HearCom

FP6–004171 HEARCOM

Hearing in the Communication Society

INTEGRATED PROJECT

Information Society Technologies

D-7-1b: Speech recognition tests for two additional languages: Polish and French

Contractual Date of Delivery:	28-2-2007 (+45 days)
Actual Date of Submission:	24-04-2007
Editor:	E. Ozimek
Sub-Project/Work-Package:	SP3/WP7
Version:	3.0
Total number of pages:	39

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)

Deliverable D-7-1b

VERSION DETAILS	
Version:	3.0
Date:	24–04-2007
Status:	Final deliverable

CONTRIBUTOR(S) to DELIVERABLE	
Partner	**Name**
PL - IAM	Edward Ozimek
PL - IAM	Dariusz Kutzner
PL - IAM	Aleksander Sęk
PL - IAM	Andrzej Wicher
BE - LEU	Jan Wouters
BE - LEU	Heleen Luts
BE - LEU	Koen Eneman

DOCUMENT HISTORY			
Version	**Date**	**Responsible**	**Description**
1.0	28-12-2006	Edward Ozimek	First draft
2.0	14-03-2007	Jan Wouters	Merge Polish and French part
3.0	24-04-2007	Edward Ozimek	Final deliverable

DELIVERABLE REVIEW			
Version	**Date**	**Reviewed by**	**Conclusion***
2.0	04-04-2007	Rob Drullman	Accept with minor modifications
2.0	17-04-2007	Kirsten Wagener	Accept with minor modifications

* e.g. Accept, Develop, Modify, Rework, Update

Table of Contents

List of Figures

List of Tables

Acknowledgement

Supported by grants from the European Union FP6, Project 004171 HEARCOM. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

Pre-Amble

Workpackage 7 of the HearCom project concentrates on the evaluation of new signal enhancement techniques for hearing aids. For this purpose standardized evaluation procedures are defined to compare and validate the algorithms in different European languages. The algorithms were developed and technically evaluated in workpackage 5.

Deliverable D-7-1 describes the speech materials and test procedures to be used for the perceptual evaluation. The present deliverable D-7-1b describes the development of two additional speech test materials for Polish and French, to complement the European inventory of speech tests and to allow better across-language comparisons in evaluations. This work is performed in close collaboration with workpackage 1. In WP1, speech tests are also being developed, but with a different goal, namely to assess the individual communication performance and the auditory impairment of a listener.

1 Executive Summary

This report summarizes the development of a Polish and a French sentence test for determining speech intelligibility in noise. Both the speech material and the standardised measurement procedure aim at the determination of the psychometric (intelligibility) functions for hearing-impaired listeners in aided conditions.

The Polish sentence material chosen allowed a preparation of three sets of sentences. The first set is composed of 25 lists, each comprising 20 sentences. It was used for sentence scoring of the speech intelligibility. The second set was composed of 22 lists of 20 sentences each. For this set, word-based intelligibility scoring was applied. The third set (sentence scoring) is composed of 37 lists of 13 sentences each. All sets were phonemically and statistically balanced, i.e. the respective lists revealed a comparable phonemic content and similar psychometric functions. The speech reception threshold (SRT) and slope of the psychometric function at the SRT (S_{50}) were determined. The test sentences were mixed digitally with a speech babble noise (masker). The masker was presented at a constant level of 70 dB SPL. It was found that for the normal-hearing subjects (treated as a control group), the mean SRT and the mean list specific S_{50list} for the first set were –6.1 dB SNR and 25.5 %/dB, respectively. The mean SRT and S_{50list} for the second set were –7.4 dB SNR and 22.4 %/dB and for the third set -6.1 dB SNR and 25.6 %/dB, respectively. Our data indicate that the steepness of the psychometric function for the Polish language is larger than that of comparable sentence tests in other languages.

The French sentence material was originally recorded and evaluated by Wable (2001). As there were no reference data available for normal-hearing listeners, the sentence material was re-evaluated. Fourteen test lists of ten sentences and one training list of 20 sentences were composed. The performance of Francophone listeners from France and Belgium was compared. The mean SRT was -7.8 dB SNR and -7.1 dB SNR for French and Belgian normal-hearing listeners respectively. The average slope of the psychometric function was 20.2 %/dB.

2 Introduction

Different methods have been proposed for measuring speech intelligibility in quiet and in noise (Kalikow et al., 1977; Hagerman, 1982; Nilsson et al., 1994; Kollmeier and Wesselkamp, 1997; Brachmański and Staroniewicz, 1999; Versfeld et al., 2000; Smits et al., 2004; Wagener et al., 1999, 2003, 2006b, 2007). They differ in aspects as: the structure of the speech material, details of the test procedure, presentation level, range of the signal-to-noise ratio (SNR), type of interfering noise and presentation mode. As far as the

structure of the speech material is concerned, one can distinguish three basic types of tests: word intelligibility tests (Runge and Hosford-Dunn, 1985; Pruszewicz et al., 1994a;b; Bosman and Smoorenburg, 1995; Martin, 1997), digit intelligibility tests (Pruszewicz at al., 1994a;b; Smits et al., 2004; Wagener et al., 2006b, 2007) and sentence intelligibility tests. The sentence tests can be divided into those using meaningful, everyday utterances (Plomp and Mimpen, 1979b; Smoorenburg, 1992; Nilsson et al., 1994; Kollmeier and Wesselkamp, 1997; Versfeld et al., 2000) and those using semantically unpredictable (nonsense) sentences (Hagerman, 1982; Wagener et al., 1999,2003).

