Speech intelligibility determined with various tests presented against noise

Edward Ozimek, Dariusz Kutzner and Anna Warzybok

Umultowska 85, 61-614 Poznan, Poland
ozimaku@amu.edu.pl

The paper describes three new Polish tests for assessment of speech intelligibility against noise. These are: Polish Sentence Test (PST), Polish Matrix Test (PMT) and Polish Digit Triplets Test (PDTT). The PST comprises sentences taken from everyday life and aims at clinical speech intelligibility measurements. The PMT contains semantically unpredictable sentences of fixed syntactical structure. This test is useful for speech intelligibility measurements focusing on a large number of experimental conditions, and is applicable to subjects with profound hearing impairment or cochlear implants users. The PDTT contains sequences of three digits that are spoken separately, and is aimed at screening measurements, especially via telephone. The statistical and phonemic properties of each test are compared and discussed in the paper. It was found that for normal-hearing subjects, the speech reception threshold (SRT, i.e. signal to noise ratio yielding 50% speech intelligibility) and steepness of the psychometric function (S_{50list}) for the PST were equal to -6.1 dB and 25.5 %/ dB, respectively. For the PMT, SRT= -8.9 dB and S_{50list}= 14.4 %/dB and for the PDTT, SRT and S_{50list} were -9.4 dB and 19.6 %/dB, respectively. The tests have been proved to provide accurate and repeatable speech intelligibility data.

1 Introduction

Many tests have been developed to determine speech intelligibility [1-15]. The speech intelligibility assessment can focus on listener's performance, when speech is presented in quiet or in noisy conditions. However, since in natural environments speech is perceived against a background of other signals, in most cases speech-in-noise measurements are of great importance. This experimental paradigm is important for clinical purposes since it is well-known that hearing loss affects speech intelligibility in acoustically adverse conditions. Furthermore, speech-in-noise tests have been proved to be very useful for fitting hearing devices, room acoustics, auditory screening and telecommunication.

This paper focuses on three new speech intelligibility tests that have been recently developed for the Polish language: the Polish Sentence Test (PST), the Polish Matrix Test (PMT) and the Polish Digit Triplets Test (PDTT). The tests have been developed within a framework of the European HearCom project (FP6). The HearCom aims at providing tools and solutions for improving communication for hard of hearing people. The Polish tests described here were created on the basis of the previously developed European speech intelligibility tests.

Despite the fact that each test is aimed at a specific measurement scenario (clinical measurements or auditory screening), they have all been optimized for the speech-in-noise paradigm. Moreover, in each test a sequence of words (either a sentence or a digit complex) is presented as a test item. PST is based on everyday sentences of unfixed syntactical structure, and is aimed at clinical intelligibility measurements. PDTT contains strongly limited lexical content, i.e. sequences of three digits, and it has been developed for auditory screening, especially via telephone or the Internet. PMT can be regarded as some kind of 'intermediate step' between PST and PDTT. This is due to the fact that this test is, on the one hand, based on limited lexical content, but on the other hand it comprises sentences of phoneme content close to the reference phoneme distribution for the Polish language. The respective sections below outline the process of test development, verification measurements and data comparison for the three Polish tests.

In each test, during a listening session, speech was presented at a background of masking noise at different signal to noise ratios (SNR), and SRT (defined as SNR that yields 50% intelligibility) was determined. In general case, SRT was obtained from psychometric (intelligibility) function, i.e. function that links a probability of correct response to SNR. The tests have been optimized to provide relatively steep intelligibility functions since the greater the steepness of psychometric function, the more accurate the SRT estimate. Both standard constant stimuli paradigm and adaptive procedure were used for SRT measurements.

2 Sentence tests

Two sentence tests have been adopted for Polish. The first one, i.e. PST, uses everyday utterances of unfixed syntactical structure. The second one, i.e. PMT, is based on semantically unpredictable sentences of fixed syntax structure (PMT).

2.1 Polish Sentence Test

This test is the Polish version of the sentence test of structure proposed by Plomp and Mimpen for the Dutch language [1]. Like in Dutch [1], German [2] and English [3] tests, PST was based on everyday sentences of unfixed syntax structure. PST was composed of 37 statistically and phonemically equivalent lists, each contaning 13 sentences.

