SPEECH INTELLIGIBILITY DEPENDENCE ON NUMBER AND POSITION OF SIMULTANEOUS SOUND SOURCES

Kristian Jambrošić, Marko Horvat, Dominik Kisić, Tin Oberman

1 Introduction

The capacity of human hearing to clearly distinguish sound sources, either by their position and/or their informational contents, was a major research topic in the field of psychoacoustics in the last couple of decades [BLAUERT 1996, SUZUKI et al 2011]. Nowadays, the findings of this research become even more important because of novel technologies of three-dimensional sound reproduction system, including also virtual reality audio-video systems, that provide means of synthesizing any virtual sound field with almost an unlimited number of sound sources around the listener [ALTMAN et al 2016, VORLÄNDER 2007, LOKKI et al 2008].

By examining such sound reproduction systems in more details, one can observe that multichannel reproduction systems, consisting of several spaced loudspeakers, are driven with usually heavily correlated audio signals, especially having in mind the amplitude panning law used in all stereophonic systems [HOLMAN 2007, ROGINSKA 2017]. Therefore, although signals from different loudspeakers could be completely uncorrelated, they are mostly quite correlated in the audio production. On the other hand, virtual reality systems and auralization systems in general that use loudspeakers or headphones for sound reproduction can support many virtual sound sources from various directions and distances. The sound sources here can be also completely uncorrelated, and these systems offer the most natural hearing experience. This is especially true for 3D ambisonics systems and binaural systems with head position tracking [OREINOS et al 2015, JAMBROSIC et al 2019].

Evaluation of sound perception tasks is done using listening tests on a statistically significant sample of listeners since there is a certain variability on how each individual person perceives sound. Therefore, many listening test procedures have been introduced in became very common in order to optimize the design and implementation of these tests [BECH et al 2006, KISIĆ 2019]. Some decisions must be made before starting with the test design. One important choice is between in-situ listening tests (in a natural environment), and tests in laboratories where the sound field is re-created using either loudspeakers in different setups, or headphones. The later spaces offer more control over all interfering parameters (heat, rain, sunlight, noise, etc.), so they are very much preferred. But, even in the tests are conducted in laboratories, one has to choose between tests in free field conditions (e.g. using anechoic chambers) where the room influence can be neglected, and tests in more reverberant spaces which might be a more natural surroundings for some sort of tests since the hearing mechanism operates non-stop in non-anechoic environments.

In this paper, a typical cocktail-party effect setup with an increasing number of simultaneous uncorrelated speech sources was examined in two acoustically different rooms in order to find out the capacity of suppressing unwanted signals and to check the room influence on the speech resolving mechanism of people. Other authors have also made research for multi speaker scenarious, but not always in a natural, reverberant environment where real-life conditions would be simulated as close as possible [HAWLEY et al 1999]. Moreover, these

experiments were rarely conducted using many simultaneous natural sound sources, thus simulating a typical cocktail party setup where more than ten simultaneous speakers from various directions can easily occur.

2 Listening test setup

2.1 Hardware and software setup

Having in mind the common situations where verbal communication is required, the listening tests in this research were made with all the sources, both useful and disturbing, located in the horizontal plane.

The test setup consisted of loudspeakers placed along the circumference of a circle with the radius of 2 meters. The listener was placed in the centre of the said circle, thereby being located at equal distance from all the loudspeakers. The listening position was set to be in the horizontal plane that contained the acoustics centres of all the loudspeakers. The height of the seated listener was adjusted accordingly, so that the ears of the listener would be exactly in the described horizontal plane. The number of active loudspeakers and their positions was changed from one listening test to the next, as described below in detail. Given the radius of the said circle and the size of the loudspeakers, the minimum possible azimuthal spacing between the adjacent loudspeakers was 30°, which corresponds to 12 possible loudspeaker positions. Since the usual communication is done in the frontal half of the horizontal plane, for the test the frontal 7 positions were used plus the position at 180° azimuth as an anchor point directly behind the listener.

