Date of publication SEP-30, 2022, date of current version AUG-16, 2022.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.21.3.2690

Impact of University Classroom Size on the Relationship between Speech Quality and Intelligibility

ARKADIY PRODEUS[1], MARYNA DIDKOVSKA[2], KATERYNA KUKHARICHEVA[1]

[1]Department of Acoustic and Multimedia Electronic Systems, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine, (e-mail: aprodeus@gmail.com, katerynakt@gmail.com)

[2]Department of Mathematical Methods of System Analysis, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine, (e-mail: maryna.didkovska@gmail.com)

Corresponding author: Arkadiy Prodeus (e-mail: aprodeus@gmail.com).

ABSTRACT In this paper, five objective measures of the quality of speech signals distorted by reverberation are compared with the Speech Transmission Index (STI). The main aim of the comparison is to further test and explain the reasons for the previously discovered phenomenon of an increase in the speech quality and intelligibility with increasing room size. The comparison is performed for three university classrooms of small, medium and large sizes. The correlation coefficients between the quality and intelligibility estimates of speech obtained for 5-6 points of each room are estimated. Speech signal quality is assessed using intrusive measures such as segmental signal-to-noise ratio (SSNR), log-spectral distortion (LSD), frequency-weighted segmental signal-to-noise ratio (FWSNR), bark spectral distortion (BSD), and perceptual evaluation of speech quality (PESQ). For BSD, high correlation coefficients (0.57-0.99) are determined for rooms of all sizes and an increase in the correlation coefficient with the room size increase is found, which can be explained by a decrease in the density of early sound reflections. For FWSNR, high correlation (0.65-0.98) is determined for medium and large rooms. For PESQ, high correlation (0.96-0.99) is obtained for large classroom. SSNR and LSD are found to be uncorrelated with STI for rooms of all sizes.

KEYWORDS binaural room impulse response; speech quality; speech intelligibility; objective measure.

I. INTRODUCTION

THE volume of the room significantly affects its acoustic properties, which is manifested in changes in the shape and parameters of the room impulse response (RIR) [1]. Indeed, increasing the size of the room usually leads to an increase in reverberation time, a decrease in the average density of sound reflections, an increase in the time border between early reflections and late reverberation [1].

Finally, the volume of the room affects such speech perception indicators as the speech quality and speech intelligibility. These indicators are usually associated with the reverberation time T60 [2], which can be explained by the destructive effect of late reverberation, which increases with increasing reverberation time [3]. In part, the popularity of reverberation time as a measure of speech intelligibility can be explained by the relative simplicity of measuring reverberation time. Since the provision of quality and intelligible speech is especially important for primary school students and students with hearing loss, it is understandable that researchers are interested in finding ways to optimize reverberation time [4, 5].

The correlation between the parameters C50, EDT, T30 values and STI values in different locations of university auditoriums of large, medium and small size, was studied in [6]. It was shown that C50 and EDT, in contrast to T30, can be used for measurement of speech intelligibility in different locations of a room. However, it was found that the use of EDT for speech intelligibility measurement can be problematic in small classrooms due to the insufficiently high level of correlation (approximately 0.5) of EDT values with STI values in different locations. Possible causes of the phenomenon were not mentioned in [6], although it would be appropriate to assume that some feature of early reflections in small rooms plays a role here.

In general, it is believed that early sound reflections are useful because they provide high speech intelligibility [4, 7]. In [4], a conclusion was even made about the equivalence of the influence of early reflections and direct sound on speech intelligibility. However, the results of research presented in [8] do not agree with this conclusion and show that early reflections play a less important role, compared to direct sound, in ensuring high speech intelligibility. These results were not thoroughly explained in [8], although in reality the explanation may not be very complicated. Indeed, given that the formation of early reflections is similar to the formation of a signal at the output of the comb filter [1], the quality of the signal generated by early reflections should be lower than the quality of the direct signal due to uneven frequency response of the virtual comb filter. Thus, the presence of early sound reflections can cause a mismatch between speech intelligibility and speech quality, when speech intelligibility increases and speech quality decreases. However, the conditions under which such disagreement takes place or under which it is violated have remained unexplored in [8].

STI score increase to 8–13% near the back wall of the room, compared to STI score in the room center, was shown in [6], in the study of speech understanding in different locations of university auditoriums. However, the statistical relationship between STI estimates and estimates of objective measures of speech quality has not been studied. It has only been suggested that there is such a relationship because of the information about the RIR is contained in both types of measures.

This issue was partially solved in [9], where the correlation level between BSD, LSD, PESQ and STI was studied. RIRs for 5-6 locations of three university classrooms, previously considered in [6], were used for experimental studies. For the BSD and STI, the correlation coefficient R was high in all auditoriums and was 0.92-0.98 for a large room, 0.8 for a medium room, and 0.6-0.99 for a small room. For the PESQ and STI, the correlation coefficient R was high (0.96-0.99) only for large classroom. LSD was found to be uncorrelated with STI for all rooms. These results indicate the existence of the phenomenon of the room size influence on the correlation level between the speech quality and speech intelligibility. However, the possible mechanism of this phenomenon was not considered in [9].

The objective of the paper is to find an explanation for the cause of the phenomenon of increasing speech quality and speech intelligibility with increasing room size. Another objective is to find quality measures competing with BSD that could simultaneously act as speech intelligibility measures in auditoriums of different sizes.

II. PROBLEM STATEMENT

Speech intelligibility is closely related to the content of the message and can therefore serve as a measure of the amount of information perceived by the listener [10, 11]. The quality of speech reflects the emotional reaction of a person to the distortion of the form of the speech signal, regardless of its content [12, 13].

The results of studies of noisy speech signals show that, as a rule, high-quality speech is simultaneously intelligible [13]. There are few exceptions to this rule. An example of such an exception is the communication line test, where the speaker's pronunciation irritates the listener, resulting in a low subjective assessment of speech quality (MOS scale). At the same time, the subjective assessment of speech intelligibility performed by the articulatory method is also likely to be low. However, an objective assessment of intrusive speech quality will result in a high DMOS score if there is no significant signal distortion in the communication channel. Another example of high-quality, but incomprehensible, speech can be found in [13].

However, intelligible speech is not necessarily good. A well-known example is vocoder communication lines, where significant distortion of the waveform is considered acceptable [11]. Another example is the preliminary high-frequency filtering of signals, which allows increasing the efficiency of automatic speech recognition systems [14]. To increase the intelligibility of speech masked by intense noise, it is possible to use algorithms for intentional distortion of speech signals in the time or spectral domain, or in both domains at once [15]. Decreased intelligibility and quality of speech when using speech enhancement algorithms is a known fact [16]. Because people with hearing impairments are most sensitive to this phenomenon, the use of algorithms for intentional distortion of speech signals can significantly improve speech intelligibility [17]. Another way to increase speech intelligibility, while deteriorating its quality, is signal clipping used in hearing aids and cochlear systems [18].

