1 are shown in Tab. 5. The mean percentage of correctly understood logatomes and the standard deviation across the entire group of listeners is indicated for each test case, both for room 1 and room 2. The mean values are also shown in graphic form in Fig. 5.

Test case	The number of distracting sources	The azimuth of the distracting sources	Room 1		Room 2	
			Mean (%)	Standard deviation (%)	Mean (%)	Standard deviation (%)
1	2	-60°, 60°	75,7	7,3	68,4	6,7
2	2 (4)	-60° (-90° and -30°), 60° (30° and 90°)	74,5	7,7	67,6	7,9
4	3	-90°, 90°, 180°	55,9	11,6	47,8	9,1
3	4	-60°, -30°, 30°, 60°	51	11,7	48,6	8,3

Tab. 5: The percentage of correctly understood logatomes for the four test cases in rooms 1 and 2 as the mean and the standard deviation for the tested group of listeners.

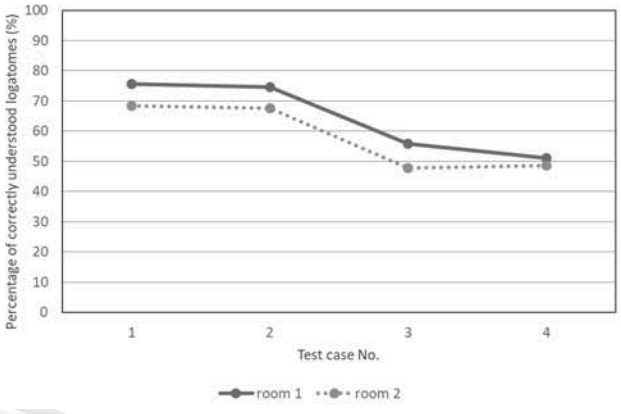

Fig. 5: The average percentage of correctly understood logatomes for four test cases, shown for test rooms 1 and 2

The data obtained in experiment 1 was statistically analyzed by means of two-way ANOVA. The independent variables (factors) on the test were the test case, which tests the influence of the number and the position of the distracting sources, and the test room, which tests the influence of acoustical properties of the room. The dependent variable is the speech intelligibility score defined as the percentage of correctly understood logatomes.

The results of the two-way ANOVA reveal that there is a statistically significant difference between the test cases examined in this experiment, i.e. between different configurations of disturbing sources, with $F(3,128) = 61.00$, $p = 1.47 \times 10^{-24}$. The post-hoc Tukey test reveals that test cases 1 and 2 result in significantly higher logatome intelligibility than test cases 3 and 4. The main difference is the number of disturbing sources, and the resulting logatome intelligibility is, on average, about 20 % lower for test cases 3 and 4 with three and four disturbing sources, compared to test cases 1 and 2 with 2 disturbing sources. It is interesting to note that there is hardly any difference in obtained intelligibility between test cases 1 and 2, which essentially represent the same configuration of sources, but in test case 1 it was achieved with real sources, whereas in test case 2 it was composed of phantom sources. The obtained results reveal that the number and the configuration of disturbing sources used in this experiment undoubtedly have an influence on the ability of the listener to single out the reference source and understand the content it produces.

The results of the ANOVA test also reveal a statistically significant difference in logatome intelligibility between rooms 1 and 2, with $F(1,128) = 15.89$, $p = 1.12 \times 10^{-4}$. This suggests that room acoustics, as examined in this experiment, also has an influence on speech intelligibility as the indicator of the ability of the listener to discriminate a reference source from the disturbing ones. As indicated above, the reverberation time in rooms 1 and 2 is 0.57 s and 1.05 s, respectively, and the corresponding hall radius is 1.14 m and 0.69 m, respectively. The direct-to-reverberant ratio at the listener position will be less favorable in room 2, resulting in lower speech intelligibility, and making it more difficult for the listener to single out the reference source that emits useful information. The marginal means calculated for rooms 1 and 2 reveal the difference in logatome intelligibility to be about 6 % on average, which is expected, given that the acoustical conditions are not drastically different in rooms 1 and 2.