The loudspeakers used in this experiment were active, bookshelf-sized near-field studio monitors with a reasonably flat (within ±3 dB) frequency response in the frequency range from 70 to 20000 Hz, thus representing high-quality sources for speech reproduction. The maximum sound pressure level these loudspeakers can provide is 101 dB at 1 meter, which was more than enough to ensure that the reproduced speech would not have audible distortions at any reproduction level used in the tests. The loudspeakers were placed on heavy and stable metal stands. The loudspeakers were connected to a multichannel sound card and through it with a personal computer with appropriate DAW software. All test signals were sampled with a standard sampling frequency of 44.1 kHz and quantized at 16-bit resolution.

The test setup included two basic configurations. In both of them, the test signal was always reproduced by the loudspeaker at the azimuth of 0°. The first configuration has the listener face the loudspeaker at azimuth 0°. The loudspeakers are distributed evenly every 30° in the forward half-circle (from the viewpoint of the listener), with an additional loudspeaker at azimuth 180°. In the second configuration, the listener is now facing the loudspeaker at the azimuth of 90°, and the loudspeakers are still evenly distributed from -90° to 90° (now the left half-circle from the viewpoint of the listener). In this case, there is no loudspeaker at azimuth 180°. Both configurations are shown in Figure 1.

2.2 Test signals

For the purpose of this research, test speech signals in form of full sentences were used in listening tests. A total of 184 different sentences were chosen from Croatian literature works. All test sentences were recorded as they were read by a reference male speaker at a uniform speed. The length of all the sentences was similar, and the time required to read each sentence never exceeded 5.5 seconds. After recording, the recorded waveforms containing individual

sentences were analysed and their level adjusted, so that the average RMS power of each waveform (sentence) was set to -24 dB(FS). The distraction speech signals were also full sentences chosen randomly from radio shows. The speakers were male, and apart from them speaking, the chosen signals contained no background music or noise. The length of these distraction signals was set to 6 seconds, with fade-in and fade-out length of 250 ms, so that the precedence effect would be eliminated. The content of the distraction sentences was not analysed. The waveforms containing the distractions sentences were adjusted for level the same way as was done with test sentences.

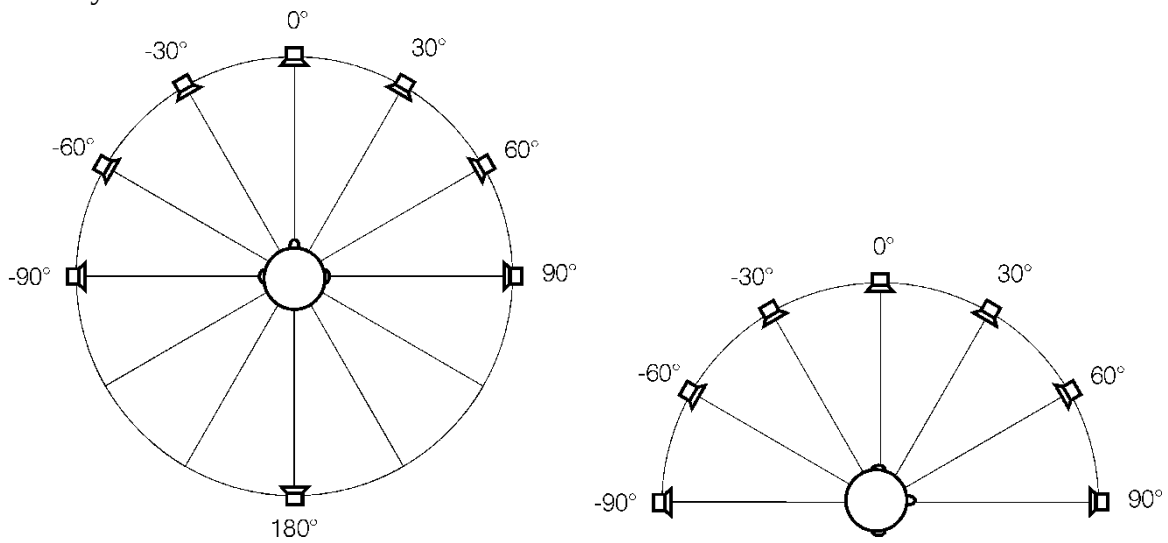

Figure 1. Loudspeaker configurations and the orientation of the listener, as used in listening tests

