PAN ARCHIVES OF ACOUSTICS
Vol. 40, No. 1, pp. 41–50 (2015)

DOI: 10.1515/aoa-2015-0006

The Influence of the Semantic Material on the Assessment of Speech Reception Threshold

Magdalena KRENZ, Andrzej WICHER, Aleksander SĘK

Institute of Acoustics, Faculty of Physics, Adam Mickiewicz University
Umultowska 85, 61-614 Poznań, Poland; e-mail: oleksek@amu.edu.pl

(*received May 9, 2014; accepted November 27, 2014*)

To determine speech intelligibility using the test suggested by Ozimek *et al.* (2009), the subject composed sentences with the words presented on a computer screen. However, the number and the type of these words were chosen arbitrarily. The subject was always presented with 18, similarly sounding words. Therefore, the aim of this study was to determine whether the number and the type of alternative words used by Ozimek *et al.* (2009), had a significant influence on the speech intelligibility. The aim was also to determine an optimal number of alternative words: i.e., the number that did not affect the speech reception threshold (SRT) and not unduly lengthened the duration of the test. The study conducted using a group of 10 subjects with normal hearing showed that an increase in the number of words to choose from 12 to 30 increased the speech intelligibility by about 0.3 dB/6 words. The use of paronyms as alternative words as opposed to random words, leads to an increase in the speech intelligibility by about 0.6 dB, which is equivalent to a decrease in intelligibility by 15 percentage points. Enlarging the number of words to choose from, and switching alternative words to paronyms, led to an increase in response time from approximately 11 to 16 s. It seems that the use of paronyms as alternative words as well as using 12 or 18 words to choose from is the best choice when using the Polish Sentence Test (PST).

Keywords: speech intelligibility, speech test, speech reception threshold.

1. Introduction

Speech audiometry is an important tool for the diagnosis of hearing loss. Its application, consisting in determining speech intelligibility, which is often presented against a background noise, allows for the assessment of the extent and location of the damage of the hearing organ and central nervous system. The result of such measurements is the so called articulation curve (psychometric function), showing the dependence of the percentage of correctly repeated verbal units on the level of the signal or the signal-to-noise ratio (SNR). On the basis of this function it is possible to determine the speech reception threshold (SRT), i.e., the signal-to-noise ratio at which the subject repeated correctly 50% of the test items, a standard deviation (SD), and its steepness in the SRT (S_{50}) point.

The tests of Pruszewicz *et al.* (1994a; 1994b) are often used to study speech intelligibility of Polish language in adults. They are composed of polysyllabic numerical lists and lists of words formed from monosyllabic, commonly used nouns. Each of the ten articulation lists of the test is phonemically, semantically, structurally, grammatically, acoustically, energetically, and dynamically balanced. The numerical tests are used for a quantitative evaluation of hearing loss, while the word tests are used for a qualitative assessment of the type of hearing loss.

Another type of test is Brachmański and Staroniewicz's (1999) pseudoword test consisting of words that have no semantic value. To properly repeat a logatome, the observer needs to hear the word's every phoneme. The intelligibility determined by this test seems to be more objective and independent of the knowledge and intelligence of the observer. The test consists of 20 series, each has three lists, and each list contains 100 logatomes. All lists are structurally and phonemically balanced.

One of the recently developed tests for determining speech intelligibility is the Polish Sentence Test, (PST) (Ozimek *et al.*, 2009), which was based on the method proposed by Plomp and Mimpen (1979). This test consists of 37 lists of 13 sentences each (a total of 481 sentences). The PST sentences have a proper logical

structure. They are grammatically and syntactically correct. These are sentences used in the everyday life and are understandable to any social group. They do not include colloquial speech, slang, or dialect. Each sentence is singular and in the indicative form. The sentences are short, containing from 3 to 6 words (no more than 9 syllables in total), and each contains a predicate. They do not include punctuation marks or proper nouns. They are semantically neutral and do not contain topics related to violence, politics, or religion. Every sentence occurs only once in the test. Sentences in each of the 37 lists are phonemically, grammatically, acoustically, energetically, and dynamically balanced.

Examination of speech intelligibility using the PST is based on the use of a masking noise called babble-noise (BN). This noise was created by summing up the waveforms of all the test sentences, where all sentences were shifted in time by a random value, and half of them were reversed in time. The result is a 15-s noise, whose random time periods were presented in the measurements of speech intelligibility (Ozimek *et al.*, 2009). BN spectrum is shown in Fig. 1 as a solid line. It significantly resembles the elevation of components in the range of 5–10 kHz, which is characteristic for the Polish language. It results from the significant content of fricatives and affricates consonants in Polish.

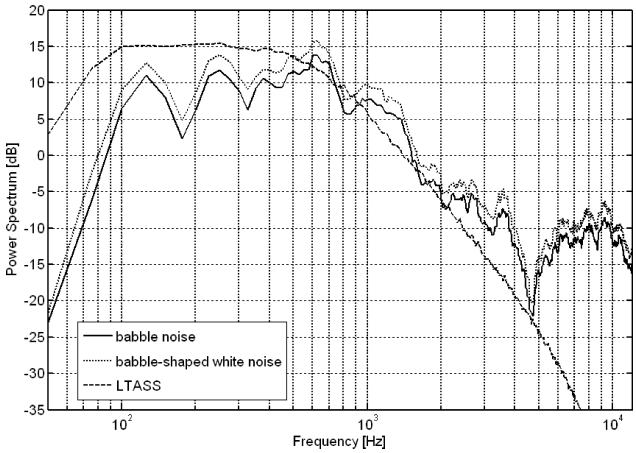

Fig. 1. Power spectral density for babble-noise (BN), babble-shaped noise (BSN), and speech shaped noise (LTASS) (Fastl, 1987). Because power spectra for BN and BSN are nearly identical, the power spectrum curve for BSN was shifted upwards by 2 dB. Adopted from (Ozimek *et al.*, 2009).

