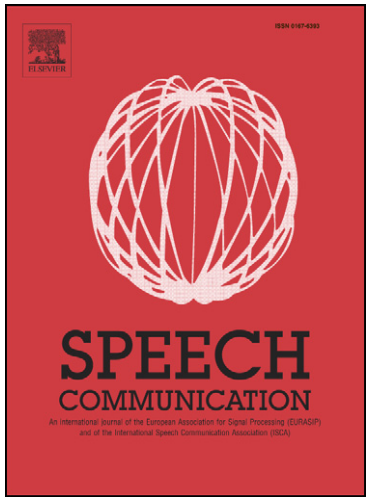
Development and evaluation of Polish digit triplet test for auditory screening

Edward Ozimek, Dariusz Kutzner, Aleksander Sęk, Andrzej Wicher



Development and evaluation of Polish digit triplet test for auditory screening

Edward Ozimek, Dariusz Kutzner, Aleksander Sęk, Andrzej Wicher

PII: S0167-6393(08)00139-8
DOI: 10.1016/j.specom.2008.09.007
Reference: SPECOM 1755

To appear in: *Speech Communication*

Received Date: 4 February 2008
Revised Date: 19 September 2008
Accepted Date: 19 September 2008

Please cite this article as: Ozimek, E., Kutzner, D., Sęk, A., Wicher, A., Development and evaluation of Polish digit triplet test for auditory screening, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.09.007

Development and evaluation of Polish digit triplet test for auditory screening

Edward Ozimek, Dariusz Kutzner, Aleksander Sęk, Andrzej Wicher

Institute of Acoustics, A. Mickiewicz University, 61-614 Poznań, ul. Umultowska 85, Poland

Corresponding author:
Edward Ozimek
Institute of Acoustics, A. Mickiewicz University
Umultowska 85
61-614 Poznań, Poland
tel.: +48 618295133, fax: +48 618295123
E-mail: ozimaku@main.amu.edu.pl

Abstract

The objective of this study was to develop and evaluate the Polish digit triplet test for speech intelligibility screening. The first part of the paper deals with the preparation of the speech material, the recording procedure and a listening experiment. In this part, triplet-specific intelligibility functions for 160 different digit complexes were determined and 100 'optimal' triplets were selected. Subsequently, 4 statistically balanced lists, each containing 25 different digit triplets, were developed. The speech material was phonemically equalized across the lists. The mean SRT and mean list-specific slope S_{50} for the Polish test are –9.4 dB and 19.4 %/dB, respectively, and are very similar to the data characterizing the German digit triplet test. The second part describes the results of the verification experiments in which reliability of the developed test were analyzed. The retest measurements were carried out by means of the standard constant stimuli paradigm and the adaptive procedure. It was found that mean SRT obtained with retest study was within the limits of standard deviation, in agreement with those obtained in the basic experiment.

Keywords: speech intelligibility; intelligibility function; digit triplet; speech-reception-threshold; auditory screening.

1. INTRODUCTION

Digits and numbers as a test material have been used in speech perception measurements for a long time. They have been used for speech intelligibility measurements and for clinical purposes (Fletcher, 1929; Pruszewicz *et al.*, 1994a, b), auditory screening (Smits *et al.*, 2004; Wagener *et al.*, 2005a) and in studies on the influence of speech context on intelligibility (Miller *et al.*, 1951; Kalikow *et al.*, 1977). They have been also used for a study devoted to pronunciation differences between native and non-native American English speakers (Schmidt-Nielsen, 1989). Digit triplets presented in a dichotic way were proved to be useful in evaluation of central auditory processing (Bellis, 1996). They were used in experiments on dichotic speech intelligibility for various signal bandwidths (Strouse and Wilson, 2000). Digit material was also used in a recent study by Smits and Houtgast (2007) on digit recognition for different kinds of masking signals.

A special category of digit test material are complexes composed of different digits, namely digit pairs like 2-5, 8-1, or digit triplets like 1-4-3, 9-2-1, pronounced as: *two-five*, *eight-one*, or *one-four-three*, *nine-two-one*, etc. Like standard words or sentence tests, the triplets are usually presented to the subject against an interfering noise at various signal-to-noise ratios (SNR). On the basis of the test results, an intelligibility function is derived and the speech-reception-threshold (SRT), i.e. SNR yielding 50% speech intelligibility, can be estimated.

Digit triplets have several advantages over single digits or numbers. First of all, they produce steep intelligibility functions, so the SRT estimate is characterized by a relatively low standard deviation (Smits *et al.*, 2004). Since a given digit triplet is composed of only three elements, it is easy to remember the successive digits in a triplet. Digit triplet tests were proved to provide accurate SRT values, without making demands on the patient's cognitive abilities (memory) (Smits *et al.*, 2004). Besides, since digit triplets have no context, it is hard

to learn them all by heart, especially when they are presented in a random order. Accordingly, a given triplet list can be presented to a subject several times without the so-called learning effect which influences the measured SRTs. Although SRTs obtained for the digit triplet test are lower than those for the sentence test, the results of both tests are correlated (Smits *et al.*, 2004). Thus, it is possible to predict the SRT for sentence intelligibility on the basis of the SRT obtained for the digit triplet material. Furthermore, since measurements of the SRT can be done via telephone or the Internet, the digit triplet test can be widely used for extensive screening of speech intelligibility.

It should be mentioned at this point that digit triplet tests are mainly used for general hearing screening. This is because they are composed of only ten words (i.e. single digits), hence they fail to reflect everyday speech accurately. Therefore, they cannot serve as a test for precise speech intelligibility measurements. Furthermore, since a digit triplet test contains a relatively large number of vowels, while certain phonemes do not occur in it at all, triplet tests do not reflect the general phoneme distribution of a given language. However, due to the aforementioned correlation between SRTs obtained for digit triplets and sentences, the digit triplet materials were proved to be very useful for screening purposes (Smits *et al.*, 2004; Wagener *et al.*, 2005a).

The digit triplet studies examined the relationship between SRTs and self-reporting auditory disability of adult listeners (Smits *et al.*, 2006). Digit triplet tests have been so far developed for Canadian English (Rudmin, 1987), Danish (Elberling *et al.*, 1989), American English (Ramkissoon *et al.*, 2002, Wilson and Weakley, 2004), English (Hall, 2006), Dutch (Smits *et al.*, 2004) and German (Wagener *et al.*, 2005a).

Apart from auditory screening, digit triplets were also used in the investigations concerning various aspects of speech perception and processing in the auditory pathway, for example: dichotic listening (Broadbent, 1954). More recently, digit triplets (and digit pairs)

have been used in a study concerning intelligibility of speech presented in a multi-talker noise for normally-hearing and hearing-impaired subjects (Wilson *et al.*, 2005).