The spectral content of all sentences was analysed and expressed as average spectrum. The goal of this analysis was to determine if there are notable differences between the speech spectrum of the speaker who read the test sentences and the speech spectrum of the speakers who provided the distraction sentences. The results of this analysis is shown in Figure 2.

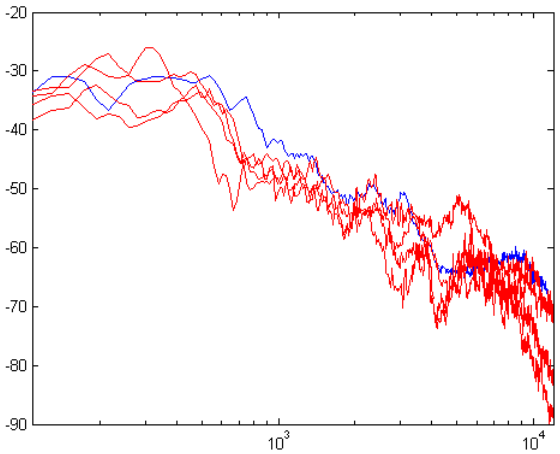

Figure 2. The average amplitude spectrum of the speaker who read the test sentences (blue) and the average spectrum of the selected four speakers who provided the distraction sentences (red) (in dB) vs frequency (in Hz)

2.3 The course of the tests

The audio signals in each test example were arranged in the following manner. A sinusoidal signal with the frequency of 250 Hz and the duration of 0.5 seconds was reproduced first as an announcement to the listener that the test example is about to commence. This "warning sign" was reproduced by the loudspeaker that also reproduces test sentences. The reproduction of the distraction sentences began 0.5 seconds after the warning signal had finished. Each distraction sentence was reproduced by a single loudspeaker. The reproduction of the test sentence began 1 second after the warning signal had finished, and the test sentence was reproduced over a separate loudspeaker. The test example was followed by a 23-second long period of silence, during which the listeners had to write down the entire test sentence as they had heard it. Three different groups of one test sentence and one or more distraction sentences were put together, so that the diversity of interaction between different speech signals would be as large as possible.

An example of a test example used in the listening tests is shown in Figure 3. The warning sinusoidal signal and the test sentence are sent to the same track in the DAW software, and, consequently, to the same loudspeaker. Four distraction sentences occupy the four remaining tracks, and are sent to four different loudspeakers.

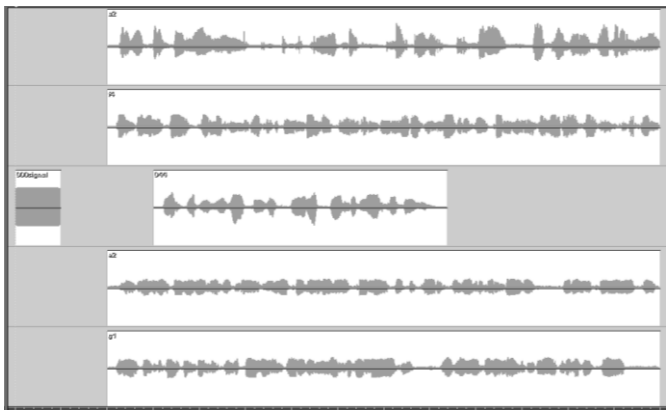

Figure 3. An example of the arrangement of the individual tracks in a test example