The ANOVA test did not reveal a statistically significant interaction between the configuration of disturbing sources and the room, as two factors in this experiment, with $F(3,128) = 0.70$, $p = 0.5567$. This suggests that the change in acoustic conditions in the room resulted in a constant change in logatome intelligibility regardless of the configuration of disturbing sources, which is clearly visible in fig. 5. The only minor deviation from this finding can be seen for test case 4, where the influence of the room acoustics is, in fact, the smallest.

3.2. Experiment 2 - keywords in sentences

The results of the tests conducted for eleven test cases in experiment 2 are shown in table 6. The mean percentage of correctly understood keywords and the standard deviation across the entire group of listeners is indicated for each test case, both for room 1 and room 3. The mean values are also shown in graphic form in Fig. 6.

Test case	The number of distracting sources	The azimuth of the distracting sources	Room 1		Room 3	
			Mean (%)	Standard deviation (%)	Mean (%)	Standard deviation (%)
1	1	-30°	96,6	5,9	83,9	15,9
2	1	-60°	98,5	3,3	92,8	9,4
3	1	-90°	95,1	7,8	91,1	8,6
4	1	-180°	94,1	10,9	86,7	11,3
5	2	-30°, 30°	72,5	17,6	53,9	12,1
6	2	-60°, 60°	79,9	12,5	62,2	25,0
7	2	-90°, 90°	80,4	14,1	61,1	24,1
8	3	-180°, -90°, 90°	58,8	17,0	30,6	12,9
9	4	-60°, -30°, 30°, 60°	37,3	20,0	20,0	10,4
10	4	-90°, -60°, 60°, 90°	44,6	15,3	22,8	13,5
11	6	-90°, -60°, -30°, 30°, 60°, 90°	10,8	12,4	5,0	5,3

Tab. 6: The percentage of correctly heard keywords in the sentences for the 11 test cases in room 1 and 3 as the mean and the standard deviation for the tested group of listeners.

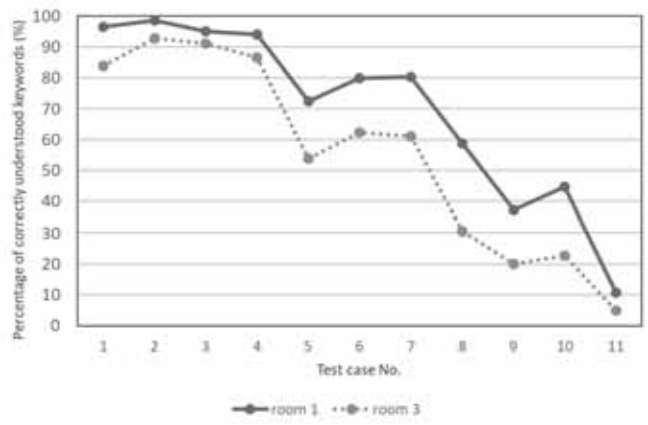

Fig. 6: The average percentage of correctly understood keywords for eleven test cases, shown for test rooms 1 and 3

The data obtained in experiment 2 was also statistically analyzed using the two-way ANOVA procedure. As before, the factors on the test were the configuration of disturbing sources, and the test room, which reflects the influence of room acoustics. The dependent variable in this test is again the speech intelligibility score, but this time is was defined as the percentage of correctly understood meaningful keywords within full sentences.

The results of the two-way ANOVA reveal a statistically significant difference in speech intelligibility scores for different configurations of disturbing sources defined in test cases examined in this experiment, with $F(10,330) = 145.67$, $p = 1.34 \times 10^{-114}$. The multiple pairwise comparison analysis conducted once again using the post-hoc Tukey test reveals that the mean intelligibility scores can be grouped together according to the number of disturbing sources. The property of each group is that the intelligibility scores for the test cases contained within a group do not significantly differ among themselves, but are significantly different than the scores of test cases in all other groups. Following this finding, test cases 1 to 4 with one disturbing source can be grouped together, test cases 5 to 7 with two disturbing sources form another group, test case 8 with three disturbing sources forms a group on its own, test cases 9 and 10 with four disturbing sources can be grouped together, and test case 11 with six disturbing sources again forms a group on its own. The mean keyword intelligibility scores decrease steadily with the increasing number of disturbing sources.