The study of speech intelligibility by Ozimek *et al.* (2009) was also conducted using a speech-like noise (Fastl, 1987) (see Fig. 1 dashed line) or the noise obtained by filtering the white noise with a filter whose frequency response was based on an averaged, long-term speech spectrum of Polish (called babble-shaped noise, BSN) (see Fig. 1 dotted line) (Versfeld *et al.*, 2000; Kollmeier, Wesselkamp, 1997). However, taking into account both the energetic and informational masking (Kidd Jr. *et al.*, 2007; Cooke *et al.*, 2008; Shinn-Cunningham, 2013; Brungart *et al.*, 2013), the BN masker seems to be the most appropriate. It maps to the highest extent both the temporal and spectral structure of the Polish language resulting in the highest speech reception threshold (i.e., worse intelligibility). The use of this particular masker is related to the masking of both energy and information, which fully reproduces most of real situations.

An important advantage of determining speech intelligibility with a background noise using PST is the use of the adaptive method (i.e., 1-up/1-down) (Levitt, 1971). The subject are presented with a sentence at a background noise (at a given SNR) and their task is to create a sentence from the 18 words presented on the screen. If the answer is correct (all words of the sentence and their order, the so-called sentence scoring), the SNR is reduced by the step value. If the answer is incorrect, the SNR is increased by the step value. The initial SNR was large enough to allowed giving a correct answer. Within a single measurement run, which was associated with presentation of one list of 13 sentences, 14 SNRs were obtained and the threshold value as well as its standard deviation (SD) were calculated as the average of the last eight values of SNR.

In each case the task of the subject was to create a sentence from 18 different words displayed in the alphabetic order on the screen (so called choice words), which were similar in sound and spelling, i.e., paronyms. The 18 words contained both the words of the played sentence and the alternative ones. For example, if the played sentence was composed of 5 words then 13 words were alternative ones.

However, both the number and type of alternative words were arbitrarily adopted by Ozimek *et al.* (2009), although assumed as optimal by the authors of the cited work. However, these arbitrarily chosen parameters raise some doubts, because each arbitrarily adopted parameter in experiments leads to an uncontrolled influence on the obtained results. Therefore, the main aim of this work was to analyze the influence of the semantic material of the Polish Sentence Test on the speech reception threshold (SRT) and the steepness of the psychometric function (S_{50}).

2. The aim

The main aim of this study was to analyze the influence of the semantic material of the Polish Sentence Test on the speech reception threshold (SRT) and determine whether the number and type of alternative words that are available to a subject in the procedure proposed by Ozimek *et al.* (2009) has a significant in-

fluence on the speech intelligibility. In addition, this work aimed at indicating the number of alternative words that would be optimal: i.e., does not significantly influence the SRT, does not contribute to an increase of the degree of difficulty and does not unduly lengthen the duration of the test.

The experiments by Ozimek *et al.* (2009) were carried out for 18 similarly sounding alternative words. Their number and type have been adopted arbitrarily and perhaps the use of a larger number of alternative words would increase the accuracy of determining the SRT and S_{50}. This can also lead to an assumption that the determination of the SRT for a smaller number of alternative words would be sufficient. The experiment aimed at checking whether the change in the number of alternative words presented to the subjects affects the speech intelligibility was also conducted. It was performed for 12, 18, 24, and 30 words to choose from.

Another, not less important aim of the study, was to determine whether the type of alternative words, i.e., additional words displayed on the screen together with the sentence words, has an impact on speech intelligibility. Ozimek *et al.* (2009) used only 18 paronyms, that is, words whose spelling and pronunciation were very similar to the words contained in the sentences. It can be expected that the use of random words can have a significant influence on the determined SRT. Therefore, the speech intelligibility measurements were carried out using both paronyms for each number of alternative words, as well as random words, which are dissimilar to the words making up the actually presented sentence. These random words were drawn (without repetition) from the set of all words used by Ozimek *et al.* (2009) with the words belonging to the assessed sentence.

Paronyms are "*similar sounding words, unrelated etymologically and semantically*" (Głowiński *et al.*, 1998; Kita, Polański, 2004). They are words, often confused with each other, of similar spelling, pronunciation, and sound articulation. These words have a similar structure and shape but a completely different meaning. In the dictionary of paronyms (Kita, Polański, 2004) closely sounding words are also called words formed from the same word base formation, occurring with various formants (e.g. *swimmer, swimming were derived from the word swim*) (See also http://www.oxforddictionaries.com/definition/american_english/paronym).

Qualitative differences between paronyms are due to the switched sound system (e.g. *rak – kra, pepita – pipeta*) or the conversion of one of the vowels into another (e.g. *kot – lot, kat – mat*). Whereas quantitative differences are the result of the presence of additional sounds or more sounds within a word (e.g., *wariat – wariant* (Kita, Polański, 2004)).

It seems that the use of paronyms as alternative words is justified, because the correct reproduction of the word is possible only when all the phonemes that make up the word are clearly heard.

3. The method

3.1. The test material

The Polish Sentence Test, PST was the sound material used in the present study. The PST consists of 37 lists and each list contains 13 sentences. The maximum number of words (and syllables) in each sentence is no more than 9 (6). There was no word in the whole test that consisted of more than 3 syllables. All the sentences were grammatically and syntactically correct and semantically neutral, i.e., political, war, or gender topics were excluded. The sentences did not contain proverbs, questions, or proper names. All the test material was recorded in a radio studio by a professional male speaker (27 years) who was a Polish radio announcer (see Ozimek *et al.*, 2009, for detials).