2.4 The investigated spaces

The listening tests were made in two acoustically very different spaces, but similar in terms of size. Both rooms have the same width and height, and differ only in length, which also reflects to a difference in volume. As for their acoustical properties, the room designated as a listening room is acoustically treated to meet the relevant requirements, and the resulting mid-frequency single-number reverberation time is 0.57 seconds. On the other hand, the acoustical properties of the room used as a lecture room have not been improved in any way, as no acoustic treatment has ever been implemented. Therefore, the acoustic situation in the room is dictated only by the existing finishing materials, and the resulting mid-frequency reverberation time has an excessive value of 1.39 seconds. The basic data on the two investigated spaces is given in Table 1.

2.5 The form of the results

To analyse the speech intelligibility, four keywords were defined in each test sentence. These keywords were compared with the ones written down by the listeners after they had heard a

test sentence. The keywords that were correctly heard and written down were counted, and the overall intelligibility was determined as the ratio of the correctly heard keywords to the total number of keywords in a given test series. The obtained value was expressed as a percentage.

Table 1. The basic data on the two investigated spaces

Room	Length (m)	Width (m)	Height (m)	Volume (m^3)	Reverberation time RT_{20} (s)	Noise level L_{EQ} (dBA)
Listening room	10.20	7.05	3.20	230	0.57	31
Lecture room	11.95	7.05	3.20	270	1.39	42

2.6 The listening test cases

For the purpose of this research, 15 different test cases were defined, as shown in Table XX. The test sentences were reproduced with a loudspeaker at azimuth 0° in all cases. The number of distraction signals and the azimuth of the loudspeakers that were reproducing them are shown in the corresponding columns of Table 2. In the first 11 cases, the listener faces the loudspeaker that reproduces the test sentences, and in the remaining four cases, the orientation of the listener is changed by 90° in the clockwise direction. The same group of listening test cases was done in both investigated spaces.

A total of 17 listeners took part in the testing in the listening room, and 15 listeners participated in the tests carried out in the lecture room.

Table 2. Test cases defined for examination by listening tests

Test case	Number of distraction signals	Distraction coming from azimuth (°)	The orientation of the listener towards azimuth (°)
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
12	2	-30, 30	90
13	2	-60, 60	90
14	2	-90, 90	90
15	4	-60, -30, 30, 60	90

3 The results

The results of the listening tests are shown in Table 3 for all 15 test cases, and for both investigated rooms. The results shown in Table 3 suggest that, on a global scale, the macro-acoustical conditions in a room, perceived and described by its reverberance, have an influence on speech intelligibility. The results obtained for a more reverberant lecture room show a consistent increase in the mean percentage of incorrectly heard keywords for all 15

test cases, compared to the results obtained for the listening room with well-controlled reverberation.

Table 3. The percentage of incorrectly heard keywords, displayed for all 15 test cases in both rooms as a) mean value, and b) standard deviation for the entire group of listeners

Test case	Listening room		Lecture room	
	Mean (%)	Standard deviation (%)	Mean (%)	Standard deviation (%)
1	3,4	5,9	16,1	15,9
2	1,5	3,3	7,2	9,4
3	4,9	7,8	8,9	8,6
4	5,9	10,9	13,3	11,3
5	27,5	17,6	46,1	12,1
6	20,1	12,5	37,8	25,0
7	19,6	14,1	38,9	24,1
8	41,2	17,0	69,4	12,9
9	62,7	20,0	80,0	10,4
10	55,4	15,3	77,2	13,5
11	89,2	12,4	95,0	5,3
12	42,2	21,3	58,9	17,7
13	21,6	16,9	40,0	17,0
14	5,4	7,8	22,8	14,9
15	78,9	12,5	83,3	17,5

As a parameter, the number of distracting signals/sounds also has an influence on speech intelligibility. Before employing the distracting speech signals, an initial test was made without them, and in this case, there were no errors in understanding any of the keywords; in other words, the percentage of incorrectly heard words was zero. With introduced distracting speech sounds, the percentage of misunderstood words rises, with its value reaching 90 % for six distracting sounds. Since all the sounds were adjusted so that their RMS level is the same, the sheer energy of the distracting sounds will increase with the number of distracting sounds, and surpass the energy of a single source that reproduces the useful test sound.