A closer look reveals that the lowest intelligibility score within such a group are attributed to the test case where the disturbing sources are located close to the reference source. The reason for this is that it is more difficult for the listener to single out and concentrate on the reference source if the sources are located close together. Specifically, in the group of test cases 1 to 4, case 1 yields a lower intelligibility score than cases 2 and 3. The same is observed for the group of cases 5 to 7, where case 5 yields the lowest score, and again in the group of cases 9 and 10, where case 9 yields the lower score of the two.

Another observation is that the intelligibility scores tend to be lower than expected if there is a disturbing source at 180°, i.e. right behind the listener, because a source at that position might cause a front-back confusion to some degree. This is visible in case 4, which yields the lowest intelligibility score of all the cases that have one disturbing source. Additionally, case 8 which has three disturbing sources, one of which is behind the listener, yields a somewhat lower intelligibility score than expected, given the sheer number of disturbing sources.

When it comes to the examination of the influence of the room itself, the results of the ANOVA test reveal a statistically significant difference in keyword intelligibility in rooms 1 and 3, with $F(1,330) = 92.70$, $p = 1.69 \times 10^{-19}$. This indicates that room acoustics again has an influence on the ability of the listener to single out and concentrate on the reference source while being disturbed by other sources, which is reflected in the obtained keyword intelligibility scores. The difference in acoustical conditions in these rooms is somewhat bigger, regarding the reverberation time, which is again 0.57 s in room 1, but rises to 1.39 s in room 3. The corresponding hall radii are 1.14 m and 0.79 m, respectively. The direct-to-reverberant ratio will be less favorable in room 3, which will in turn lead to decreased speech intelligibility. The marginal means calculated for rooms 1 and 3 reveal the difference in logatome intelligibility to be about 14 % on average.

As opposed to experiment 1, the ANOVA test reveals a statistically significant interaction between the configuration of disturbing sources and the room, with $F(10,330) = 2.50$, $p = 0.0066$. This suggests that the change in acoustic conditions in the room will have an effect on keyword intelligibility, but the effect size will depend on the configuration of disturbing sources. The average keyword intelligibility scores were examined, as obtained for all configurations of disturbing sources in both rooms, as shown in fig. 6. The examination reveals that moving the experiment from room 1 (good acoustics) to room 3 (fairly bad acoustics) will result in the largest decrease of keyword intelligibility if the disturbance to the listener is moderate, i.e. in cases when there are 2, 3 or 4 disturbing sources. For these cases, the average reduction of the mean keyword intelligibility scores is 20 %, and on the case-by-case basis this reduction ranges from 17 % to 28 %. If the disturbance is low, i.e. with only one disturbing source, the listener seems to be able to discriminate the reference source quite well, and for these four cases, moving from room 1 to room 3 results in the average decrease of the mean keyword intelligibility score of only 7.5 %. On the case-by-case basis, this decrease ranges from 4 % to 13 %, the worst case being when the disturbing source is close to the reference source. In case of severe disturbance, i.e. test case 11 with 6 disturbing sources, the average keyword intelligibility score is already

very low even in room 1 with good acoustics, due to the sheer magnitude of the disturbance. In this case, moving to room 3 with bad acoustic has little influence on the mean intelligibility score, reducing it from 11% in room 1 to barely 5 % in room 3.

Figs. 7 and 8 show the number of correctly understood keywords in test room 1 and 3 respectively, depending on the number of distracting sources. The curves connect test cases with the same minimum angular distance between the reference source and the closest distracting source. This graphical representation of the results shows that the increase in the number of distracting speakers results in a virtually linear decrease of the average keyword intelligibility score, regardless of the minimum angular distance between the reference and distracting sources, although the intelligibility scores are slightly higher when this angular distance is bigger. The comparison of the results obtained in two test rooms reveals that the reverberance in room 3 results in a deviation from the linear decrease of intelligibility scores observed in room 1. It can be argued that there is a certain saturation point for these curves since 0 % of correctly understood keywords is, in fact, the maximum possible error rate. Moreover, for all test cases the percentage of correctly understood keywords is lower in the more reverberant room, as clearly shown in Tab. 6.