As far as Polish materials for speech intelligibility are concerned, several tests have been proposed so far. For example, the Polish CVC and number test (Pruszewicz *et al.*, 1994a,b), the logatome test (Brachmański and Staroniewicz, 1999), the Corpora database (Grocholewski, 2001) and the recently developed Polish sentence test (Ozimek *et al.*, 2006). There is, however, no digit triplet test for the Polish language. This fact prompted us to develop such a test based on a statistically equivalent list of comparable phoneme distributions.

2. DEVELOPMENT OF THE TEST MATERIAL

The development of the Polish digit triplet test progressed through the following stages: initial selection of digit triplets, recording session, triplet intelligibility measurements, selection of a homogeneous group of triplets and composition of statistically and phonemically equivalent lists containing different digit complexes.

2. 1. Preliminary selection of digit triplets and recording session

A single digit triplet is a digit complex composed of 3 digits of values from 0, 1…to 9. Thus, a list containing the total number of possible combinations, i.e. 10^3, was prepared. It has been generally suggested that for the sake of homogeneity the test should contain only monosyllabic digits (Smits *et al.*, 2004; Wagener *et al.*, 2005a). It results from the fact that when presented digits, subjects first discriminate the number of their syllables and then discriminate on the basis of phonemes. If there is a disproportion between the number of monosyllabic and disyllabic digits, certain words can deviate perceptually from others and, consequently, the homogeneity of speech material might be reduced. This must have an effect

on the intelligibility function. This problem was considered thoroughly and it was found that when proportions of disyllabic and monosyllabic digits were similar, the difference between SRTs for two types of digits was about 0.2 dB, thus indicating that the type of digit (either disyllabic or monosyllabic) did not have any significant effects on speech material homogeneity. However, if the proportion of monosyllabic and disyllabic digits is high, for example 8 to 2, the difference between SRTs for two types of digits amounts to about 1 dB. In this case it is reasonable to reject the digits that deviate perceptually from others and might reduce material homogeneity. This is why the disyllabic digit '7' and the disyllabic digits '7' and '9' were excluded from German (Wagener *et al.*, 2005a) and Dutch (Smits *et al.*, 2004) digit triplet tests, respectively. However, in the Polish language as many as six out of ten digits are disyllabic. Therefore, if the disyllabic digits were to be excluded, the test would consist of four digits only. Accordingly, the prepared test has been composed of all monosyllabic as well as disyllabic digits, i.e. all digits from 0 to 9. From the possible 10^3 combinations of different triplets, complexes comprising repeated digits were excluded (i.e. 1-1-2, 4-5-4 or 1-1-1)[1]. The number of available combinations was therefore reduced to 720. In the next step of triplet selection, digit distribution across triplets was considered. The distribution should be as uniform as possible (i.e. the probability that a given digit occurs in available position should be approximately the same). This pre-selection reduced the number of digit triplets to 160 which is a somewhat arbitrary number[2].

[1] It should be mentioned that in the German digit triplet test, complexes containing repeated digits are permitted (Wagener *et al.*, 2005a).

[2] Our intention was to include 100 different triplets in the test, all of them yielding the same intelligibility. Since we had expected that some triplets would not meet this condition, it was decided that more than 100 triplets had to be recorded. Since the measurements of the triplet-specific psychometric functions for 50 subjects were expected to be time-consuming, the number of 160 triplets was chosen as a reasonable number.

The 160 triplets were read out in a recording studio by a male Polish native speaker. The speaker was asked to keep a natural intonation, with approximately the same loudness level and vocal effort over time. Each triplet was read out and recorded at least twice with the Neumann U87 capacitor microphone. To avoid the so-called proximity effect (i.e. enhancement of speech low-frequency components), the microphone functioned in the omni-directional mode. The microphone output fed one of the input channels of the Yamaha 02R mixer, in which the signal was pre-amplified and converted into the digital domain at a sampling rate of 44.1 kHz and with a resolution of 24 bits. It was also digitally high-pass filtered at a cut-off frequency of 80 Hz. The signals were then sent via an optical connection (ADAT-type) to a PC and stored on a hard disc using the Samplitude Pro v.8.2. software. After the recording, all the signals were edited, labelled unambiguously and stored on a PC hard disc as separate sound files (*.wav).

2. 2. Measurements of triplets intelligibility against noise

2. 2. 1. Equipment, procedure and subjects

The signals were generated by means of the Tucker-Davis Technologies (TDT) system 3 comprising: the RP2 Real Time Processor and the HB7 headphone amplifier. They were presented monaurally to the subjects via the Sennheiser HD 580 headphones. The system and measurements were controlled by experiment-devoted software implemented in Matlab 6.5 (*MathWorks*). The digit triplets were mixed with an interfering signal called the digit-shaped noise, i.e. signal generated by means of superimposing all the recorded triplets. Before the summation, half of the triplets (randomly chosen) were reversed in the time domain. Moreover, in order to obtain a 10-second duration of the digit-shaped noise, the successive triplets were shifted in the time domain. The power spectrum of the digit-shaped noise matched power spectra of the triplets. During the listening experiment, some portion of

the masker was randomly taken from the 10-second digit-shaped noise. Therefore, one can say that the noise used in the study was not a frozen noise. Fig. 1 presents the power spectrum density of the digit-shaped noise used in the present study.

INSERT FIG. 1

For each triplet, speech intelligibility was measured by means of the constant stimuli paradigm. The masking noise started 300 ms before the speech signal and ended 300 ms after the signal. By integrating all samples of a signal, the root mean square (rms) of speech and noise was determined. Each out of 160 digit triplets was presented to a subject at 7 values of SNR: -14.5; -13.0; -11.5; -10.0; -8.5; -7.0 and -5.5 dB. The SNR was defined as the difference between 20log of speech rms_s and 20log of noise rms_n. Similar rms definition was used in studies by Versfeld *et al*. (2000), Kollmeier and Wesselkamp (1997), Smits *et al.* (2004) and Wagener *et al*. (2005a,b). The SNR values had been chosen on the basis of preliminary measurements and turned out to encompass the expected range of SRT values. The level of noise was kept constant at 70 dB SPL, so the SNR value was determined by the triplet level. The signal level was calibrated with B&K instruments (the artificial ear type 4153 connected to the microphone type 4134, the preamplifier type 2669 and the amplifier type 2610). The order of triplet presentation and that of the SNRs were both randomized according to a uniform distribution. It should be stressed that each triplet was allowed to be presented to a given listener 7 times (i.e. at 7 different SNRs). Digit triplets have hardly context, therefore it has been assumed that it would be difficult to memorize them[3]. Like in the case of the Dutch triplet test (Smits *et al.*, 2004), the signals were not presented in any carrier phrase (in a telephone test, a precise test instruction is planned to be presented to a subject prior to the measurement).

[3] In the case of sentence tests, each utterance is presented to a given listener only once in order to avoid the learning effect (Versfeld *et al.*, 2000; Ozimek *et al.*, 2006).