To determine a single SRT value, the subjects were presented with a sentence from a randomly selected list. After the sentence presentation, 12, 18, 24, or 30 words to choose from appeared on the screen. The subject had to create the presented sentence (see Fig. 2, which is an example of a set of 24 paronyms – including the words of the presented sentences – which the subject heard in the experiment). Amongst the words, displayed alphabetically, were all the words occurring in the sentence and the alternative words. In the first part of the experiment the alternative words were paronyms, i.e., closely sounding words, matched uniformly to the words in each sentence. In the second part of the experiment the alternative words were the words randomly selected from all the words in the test. For each number of words to choose and for each type of these words (paronyms or random words) the subject heard three different articulation lists, which gave a total of 24 lists to one subject. Each list applied

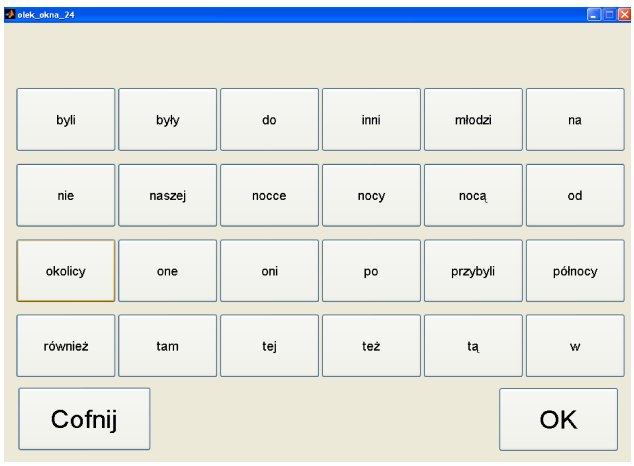

Fig. 2. Example of a screen view containing 24 words to choose from, where paronyms are alternative words.

to one subject was presented only once. When choosing the alternative words the intention was to assign the same number of alternatives for every word occurring in a sentence. These words contain no more than 3 syllables and were semantically neutral. No closely sounding word appeared twice among the alternative words assigned to a given sentence.

The main goal of a single measurement (a single experimental run) was to determine the speech reception threshold (SRT), i.e., such a signal-to-noise ratio at which the listener correctly repeated 50% of the presented sentences. In an adaptive procedure 1-up/1-down, sentences were presented monaurally at a background of the babble-noise (BN) with a fixed level of 65 dB SPL, with the initial value of the SNR in the range (−3 to 0) dB. A correct answer of the subject (i.e., when the whole sentence was reproduced correctly) caused a reduction in the SNR by the step value and an incorrect answer increased the value by a step. Initially, the SNR step change had a value of 2 dB. However, the first incorrect answer changed the step value to 1 dB, which remained constant to the end of a single measurement. The SRT was determined based on the average of the last 8 SNRs. An exemplary course of changes of the SNR as a function of the sentence number for a single list is shown in Fig. 3. The average SRT values presented later in this work were based on three independent measurements, i.e., for three different lists.

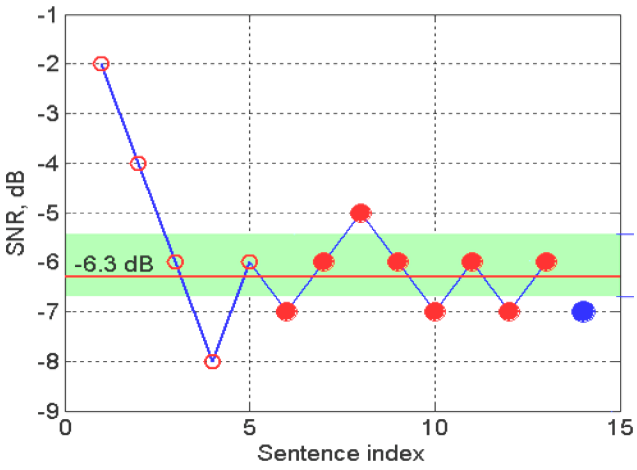

Fig. 3. Signal-to-noise ratio as a function of a sentence number in a list. The SRT was calculated based on 8 last SNRs (filled circles) and is equal to −6.3 dB. The last turnpoint, which is not connected with the other ones, is a virtual turnpoint and the shaded area presents +/− SD. Adopted from Ozimek (2009).

During the study, the response time of each subject was also measured, and the subjects participating in the experiment were not informed of this. This time was measured from the moment when the window with the words to choose appeared on the monitor, until the listener gave a response. The average time obtained for individuals, for the type, and number of alternative words was calculated as the average of all the times obtained for the three lists.

3.2. The apparatus

The previously recorded sentences were played by means of the TDT III system at a background of the babble (24414.0625 Hz sampling rate), which was controlled from a PC running a Matlab based script. This set was completed by a programmable attenuator (PA5) and the headphone buffer (HB7) which were connected to Sennheiser HD 580 headphones. Babble noise began 300 ms before a sentence and ended 300 ms after its end. Additionally, the Hanning window with 20 ms of rise/fall time was used.

During the experiment, the subject was in a sound proof cabin equipped with a touch sensitive screen monitor. Prior to testing, the subject was informed of the course of the experiment. A single series of measurements lasted about 20 minutes, which allowed presenting 3 lists of the test.

3.3. The subjects

10 normal hearing subjects (with the hearing threshold lower than 20 dB HL up to 8 kHz) took part in the investigations. They were students of the Institute of Acoustics, Adam Mickiewicz University and they participated as volunteers. Each subject took part in a training session in which they were presented with sets of sentences which were not used in the main experiment.

3.4. The psychometric function for speech intelligibility

The speech reception threshold (SRT) determines the signal-to-noise ratio at which the identification of 50% of the presented elements of language occurs while speech is presented against a background noise. Therefore, to determine the SRT it is necessary to present words or sentences for several different SNR values.

The psychometric function (articulation function) binds speech intelligibility [%] and the sound pressure level [dB SPL] or the signal-to-noise ratio SNR [dB], and is defined as follows (1):

$$\Phi(\mathrm{SNR}) = \frac{100}{\sqrt{2\pi}} \int_{-\infty}^{\mathrm{SNR}-\mathrm{SRT}/\sigma} e^{-t^2/2}\,\mathrm{d}t, \qquad (1)$$

where t expresses the current value of the SNR.