2.1.1 Signal preparation

The procedure of test development consisted of several stages. First, 1,200 sentences (in a written form) were selected from a large database containing several million of Polish utterances taken from newspapers, TV, literature etc. The chosen group of 1,200 sentences met the criteria used by Plomp and Mimpen [1] and Versfeld [4]: sentences contained 8 or 9 syllables; words contained up to 3 syllables, contained no proper names and professionally-issued expressions. Finally, words of offending nature or revealing non-PC content were excluded (an example of selected sentence: *Teraz nie mają na to czasu* 'They have no time now', *Listonosz wszedł do werandy* 'A postman stepped into a veranda').

The sentences were recorded in a professional studio. They were read out by a male radio speaker at natural intonation and tempo. The recording was performed using the Neumann U87 capacitor microphone. The microphone output fed one of the input channels of the Yamaha 02R mixer. In the mixer, the microphone 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. The signals were then sent via an optical connection (ADAT-type) to a PC and stored on a computer hard disc.

The first step in achieving high accuracy of speech-in-noise measurement was to use interfering noise that matched power spectra of masked utterances [11, 16]. In order to generate such a masker, multiple superposition of all recorded sentences was performed, resulting in the so-called babble noise that was characterized by the average power spectrum of the sentences presented.

2.1.2 Measurements of sentence-specific psychometric functions

The recorded sentences were presented to 35 normal-hearing subjects at 5 different SNRs ranging from -9 dB to -1 dB. The sound pressure level of masking noise was kept constant at 70 dB SPL, thus SNR was determined by the speech level. In order to avoid the learning effect, each sentence was presented to a given listener only once.

The signals were generated by means of the TDT 3 equipment (the 24-bit real time signal processor RP2, the headphone buffer HB7) and presented monaurally via the Sennheiser HD 580 headphones. During the experiment, subjects were seated in an acoustically-insulated booth. The subjects' task was to repeat what they have heard. Their responses were recorded on a computer hard disc as .wav files.

Subsequently, all responses were scored on the basis of sentence intelligibility [4], i.e. a subject's response was considered correct when all the words constituting the just-heard sentence were repeated correctly (scored 1), otherwise it was scored 0. The total number of gathered and analyzed data was 42,000. The proportion of correct responses at each SNR and for each sentence was calculated. Finally, psychometric functions (modeled by cumulative standard distributions, CDF) were fitted to the data using maximum-likelihood (ML) method and sentence-specific SRT and S_{50} were obtained.

2.1.3 Selection of speech material and composition of final lists

It is well-known that high reliability and accuracy of speech intelligibility measurement requires test items yielding comparable intelligibility data and steep psychometric functions [2, 4, 17]. Thus, only sentences of SRT falling into the range ±1.5 dB with respect to the mean SRT and of S_{50} larger than 15 %/ dB were included into final lists. These criteria resulted in a selection of 500 optimal test items (700 sentences were rejected). The selected optimal sentences were split into 37 statistically and phonemically equivalent lists, each of 13 sentences. The lists were composed by means of Monte Carlo simulations that found list configuration meeting the following critera: list-specific SRT did not deviate from the mean SRT by more than 0.1 dB for all the lists, whereas deviation of phoneme distribution for each list did not exceed 2.5 percent point with respect to the reference phoneme distribution of the Polish language. Finally, the statistical equivalence of the lists was analyzed by means of subjecting list specific SRT and S_{50} to two one-way ANOVAs. It turned out that neither SRT {$F(36,480)=0.4$, $p=0.99$} nor S_{50} {$F(36,480)=0.34$, $p=0.99$} depended on list index. The retest measurements were carried out to confirm statistical equivalence of the lists. The SRT for each list was determined be means of constant stimulus paradigm as well as by the adaptive procedure with 1-up/1-down decision rule converging to 50%-intelligibility. Re-measured SRT was –6.2 dB, standard deviation across lists was 0.3 dB.