During the intelligibility measurements, the subject was seated in an acoustically-insulated booth and asked to type on a keyboard what he/she had just heard. In order to avoid typing in mistakes resulting from inattentiveness or fatigue, the subject was asked to confirm his/her response by typing it in once again. The response was stored on a hard disc only if both the first and the second response were identical ('double response'). If the subject did not duplicate the first response correctly, he/she was asked to type it in again until a correct confirmation was obtained. The main advantage of such a procedure over typing the responses in without further confirmation was that the subjects were forced to pay attention to what they were typing on a keyboard. Furthermore, the so-called 'lapse rate' (i.e. the probability of mistakes made for intelligible speech signals) is reduced and, consequently, bias in intelligibility data is minimized (Klein, 2001). The subject's response was scored 1 if the entire triplet was repeated correctly, otherwise the response was scored 0 (this method is called triplet scoring[4], which produces steep intelligibility functions, resulting in more accurate measurement). A similar scoring method was used in the Dutch triplet test aimed at screening via telephone by means of adaptive procedure (Smits *et al.*, 2004). Each subject was presented with 1,120 triplets (7 SNRs*160 triplets). The experimental session lasted about 4 hours for each subject, including breaks that the subjects were allowed to take whenever they wished. Fifty normal-hearing subjects took part in the experiment (22 female and 28 male). The hearing level of all the subjects at audiometric frequencies from 250 to 8000 Hz did not exceed 15 dB HL. The subjects were paid for their participation in the

[4] Since the measurements were not aimed at determining digit-specific psychometric functions, intelligibility score for separate digits was not considered in this study, although digit scoring might provide some additional information. Apart from that, to shorten measurement time in practical application of the test, it was assumed that a fast adaptive 1-up/1-down procedure will be used. This procedure can be applied for triplet scoring but not for digit scoring

experimental sessions. In total, 56,000 digit triplets were presented during the measurements (excluding training sessions).

2. 2. 2. Determination of the intelligibility function parameters: SRT and slope (S_{50})

The obtained intelligibility scores were pooled across subjects for each triplet and each SNR, and proportions of correct responses were computed. Since a single triplet was presented at a given SNR 50 times (for 50 subjects), the intelligibility score was always an integer multiple of 2% (100%/50 presentations). As expected, the intelligibility score increased monotonically with the SNR and turned out to depend significantly on this parameter {F(159,1119)=1410.12, p<<0.001}.

Using the maximum likelihood method (ML), the intelligibility functions were fitted to the obtained data. Due to data-pooling across subjects, the intelligibility functions exhibit properties for respective triplets obtained for normal-hearing population; thus, these functions will be called triplet-specific intelligibility functions. A triplet-specific intelligibility function links the probability of a correct triplet response to the SNR value at which the triplet is presented. As far as this study is concerned, the cumulative distribution function (CDF), expressed by formula (1), was fitted to the experimental data.

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

This function is characterized by two parameters: SRT (i.e. the signal-to-noise ratio that produces 50% correct responses) and slope S_{50} (determined at the SRT point). For a single digit triplet, the slope of CDF function and the standard deviation σ of corresponding probability density function (PDF) are easily convertible by formula (2):

$$S_{50} = \frac{100}{\sigma\sqrt{2\pi}} \qquad (2)$$

With data points describing each triplet-specific intelligibility function it was possible to apply an iterative procedure based on the *maximum likelihood* (ML) criterion for finding the triplet-specific SRT and the triplet-specific slope S_{50} (SRT and S_{50} were adjusted iteratively in such a way that the negative log-likelihood ratio was minimized). Finally, 160 triplet-specific intelligibility functions and 160 corresponding SRT and S_{50} values were determined.

3. COMPOSITION OF THE FINAL DIGIT TRIPLET TEST

3. 1. Final selection of triplets

It is well known that reliable and accurate tests for speech intelligibility measurements must be composed of statistically equivalent lists of speech elements, i.e. characterized by similar and steep intelligibility functions for each list regardless of its index (Plomp and Mimpen, 1979; Nilsson *et al.*, 1994; Kollmeier and Wesselkamp, 1997; Versfeld *et al.*, 2000; Wagener, 2003; Smits *et al.*, 2004; Wagener *et al.*, 2005a; Ozimek *et al.*, 2006). The similarity of SRT values and large S_{50} implies low intra- and inter-variability of the lists, i.e. high accuracy of speech intelligibility measurements. The accuracy of measurement is related to the so-called list-specific intelligibility function, $S_{50list.}$ The SRT of this function, i.e. list-specific SRT, is determined by means of averaging SRT values of elements in the lists; in this study triplet-specific SRTs. The list-specific intelligibility function can be determined according to the probabilistic model proposed by Kollmeier (1990). The model states that the list-specific intelligibility function can be described as a convolution of two functions: a distribution of word-specific SRT values and intelligibility function having slope S_{50mean}, which is equal to average slope of word-specific intelligibility functions. Moreover, the list-specific slope S_{50list} can be determined according to formula (3) (Kollmeier, 1990):

$$S_{50list} \approx \frac{S_{50\,mean}}{\sqrt{1+\frac{16 S_{50mean}^2 \sigma^2{}_{SRT}}{(\ln(2e^{1/2}-1+2e^{1/4}))^2}}} \qquad (3)$$

where: S_{50mean} is average S_{50} across digit triplets (expressed in dB^{-1}, i.e. 21.4 %/dB–0.214 dB^{-1}) and σ_{SRT} is the standard deviation of SRTs across digit triplets. Formula (3) indicates that large slope of the list-specific intelligibility function, i.e. high accuracy of SRT measurement, requires a large S_{50mean} value and minimal spread of SRTs, i.e. minimization of σ_{SRT}. As a rule of thumb, the more steep and more similar intelligibility functions characterizing the test items are, the more accurate and repeatable measurement is possible. In the ideal case there is no spread of SRTs across digit triplets, i.e. $\sigma_{SRT}=0$ and $S_{50list}=S_{50mean}$. Nevertheless, in real measurements $\sigma_{SRT}>0$, therefore $S_{50list}<S_{50mean}$ and, according to formula (2), the accuracy of SRT estimation decreases.

Two main approaches to obtaining relatively large S_{50list} values have been proposed in the literature. In the first approach, some portion of speech material is recorded and intelligibility functions for all elements are determined (for example, for the whole sentences or digit triplets as well as for individual words or digits, respectively). Subsequently, S_{50list} is maximized by means of equalization of respective speech elements with respect to intelligibility, i.e. σ_{SRT} is minimized (Kollmeier and Wesselkamp, 1997; Wagener *et al.*, 1999a,c,b; Wagener *et al.*, 2005b).