The classroom, which is a kind of communication channel, can be roughly considered a linear system with a RIR, the envelope of which is similar to the descending exponent [1]. The presence of rather powerful discrete bursts in the initial RIR section, caused by early reflections of sound from different reflective surfaces of the room (walls, ceiling, floor, furniture, etc.), allows us to compare the room with a comb filter, which distorts the signal shape and reduces its quality and intelligibility. On the other hand, early sound reflections enhance the signal-to-noise ratio, which leads to increased speech intelligibility, especially noticeable in large audiences [4]. As can be seen, the question of the impact of early reflections on the relationship between speech quality and intelligibility in different sizes rooms is relevant. It should be noted that this issue has not yet received proper coverage in the scientific literature.

The simplest and most direct way to answer this question is to compare objective estimates of speech quality and speech intelligibility, obtained using RIRs estimates of actual premises. Given that RIR estimates are different in different locations of the room, as well as taking into account the random nature of the measurement results, such a comparison is appropriate to perform by calculating the Pearson correlation coefficient. By comparing the correlation coefficient estimates obtained for rooms of different sizes, one can not only investigate the phenomenon of room size influence on the relationship between speech quality and speech intelligibility estimates, but also try to find a valid explanation for this phenomenon.

III. EXPERIMENTAL SETUP

A. ROOMS FEATURES

There are various proposals for the classification of university classrooms by size. For example, in [19] it is proposed to consider small rooms with a volume of less than 230 m^3, medium-sized rooms including those with a volume of 230-350 m^3, and large rooms with a volume of more than 350 m^3.

This paper presents the results of research that used binaural room impulse responses (BRIRs) estimates of small, medium and large auditoriums which volumes are 177 m^3, 270 m^3 and 370 m^3, respectively. In the future, these classrooms will be designated as rooms №1, №2 and №3, respectively. All classrooms are in the shape of a parallelepiped, which is typical of university premises. Auditoriums №1 and №2 belongs to NTUU "Igor Sikorsky Kyiv Polytechnic Institute" (Ukraine). Auditorium №3 belongs to RWTH Aachen University (Germany) [20]. The plans of these rooms are shown in Fig. 1. These plans indicate the locations of sound source and artificial head. Two microphones were attached to the ears of the artificial head, which allowed to make two-channel recordings of test signals and use these recordings to assess the BRIRs. All rooms had parquet floors and brick plastered walls. The height of all rooms was close to 3.1 m.

The influence of furniture on the results of acoustic measurements was not taken into account in this paper. An assessment of the degree of such influence was beyond the scope of this paper and should be done in the future.

EDT and T30 reverberation time values were estimated according to ISO 3382-1 [2]. The values of T30, averaged over the channels and all locations of the artificial head, were close and amounted to 0.82 s, 0.91 s and 0.81 s for rooms №1, №2 and №3, respectively. The values of EDT, similarly averaged, differed more significantly and were 1.11 s, 0.82 s and 0.68 s for rooms №1, №2 and №3, respectively [6].

B. TEST SIGNALS AND RECORDING EQUIPMENT

The basic element in the construction of the test signal was a maximum length sequence (MLS) signal containing 2^{16} samples [20]. The test signal was created by repeating this basic element 17 times. The RIR was calculated by averaging the last 16 bursts of the cross-correlation function between the microphone output signal and the test signal, which increased the signal-to-noise ratio by 12 dB [20].

Recording of test mls-signals in classrooms №1 and №2 was performed with a sampling frequency of 44.1 kHz and a quantization depth of 24 bits using household and semi-professional electroacoustic equipment. The equipment included an active speaker Genius SP-HF 2.0 500 and measuring condenser microphones Superlux ECM-999, which were attached to the ears of a homemade artificial head. At the stage of BRIR assessment, the frequency response non-uniformity of the measuring system was compensated.

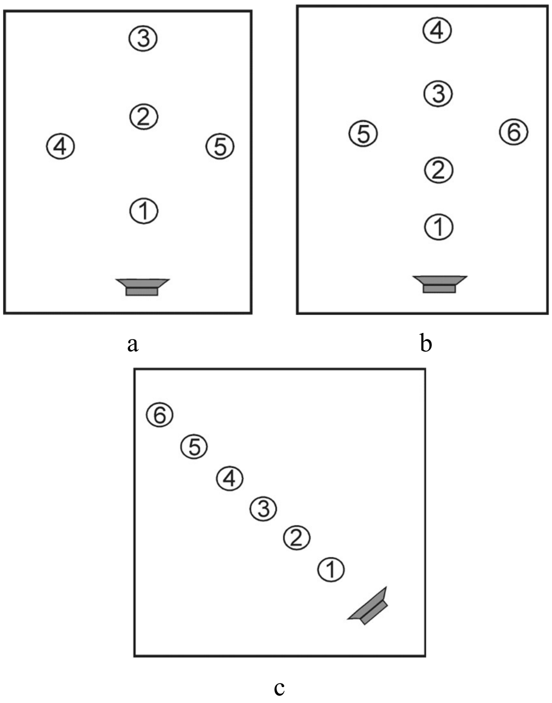


Figure 1. Locations of sound source and artificial head for small (a), medium (b) and large (c) auditoriums

BRIR estimates for auditorium №3 are borrowed from [20]. In this case, the measurements of BRIRs were performed using professional audio equipment.

C. SPEECH SIGNALS

The records of 8 speakers (4 women and 4 men), who read the same legal text, were used for objective assessment of the signal quality. The duration of the analyzed signals was 15 s, the sampling frequency was 22050 Hz, the bit depth was 16 bits. Note that this duration was chosen taking into account the fact that it was sufficient to have a signal duration of 10-15 s for subjective evaluation of signal quality on the DMOS scale [12]. However, other combinations of the speaker's number N and the signal duration L are additionally considered in this paper:

- N = 8, L = 60 s;
- N = 12 (6 women and 6 men), L = 15 c.

The purpose of considering such combinations is to assess the sensitivity of research results to changes in research conditions.

D. SIGNAL PROCESSING

Signal processing was performed in three stages:

- for each location of each of the rooms, the values of speech quality measures were evaluated, such as SSNR, BSD, LSD, PESQ and FWSNR;
- for each location of each of the rooms, the values of the STI were assessed;
- for each of the rooms, Pearson's correlation coefficients were calculated between SSNR, BSD, LSD, PESQ, FWSNR and STI values, which corresponded to the sets of locations of each classroom.

Evaluation of the quality of the speech signals distorted by reverberation was performed in two stages:

- the signal $y(t)=x(t)\otimes h(t)$ was formed, $\otimes$ is convolution symbol, $x(t)$ is clear speech signal, $h(t)$ is BRIR;
- the quality measures SSNR, BSD, LSD, PESQ and FWSNR values were calculated.