2.2 Polish Matrix Test

Although Plomp-type sentence tests are very important in clinical practice, they can be insufficient for measurements in which a number of SRTs to be determined is greater than a number of lists available. Due to the learning effect, the presentation of the same sentence list is not allowed since the measurement results might be influenced by a listener's memory. This prompted us to develop a matrix sentence test that enables SRT measurements in experiments in which speech intelligibility is analyzed in a variety of conditions. The Polish Matrix Test is based on the German OLSA [12-14] or Danish DANTALE II [11] tests. Unlike Plomp-type test, in the case of matrix test, the sentence to be presented is generated by a random permutation of adequately selected and level-corrected words.

2.2.1 Structure of speech material

The test is based on a 50-word matrix containing words which are most frequently used in Polish. What is more, the phoneme content of words was analyzed during their selection in order to provide phoneme distribution as close as possible to the reference phoneme distribution for Polish. The sentences of fixed syntactical structure (name, verb, numeral, adjective, object) were generated by a permutation of randomly chosen words from the respective columns of Table 1.

Name	Verb	Numeral	Adjective	Object
Tomasz (Thomas)	*nosi (carries)*	*pięć (five)*	*dobrych (good)*	*piłek (balls)*
Paweł (Paul)	***woli (preferes)***	*sześć (six)*	*tanich (cheap)*	*gazet (papers)*
Adam (Adam)	*widzi (sees)*	*siedem (seven)*	*drogich (expensive)*	*soków (juices)*
Maciej (Mathias)	*bierze (takes)*	***osiem (eight)***	*pięknych (beautiful)*	*dzwonów (bells)*
Michał (Michael)	*daje (gives)*	*dziewięć (nine)*	*nowych (new)*	***opon (tyres)***
Anna (Anne)	*ma (has)*	*dużo (a lot of)*	*starych (old)*	*stołów (tables)*
Ewa (Eva)	*robi (makes)*	*sto (hunderd)*	***białych (white)***	*klocków (bricks)*
Maria (Maria)	*kupi (will buy)*	*tysiąc (thousand)*	*żółtych (yellow)*	*toreb (bags)*
Zofia (Sophie)	*wygra (will win)*	*wiele (many)*	*czarnych (black)*	*okien (windows)*
Julia (Julia)	*sprzeda (will sell)*	*kilka (several)*	*dziwnych (strange)*	*koszy (boxes)*

Table 1. 50-word matrix for PMT. Bold font denotes randomly selected words corresponding to the sentence *Thomas prefers eight white tyres*

Since each column contains 10 words, this approach yields 10^5 grammatically correct, yet semantically unpredictable sentences.

2.2.2 Stimuli

On the basis of the matrix, 100 sentences were generated in a way described in [11, 12]. In order to preserve a natural intonation of the synthesized sentences, each word was recorded in coarticulation with the following words (i.e. 10 times). Thus 500 words were recorded, cut and labelled unambiguously. The signals were recorded by the AKG C 1000 S microphone, amplified in the RME QuadMic 4-Channel microphone amplifier and A/D-converted by RME Hammerfall DSP Multiface. Finally, the signals were stored on a PC hard disc.

2.2.3 Measurement of word-specific psychometric functions

According to the probabilistic model proposed by Kollmeier [18], the high steepness of test-specific intelligibility function demands equalization of intelligibility of the respective test items. The measurements aimed at determination of word-specific psychometric functions for 500 recorded words. 100 sentences were synthesized in such a way that each of the 500 words occurred once. The sentences were presented in masking noise. The masker level was 65 dB SPL. Each sentence (thus each word) was presented to the subject at 11 SNRs ranging from -16.5 dB to -1.5 dB. The order of sentences and SNRs was randomized. 30 normal-hearing subjects participated in the investigations. The data obtained for each word and each SNR were polled across subjects and proportions of correct response were computed. Subsequently, the 500 psychometric functions were determined be means of fitting logistic functions to empirical data (ML criterion). The same experimental equipment as in the case of PST was used in the measurements. The total number of collected and analyzed responses was 165,000.