The second approach appears to be less complicated, and it was employed for the purpose of this study. In this case, a relatively large portion of speech material is recorded, and corresponding intelligibility functions are determined. Subsequently, 'optimal' speech elements, i.e. of relatively large S_{50} and close SRT values are selected, while elements with small S_{50} and SRT that deviate considerably from the remaining ones are rejected (Versfeld *et al.*, 2000; Ozimek *et al.*, 2006). This method also preserves the natural intonation and co-articulation.

To minimize the spread of the SRT values across digit triplets and to obtain a relatively large slope of the list-specific intelligibility functions, the following conditions of triplet selection were set:

- triplet-specific SRT values should fall into the range of ± 1.5 dB with respect to the average SRT obtained for 160 triplets (i.e. minimization of σ_{SRT} across digit triplets),
- triplet-specific S_{50} values should be at least 13 %/dB (i.e. maximization of S_{50mean}).

In this way the number of triplets fulfilling these conditions was reduced from 160 to 103. Finally, it was decided to end up this selection with 100 triplets. Tab. I. presents statistics of the 160 'original' triplets and the group of 100 selected test items.

INSERT TAB. I

As can be seen, the mean SRT of the selected 100 digit triplets was –9.4 dB, i.e. remained the same like for the pre-selected 160 triplets. However, the rejection of triplets that considerably deviated from 'average intelligibility' resulted in a reduction of σ_{SRT} from 3.5 dB to 0.8 dB. Furthermore, due to the fact that the selection of triplets with respect to SRT was additionally followed by the selection of test items of relatively large S_{50}, mean S_{50} increased from 15%/dB to 21.4 %/dB. As far as the probabilistic model is concerned, the numerator of formula (3) was therefore increased, whereas the denominator was decreased. Hence, the S_{50} for group selected triplets was 19.4 %/ dB (please note that $\sigma_{SRT}>0$, thus $S_{50lgroup}$ does not correspond strictly to S_{50mean}), instead of 9%/dB for all the 160 triplets.

The next step of test development was just to split the group of 100 best test items into statistically and phonemically equivalent lists.

3. 2. Composition of statistically and phonemically equivalent lists

On the basis of the 100 triplets selected, 4 triplet lists containing 25 different triplets were composed. A dedicated algorithm was used to generate statistically equivalent (i.e.

yielding similar SRT and S_{50list}) and phonemically equivalent (i.e. yielding comparable phoneme distributions) lists. The algorithm performed the following operations:

- generation of a random permutation of triplet indexes and formation of 4 preliminary lists, each comprising 25 triplets,
- analysis of the mean SRT and S_{50} characterizing each list,
- analysis of phonemic content of the lists.

These operations were repeated until 4 lists, composed of different digit triplets, met the following conditions:

- the mean SRT and S_{50mean} characterizing each list fell into the range of ±0.1 dB and ±1 %/dB, respectively, of the mean SRT and mean S_{50} for the 100 selected triplets, i.e. –9.4 dB and 21.4 %/dB, respectively (i.e. the lists are composed of different triplets, but yielding similar intelligibility),
- for each list, a distribution of each phoneme fell into the range of ±1.5 percentage points with respect to the reference phoneme distribution characterizing phonemic content of all the recorded triplets (i.e. the lists are composed of different triplets, but revealing comparable phoneme distributions).

The lists meeting these conditions revealed very close phonemic content and produced similar list-specific intelligibility functions. Fig. 2 shows triplet-specific intelligibility functions fitted to experimental data for 25 digit triplets constituting list no. 1.

INSERT FIG. 2

Fig. 3 presents the triplet-specific (thin lines) and the list-specific intelligibility functions (solid lines) for the digit triplet lists composed, while Fig. 4 shows a comparison of only the list-specific intelligibility functions. The slope of the list-specific intelligibility functions was determined according to equation (3).

INSERT FIG. 3

INSERT FIG. 4

As can be seen, the lists produce very similar SRT values. Standard deviation across SRT values in each list does not exceed 1 dB. The statistical equivalence of the respective lists was confirmed by the results of ANOVA test applied to the baseline data[5]. The results of the test proved that SRT did not depend statistically on the list index {F(3,99)=0.19, p=0.9}.

4. RETEST MEASUREMENTS

In order to verify the statistical balance of the Polish digit triplet lists, three retest experiments have been carried out. In the first one, individual intelligibility functions were derived for each list using a standard constant stimuli paradigm, i.e. intelligibility scores were measured, and SRT and S_{50} parameters were determined by means of fitting CDF to the intelligibility data. In the second one, statistical parameters characterizing the respective lists were analyzed by application of the adaptive staircase procedure with the 1-up/1-down decision rule, and SRTs of the lists were compared. The first and second retest measurements were performed using the same apparatus as in the main investigations (see p. 2.2.1). During these measurements the subject was seated in an acoustically-insulated booth.

Since in the Internet hearing screening scenario, measurements were carried out via a commonplace equipment, i.e. PC computers equipped with a standard sound card applying the adaptive *up/down* method, the third retest experiment was conducted using a standard PC notebook computer. In these measurements, the subject was seated in a room (office) characterized by normal everyday acoustical conditions. In the first and the second experiments, the sound pressure level was kept constant at 70 dB SPL, while in the third one, the subject was allowed to adjust sound loudness to comfort level prior to the measurement.

[5] Statistical properties of the test have been confirmed in a further retest study carried out for another group of subjects (chapter 4).

The main purpose of these three verification experiments was to find out whether data obtained in these conditions are comparable. During the measurements, the signals were presented monaurally via the Sennheiser HD 580 headphones. For each of these three experiments, a new group of twenty normal-hearing, young subjects was recruited. The absolute threshold of all the subjects at audiometric frequencies from 250 to 8000 Hz was less than 15 dB HL. The subjects were paid for their participation in the experiments.

4. 1. Retest experiment 1. Constant stimuli paradigm

In the first retest experiment, the triplets for respective lists were presented at the following SNR values: –13.0; –11.5; –10.0; – 8.5 and –7.0 dB. These SNRs have been chosen to encompass optimally the expected SRT (i.e. about –9.4 dB). During the experiments, a randomly chosen list was presented at a randomly chosen SNR value. The order of the triplets presented was also randomized. Each list was presented to the subject at each SNR. Fig. 5 presents an example of intelligibility data and a fitted individual intelligibility function obtained in this experiment.

INSERT FIG. 5

The SRTs obtained for the respective lists and subjects were very close, i.e. the results of the measurements confirmed the expected reliability of the developed lists. The mean data is presented in Tab. II (data in row R1).

Apart from test reliability, the statistical balance across the lists was re-analyzed according to the retest data. Statistical analysis was performed to examine equivalence of the four lists constituting the Polish digit triplet test. The intelligibility scores obtained for the respective lists and SNRs were analyzed by means of a two-way analysis of variance, ANOVA. It turned out that the 'list' factor was statistically insignificant {$F(3,399)=1.35$, $p=0.25$}, while SNR was, as expected, highly statistically significant {$F(4,399)=884.94$,

p<<0.001}. Therefore, the ANOVA results revealed that triplet intelligibility score was determined entirely by SNR value and did not depend on the list index used in the measurements. Thus, the lists may be considered as fully equivalent, yielding similar intelligibility data.