SSNR value was calculated as follows [13]:

$$SSNR=\frac{1}{M}\sum_{m=1}^{M}10\lg\left[\frac{\sum\limits_{n=S(m-1)+1}^{Sm}x^2(n,m)}{\sum\limits_{n=S(m-1)+1}^{Sm}[x(n,m)-y(n,m)]^2}\right] \qquad (1)$$

$x(n,m)$ and $y(n,m)$ are n-th samples of m-th frame of clear signal $x(n)$ and distorted signal $y(n)$, respectively, M is frames quantity, S is frame samples quantity.

LSD value [13] was calculated using the amplitude spectra of the signal frames

$$LSD=\frac{2}{JM}\sum_{m=1}^{M}\sum_{j=1}^{J}\left|G\{X(j,m)\}-G\{Y(j,m)\}\right|, \qquad (2)$$

$$G\{X(j,m)\}=\max\{20\lg(|X(j,m)|),\delta\},$$

$$\delta=\max_{j,m}\{20\lg(|X(j,m)|)\}-50,$$

$|X(j,m)|$ and $|Y(j,m)|$ are amplitude spectra of m-th frames of signals $x(n)$ and $y(n)$, respectively, $|\cdot|$ is module symbol, j is spectrum sample number, J is quantity of spectrum samples.

FWSNR value [13] was also calculated using the amplitude spectra of the signal frames

$$FWSNR = \frac{10}{M}\sum_{m=1}^{M}\frac{\sum_{k=1}^{K} W(k,m)\lg\frac{|X(k,m)|^2}{(|X(k,m)|-|Y(k,m)|)^2}}{\sum_{k=1}^{K} W(k,m)}, \quad (3)$$

where k is critical band number, K is critical band quantity, $|X(k,m)|$ is amplitude spectrum of m-th frame of clear signal, calculated using a Gaussian window, $W(j,m) =| X(j,m)|^{0.2}$ is weighting factor.

BSD value [13] was calculated as follows

$$BSD = \frac{\sum_{m=1}^{M}\sum_{k=1}^{K}\left[B_x(k,m)-B_y(k,m)\right]^2}{\sum_{m=1}^{M}\sum_{k=1}^{K}\left[B_x(k,m)\right]^2}, \quad (4)$$

where $B_x(k,m)$ and $B_y(k,m)$ are bark spectra of m th frames of $x(n)$ and $y(n)$, respectively.

The description of the PESQ calculation algorithm in this paper is not given due to its cumbersomeness. Note that the PESQ wideband version was used [13, 21].

STI evaluation [22] was performed according to

$$STI = \sum_{k=1}^{7}\alpha_k \cdot MT_k - \sum_{k=1}^{6}\beta_k \cdot \sqrt{MT_k \cdot MT_{k+1}}, \quad (5)$$

$$MT_k = \frac{1}{14}\sum_{i=1}^{14} T_{ki},$$

$$T_{ki} = \begin{cases} 0, & E_{ki} < -15; \\ (E_{ki}+15)/30, & -15 \le E_{ki} \le +15; \\ 1, & E_{ki} < -15, \end{cases}$$

$$E_{ki} = 10\lg\frac{m_{ki}}{1-m_{ki}},$$

$$m_{ki} = \left|\int_0^{\infty} h_{rk}^2(t)\exp(-j2\pi F_i t)dt\right| \Big/ \int_0^{\infty} h_{rk}^2(t)dt,$$

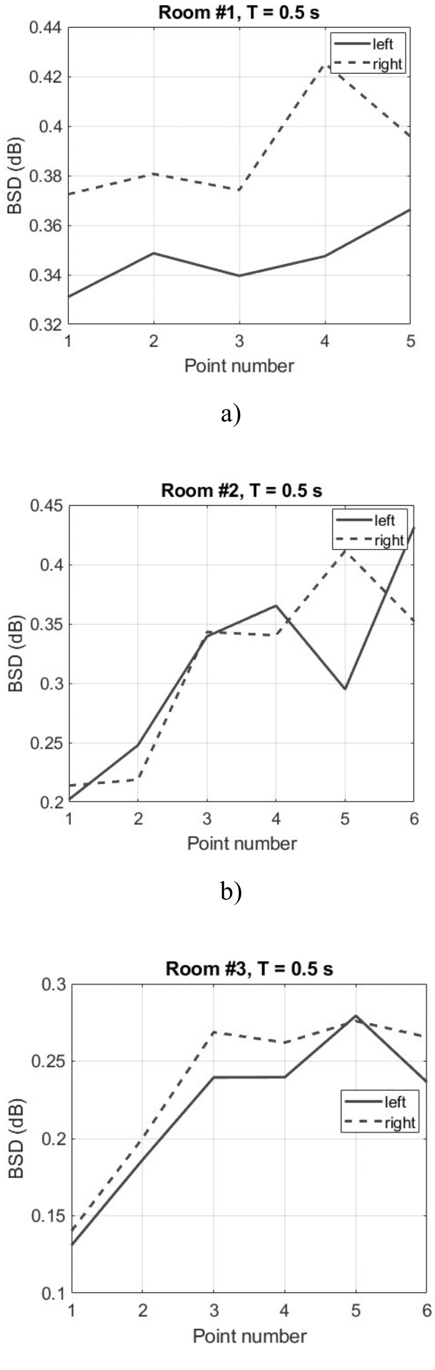
where α_k and β_k are weight and redundancy coefficients [22], respectively, $h_k(t)$ obtained by octave filtering BRIR $h(t)$, $k = \overline{1,7}$ is octave filter number, F_i = 0.63–12.5 Hz. Octave filters with center frequencies $f_0 = 125,\ldots,8000$ Hz are used here.

Pearson's correlation coefficients R between SSNR, BSD, LSD, PESQ, FWSNR and STI were calculated according to standard rules [23].

IV. EXPERIMENTAL RESULTS

A. QUALITY, INTELLIGIBILITY AND CORRELATION ESTIMATES

The results of the evaluation of the quality measures BSD, FWSNR and PESQ, provided that the duration of the speech signal is L = 15 s, and the total number of speakers is N = 8, are shown in Figs. 2, 3 and 4, respectively. These results are graphs of the above parameters estimates dependence on the numbers of points where the artificial head was located. Graphs of SSNR and LSD estimates are not given in this paper because they do not correlate with STI estimates (Table 1) and are therefore not informative. STI estimates are shown in Fig. 5.