The position of the distracting sound sources relative to the position of the source of useful sound, and relative to the listener has proved to have an impact on the resulting speech intelligibility. The general conclusion that can be drawn from the data is that speech intelligibility will be higher if the hearing system is able to make a clear distinction between individual sources. The said distinction is viewed in terms of the interaural time and level differences as important cues in binaural listening. In terms of the obtained results, the effect is most visible with one or two distracting sources. The general observation is that the distracting sound sources should be placed far enough apart in space from the source of useful sound, so that each source would produce different binaural cues at the listener position. A bad case would be to place the distracting source(s) close to the source of useful sound, but

also behind the listener. In all these cases, the binaural cues will be similar for all the sources, making it difficult for the hearing system to distinguish between them.

The orientation of the listener relative to the source of useful sound and to distracting sources also plays a significant role in understanding speech. A comparison of results obtained for test cases 5, 6, and 7 with 12, 13 and 14, respectively, reveals that the rate of improvement of speech intelligibility is greater for the latter group of cases. In other words, when the listener is not actually facing the source of useful sound, but that sound comes from the lateral direction (from the left in this case), moving the distracting sources apart from each other and from the source of useful sound will lead to a fast and great improvement when it comes to understanding speech.

Figure 4 shows these results also graphically. It becomes clear that the increase in the number of distracting speakers is almost linearly increasing also the number of incorrect keywords, until a certain saturation point for these curves since 100% is the maximum error rate. Moreover, the difference in keyword errors between the listening room and the lecture room is shown in Figure 5. For all test cases, the keyword error is higher in the more reverberant room, as it was obvious from the results shown in Table 3.

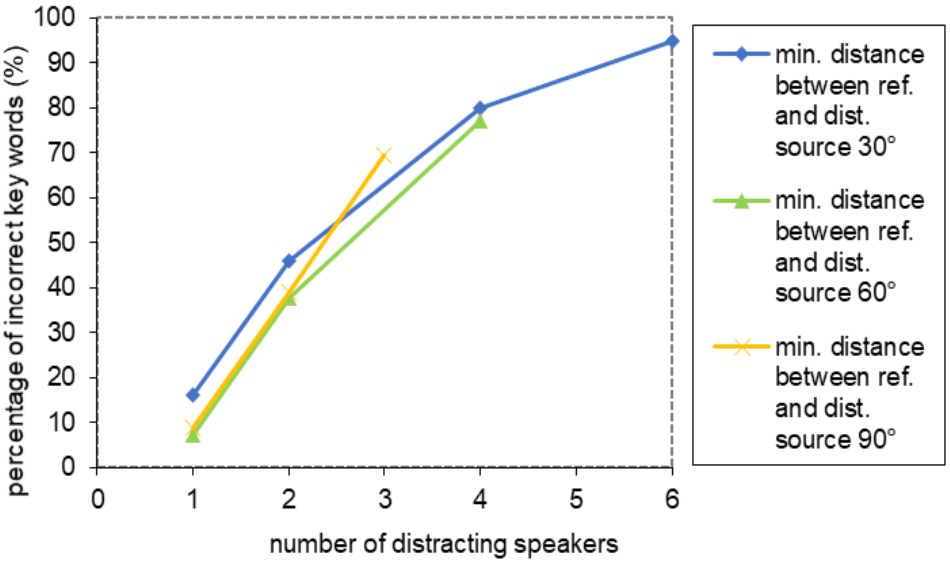

Figure 4. Average keyword error vs. number of distracting speakers in the listening room. Different curves connect cases with same minimum angular distance between the referent and closest distracting source.