The statistical equivalence of respective lists was also tested by means of analysis of SRT and S_{50} values of intelligibility functions fitted to the individual intelligibility scores. The obtained individual SRT and S_{50} were pooled across subjects and two separate one-way ANOVAs were carried out, which revealed that neither SRT {F(3,79)=0.79, p=0.50} nor S_{50} {F(3,79)=1.32, p=0.27} depended on the list index. The mean SRT and S_{50} for present measurements are –9.5 dB and 20.4 %/dB, respectively, and are close to the expected values. The minimal and maximal SRT across the lists was -9.7 dB and -9.4 dB, respectively. The minimal and maximal SRT across subjects was -10.0 dB and -9.0 dB, respectively.

4. 2. Adaptive measurements

At this stage of retest measurements, an adaptive procedure with 1-up/1-down decision rule was employed to determine speech intelligibility. The adaptive procedure with this decision rule converges to the 50% correct equilibrium point of an intelligibility function, i.e. SRT value. In this experiment, the SNR was changed adaptively. The masker level was kept at 70 dB SPL, so SNR was determined by the speech signal level. The SNR was either decreased or increased by a certain value (the so-called step) depending on whether the most recent subject response was correct (triplet scoring) or incorrect. The initial value of SNR was –3 dB, which was far above the expected SRT, i.e. speech was perfectly understood at the beginning of the measurement. In order to make the convergence procedure to SRT point as quick as possible, the initial step was set to 2 dB and was reduced to 1 dB after the third incorrect response. During measurements, 25 triplets were presented, while SRT was calculated as a mean of the last 15 SNRs at which the signals were presented, including the

so-called 'virtual SNR', i.e. SNR determined on the basis of the last-presented digit triplet (Smits *et al.*, 2004). The order of triplets within the list as well as the list index were randomized.

4. 2. 1. Retest experiment 2. Adaptive procedure with laboratory equipment

At this stage of the retest study, the same apparatus as in the development of the lists was used in measurements (p. 2.2.1). Like in the case of retest measurements with the constant stimuli method, the SRTs obtained for the adaptive procedure were very close to the expected values. The obtained SRTs were pooled across listeners and subjected to one-way ANOVA with respect to the 'list' factor, which revealed that the differences in SRT across the lists were statistically insignificant {$F(3,79)=0.77$, $p=0.51$}. The difference between the maximal and minimal SRT across lists was only 0.3 dB, while minimal and maximal SRTs across subjects were -10.5 dB and -8.7 dB. The determined SRTs (averaged across subjects) are presented in Tab. II. (row R2).

4. 2. 2. Retest experiment 3. Adaptive procedure with PC and standard sound card

Since the digit triplet tests will be mainly used via the Internet, the retest measurement was conducted with a notebook equipped with a standard (nonprofessional) sound card (Realtek HD). The signals were presented via the Sennheiser HD 580 headphones. To make the measurement environment more natural, the experiments took place in a standard room (office). Prior to the measurement session, the listeners were allowed to adjust loudness of the sound for their comfort.

The gathered SRTs were pooled across subjects and analyzed by one-way ANOVA. Again, it turned out that the 'list' factor was not statistically significant {$F(3,79)=0.71$, $p=0.55$}. The minimal and maximal SRTs for lists were -9.7 dB and -9.3 dB, respectively,

whereas corresponding values across subjects were -10.2 dB and -8.5 dB, respectively. The determined SRTs (averaged across subjects) are presented in Tab. II (row R3).

This indicates that performing intelligibility measurement in an ordinary room (background level below 55 dB(A)) by using a PC equipped with a standard sound card yields similar SRTs to those obtained in laboratory conditions. The results of the retest measurements, with commonplace PC equipment, are in line with the basic test values. The statistical balance of the developed triplet materials were also confirmed in this experimental scenario.

4. 3. Retest experiments. Data comparison and derivation of normative data.

Tab. II presents mean data obtained during test development and gathered in the retest experiments: constant stimuli (R1), adaptive procedure in laboratory conditions (R2) and using PC with a nonprofessional sound card (R3).

INSERT TAB. II

In order to examine reliability of the speech intelligibility assessments by means of comparison of data gathered during test development and evaluation experiments, the SRTs were subjected to a two-way ANOVA with respect to the 'list factor' and the 'experiment' factor. The ANOVA showed that both factors were statistically insignificant, i.e. SRT did not depend on digit triplet list used {$F(3,339)=0.52$, $p=0.66$} and there were no differences across the respective measurements {$F(3,339)=0.99$, $p=0.39$} (including test development and evaluation experiments). Thus, the retest measurements confirmed reliability of the speech assessments carried out by means of the Polish digit triplet test.

Since it had been shown that there were no statistical differences both across lists developed and experimental methods, all SRTs were pooled across lists and experiments. Subsequently, mean SRT was calculated and turned out to be -9.4 dB. The upper limit of the 95%-confidence interval was -8.0 dB, which means that if the test is performed, the SRT above

-8.0 dB is significantly worse than the average.

5. COMPARISON OF INTELLIGIBILITY FUNCTION FOR DIGIT TRIPLETS ACROSS LANGUAGES

Fig. 6 presents a comparison of the mean list-specific intelligibility functions for the Polish digit triplet test (solid lines) with respect to the mean list-specific intelligibility functions for digit triplet tests developed for several European languages. It should be emphasized that due to certain differences in linguistic structure and measurement methods used in different laboratories, only a general comparison is possible. All the intelligibility functions were obtained in measurements via headphones.

INSERT FIG. 6

As can be seen, SRT varies across languages from –12.0 dB (English test) to –9.4 dB (German test); the smallest slope is 16.0 %/dB (the Dutch test), while the largest slope is observed for the Polish materials. Functions characterizing the German and Polish tests are almost identical.

It is worth noting that mean SRT obtained for the Polish digit triplets, i.e. –9.4 dB, is considerably lower than that for the Polish sentence test, where mean SRT= –6.1 dB (Ozimek *et al.*, 2006). In other words, in masking conditions the listeners' performance turned out to be much better when they were presented with digit triplets rather than meaningful sentences. This is because the triplet material itself is actually composed of 10 words, which markedly improves speech intelligibility (Miller *et al.*, 1951; Smits *et al.*, 2004). The difference between SRTs obtained for digit triplets and sentence material is in agreement with the results of measurements for the Dutch language for which SRTs obtained for digit triplet test (Smits *et al.*, 2004) and sentence test (Plomp and Mimpen, 1979) are –11.2 dB and –5.5 dB, respectively. The difference is also noticeable if the SRT obtained for the German digit triplet

test (Wagener *et al.*, 2005a) and the German Göttingen sentence test (Kollmeier and Wesselkamp, 1997) are compared. The SRT values amount to –9.3 dB and –6.2 dB, respectively.