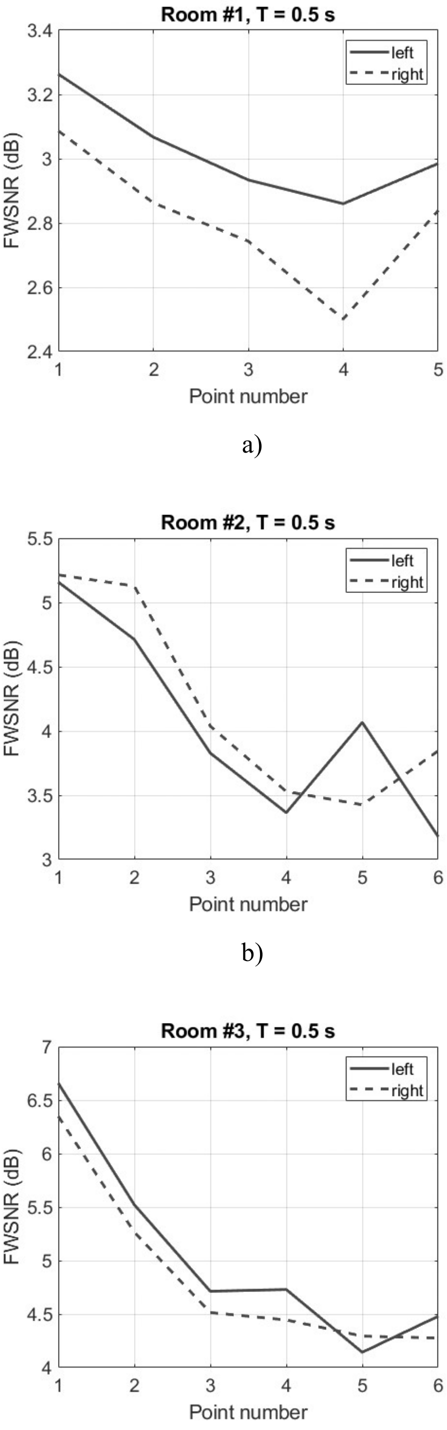


a)

b)

c)

Figure 2. BSD score vs point number: small (a), medium (b) and large (c) classrooms



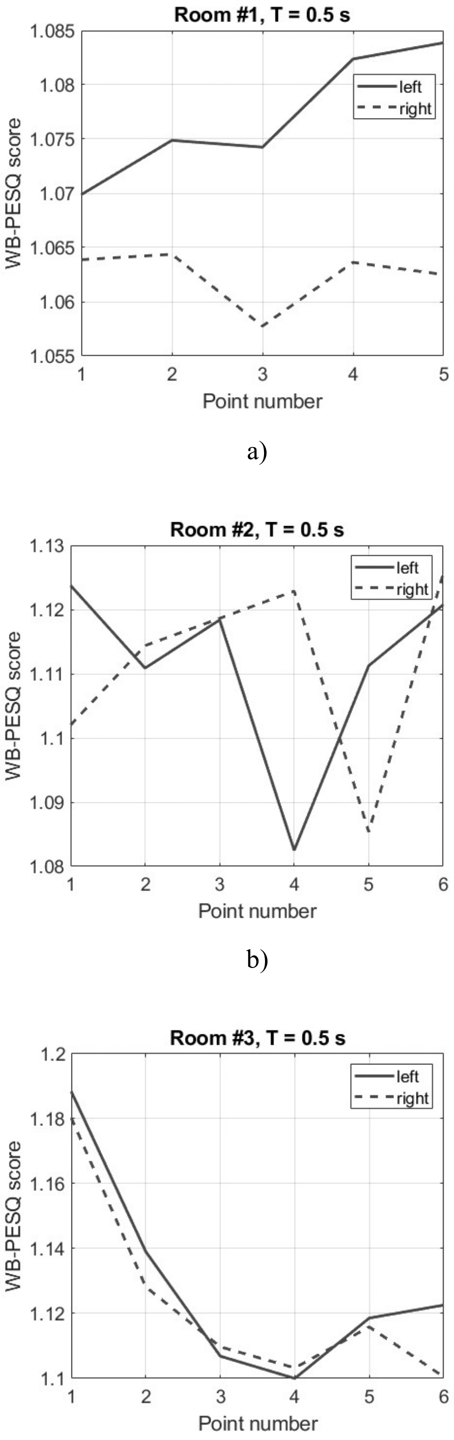
a)

b)

c)

Figure 3. FWSNR score vs point number: small (a), medium (b) and large (c) classrooms

Estimates of the correlation coefficients R between SSNR, BSD, LSD, PESQ, FWSNR and STI are given in Fig. 6 and in Table 1. These results indicate a close correlation between STI and BSD estimates for all studied rooms. Besides, the correlation level increases with the auditorium size. As can be seen, $R \approx -0.57$ for left channel and $R \approx -0.99$ for right channel in the case of small classroom.



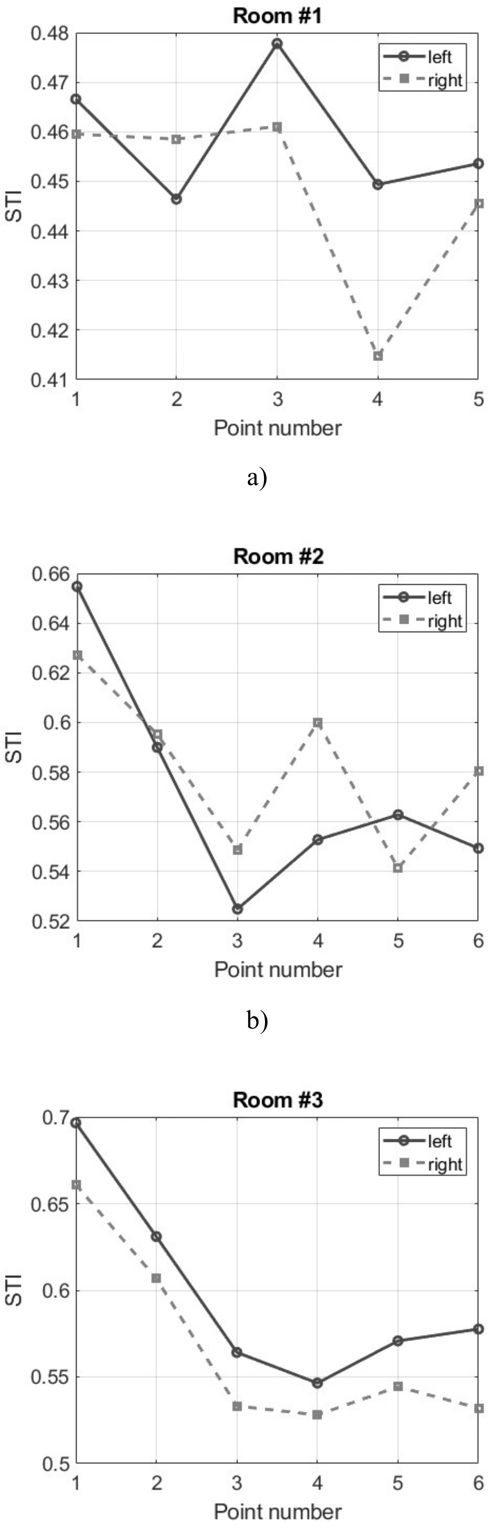
a)

b)

c)

Figure 4. PESQ score vs point number: small (a), medium (b) and large (c) classrooms

For medium and large rooms, R values for pair STI-BSD are close to minus 0.8 and minus 0.92-0.98, respectively. The correlation between FWSNR and STI is also high, except for small rooms (0.17 for left channel and 0.79 for right channel).