SRT, which is the threshold of speech intelligibility, and its standard deviation (SD) are the most important parameters of the psychometric function. The steepness of a psychometric function (S_{50}) at the SRT

point is inversely proportional to the standard deviation (SD), which is expressed by the following equation (2):

$$S_{50} = \frac{100}{SD\sqrt{2\pi}}. \tag{2}$$

Figure 4 contains an example of the relationship of the psychometric function of speech intelligibility [%] and the signal-to-noise ratio [dB]. In the figure, the SRT is 9.7 dB and the steepness of the psychometric function is 24.9%/dB. A change in the SRT, for example by 1 dB, means a reduction/increase of intelligibility by nearly 25%, which is a significant value.

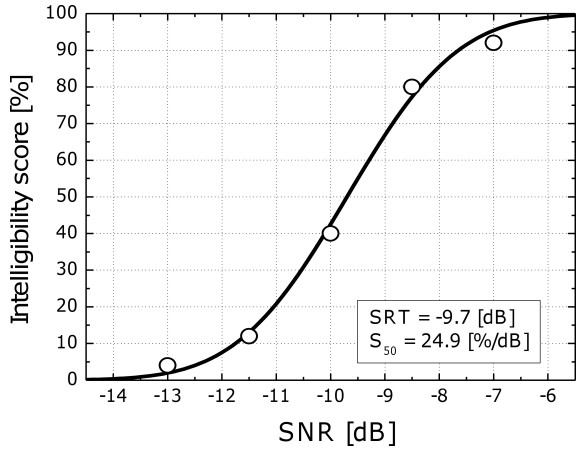

Fig. 4. Example of a psychometric function gathered with the Polish Sentence Test (Ozimek *et al.*, 2009).

4. The results

4.1. The influence of the number of alternative words on the SRT

As it was previously reported, the main aim of this study was to analyze the influence of the semantic material of the PST on the speech reception threshold (SRT) and to investigate an influence of the number of alternative words on the threshold. It was tested, whether the value of SRT changes with an increase in the number of words to choose from. Each of the ten subjects with normal hearing listened to three lists of sentences. They had a choice of size of words: 12, 18, 24, or 30, from which they arranged the heard sentence. From the results obtained for each subject, the SRT was calculated for a different number and type of alternative words, and then the SRT values were averaged with respect to all subjects.

The results of these tests are shown in Figs. 5 and 6. The columns in Fig. 5 illustrate the SRT for different numbers of words to choose from, while the alternative words were paronyms. Vertical bars represent +/− one standard deviation, while the numbers within the columns represent the SRT values (dB).

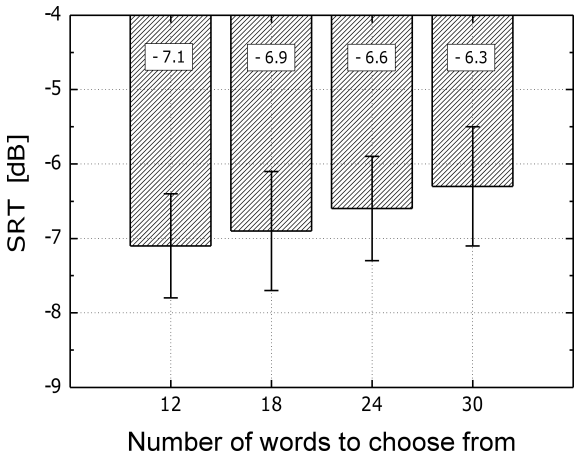

Fig. 5. Influence of the number of words on the speech reception threshold (SRT) for paronyms as alternative words. Vertical bars present +/− the standard deviation (SD), while numbers within the columns present the SRT values.

As it is clear from the data contained there, the obtained results vary with respect to the number of words to choose from. The lowest threshold was obtained for the 12 words (−7.1 dB). Increasing in the number of paronyms (or words to choose from) resulted in a significant increase in the SRT, which reached −6.9, −6.6, and −6.3 dB for 18, 24, and 30 words to choose from respectively. Generally, an increase in the number of alternative words (paronyms) systematically contributed to the deterioration of the speech reception threshold (about 0.2–0.3 dB/6 words).

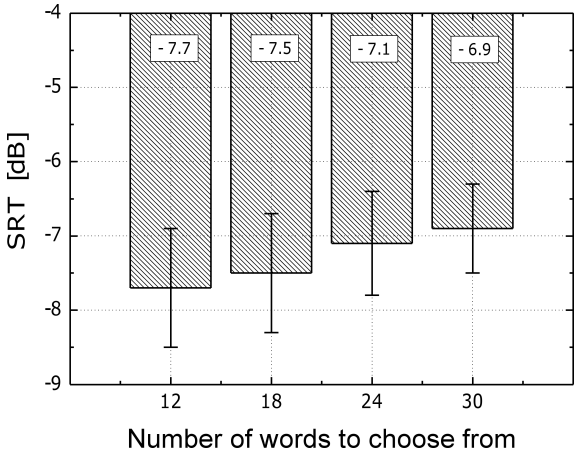

Fig. 6. Influence of the number of words on the speech reception threshold (SRT) for alternative words randomly chosen from the PST. Vertical bars present +/− the standard deviation (SD), while numbers within the columns present the SRT values.

Figure 6 shows similar data obtained when the alternative words were random words, drawn from the set of the PST test words (randomly selected without the words present in the sentence). As in the case

of paronyms as alternative words, increasing the number of words to select from caused significant variations in the speech intelligibility threshold. The lowest values were obtained for 12 words to choose from (-7.7 dB). However, these thresholds reached -7.5, -7.1, and -6.9 dB for 18, 24, and 30 words to choose from, respectively.