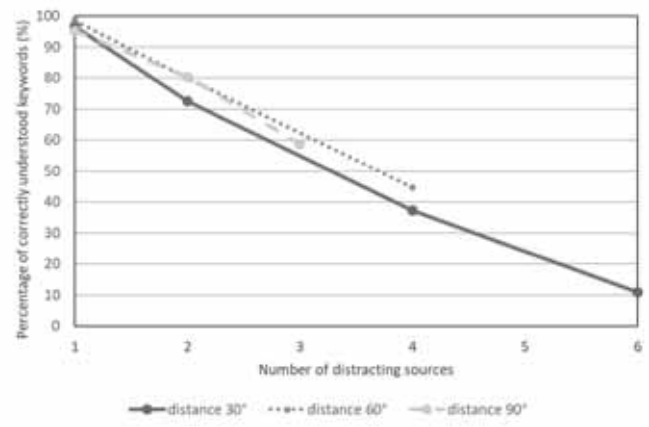

Fig. 7: The average percentage of correctly understood keywords in test room 1 depending on the minimum angular distance between the reference source and the closest distracting source

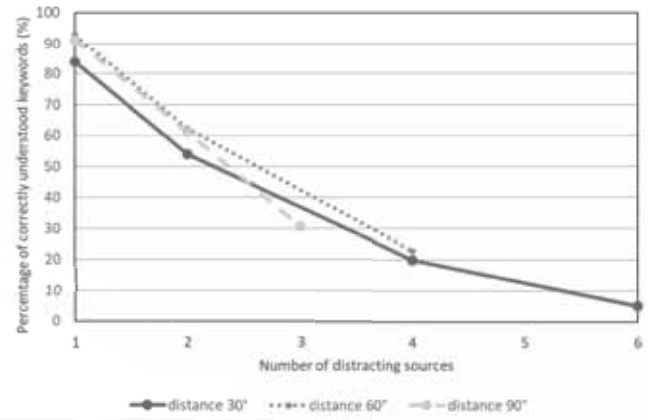

Fig. 8: The average percentage of correctly understood keywords in test room 3 depending on the minimum angular distance between the reference source and the closest distracting source

4. CONCLUSION

This paper presents several experiments designed around the cocktail-party problem. The incentive for this work was the development of new audio systems and technologies in recent years, when it became very easy to design any virtual sound environment with an almost unlimited number of sound sources around the listener. A particularly important part of such scenarios is a multi-talker environment, in which several persons talk simultaneously, acting as distracting sound sources for one another. Such scenarios are nowadays common in video conference calls with multiple participants, computer games, various VR/AR/MR applications, etc. Moreover, home cinema or cinema multichannel sound reproduction systems have the same ability of designing multisource sound environments. At the same time, they are often installed in acoustically unfavorable rooms where excessive reverberation negatively affects the intelligibility of the speakers, even more if there is more than one source active at the same time.

A multichannel loudspeaker setup was used to investigate the influence of the number of distracting speakers on the speech intelligibility of a reference speaker in a typical cocktail party effect, and the test was made in three acoustically different rooms.

The experiments conducted in this research examined speech intelligibility as the measure of the ability of the listeners to concentrate on a single source and understand the information it conveys, while being distracted by other sources that surround them, both in favorable and unfavorable acoustic conditions.

The results of the conducted experiments indicate that the number of distracting sources is the main factor that has an influence on the tested speech intelligibility in this particular setup. The distraction was tested with one, two, three, four, and six distracting sources, with different spatial layout in terms of the positions of the distracting sources. The resulting speech intelligibility suffered a steady and statistically significant decrease as the number of distracting sources was increased. The distraction by one distracting source can be declared low. Two to four sources will create a moderate distraction, while six distracting sources will result in severe distraction.

In terms of the position of the distracting source, the greatest distraction will be caused if the distracting source(s) is close to the source of interest, or if it is positioned directly behind the listener.

When it comes to comparing the distraction caused by real sources and the one caused by phantom sources in the exact same configuration, the results reveal that there is virtually no difference between these two situations.

The results of the tests also indicate that the influence of acoustical properties of the rooms in which the tests were made is also statistically significant. As none of the three rooms had extremely bad acoustics, the decrease of speech intelligibility due to worsening of the acoustical conditions is considerably smaller than the one observed for the number of distracting sources.

A detailed analysis reveals that the influence of the acoustical properties of the room is the greatest if the distraction caused by distracting sources is moderate. Otherwise, the distraction is either too low or too severe to be significantly influenced by the room itself.