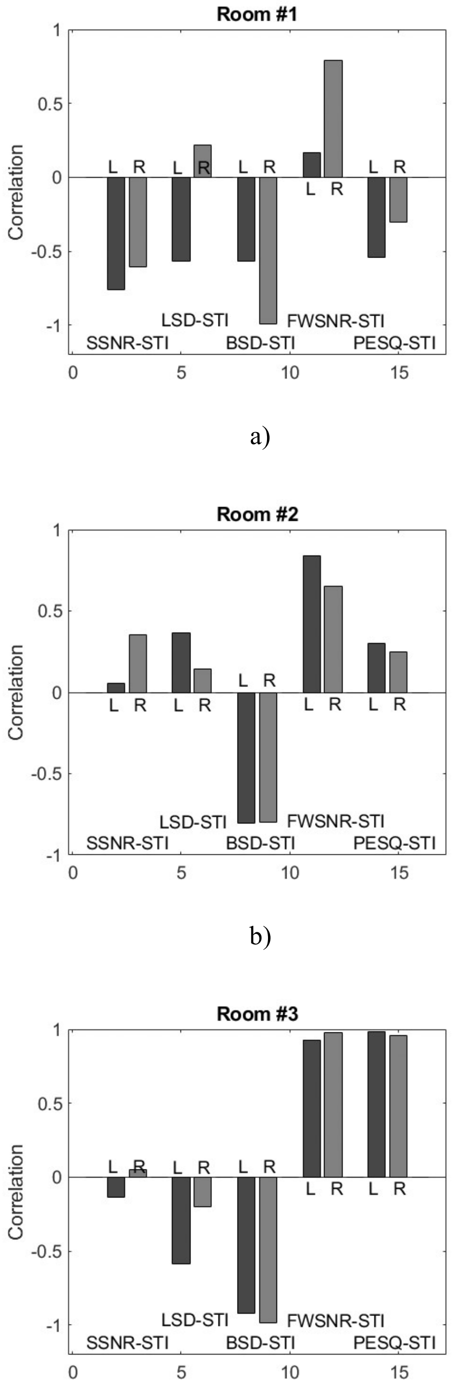
a)

b)

c)

Figure 5. STI score vs point number: small (a), medium (b) and large (c) classrooms

A high level of correlation for the PESQ-STI pair occurs only for a large auditorium.



a)

b)

c)

Figure 6. R values for small (a), medium (b) and large (c) classrooms

As can be seen, the BSD can be used as a measure of speech intelligibility in auditoriums of various sizes, provided that the effect of noise is smaller compared to the effect of reverberation. The FWSNR measurement can also be used to evaluate speech intelligibility in large and medium-sized rooms. However, the suitability of FWSNR for use in small classrooms is questionable and should be further tested. The use of PESQ as a measure of speech intelligibility is only possible in large classrooms. The SSNR and LSD measures turned out to be practically unsuitable as indicators of speech intelligibility.

Table 1. Values of correlation coefficient R

Pair	Channel	Room 1	Room 2	Room 3
SSNR-STI	**L**	-0.760	0.055	-0.136
	R	-0.604	0.351	0.050
LSD-STI	**L**	-0.564	0.367	-0.585
	R	0.216	0.146	-0.198
BSD-STI	**L**	-0.569	-0.801	-0.922
	R	-0.990	-0.795	-0.983
FWSNR-STI	**L**	0.168	0.841	0.929
	R	0.788	0.651	0.977
PESQ-STI	**L**	-0.542	0.302	0.986
	R	-0.304	0.250	0.958

Analysis of the mapping from BSD to STI estimates (Fig. 7) for the left channel (the mapping is similar for the right channel) shows that the linear approximation of the relation between BSD and STI is acceptable for practical application. Indeed, an error er_2 of the second-order polynomial STI (BSD) dependency approximation is of the same order as the linear approximation error er_1 (Fig. 7). If the linear dependence STI (BSD) is described by the expression $STI = a_0 + a_1 \cdot BSD$, then the coefficient a_0 values are 0.65, 0.71 and 0.81 for the left channel, and coefficient a_1 values are minus 0.57, minus 0.44 and minus 1 for rooms №1, №2 and №3, respectively. Similarly, for right channel, coefficient a_0 values are 0.79, 0.68 and 0.88, and coefficient a_1 values are minus 0.88, minus 0.33 and minus 1 for rooms №1, №2 and №3, respectively. It is clear that these values are approximate due to the small number of seats in each of the premises, as well as due to the small number of considered premises. Therefore, in the future it will be appropriate to clarify the values of these coefficients by increasing the amount of statistics.

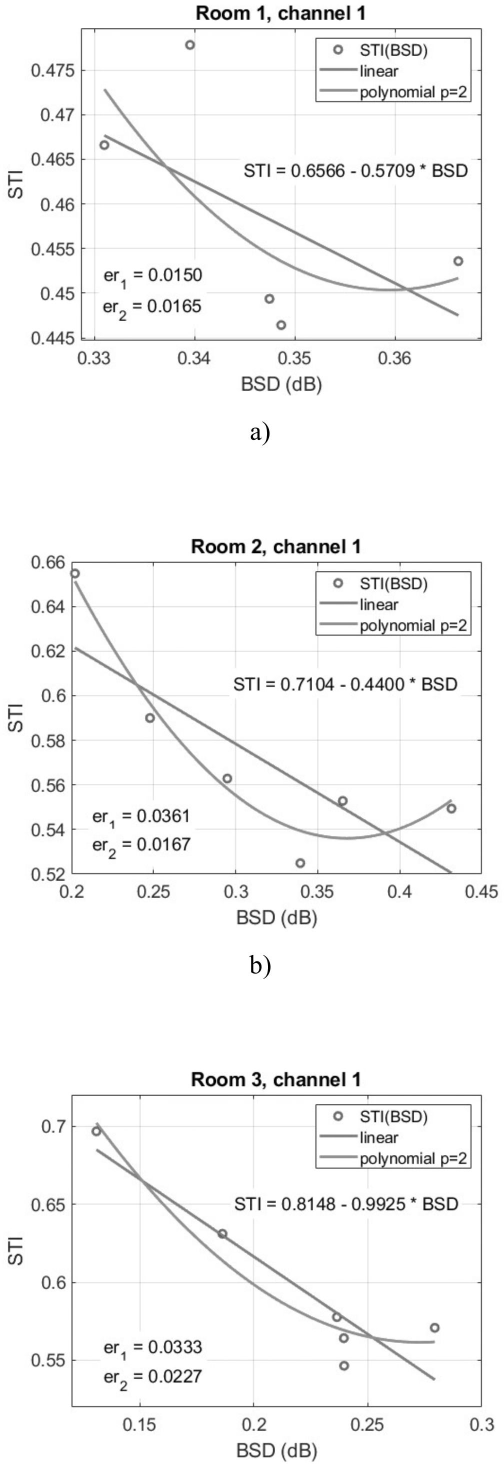
B. ROBUSTNESS OF OBTAINED RESULTS

Since the above estimates of R were obtained for a certain amount of data, it is advisable to check the stability of R estimates to the increased volume of statistics.