The dependencies shown in Figs. 5 and 6 unambiguously suggest that speech intelligibility is dependent on the number of words to choose from and it is higher with a fewer number of words. It is also worth mentioning that a systematic difference between the speech intelligibility for different types of alternative words occurs. For paronyms, regardless of the number of words to choose from, the speech intelligibility threshold reached values greater than for random words.

The differences of the obtained speech reception threshold, both for a different number of words to choose from and for two types of alternative words are rather small. Their significance may raise some doubts, especially when comparing them with their standard deviations, whose doubled values are indicated in each case in the columns illustrating the experimental data. This can make it difficult to draw firm conclusions on the influence of the analyzed parameters on speech intelligibility.

Therefore, a within-subject ANOVA was carried out to the collected data, in which the results for the individual subjects were treated as repeating the same measurement. This analysis was done with respect to two main factors: the type of alternative words (paronyms or random words) and the number of words to choose from (12, 18, 24, and 30) using Genstat statistical package (Lane *et al.*, 1987). The type of alternative words (random, paronyms) and their number proved to be highly statistically significant [$F(1,9) = 23{,}21$; $p < 0.001$] and [$F(3,27) = 28.21$; $p < 0.001$] for the type and number of words to choose from, respectively. However, it is worth noting that the interaction between the type and number of words to choose from was not statistically significant [$F(3,27) = 0.45$; $p = 0.720$]. This means that the differences in the SRT obtained for both types of alternative words were approximately the same, regardless of the number of alternative words.

Taking into account both types of alternative words, Tukey's post-hoc test was conducted. Multiple comparisons of the average SRT values were made due to the number of words to choose from. The significance level $p = 0.05$ was assumed. Tukey's test showed statistically significant differences between the mean SRT for the number of words 12 and 24, 12 and 30, and 18 and 30. In other words, it was possible to identify three homogeneous groups of mean SRTs due to the number of alternative words, i.e., 12–18, 18–24, and 24–30. Post-hoc analyses of subgroups were also conducted: with paronyms (Fig. 5) and random words (Fig. 6). Tukey's test showed that for randomly chosen alternatives words, three homogeneous groups of mean-SRT exist, as in the previous analysis. In turn, when using paronyms, SRT did not show significant differences when using 12, 18, or 24 words to choose from, as well as when the number of words was 24 or 30.

Direct observations of people involved in the experiment and the obtained data suggest that the most difficult task for the subjects was to recreate the sentences when a large number of alternative words appeared. For 30 words to choose from, the subjects had the most problems with the correct reconstruction of sentences. It seems that the word search, even though they were arranged in alphabetical order, was burdened with some difficulty associated with partial forgetting of the heard content. However, when the monitor presented only 12 words, the subjects recreated the sentence more correctly. It could be also assumed that for the subjects, even though they did not hear the whole speech, sentence recreation was easier on the basis of a limited number of words on a screen. While analyzing the data from Fig. 5 and Fig. 6 it is worth noting that regardless of the type of words (paronyms or random words) the increase in the SRT with a change in the number of alternative words from 12 to 30 is equal 0.8 dB (approx. 0.2–0.3 dB/6 words). It is not a large value (less than 1 dB). However, taking into account the slope of the psychometric function (30%/dB) (Ozimek *et al.*, 2009) it indicates that the change in the number of alternative words by, for example, 9, will change intelligibility by about 12–15 percentage points, which is a significant value. It seems that despite minor SRT changes (in the absolute scale), together with an increase in the number of words to choose from, the number of words should not result for an arbitrary choice. It has a significant influence on the final result, which was fully confirmed by the analysis of variance and post-hoc analysis.

The outcome of the presented relations and their preliminary analysis does not allow for specification of the optimal number of words to choose from which should be presented to the subjects. This relationship is characterized by a monotonic course, i.e., systematic increase in the threshold caused by an increasing number of alternative words. It seems that the choice of the optimal number of words should be also linked with other parameters associated with the test, such as the time needed to answer, as shown later in this work.

4.2. Influence of the type of alternative words on the SRT

A direct comparison of the SRT for the two types of words used as alternative is shown in Fig. 7. Subsequent panels of the figure depict the averaged data