In future work the experiment might be recreated in extreme acoustical conditions, i.e. in a room with very poor acoustical properties in terms of excessive reverberation.

The results of this research can be used for developing guidelines for sound engineers when designing and installing multichannel loudspeaker setups in rooms, or for designers of audio in computer games or in VR systems for headphones--based playback.

ACKNOWLEDGEMENT

The authors acknowledge financial support by the Croatian Science Foundation, (HRZZ IP-2018-01-6308, „Audio Technologies in Virtual Reality Systems for Auralization Applications (AUTAURA)“ for this paper.

REFERENCES

[1] Blauert, J.: Spatial hearing, MIT Press, ISBN 978-0262024136, 1996
[2] Suzuki, Y., Brungart, D., Iwaya, Y.: Principles and Applications of Spatial Hearing. World Scientific Publishing Company, ISBN 978-9814313872, 2011
[3] Altman, M., Krauss, K., Susal, J, Tsingos, N.: Immersive Audio for VR, In Proceedings of the Audio Engineering Society Conference on Audio for Virtual and Augmented Reality, 2016
[4] Vorländer, M.: Auralization - Fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality, Springer-Verlag, ISBN 978-3540488293, 2007
[5] Lokki, T., Savioja, L.: Virtual Acoustics, Handbook of Signal Processing in Acoustics (Ed. Havelock, D, Kuwano, S, Vorländer, M), Springer Verlag, ISBN 978-0387776989, p. 761-771, 2008
[6] Holman, T.: Surround Sound, Up and Running, Focal Press, ISBN 978-0240808291, 2007
[7] Roginska, A., Geluso, P., Immersive Sound, Routledgek, ISBN 978-1138900004, 2017
[8] Oreinos, C., Buchholz, J. M.: Objective analysis of ambisonics for hearing aid applications: Effect of listener's head, room reverberation, and directional microphones, J. Acoust Soc Am, Volume 137(6), p. 3447-3465, 2015
[9] Jambrošić, K., Krhen, M., Horvat, M., Oberman, T.: The use of inertial measurement units in virtual reality systems for auralization applications. In Proceedings of the ICA 2019 conference, Aachen, p. 2611-2618, 2019
[10] Stitt, P., Hendrickx, E., Messonnier, J.-C., Katz, B. FG.: The Role of Head Tracking in Binaural Rendering, 29th Tonmeistertagung – VDT International convention, p. 350-355, 2016
[11] Zhong, X., Sun, L., Yost, W.: Active binaural localization of multiple sound sources, Robotics and Autonomous Systems, ISSN 0921-8890, Elsevier, Volume 85, p. 83-92, 2016
[12] Benaroya, E. L., Obin, N., Liuni, M., Roebel, A., Raumel, W., Argentieri, S.: Binaural Localization of Multiple Sound Sources by Non-Negative Tensor Factorization, IEEE/ACM Transactions on Audio, ISSN: 2329-9290, Speech and Language Processing, IEEE, Volume 26, No. 6, 2018
[13] Fang, Y., Xu, Z.; Multiple Sound Source Localization and Counting Using One Pair of Microphones in Noisy and Reverberant Environments, Mathematical Problems in Engineering, ISSN 1563-5147, Hindawi, Volume 2020, Article ID 8937829, 2020
[14] Valin, J.-M., Michaud, F., Hadjour, B., Rouat, J.: Localization of Simultaneous Moving Sound Sources for Mobile Robot Using a Frequency-Domain Steered Beamformer Approach, Proceedings of IEEE International Conference on Robotics and Automation (ICRA), p. 1033-1038, 2004
[15] Jia, M., Wu, Y., Bao, C., Wang, J.: Multiple Sound Sources Localization with Frame-by-Frame Component Removal of Statistically Dominant Source, Sensors, SSN 1424-8220, MDPI, Volume 18(11), ID 3613, 2018
[16] Bech, S., Zacharov, N.: Perceptual Audio Evaluation - Theory, Method and Application, John Wiley & Sons, ISBN 978-0470869239, 2006
[17] Kisić, D., Horvat, M., Jambrošić, K., A methodology and a tool for interchangeable reproduction of sound samples in listening tests through different sound reproduction systems, In Proceedings of the ICA 2019 conference, Aachen, p. 6145-6149, 2019
[18] Hawley, M. L., Litovsky, R. Y., Colburn, H. S.: Speech intelligibility and localization in a multi-source environment, Journal of the Acoustical Society of America, Volume 105, p. 3436-3448, 1999
[19] Hawley, M. L., Litovsky, R. Y.: The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer, Journal of the Acoustical Society of America, Volume 115, p. 833-843, 2004
[20] Jambrošić, K., Krhen, M., Horvat, M., Jagušt, T.: Measurement of IMU sensor quality used for head tracking in auralization systems, In proceedings of the Forum Acusticum 2020 Lyon, 2020