Graphs of estimates R for conditions N = 8, L = 60 s are shown in Fig. 8, and similar graphs for conditions N = 12, L = 15 s are shown in Fig. 9.

Comparing Figs. 6, 8 and 9, it is easy to conclude that the initial conclusions about the nature and degree of correlations between objective assessments of the quality and intelligibility of speech remain unchanged:

- intrusive estimates of speech quality BSD, formed taking into account the peculiarities of the human auditory system (critical frequency bands, Weber-Fechner law, etc. [24]), correlate with STI estimates for classrooms of all sizes, and the degree of correlation increases with increasing room size;
- the FWSNR and STI estimates does not correlate in the case of small premises, and the correlation of STI and PESQ scores is high only in large rooms;
- intrusive estimates of speech quality in the form of the mean squared error in the time (SSNR) or spectral (LSD) domains, almost do not correlate with STI estimates.

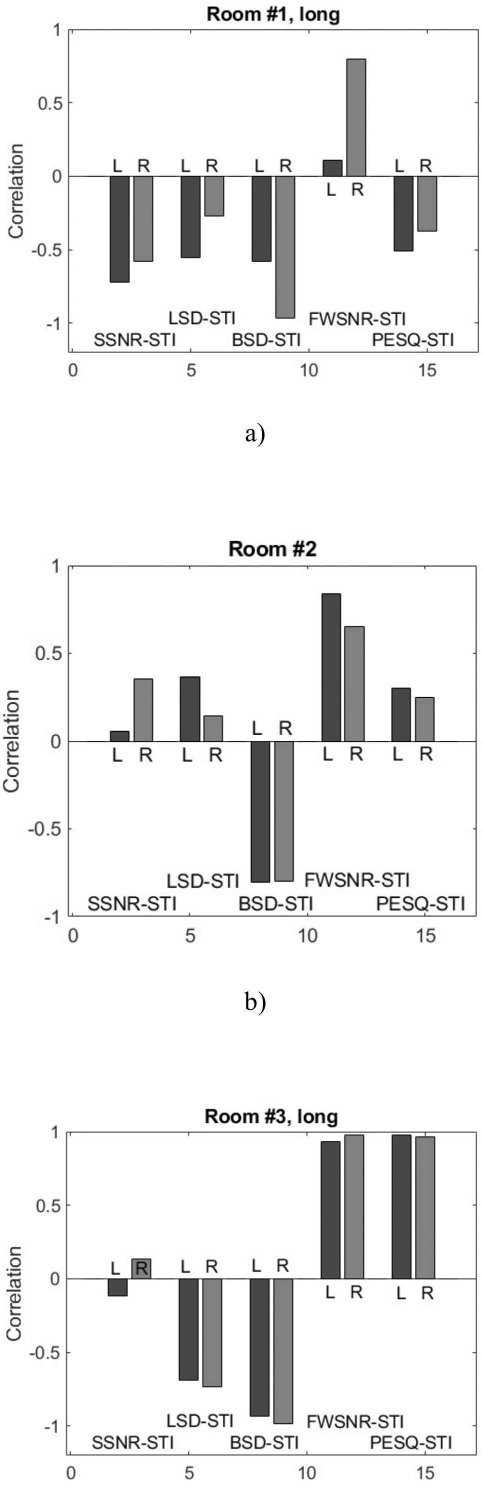


a)

b)

c)

Figure 7. Mapping from BSD to STI for small (a), medium (b) and large (c) classrooms



a)

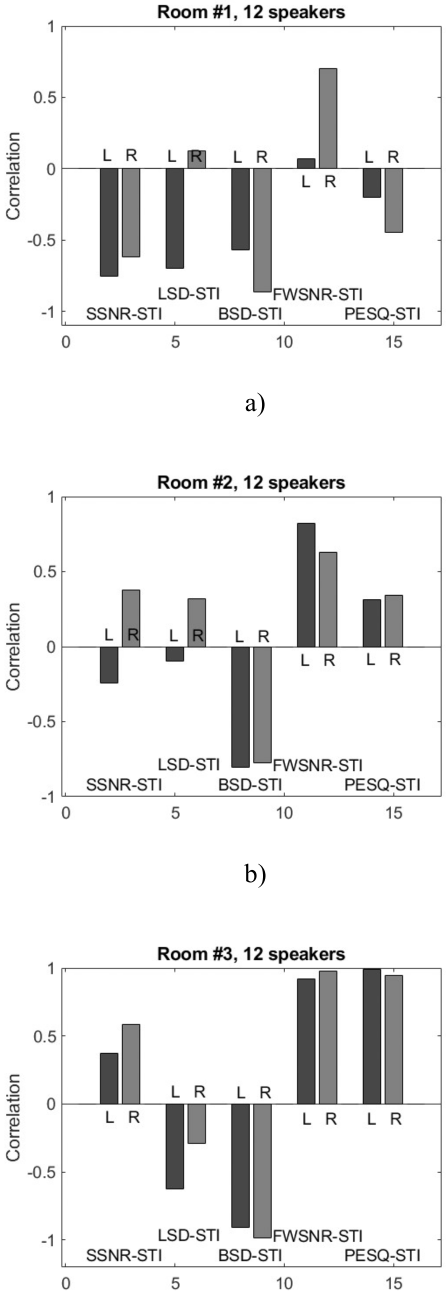
b)

c)

Figure 8. R values for small (a), medium (b) and large (c) classrooms for N = 8, L = 60 s

V. DISCUSSION

As noted above, SSTI or LSD estimates, unlike BSD, FWSNR, and PESQ estimates, do not correlate with STI estimates. This fact agrees well with the well-known criticism of the use of a mean squared error as a speech quality measure [13].



a)

b)

c)

Figure 9. R values for small (a), medium (b) and large (c) classrooms for N = 12, L = 15 s

Insufficiently high efficiency of PESQ as an indoor speech intelligibility measure compared to BSD and FWSNR estimates can be explained by the fact that the PESQ measure was designed originally to solve another problem, namely, end-to-end quality testing of speech codecs and telephone networks [21]. However, it should be noted that the developers of PESQ have shown the ability to use PESQ as a speech intelligibility measure in low bit rate vocoders [25].

As it turns out, BSD is slightly better than FWSNR at estimating speech intelligibility in rooms of any size. This fact can be explained by a more complete consideration of the human auditory system properties in the BSD calculations algorithm [13, 26].

It is more difficult to explain the phenomenon of increasing correlation coefficient R between BSD and FWSNR scores and STI score with increasing room size.

Previous studies have shown a similar behavior of the correlation coefficient between EDT and STI estimates when R estimates were minus 0.52, minus 0.94 and minus 0.93 for rooms №1, №2 and №3, respectively [6]. The average, by channels and locations of each room, EDT values were 1.11 s, 0.82 s and 0.68 s, and STI values were 0.4, 0.52 and 0.58 for rooms № 1, №2 and №3, respectively [6].