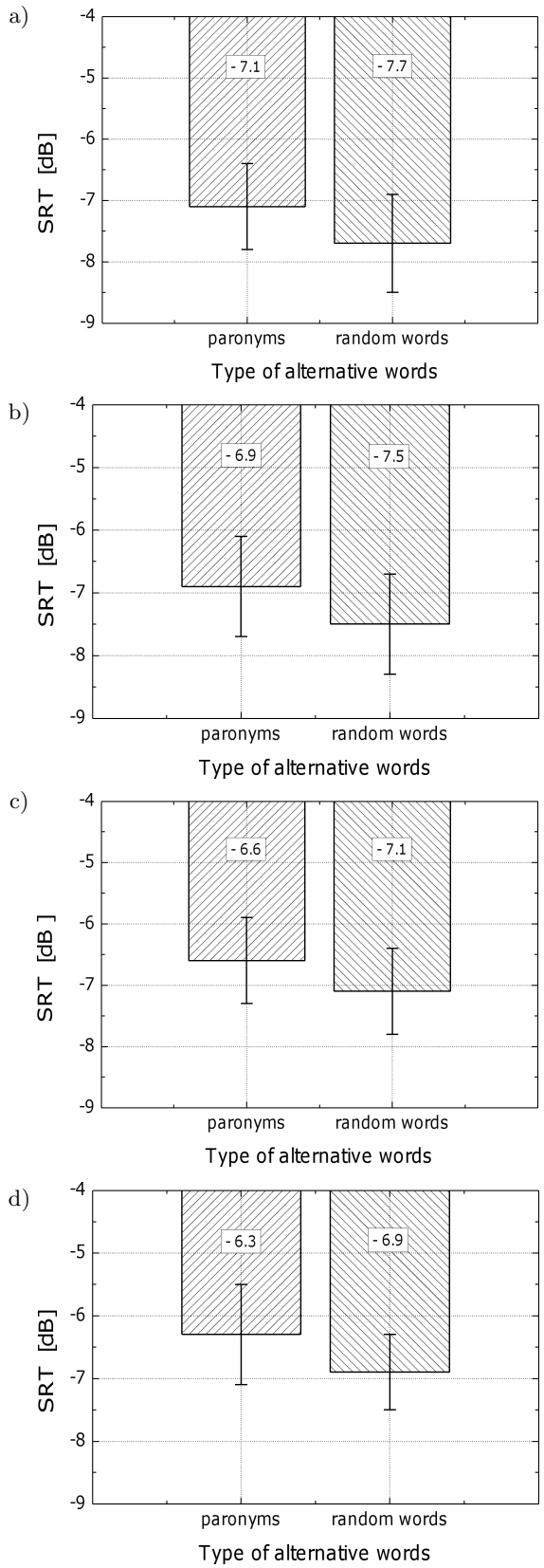

Fig. 7. Influence of the alternative word type on the STR for: a) 12, b) 18, c) 24, and d) 30 words to choose from. Vertical bars present +/− the standard deviation (SD), while numbers within the columns present the SRT values.

for the subjects obtained respectively for 12, 18, 24, and 30 words. Such a presentation of the experimental data obtained highlights the crucial differences of the SRT for an equal number and various types of alternative words. The difference in most cases amounts to 0.6 dB and is statistically significant ($[F(3,27) = 28.21$, $p < 0.001]$). It demonstrates that the correct reproduction of the sentence by the subjects was more difficult when the task was to recreate the sentence from the set of words containing paronyms. Mutual similarity of these words as well as the identical beginnings or endings of words definitely hampered the correct reproduction of sentences. When the words included the ones randomly chosen from the PST, the subjects gave correct answers more often.

The above statements were fully confirmed in the statistical analysis. Analysis of variance showed that regardless of the number of alternative words, randomly selected alternative words gave statistically significant, smaller values of the SRT than for paronyms. For 12 words, the analysis showed that $[F(1,9) = 17.82$; $p < 0.002]$, for 18 words $[F(1,9) = 10.63$; $p < 0.01]$, for 24 words $[F(1,9) = 12{,}78$; $p < 0.006]$, and for the 30 words $[F(1,9) = 22.83$; $p < 0.001]$.

The influence of the type of words (paronyms or random words) on the SRT should be included in the final version of the PST, which should include both types of alternative words. Exchanging paronyms for random words varies the SRT threshold by 0.6 dB, which is equivalent to changing the intelligibility by more than 15 percentage points, when speech intelligibility is analyzed based on psychometric functions. This value is significant, not without an impact on the final estimate of the SRT.

4.3. Response time

The time needed to respond to the individual sentence is essential in the analysis of speech, because it suggests the degree of difficulty of the test. Therefore an analysis of the relationship between changes in the number and type of words to choose from, and the response time of the subject's answers was performed. This time was measured from the moment when the words to choose from appeared on the screen, up to the time of acceptance of the answer. For each type and the number of alternative words, the collected response times were averaged across the subject and are presented in Fig. 8.

As shown in Fig. 8, increasing the number of words to choose from 12 to 24 causes a steady increase in the time required to respond. These times are specified in each column with numbers without parentheses, For paronyms, the growth also occurs with the increase of words to choose from 24 to 30. However, in the case of random words from the words of the test, a slight reduction of the time occurs.

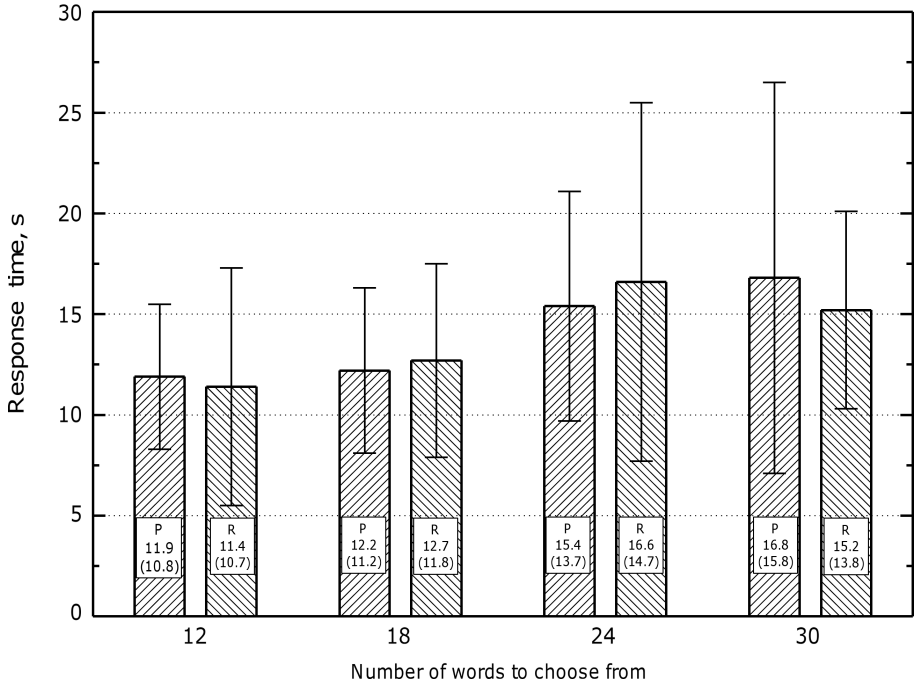

Fig. 8. Response time as a function of the number of words to choose from: for alternative words randomly chosen for the PST (R) and paronyms (P). Vertical bars present +/− the standard deviation (SD). The numbers within the columns in brackets express the averaged response time, while the numbers without brackets present the averaged response time excluding the shortest and longest response times (see the text for details).