Kristian Jambrošić *is a full-time professor at the Department of Electroacoustics of the Faculty of Electrical Engineering and Computing, University of Zagreb, where he obtained his PhD degree in the field of psychoacoustics. His main research activities are in the field of architectural acoustics, acoustic measurements, soundscape research, auralization, perception of sound and noise abatement techniques. He participates in several international and national research projects and has published as author or co-author more than one hundred and twenty papers in scientific journals and in conference proceedings. He supervised more than ninety bachelor and master thesis and teaches courses on acoustics both on the Faculty of Electrical Engineering and Computing and Academy of Music. He is a member of the executive council of the European Acoustics Association where he served as General Secretary in two terms. He has also active in the board of the Acoustical Society of Croatia. Currently he is leading a research group of the Auralization laboratory.*

Marko Horvat *is an assistant professor at the Department of Electroacoustics of the University of Zagreb Faculty of Electrical Engineering and Computing in Zagreb, Croatia, where he received his Masters and PhD degree in electrical engineering. His scientific, teaching and professional interests are focused on problems in the fields of room and building acoustics, psychoacoustics, noise and environmental acoustics, electroacoustic devices and systems, product sound quality, and the characterization of ultrasonic elements and transducers. The emphasis in his research activities is put on devising and implementing new measurement procedures, and investigating the perceptual dimension of acoustic phenomena. He teaches or has taught several courses on all education levels, such as "Sound reinforcement", "Electroacoustics and audiotechnics", "Audio recording and processing", etc. He has worked on several national and international scientific projects. He is the author or co-author of about 80 scientific publications, as well as about 50 technical projects focused on device and system calibration, measurements in room acoustics, noise and vibrations, as well as design solutions of room acoustics, sound reinforcement systems, and noise and vibration control.*

Dominik Kisić *was born in Zagreb, Croatia where he finished a mathematical gymnasium while studying in parallel violin at music high school Elly Bašić. He graduated with a master's degree at the Faculty of Electrical Engineering and Computing in Zagreb in 2017. During his studies he specialized in electronics and acoustics and also spent one semester of his education at HfM Detmold, Germany learning about the field of music acoustics. After completing his studies, he worked at a company Xylon as a hardware development engineer, and since 2018 he is working as a PhD candidate and a teaching assistant at the Department of Electroacoustics at the Faculty of Electrical Engineering and Computing in Zagreb. He is also active in music as a street musician, sound engineer, and producer. So far, he has participated in numerous performances and produced several albums in a small studio that he designed and set up himself.*

Tin Oberman *is a researcher, urban planner, architect and a musician. He got is M.Arch. and Ph.D. (Architecture and Urbanism) degrees at the University of Zagreb, in Croatia. He was awarded the Assistant Professor title at the Faculty of Architecture, University of Zagreb in 2019. He is currently working as a Research Fellow at the UCL Institute for Environmental Design and Engineering, the Bartlett, on the ERC – funded research project Soundscape Indices, led by Prof Jian Kang. During his studies in architecture and urbanism and following his education in music theory, he started working on several music projects. During his years working at the University of Zagreb, he acquired significant teaching experience in assisting undergraduate students through courses covering urban design, spatial planning, landscape architecture and history of landscape architecture at the Faculty of Architecture and the Faculty of Forestry at the University of Zagreb. At the Faculty of Architecture, he was involved in organisation of research projects and contributed to several urban planning teams. His research work included collaboration with the Institute of Electroacoustics at the Faculty of Electrical Engineering and Computing, where he gained experience in auralisation and soundscape field research, focused on the enhancement of soundscape of urban open spaces. He is author of research and conference papers.*