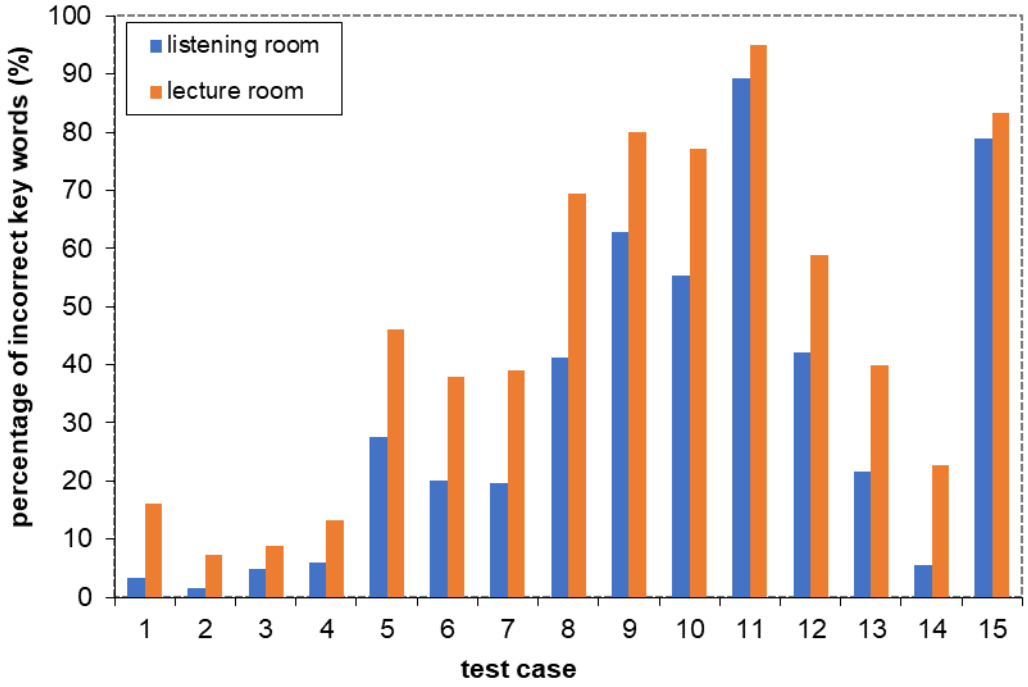

Figure 5. Comparison of average keyword error between listening room and lecture room for the same test cases.

4 Conclusions

Multichannel sound reproduction systems became very common in the last two decades and are used widely. It is well known that different rooms have different speech intelligibility depending on their characteristics, thus influencing the listening experience. A multichannel loudspeaker setup was used to investigate the influence of the number of distracting speakers on the speech intelligibility of a referent speaker in a typical cocktail party effect test. The complete set of tests were repeated in two rooms of similar size, but different acoustic finishing, thus having different reverberation times and objective speech intelligibility parameters.

It was found that speech intelligibility, measured as the number of correctly heard key words in spoken sentences, falls below 50% for two or three distracting sources which were simultaneously emitting speech signals. This effect depends on the azimuth and angular distance of the distracting sources. Moreover, since the two test rooms had different acoustic finishing of their walls, although of similar shape and size, there was a difference in the direction, amplitude and amount of early reflections reaching the test persons in both rooms. It became obvious that this influenced the sound source perception in increasing the sound image shift and was directly influencing the overall speech intelligibility in both rooms. Generally, rooms with bigger reverberation time give worse intelligibility.

All these outcomes have been additionally proven by statistical tools, thus giving significance to the findings. Therefore, the results of this research can be used as a recommendation for adding multiple speaking signals in a listening environment, both in the real and the virtual world.

Acknowledgements

The authors acknowledge financial support of the EU H2020-MSCA-RISE-2015 "papabuild„ project (grant agreement No. 690970) for this paper.