These results were subjected to a within-subject ANOVA in which alternative type of words and number of words to choose from were analyzed. Type of the words turned out to be statistically non-significant [$F(1,9) = 0.03$; $p = 0.856$]. The average times obtained for the two types of alternative words were equal to 14.08 and 13.97 for paronyms and random words, respectively. However, the number of alternative words proved to be a statistically significant factor [$F(3,27) = 4.39$; $p = 0.012$] reproducing the observed increase in time with the increasing number of words to choose from. Nevertheless, the average times obtained for 24 and 30 alternative words are very close to each other (i.e., 16,02 and 16 seconds respectively). The correlation between the type and number of words was not statistically significant [$F(3,27) = 0.62$; $p = 0.611$]. This suggests that regardless of the number of words to choose from, the differences in the time needed to respond in the case of paronyms and random words were approximately constant. It is worth noting, however, that for 30 words to choose from and for random words, this rule is not met: increasing the number of words to choose from up to 30 caused a reduction in the time needed to respond.

Tukey test conducted on the results obtained for paronyms showed that the differences in the time needed to respond in the case of 12 or 18 alternative words were not statistically significant. The situation was similar when applied to 24 or 30 alternative words. This means that the relationship is analogous to the SRT changes. In turn, using randomly selected alternative words has shown that differences in the time needed to answer were not statistically significant when using 12 or 18 alternative words, 18 or 30, and 24 or 30 words to choose from. The results of these analyses showed that the change in the number of alternative words brought about a similar effect regardless of the type of alternative words to choose from.

The obtained results revealed that the time needed to answer, contained in the histogram for all experimental conditions shown in Fig. 9, was varied and ranged from 1 to 100 s. This scatter is partly due to

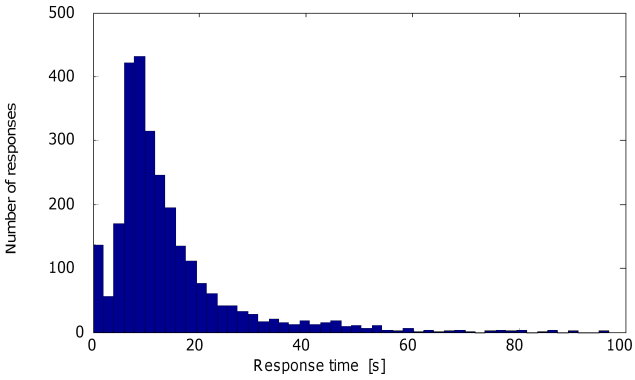

Fig. 9. Histogram of the response time ($\Delta t = 2$ s).

a different number of words to choose from. For example, with 30 words to choose from, this time lengthened, as shown above. However, there were also cases in which the subjects were unable to reproduce the heard sentence and immediately switched to the next sentence. This was especially true when the signal to noise ratio was lower than the speech reception threshold. The resulting response time was very short. It should be noted, however, that in general the longest and shortest time to respond was approximately the same for both types of alternative words used and for different number of words to choose from.

Therefore, to determine a more realistic average time needed to answer and analyze the differences between them for all experimental situations, further analysis of the collection to the times was conducted. However, this collection is limited to values greater than 5 s, and those that are no more than two times greater than that achieved global mean value. The mean values limited in the above manner of the collection of response time are shown in Fig. 8 by numbers in parentheses. This limited collection of data was also subjected to a similar within-subject ANOVA, in which statistical significance of the impact of the type and number of alternative words for the response time was tested. The analysis resulted in similar results as in the case of all the times collected. The number of alternative words was statistically significant [$F(3,27) = 4.76$; $p = 0.009$]. However, both the type of alternative words and correlation of type and number of the words to choose from were not statistically significant: [$F(1,9) = 0.34$; $p = 0.576$] and [$F(3,27) = 1.18$; $p = 0.334$].

5. Discussion

Speech intelligibility measurements play an extremely important role in the diagnosis of hearing loss, hence the need to develop the appropriate diagnostic tests. It seems that the Polish Sentence Test proposed by Ozimek *et al.* (2009), is an important step towards the full development of such a tool for the Polish language and can be easily transferred to other Slavic languages. The compactness of this test and the analysis carried out so far suggest it is a very good application in clinical trials, taking into account the duration of a single measurement, bringing meaningful results. The development of this type of test, in spite of numerous papers available on this subject, requires the adoption of a number of arbitrary assumptions. Therefore, Ozimek *et al.* (2009) adopted some rational and irrefutable assumptions (i.e., number of sentences within the test), some of which are thoroughly analyzed. However, among these assumptions there are those that require further evaluation and careful examination. For example, the assumption of alternative words, which in the case of the original test were paronyms, the assumption of a predetermined number of words to choose from which each subject had at his disposal, or how to respond (i.e., a computer screen with the words to choose from and not repeated sentence by the subject). Therefore, the study on these aspects of the application and reliability of the PST seems to be fully justified and necessary.

The main aim of this study was to analyze the influence of the semantic material of the PST on the SRT and verify the validity of the applicability of 18 words to choose from and type of these words. The study showed a constant, small but statistically significant, difference between the speech intelligibility presented against the background noise for the two types of words, i.e., random words from the PST and paronyms. This means that the type of alternative words, i.e., words added to the presented sentences which are presented to the subject is important. A constant difference of speech intelligibility for these types of words means that both of these types can be successfully used. However, the use of paronyms – words being very similar to the words contained in the sentence – makes the identification of sentence words a bit more difficult. The subject must be able to identify and reproduce correctly all the phonemes contained in these words, as it is the case of the logatome tests (Kluk, Moore, 2001; Brachmański, Staroniewicz, 1999). The subject cannot base on the heard words as paronyms often differ from one another with only one phoneme. This additional difficulty contributes to slightly higher SRTs which are observed in the experiment. The use of paronyms includes a significant masking not only energy, but also informational masking, which plays a significant role in the perception and intelligibility of speech. Therefore, the using of alternative paronyms as words in the PST seems fully justified.