2.2.4 Optimization

Intelligibility of the respective words will be equalized in order to provide maximal homogenization of the speech material. The steepness of test-specific psychometric functions depends on a degree of the intelligibility equalization according to Eq. (1) [18]:

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

To equalize intelligibility of the test items, sound pressure level of the words that are ‘easy’ to understand (i.e. of relatively low SRTs) will be adequately decreased, while the level of ‘difficult’ words (i.e. of high SRTs) will be increased. The level corrections lead to a reduction in σ_{SRT} and, consequently, result in an increase of test-specific steepness (Eq.(1)). However, the level modification might affect naturalness (prosody) of synthesized utterances. Therefore, an optimal degree of level correction has to be determined. Preliminary data shows that S_{50mean} for the words is 20.9 %/dB (which corresponds to S_{50mean} obtained for OLSA test [14], i.e. 20.2 %/dB). The mean SRT=-8.9 dB with σ_{SRT}=2 dB, thus test-specific steepness determined according to Eq. (1) is 14.4 %/dB (the corresponding parameters for OLSA test are: SRT=-8.4 dB, σ_{SRT}=2.7 dB and steepness before level corrections S_{50test} is 11.3 %/dB). If one assumes that level correction reduces σ_{SRT} to 1 dB, the list specific steepness will increase to 18.5 %/dB (OLSA: 15.9 %/dB).

3 Polish Digit Triplets Test

PDTT is an auditory screening test based on limited vocabulary (ten digits). It has been developed mainly for self-examination of hearing, especially through the Internet or telephone.

The digit triplets test contains sequences of three digits spoken separately (e.g. 2-6-1 is pronounced as *two-six-one*, and not as two hundred sixty one). The PDTT has been composed of all monosyllabic as well as disyllabic digits, i.e. all digits from 0 to 9. From all possible 10^3 combinations, the triplets including repeated digits were excluded (i.e. 3-3-2, 7-5-7 or 5-5-5).

Subsequently, 160 triplets were selected in such a way that for each digit the probability of occurrence in any position in triplets was approximately equal. The 160 triplets were read out in a radio studio by a male Polish native speaker. The same recording equipment as for PST was used during recordings.

3.1 Measurements of triplet-specific psychometric functions

For each triplet, speech intelligibility was measured by means of the constant stimuli paradigm. Each of the 160 digit triplets was presented to a subject (50 normal-hearing subjects took part in the experiment) at 7 values of SNR: -14.5; -13.0; -11.5; -10.0; -8.5; -7.0 and -5.5 dB. The level of noise was kept constant at 70 dB SPL, so SNR value was determined by the triplet sound pressure level. The order of the triplets presentation and that of the SNRs were randomized. During the intelligibility measurements the subject was asked to type on a keyboard what he/she had just heard. The subject’s response was scored 1 if the entire triplet was repeated correctly, otherwise the response was scored 0 (the so-called triplet scoring). Each subject was presented with 1,120 triplets (7 SNRs*160 units). The data was polled across subjects and the proportion of correct responses was determined for each triplet and each SNR. The total number of gathered and analyzed responses was 56,000. Finally, the CDF functions were fitted to the intelligibility data using the ML criterion and the digit triplet-specific SRT and S_{50} were obtained.

It was decided that the ‘optimal’ digit triplets should meet the following criteria: SRT values should fall into the range of ± 1.5 dB with respect to the average SRT obtained for 160 triplets; S_{50} values should be at least 13 %/dB. As a result, 100 triplets fulfilling the above conditions were selected. Four statistically- and phonemically-equivalent triplet lists containing 25 different triplets were composed.

3.2 PDTT – retest measurements

Three retest experiments were carried out. In the first and the second experimental session, the list equivalence was analysed in a laboratory using constant stimuli paradigm or the staircase procedure with a 1-up/1-down decision rule. The last experiment focused on SRT measurements in non-laboratory conditions (office) using a notebook with a non-professional sound card. 20 listeners took part in each experiment. The obtained results revealed there were no differences across lists and experimental conditions.

4 Comparison across tests and discussion

Table 2 presents details of the Polish tests described in this paper. Each test 'produces' relatively steep intelligibility function. The tests differ in structure and are 'suited' for different applications.