The largest slope of the intelligibility functions can be observed for the Polish and German digit triplet tests (Wagener *et al.*, 2005a), whereas the smallest slope is observed for the Dutch digit triplet test (Smits *et al.*, 2004). This difference might result from the fact that in the Dutch test, the so-called speech-shaped noise (white noise passed through an appropriate filter) was used as a masking signal; it generally produces slightly less steep intelligibility functions than noise generated by superposing speech material (Ozimek *et al.*, 2006). Moreover, the slope of intelligibility functions for the Dutch material might be biased[6] since in this case the intelligibility function was derived from adaptive data (Kaernbach, 2001). And finally, during the measurement using the Dutch test, subjects had no possibility of correcting their mistakes if they accidentally pressed an inappropriate key, and consequently the upper asymptote of the intelligibility function did not reach 100 %, which might have decreased the intelligibility function slope (Klein, 2001). Further factors affecting the inconsistency between parameters of the intelligibility functions across languages might result from differences in their linguistic structure and differences across speakers.

6. CONCLUSIONS

The main conclusions resulting from this study are as follows:

- For Polish digit triplets, the average SRT obtained in all the experiments is -9.4 dB. Apart from some differences in test material and development, this value is close to SRT for German and higher than SRTs obtained for the Dutch and English digit triplet tests.

[6] Both SRT and S_{50} can be biased when using adaptive procedure. SRT might be biased when it is computed by averaging SNR and when initial SNR considerably deviates from SRT. On the other hand, when fitting intelligibility function for adaptive data, there is no bias in SRT (Versfeld *et al.*, 2000), but slope of intelligibility function might be biased (Kaernbach, 2001).

- Apart from some differences in the development of tests, the mean list-specific slope of intelligibility function for the Polish digit triplet test is close to that characterizing the German test and slightly larger than the slope of intelligibility functions for other languages.

- The retest measurements carried out for other groups of normal-hearing subjects confirmed reliability of speech assessments carried out with the Polish digit triplet test.

Acknowledgements

This work was supported by the grant from the European Union FP6, Project 004171 HEARCOM and the State Ministry of Education and Science.

Tables

Tab. I. *Parameters of intelligibility function for the pre-selected group of 160 triplets and the selected group of 100 'best' triplets*.

	mean SRT [dB]	σ_{SRT} [dB]	mean S_{50} [%/dB]	$S_{50group}$ [%/dB]
before selection (160 triplets)	-9.4	3.5	15.0	9.0
after selection (100 triplets)	-9.4	0.8	21.4	19.4

Tab. II *Comparison of the mean SRT obtained during test development (i.e. expected values) and in the respective retest measurements R1 (constant stimuli), R2 (adaptive measurement – laboratory equipment) and R3 (adaptive measurement – nonprofessional sound card). All values present means across subjects and are expressed in dB.*

	list			
	1	2	3	4
expected	-9.2	-9.4	-9.5	-9.4
R1	-9.4	-9.5	-9.7	-9.6
R2	-9.7	-9.7	-9.4	-9.5
R3	-9.4	-9.2	-9.3	-9.3
mean	-9.5	-9.4	-9.5	-9.4

References

Bellis, T. J., 1996. Assessment and management of central auditory processing disorders in the educational setting. San Diego, Singular Publishing.

Brachmański, S., Staroniewicz, P., 1999. Fonetyczna struktura materiału testowego stosowanego w subiektywnych pomiarach jakości mowy. Speech and Language Technology. Poznań. **3,** 71-80.

Broadbent, D. E., 1954. The role of auditory localization in attention and memory span. Journal of Experimental Psychology **47**, 191-196.

Elberling, C., Ludvigsen, C., Lyregaard, P. E., 1989. DANTALE: a new Danish speech material. Scandinavian Audiology **18**(3), 196-75.

Fletcher, H., 1929. Speech and hearing. New York, Van Nostrand.

Grocholewski, S., 2001. Statystyczne podstawy systemu ARM dla języka polskiego. Wydawnictwo Politechniki Poznańskiej.

Hall, S.J., 2006. The development of a new English sentence in noise test and an English number recognition test. M.Sc. Thesis, Faculty of Engineering, Science and Mathematics, University of Southampton.

Kaernbach, C., 2001. Slope bias of psychometric functions derived from adaptive data. Perception & Psychophysics **63**(8), 1389-1398.

Kalikow, D. N., Stevens, K. N., Elliot, L. L., 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. Journal of Acoustical Society of America **61**(5), 1337-1351.

Klein, S. A., 2001. Measuring, estimating and understanding of psychometric function: A commentary. Perception and Psychophysics **63**(8), 1421-1455.

Kollmeier, B., 1990. Messmetodik, Modellierung und Verbeserung der Verstandlichkeir von Sprache. Gottingen, Georg-August-Universtat.

Kollmeier, B., Wesselkamp, M., 1997. Development and evaluation of a sentence test for objective and subjective speech intelligibility assessment. Journal of Acoustical Society of America **102**(4), 1085-1099

Miller, G. A., Heise, G. A., Lichten, W., 1951. The intelligibility of speech as a function of the context of the test materials. Journal of Experimental Psychology **41**, 329-335.

Nilsson, M., Soli, S. D., Sullivan, J. A., 1994. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. Journal of the Acoustical Society of America **95**, 1085-1099.

Ozimek, E., Kutzner, D., Sęk, A., Wicher, A., Szczepaniak, O., 2006. The Polish sentence test for speech intelligibility evaluations. Archives of Acoustics **31**(4), 431-438.

Plomp, R., Mimpen, A. M., 1979. Improving the reliability of testing the speech reception threshold for sentences. Audiology **18**, 43-53.

Pruszewicz, A., Demenko, G., Richter, L., Wika,T., 1994a. New articulation lists for speech audiometry. Part I. Otolaryngologia Polska **48**, 50-55.

Pruszewicz, A., Demenko, G., Richter, L., Wika, T., 1994b. New articulation lists for speech audiometry. Part II. Otolaryngologia Polska **48**, 56-62.

Ramkissoon, I., Proctor, A., Lansing, C. R., Bilger, R.C., 2002. Digit speech recognition

thresholds (SRT) for non-native speakers of English. American Journal of Audiology. **11,** 23-28.

Rudmin, F., 1987. Speech perception thresholds for digits. J. Audiol. Res. **27**, 15-21.

Schmidt-Nielsen, A., 1989. The intelligibility of native and non-native speakers of American English using spelling alphabet test materials. The Journal of the Acoustical Society of America **86**(S1), S76-S77.

Smits, C., Houtgast, T., 2007. Recognition of digits in different types of noise by normal and hearing-impaired listeners. International Journal of Audiology **46**, 134-144.