The study also showed the possibility of shortening the duration of the test and its simplification by reducing the number of words to choose from the 18 used by Ozimek *et al.* (2009) to 12. Regardless of the type of alternative words the difference in the SRT for 18 and 12 words was equal to only 0.2 dB, and was not statistically significant. Also, the average time to respond in a test of 12 words to choose from was shorter than in case of 18 words. Differences in the average response time are perhaps not significant, but when one takes into account the need to carry out several tests for one subject (i.e., to obtain repeatability of measurement or the need to use several different noise levels), they begin to take on a greater significance. The use of words to choose from 24 or 30 contributes to a statistically significant increase in the SRT as well as to extending of the duration of the response, so that their further use does not appear to be justified.

In conclusion, it is worth noting that the parameters proposed by Ozimek *et al.* (2009) are not in any way challenged and this test provides reliable results on

the SRT. The method of presenting the words to choose from on a screen, and the task, lead to a unique determination of the intelligibility threshold and do not pose any difficulty. Test parameters, the number of words to choose from, as well as their type (paronyms or words randomly selected from the test) are optimal and do not pose a doubt. However, to facilitate the task for the subjects, the reduction of the number of words to choose from to 12, which leads to a small reduction in the duration of the test, is also possible, as shown in this work. This does not affect the speech threshold obtained.

6. Conclusions

The results of this study allow the following conclusions:

1. The increase in the number of words to choose from 12 to 30 significantly increases the speech reception threshold. This growth was approximately the same for both paronyms and random words from the test, and is approximately 0.3 dB/6 words.
2. The use of paronyms as alternative words as opposed to the words chosen at random from the test, leads to an increase of the speech intelligibility threshold by about 0.6 dB, for each number of words to choose from. Considered the psychometric function this increase is equivalent to the intelligibility decline by 15 percentage points.
3. Enlarging the number of words to choose from, and exchanging the alternative words from random words from the test to paronyms leads to a longer response time, approximately from 11 to 16 s.
4. The use of paronyms as alternative words, as well as 12 or 18 words to choose from, seems to be the optimal choice when using the Polish Sentence Test for determining speech intelligibility.

Acknowledgments

The authors would like to thank two anonymous reviewers for helpful comments on an earlier version of this paper.

References

1. Brachmański S., Staroniewicz P. (1999), *Phonetic structure of a test material used in speech quality measurements* [in Polish: *Fonetyczna struktura materiału testowego stosowanego w subiektywnych pomiarach jakości mowy*], Speech and Language Technology, **3**, 71–80.
2. Brungart D., Iyer N., Thompson E.R., Simpson B.D., Gordon-Salant S., Schurman J., Vogel C., Grant K. (2013), *Interactions between listening effort and masker type on the energetic and informational masking of speech stimuli*, Acoustical Society of America Meeting, Montreal, **19**, 060146, 1–9.
3. Cooke M., Garcia Lecumberri M.L., Barker J. (2008), *The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception*, The Journal of the Acoustical Society of America, **123**, 414–427.
4. Fastl H. (1987), *A background noise for speech audiometry*, Audiological Acoustics, **26**, 2–13.
5. Głowiński M., Kostkiewiczowa T., Okopień-Sławińska A., Sławiński J. (1998), *Dictionary of literary terms* [in Polish: *Słownik terminów literackich*], (Ossolineum, Wrocław).
6. Kidd Jr. G., Mason C.R., Richards V.M., Gallun F.J., Durlach N.I. (2007), *Informational masking*, [in:] *Auditory Perception of Sound Sources*, Yost W.A., Popper A.N. Popper, Fay R.R. [Eds.], Springer Handbook of Auditory Research, **29**, 2008, 143–189.
7. Kita M., Polański E. (2004), *Paronyms dictionary* [in Polish: *Słownik paronimów czyli wyrazów mylonych*], PWN, Poznań.
8. Kluk K., Moore B.C.J. (2001), *Logatome intelligibility presented at a background of a speech-shaped noise for persons with "dead regions" in a high frequency range* [in Polish: *Wyrazistość logatomów prezentowanych na tle szumu mowopodobnego dla osób z niedosłuchem typu 'martwe pola' w zakresie wysokich częstotliwości*], Proceedings of Open Seminar on Acoustics, Wrocław-Polanica Zdrój.
9. Kollmeier B., Wesselkamp M. (1997), *Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment*, The Journal of the Acoustical Society of America, **102**, 2412–2421.
10. Lane P., Galwey N., Alvey N. (1987), *Genstat 5. An Introduction*, (Clarendon Press, Oxford).
11. Levitt H. (1971), *Transformed up-down methods in psychoacoustics*, J. Acoust. Soc. Am., **49**, 467–477.
12. Ozimek E., Kutzner D., Sęk A., Wicher A. (2009), *Polish sentence tests for measuring the intelligibility of speech in interfering noise*, International Journal of Audiology, **48**, 433–443.
13. Plomp R., Mimpen A.M. (1979), *Improving the reliability of testing the speech reception threshold for sentences*, Audiol., **18**, 43–53.
14. Pruszewicz A., Demenko G., Richter L., Wika T. (1994a), *New articulation lists for speech audiometry. Part I*, Otolaryngol. Pol., **48**, 50–55.
15. Pruszewicz A., Demenko G., Richter L., Wika T. (1994b), *New articulation lists for speech audiometry. Part II*, Otolaryngol. Pol., **48**, 56–62.
16. Shinn-Cunningham B. (2013), *Understanding informational masking from a neural perspective*, Acoustical Socitety of America Meeting, Montreal, **19**, 060143, 1–3.
17. Versfeld N.J., Daalder L., Festen J.M., Houtgast T. (2000), *Method for the selection of sentence materials for efficient measurement of the speech reception threshold*, The Journal of the Acoustical Society of America, **107**, 1671–1684.