Test	SRT [dB]	S_{50} [%/dB]	Scoring method	Test structure	Applica-tion
PST	-6.1	25.5	sentence	37 lists 13 sent. each	D/R
PMT	-8.9*	18.5*	word	50-word matrix	D/R
PDTT	-9.4	19.6	triplet	4 lists 25 triplets each	AS

Table 2 Details of new Polish speech intelligibility tests (* expected data, D/R – diagnostics/rehabilitation, AS – auditory screening)

4.1 Normative speech intelligibility functions

Figure 1 presents test-specific psychometric functions for the tests considered in this article: PDTT (solid line), PMT (before optimization - dashed line; after optimization - dashed dot line) and PST (dashed dot dot line). Each function represents mean intelligibility function obtained for normal-hearing subjects.

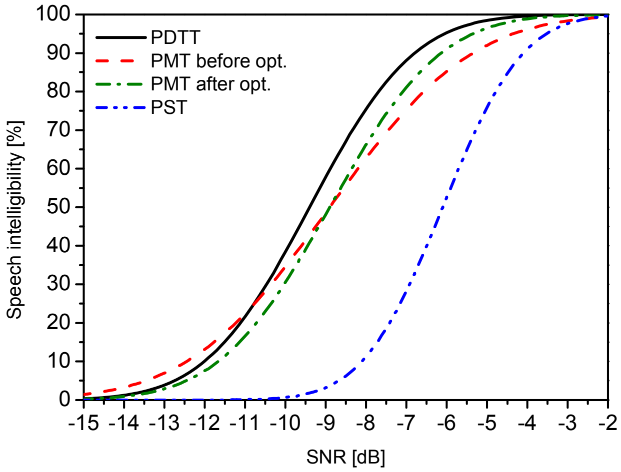

Fig.1 Normative psychometric functions for the new Polish speech-in-noise tests.

The speech tests are characterized by relatively steep psychometric functions, i.e. provide accurate speech intelligibility data. They differ in SRT values. The lowest SRT is observed for PDTT, whereas the highest SRT was obtained for PST, as expected. This is due to high redundancy and limitation of speech material used in PDTT that leads to 'better' (i.e. lower) SRTs. As far as PMT is concerned, two intelligibility functions are depicted. The dashed line depicts function for tests items of unbalanced speech intelligibility (i.e. before level modifications). The dashed dot line depicts expected intelligibility function for PMT after reduction of σ_{SRT} to 1 dB. As can be seen, the SRT for Polish matrix test is very close to SRT obtained for PDTT. This is due, among others, to the scoring method used in PMT (word-scoring) that yields lower SRTs than sentence scoring [19].

4.2 Phoneme distributions

Figure 2 shows phoneme distributions for examined tests (PST – circles, PMT – triangles and PDTT – unfilled squares) and the reference phoneme distribution for Polish (filled squares).

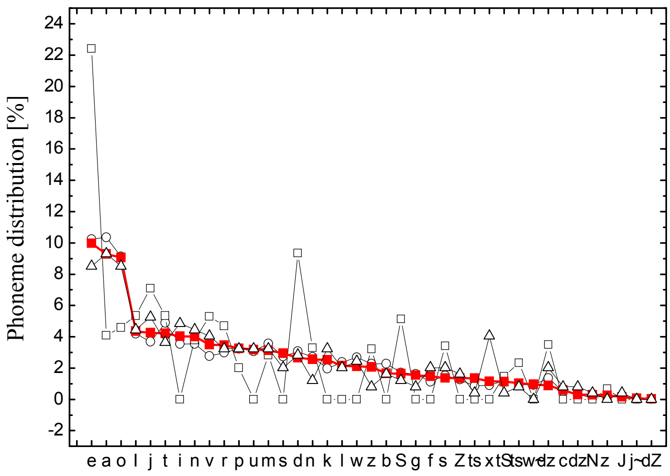

Fig. 2 Reference distribution for Polish language (filled squares) and mean distributions for mean PST (circles), PMT (triangles) and PDTT (unfilled squares).

As can be seen from Figure 2, the average phoneme distribution of PST is almost identical as the reference distribution for Polish. The maximal deviation from mean distribution across lists and across phonemes does not exceed 2.5 percent point.

As far as the phoneme distribution for PMT is concerned, it is also close to the reference distribution. However, one phoneme (x) is overrepresented. This is due to the fact that in the Polish version of 50-word matrix any adjective in the accusative form will always end in -ch (x), for example ***pięknych*** (beautiful) or ***drogich*** (expensive). However, in the matrix test each list is generated by a permutation of the same 50 words, thus each possible combination will reveal the same phoneme content, thus – unlike in PST – no deviation from mean distributions will be observed for different test realizations.