Smits, C., Kapteyn, T., Houtgast, T., 2004. Development and validation of an automatic speech-in-noise screening test by telephone. International Journal of Audiology **43**, 15-28.

Smits, C., Kramer, S. E., Houtgast, T., 2006. Speech Reception Thresholds in Noise and Self-Reported Hearing Disability in a General Adult Population. Ear and Hearing **27**(5), 538-549.

Strouse, A., Wilson, R. H., 2000. The effect of filtering and inter-digit interval on the recognition of dichotic digits. Journal of Rehabilitation Research and Development **37**(5), 599-606

Versfeld, N. J., Daalder, L., Festen, J. M., Houtgast, T., 2000. Method for the selection of sentence material for efficient measurement of the speech reception threshold. Journal of Acoustical Society of America **107**, 1671-1684.

Wagener, K., Brandt, T., Kollmeier, B., 1999a. Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test (in German). Z. Audiol. **38**, 4-15.

Wagener, K., Brandt, T., Kollmeier, B., 1999b. Development and evaluation of a German

sentence test II: Optimisation of the Oldenburg sentence tests (in German). Z. Audiol **38**, 44-56.

Wagener, K., Brandt, T., Kollmeier, B., 1999c. Development and evaluation of a German sentence test III: Evaluation of the Oldenburg sentence test (in German). Z. Audiol. **38**, 86-95.

Wagener, K., Eeenboom, F., Brand, T., Kollmeier, B. 2005a. Ziffer-Tripel-Test: Spracherverstandlichkeitstest uber das Telefon. Tagungs-CD der DGA Jahrestagung.

Wagener, K., Josvassen, J. L., Ardenkjaer R., 2005b. Design, Optimization, and Evaluation of a Danish Sentence Test in Noise. International Journal of Audiology **42**(1), 10-17.

Wilson, R. H., Burks, A. B., Weakley, G. W., 2005. A comparison of word-recognition abilities assessed with digit pairs and digit triplets in multi-talker babble. Journal of Rehabilitation Research and Development **42**(4), 499-510.

Wilson, R. H., Weakley D. G., 2004. The use of digit triplets to evaluate word-recognition abilities in multi-talker babble. Sem. Hear. **25**, 93-111.

Figure captions

Fig. 1. *Power spectrum density of the digit-shaped noise used in the investigation (reference – total rms value).*

Fig. 2. *Intelligibility scores (symbols) and fitted triplet-specific intelligibility functions (lines) for 25 digit triplets constituting list no. 1. For the sake of visibility, each panel presents data for succeeding groups of 5 digit triplets. Panel in the right bottom corner depicts a comparison of all fitted triplet-specific intelligibility functions; the solid line presents the list-specific intelligibility function for list no.1.*

Fig. 3. *Triplet-specific intelligibility functions for the respective digit triplets (thin lines) and the list-specific intelligibility functions (solid lines) of the lists. The intelligibility data for each SNR is not shown for the sake of figure clarity.*

Fig. 4. *The list-specific intelligibility functions for four respective digit triplet lists.*

Fig. 5. *Results of retest measurements using the constant stimuli method: intelligibility scores (circles) and fitted intelligibility function (solid line).*

Fig. 6. *Comparison of the mean list-specific intelligibility functions for digit triplet tests across languages.*