Since EDT characterizes the behavior of RIR in the initial time interval, the above results can be interpreted as an increase in R with increasing room size caused by the action of early sound prints.

Indeed, in box-shaped rooms, the density of reflections perceived by the listener at a certain point in time t is inversely proportional to the volume of the room V [1]

$$d(t) = \frac{4\pi c^3 t^2}{V}, \qquad (6)$$

$c \approx 340$ m/s is speed of sound in the air. A simplified model of early reflections in the form of a pulsed flow was considered in [26]

$$h(t) = \left[\delta(t) + \sum_{n=1}^{P} a_n \delta(t - t_n) \right] e^{-\alpha t}, \qquad (7)$$

where a_n and t_n are random variables distributed according to a uniform law ($0 \le a_n \le 1$, $0 \le t_n \le T_{60}$), P is the number of pulses in the time interval $[0, T_{60}]$, $\alpha = \ln(10^3)/T_{60}$ is parameter that determines the rate of attenuation of RIR. Shown in Fig. 10 graph of STI dependence on early reflections mean density $Density = P/T_{60}$ indicates a decrease in intelligibility with increasing number of early reflections from 2 to 16 in the time interval from 0 to 50 ms [26].

Indeed, it is easy to calculate that on the segment of 50 ms (interval of existence of early reflections) the reflections density values of 40, 80, 160 and 320 Hz, shown in Fig. 10, corresponds to the number of reflections 2, 4, 8 and 16, respectively. Further increase in the density of early reflections, as shown in [26], does not change the value of STI. Given (6), (7), the graph of Fig. 10 and the proximity of T30 values in rooms №1, №2 and №3, one can conclude that the increase in R with increasing room volume is caused by a decrease in the density of early reflections. Physically, this phenomenon can be explained by the fact that with a decrease in the early reflections density, the shape of the speech signal is less distorted, which means that both intelligibility and signal quality increase.

It follows that the decorrelation of speech quality and intelligibility assessments in small audiences is due to the high density of early reflections. Physically, such a decorrelation is quite difficult to explain. It can only be noted that its presence is consistent with many of the above examples, when the speech signal may be intelligible, but poor quality. However, the correctness of the assumption about the primary role of early reflections can be verified by conducting the following experiment.

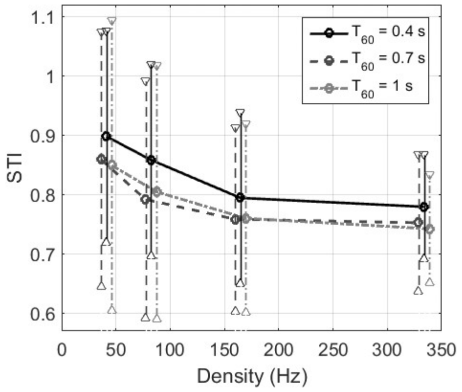


Figure 10. STI vs mean density of early reflections [26]

Let us compare the values of the correlation coefficient R, calculated for complete BRIR, with the values of R for early reflections (the initial part of the BRIR lasting 50 ms) and the values of R for late reverberation (only BRIR's tail after 50 ms). The results of R calculations for these cases are presented in Table 2 for the most informative measures BSD and FWSNR.

Table 2. Values of correlation coefficients for BRIR and its parts

Pair		BSD-STI			FWSNR-STI		
Room		1	2	3	1	2	3
Full	L	-0.57	-0.80	-0.92	0.17	0.84	0.93
	R	-0.99	-0.79	-0.98	0.79	0.65	0.98
Early	L	-0.23	-0.77	-0.88	-0.16	0.77	0.91
	R	-0.56	-0.72	-0.95	0.37	0.63	0.96
Late	L	0.66	-0.38	0.21	-0.28	0.27	-0.04
	R	0.21	-0.39	-0.21	0.15	0.30	0.20

Comparing the values of R in the rows "Early" and "Late" with those in the row "Full", one can see that the main role in ensuring the correlation of quality and intelligibility estimates is played by early reflections.

It will also be useful to search a connection between the results of this paper and the results of the analysis of the hearing system [27] and its model [28].

VI. CONCLUSIONS

The main objective of the paper is to test and explain the reasons for the previously discovered increase in the quality and intelligibility of speech with an increase in the size of the room. The degree of correlation between the quality estimates

SSNR, BSD, LSD, PESQ and FWSNR and intelligibility estimate STI is determined for different sizes university auditoriums. It is shown that correlation is high between BSD and STI estimates in all sizes premises. For FWSNR and STI estimates, the correlation level is high for large and medium auditoriums and is moderately high for small auditoriums. For PESQ and STI estimates, the correlation level is high only for large auditoriums. SSNR and LSD measures are non-correlated with STI estimates for all sizes auditoriums. The possibility of linear approximation of STI(BSD) dependence is shown, and also approximate values of coefficients of such approximation are specified that is useful for engineering applications.

The revealed phenomenon of increasing speech quality and speech intelligibility and also increasing correlation between quality and intelligibility estimates with increasing room size can be explained by the decrease in the density of early sound reflections.

Due to the limited number of auditoriums studied, it is advisable to further refine the results by increasing the number of rooms in which BRIR measurements are taken.

References

[1] H. Kuttruff, *Room Acoustics*, fifth ed., Spon Press, London and New York, 2009, 374 p. https://doi.org/10.1201/9781482266450.

[2] ISO 3382-1:2009. Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces. Available at: https://www.iso.org/obp/ui/#iso:std:iso:3382:-1:ed-1:v1:en

[3] P. Naylor, N. Gaubitch (Eds.), *Speech Dereverberation*, Springer-Verlag, London, 2010, 388 p. https://doi.org/10.1007/978-1-84996-056-4.

[4] W. Yang, J. Bradley, "Effects of room acoustics on the intelligibility of speech in classrooms," *Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1-12, 2009. https://doi.org/10.1121/1.3058900.

[5] Y. Hu, K. Kokkinakis, "Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners," *Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. EL22–EL28, 2013. https://doi.org/10.1121/1.4834455.

[6] A. Prodeus, M. Didkovska, "Assessment of speech intelligibility in university lecture rooms of different sizes using objective and subjective methods," *Eastern-European Journal of Enterprise Technologies*, vol. 3, no. 5(111), pp. 47–56, 2021. https://doi.org/10.15587/1729-4061.2021.228405.

[7] J. Lochner, J. Burger, "The influence of reflections on auditorium acoustics," *J. Sound Vib.*, vol. 1, issue 4, pp. 426–454, 1964. https://doi.org/10.1016/0022-460X(64)90057-4.

[8] I. Arweiler, J. Buchholz, T. Dau, "Speech intelligibility enhancement by early reflections," *Proceedings of the International Symposium on Auditory and Audiological Research*, Elsinore, Denmark, 2009, vol. 2, pp. 289-298. Available at: https://proceedings.isaar.eu/index.php/isaarproc/article/view/2009-29.