Among the considered speech materials, PDTT is the least lexically and phonemically representative for Polish. Thus, it cannot be used as a diagnostic test since it is based on 10 words only. In this case some phonemes are strongly overrepresented (/e/), whereas some others do not occur in the speech material at all (/u/).

5. Applications

The PST has been developed mainly for clinical diagnostics. This is due to the richness of speech material and unfixed syntactical structure of the utterances. It will be disseminated mainly among Polish audiologists, ENT doctors and hearing aid professionals. However, it should be emphasized that the Plomp-type lists can be presented to any given subject only once. This is due to the possibility of familiarization with the speech material which might affect intelligibility data. Hence, when more than 37 SRTs are to be determined (i.e. more than the total number of lists available), PST seems to be inapplicable.

Though PMT is based on strongly limited vocabulary, it has such an advantage over PST that it is possible to generate 10^5 sentences by means of a permutation of the elements in the 50-word based matrix. Accordingly, PMT is more suited for experiments in which speech intelligibility is analyzed in a variety of conditions. However, this approach has also some disadvantages. Because of the training effect, the measurements have to be preceded by a training session until stable SRTs are observed [13]. What is more, the fixed syntactical structure is not representative for everyday speech communication. PMT will be disseminated mainly among audiologists and acoustic centers in Poland.

Finally, PDTT is intended for general screening measurements. Subject’s responses can be easily typed on a telephone keyboard and, subsequently, collected and analyzed by a dedicated system connected to telephone lines. This kind of measurement scenario was used, for example, in the National Hearing Test carried out in the Netherlands [15].

Acknowledgments and references

This work received support from the European Union FP6 grant, Project 004171 HEARCOM and the State Ministry of Education and Science.

[1] R. Plomp, A.M. Mimpen, "Improving the reliability of testing the speech reception threshold for sentences", *Audiol.* 18, 43-53 (1979)

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

[3] M. Nilsson, S.D. Soli, J.A. Sullivan, "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 (1994)

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

[5] A. Pruszewicz, G. Demenko, L. Richter, T. Wika, "New articulation lists for speech audiometry. Part I", *Otolaryngol. Pol.* 48, 50-55 (1994)

[6] A. Pruszewicz, G. Demenko, L. Richter, T. Wika, "New articulation lists for speech audiometry. Part II", *Otolaryngol. Pol.* 48, 56-62 (1994)

[7] G.F. Smoorenburg, "Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram". *Journal of the Acoustical Society of America* 91(1), 421-437 (1992).

[8] A.J. Bosman, G.F. Smoorenburg, "Intelligibility of Dutch CVC syllables and sentences for listeners with normal hearing and with three types of hearing impairment", *Audiology* 34, 260-284 (1995)

[9] B. Hagerman, "Sentences for testing speech intelligibility in noise", *Scandinavian Audiology* 11, 79-87 (1982)

[10] K. Wagener, F. Eeenboom, T. Brand, and B. Kollmeier,"Ziffer-Tripel-Test: Spracherverstand-lichkeitstest uber das Telefon", *in Tagungs-CD der DGA Jahrestagung* (2005)

[11] K. Wagener, J.L. Josvassen, R. Ardenkjaer, "Design, Optimization, and Evaluation of a Danish Sentence Test in Noise". *Journal of International Audiology* 42(1), 10-17 (2003).

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

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

[14] K. Wagener, T. Brandt, B. Kollmeier, "Development and evaluation of a German sentence test II:Optimalization of the Oldenburg sentence tests". *Z.Audiol* 38, 44-56 (1999)

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

[16] S. Prosser, M. Turrini, E. Arslan, "Effects of different noises on speech discrimination by the elderly". *Acta Otolaryngologica* 476, 136-142 (1991).

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

[18] B. Kollmeier, "Messmetodik, Modellierung und Verbeserung der Verstandlichkeir von Sprache", Georg-August-Universtat: Gottingen, (1990)

[19] K. Wagener, "Factors influencing sentence intelligibility in noise", *in PhD Disertation, University of Oldenburg* (2003).