References

ALTMAN, M, KRAUSS, K, SUSAL, J, TSINGOS, N. 2016. Immersive Audio for VR, In *Proceedings of the Audio Engineering Society Conference on Audio for Virtual and Augmented Reality*, 2016. Los Angeles

BECH, S., ZACHAROV, N. 2006. *Perceptual Audio Evaluation - Theory, Method and Application*, John Wiley & Sons, 2006. ISBN 978-0470869239

BLAUERT, J. 1996. *Spatial hearing*. MIT Press, 1996. ISBN 978-0262024136

HAWLEY, M. L., LITOVSKY, R. Y., COLBURN, H. S. 1999. *Speech intelligibility and localization in a multi-source environment*, Journal of the Acoustical Society of America, vol. 105, 1999. p 3436-3448.

HOLMAN, T., 2007. *Surround Sound, Up and Running*, Focal Press 2007. ISBN 978-0240808291

JAMBROŠIĆ, K., KRHEN, M., HORVAT, M., OBERMAN, T. 2019. The use of inertial measurement units in virtual reality systems for auralization applications. In *Proceedings of the ICA 2019 conference*, 2019. Aachen, p. 2611-2618.

KISIĆ, D., HORVAT, M., JAMBROŠIĆ, K., 2019. 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*, 2019. Aachen, p. 6145-6149.

LOKKI, T, SAVIOJA, L. 2008. *Virtual Acoustics, Handbook of Signal Processing in Acoustics* (Ed. Havelock, D, Kuwano, S, Vorländer, M, in *Handbook of Signal Processing in Acoustics*). New York: Springer Verlag, 2008. ISBN 978-0387776989, p. 761-771.

OREINOS, C., BUCHHOLZ, J. M. 2015. *Objective analysis of ambisonics for hearing aid applications: Effect of listener's head, room reverberation, and directional microphones*, J. Acoust Soc Am, 2015; 137(6), p. 3447-3465.

ROGINSKA, A., GELUSO, P. 2017. *Immersive Sound*. Routledgek 2017. ISBN 978-1138900004

SUZUKI, Y., BRUNGART, D., IWAYA, Y. 2011. *Principles and Applications of Spatial Hearing*. World Scientific Publishing Company, 2011. ISBN 978-9814313872

VORLÄNDER, M. 2007. *Auralization - Fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Berlin: Springer-Verlag, 2007. ISBN 978-3540488293

Summary

Speech Intelligibility Dependence on Number and Position of Simultaneous Sound Sources. The capacity of human hearing to clearly differentiate the position and/or informational contents of a referent sound source is limited when other, masking sound sources are simultaneously active. It is important to understand these limitations since nowadays novel technologies of virtual acoustics can simulate an almost unlimited number of

sound sources in the created sound field, for example in the gaming industry. In this paper, a variation of the cocktail party effect was researched by performing speech intelligibility tests of a referent male speaking voice, but with the addition of up to six simultaneous, masking speaking voices from different direction around the listeners. The tests were repeated with the same setup in two rooms with different acoustic characteristics, thus having different speech transmission index values. The test results are compared and the decrease in speech intelligibility is shown in relation to the number of masking sources and their position.

Keywords

speech intelligibility; multiple sound sources; room acoustics; cocktail party effect

Contact Address

Kristian Jambrošić
Department of Electroacostics, Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
kristian.jambrosic@fer.hr
https://www.fer.unizg.hr/

Marko Horvat
Department of Electroacostics, Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
marko.horvat@fer.hr
https://www.fer.unizg.hr/

Dominik Kisić
Department of Electroacostics, Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
dominik.kisic@fer.hr
https://www.fer.unizg.hr/

Tin Oberman
Faculty of Architecture
University of Zagreb
Fra Andrije Kačića Miošića 26, 10000 Zagreb, Croatia
tin.oberman@arhitekt.hr
https://arhitekt.hr/ /