[9] A. Prodeus, K. Kukharicheva, M. Didkovska, "Comparison of speech quality and intelligibility assessments in university classrooms," *Int. J. Archit. Eng. Technol.*, vol. 8, pp. 52-60, 2021. https://doi.org/10.15377/2409-9821.2021.08.5.

[10] G. Fant, *Acoustic theory of speech production*, The Hague, The Netherlands, Mouton, 1960, 326 p. https://doi.org/10.1515/9783110873429.

[11] J. Flanagan *Speech communication*. In: Crocker M.J. (ed.), Encyclopedia of Acoustics. John Wiley, New York, 1997, 2017 p.

[12] N. Cote, *Integral and Diagnostic Intrusive Prediction of Speech*. Springer-Verlag Berlin Heidelberg, 2011, 267 p. https://doi.org/10.1007/978-3-642-18463-5.

[13] P. Loizou, *Speech Enhancement: Theory and Practice*, second ed., Boca Raton: CRC Press, Taylor & Francis Group, 2013, 705 p. https://doi.org/10.1201/b14529.

[14] S. Young, G. Evermann, M. Gales, et al. *The HTK Book*. Cambridge: University Engineering Department, 2009, 355 p. Available at: https://www.researchgate.net/publication/236023819_The_HTK_book.

[15] Y. Tang, C. Arnold, T. Cox, "A study on the relationship between the intelligibility and quality of algorithmically-modified speech for normal hearing listeners," *J. Otorhinolaryngol. Hear. Balance Med.*, vol. 1, no. 5, pp. 1-10, 2018. https://doi.org/10.3390/ohbm1010005.

[16] X. Xu, R. Flynn, M. Russell, "Speech intelligibility and quality: A comparative study of speech enhancement algorithms," *Proc. 28th Irish Signals and Systems Conference (ISSC)*, June 20-21, 2017, pp. 1-6, https://doi.org/10.1109/ISSC.2017.7983599.

[17] M. Keshavarzia, "Comparison of effects on subjective intelligibility and quality of speech in babble for two algorithms: A deep recurrent neural network and spectral subtraction," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. 1493–1503, 2019, pp. 1-5. https://doi.org/10.1121/1.5094765.

[18] J. Ma, Y. Hu, P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387-3405, 2009. https://doi.org/10.1121/1.3097493.

[19] C. Nestoras, S. Dance, "The interrelationship between room acoustics parameters as measured in university classrooms using four source configurations," *Building Acoustics*, vol. 20, no. 1, pp. 43–54, 2013. https://doi.org/10.1260/1351-010X.20.1.43.

[20] M. Jeub, M. Schäfer, P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *Int. Conf. Proc. on Digital Signal Processing (DSP)*, Santorini, Greece, July 5-7, 2009, pp. 1-5. https://doi.org/10.1109/ICDSP.2009.5201259.

[21] Perceptual Evaluation of Speech Quality (PESQ) ITU-T Recommendations P.862, P.862.1, P.862.2. Version 2.0. October 2005.

[22] H. J. M. Steeneken, T. Houtgast, "Validation of the revised STIr method," *Elsevier Speech Communication*, vol. 38, pp. 26-37, 2002. http://resolver.tudelft.nl/uuid:d12e5f8b-19ed-4ad3-b78d-6f6403aaf10c.

[23] D. R. Cox, D. V. Hinkley, *Theoretical Statistics*. Chapman and Hall/CRC, 1974, 528 p. https://doi.org/10.1007/978-1-4899-2887-0.

[24] E. Kandel, T. Jessell, J. Schwartz, S. Siegelbaum, A. Hudspeth, *Principles of Neural Science*, fifth ed., A. Hudspeth (ed.). McGraw-Hill, New York, 2013, 451 p. Available at: https://www.amazon.com/Principles-Neural-Science-Fifth-Kandel-ebook/dp/B009LHFYNG

[25] J. Beerends, S. Wijngaarden, R. Buuren, "Extension of ITU-T Recommendation P.862 PESQ towards Measuring Speech Intelligibility with Vocoders. In New Directions for Improving Audio Effectiveness," *Proceedings of the RTO-MP-HFM-123*, Neuilly-sur-Seine, France, 2005, pp. 10-1–10-6. https://doi.org/10.14339/rto-mp-hfm-123-10-pdf.

[26] A. Prodeus, M. Didkovska, K. Kukharicheva, D. Motorniuk, "Modeling the influence of early sound reflections on speech intelligibility," *Proceedings of the 2020 IEEE 6th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC)*, Kyiv, Ukraine, October 20-23, 2020, pp. 47-50. https://doi.org/10.1109/MSNMC50359.2020.9255657.

[27] S. Naida, V. Didkovskyi, O. Pavlenko, N. Naida, "Spectral analysis of sounds by acoustic hearing analyzer," *Proceedings of the IEEE 39th International Conference on Electronics and Nanotechnology (ELNANO)*, Kyiv, Ukraine, April 16-18, 2019, pp. 421-424. https://doi.org/10.1109/ELNANO.2019.8783915.

[28] S. Naida, V. Didkovskyi, O. Pavlenko, N. Naida, "Objective audiometry based on the formula of the middle ear parameter: A new technique for researches and differential diagnosis of hearing," *Proceedings of the 2019 IEEE 39th International Conference on Electronics and Nanotechnology (ELNANO)*, Kyiv, Ukraine, April 16-18, 2019. https://doi.org/10.1109/ELNANO.2019.8783502.

***Arkadiy PRODEUS**, received the BSc, MSc, PhD, and DSc degrees in electrical engineering from the NTUU "Kyiv Polytechnic Institute), Ukraine, in 1970, 1972, 1982, and 2012, respectively. Now he is professor at the Acoustic and Multimedia Electronic Systems Department, NTUU "Igor Sikorsky KPI". His current interests include digital signal processing, modeling and simulation, pattern recognition, speech and music signals processing.*

***Maryna DIDKOVSKA,** received the BSc, MSc, and PhD degrees in electrical engineering from the NTUU "Kyiv Polytechnic Institute), Ukraine, in 1970, 1972, and 1982, respectively. Now she is senior lecturer at the Department of Mathematical Methods of System Analysis, NTUU "Igor Sikorsky Kyiv Polytechnic Institute". Her current interests include software reliability, project management, artificial intelligence, digital signal processing.*

***Kateryna KUKHARICHEVA** B.Sc., M.Sc., in Electronics and Acoustics Equipment (2016 and 2018 years respectively) in NTUU "Igor Sikorsky Kyiv Polytechnic Institute". Currently is a 3-year student of PhD program in electronics at the Department of Acoustic and Multimedia Electronic Systems of NTUU "Igor Sikorsky Kyiv Polytechnic Institute". During university time, published scientific articles, organized elective about sound engineering and sound design.*