Fig.1

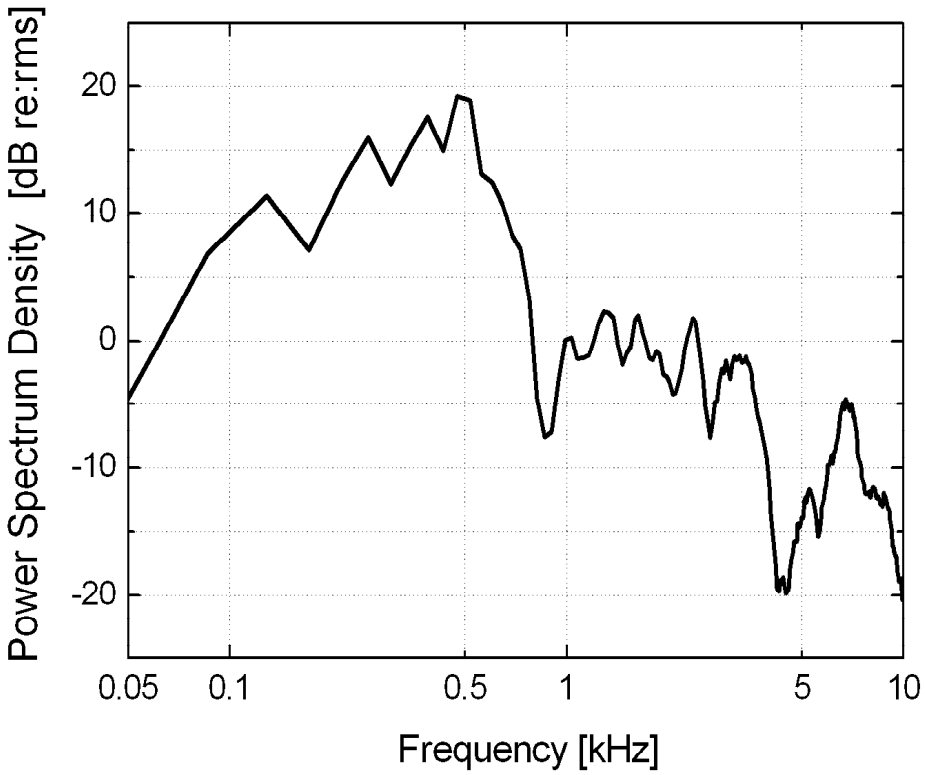


Fig.2

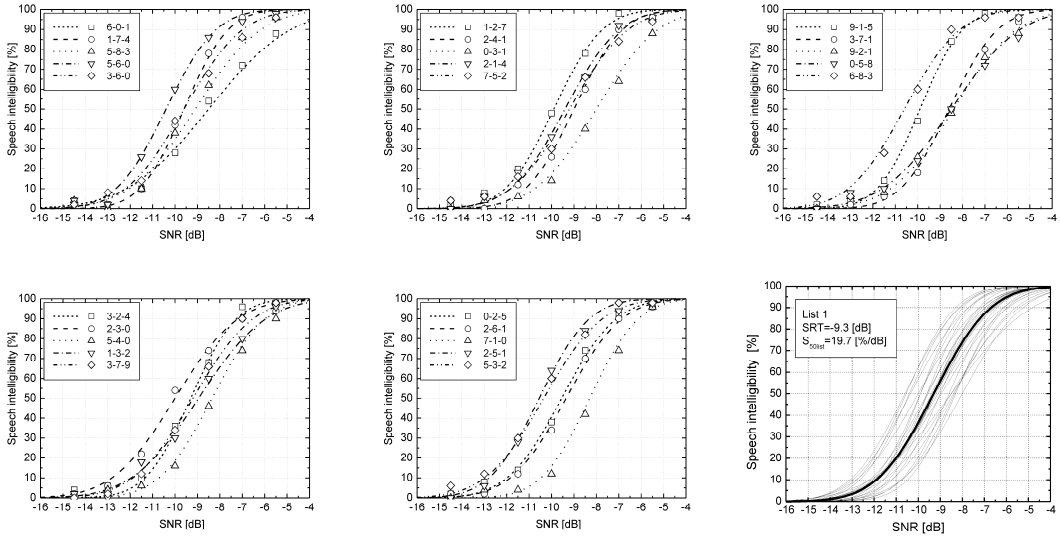


Fig.3

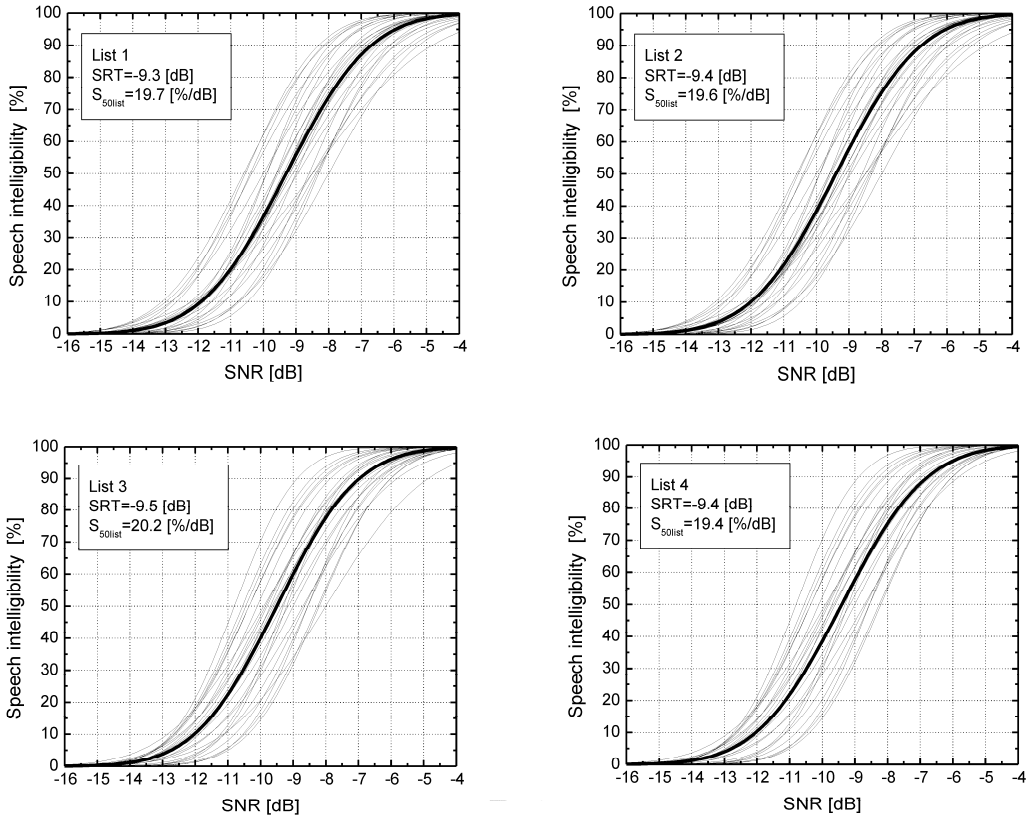


Fig.4

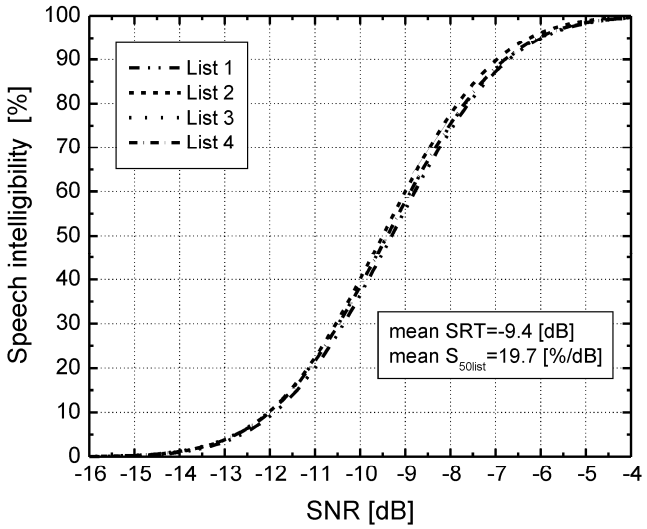


Fig.5

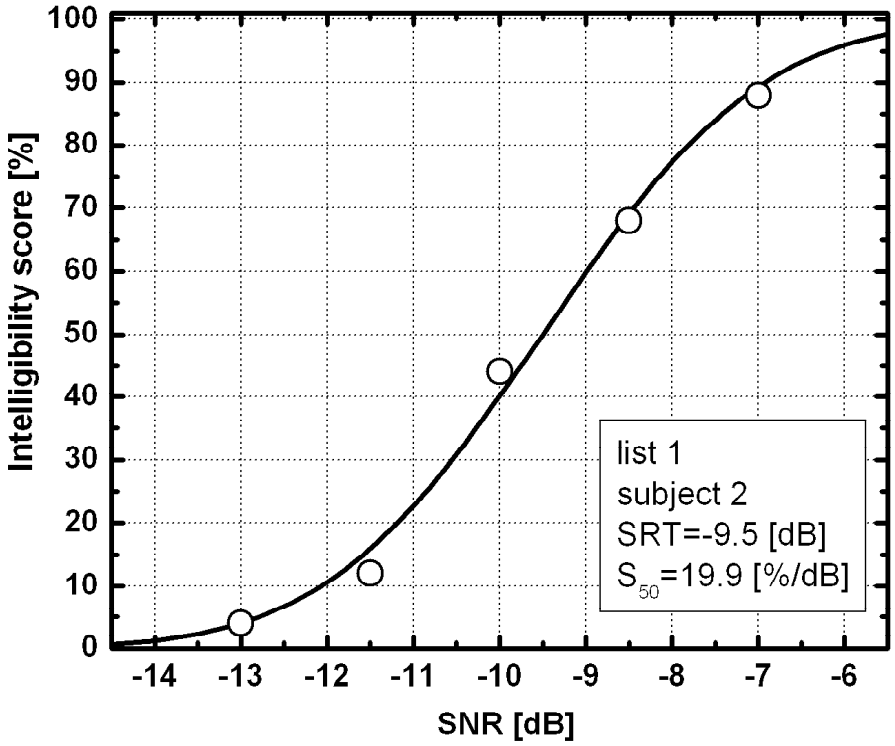


Fig.6