The main disadvantage of word intelligibility tests, as used in traditional speech audiometry, is their difficulty to determine a precise SRT value in a time-efficient manner. Sentence intelligibility tests, on the other hand, have been shown to be much more accurate (Plomp and Mimpen, 1979a; Hagerman, 1982; Nilsson et al., 1994; Kollmeier and Wesselkamp, 1997; Versfeld et al., 2000; Brandt and Kollmeier, 2002; Wagener et al., 1999, 2003). They have some advantages over the word tests in testing the quality of hearing aids with different signal processing algorithms since they provide a reliable distinction between unaided and aided patient performance. Furthermore, unlike words and digits, sentence materials reflect natural communication processes.

Many sentence tests have been developed for measuring the speech reception threshold (SRT) in noise (defined as SNR that yields 50 % intelligibility) (Kalikow et al., 1977; Plomp and Mimpen, 1979a; Nilsson et al., 1994; Kollmeier and Wesselkamp, 1997; Versfeld et al., 2000; Wagener et al., 1999, 2003). Some papers have shown that the SRT depends only on the SNR (Smoorenburg, 1992; Wagener and Brandt, 2005), but according to others the SRT depends both on the SNR and on the presentation level (Hagerman, 1982; Studebaker et al., 1999). The intelligibility of speech in a stationary noise is mainly determined by the SNR.

The present study deals with the preparation and evaluation of a Polish and a French sentence test to be used for measurement of the SRT in noise. The tests are structurally similar to the meaningful sentence tests (Dutch: Plomp and Mimpen, 1979a; American: Nilsson et al., 1994; German: Kollmeier and Wesselkamp, 1997). The developed speech tests are designed for evaluation of speech intelligibility in clinical conditions as well as for laboratory measurements.

3 The Polish sentence test

To our knowledge, the Polish sentence material is the first test designed for the Slavonic languages that are used by more than 200 million people. The description of the development of the Polish sentence test contains two parts. The first part deals with the preparation of the sentence tests, the recording procedure and a listening experiment. The second part addresses results of the intelligibility scoring of the sentence tests and a verification of the final tests.

3.1 Speech material, measuring procedure and evaluation

Till recently, there has been no Polish sentence test for measuring speech intelligibility in noise. This fact has prompted us to develop a test based on Polish sentence materials. It is known from literature that speech material used for measuring intelligibility should produce very steep intelligibility functions to permit detection of changes in intelligibility resulting from small differences in SNR. Moreover, to keep high measurement accuracy the intelligibility across different lists should not vary significantly. Therefore, the test lists should show high comparability.

3.1.1 Preparation and recording of the sentences

The test proposed in this project is a Polish version of the test presented by Plomp and Mimpen (1979a) and it will be called shortly the Plomp-type test. The test was prepared in two stages. In the first stage, about 3500 sentences were selected automatically from a large digitized database containing about 16 million sentences taken from everyday speech, literature, TV and theatre. All of them fulfilled the fundamental definition of a sentence, i.e. they included a subject, an object and a verb. They contained normal everyday contexts. The following criteria were used in the automatic selection of the sentences (similar to Versfeld et al., 2000): the total number of syllables in a sentence should be equal to eight or nine (number of words fell into a range from 3 to 7; mean 4.6 words/per sentence); the words in the sentences should not contain more than three syllables each; the sentences should not contain punctuation characters and capitals (excluding the initial capital). No duplicate sentences were selected. The second stage of sentence selection was realized manually on the basis of the following criteria: the sentences should be grammatically and syntactically correct and semantically neutral, which excluded political, war or gender-related topics. Questions, proverbs, proper names and exclamations were eliminated. This process reduced the set of sentences to 2000.

In Polish language, unlike in English, all verbs are subject to declination with respect to the speaker's gender. Besides, phonemes distinguishing the present and the past tense are usually characterized by a relatively low energy. Hence, they might contribute to increase in perception ambiguities of the sentences and, consequently, increase in the scatter of the data both within a given

sentence list and between lists. Therefore verbs that might lead to any ambiguities were rejected. This limited the final set of sentences to 1200.

Fig. 3.1 presents a comparison between the percent distributions of phonemes estimated for the set of 1200 sentences (bars) and average Polish speech (solid line) presented by Jassem (1973).

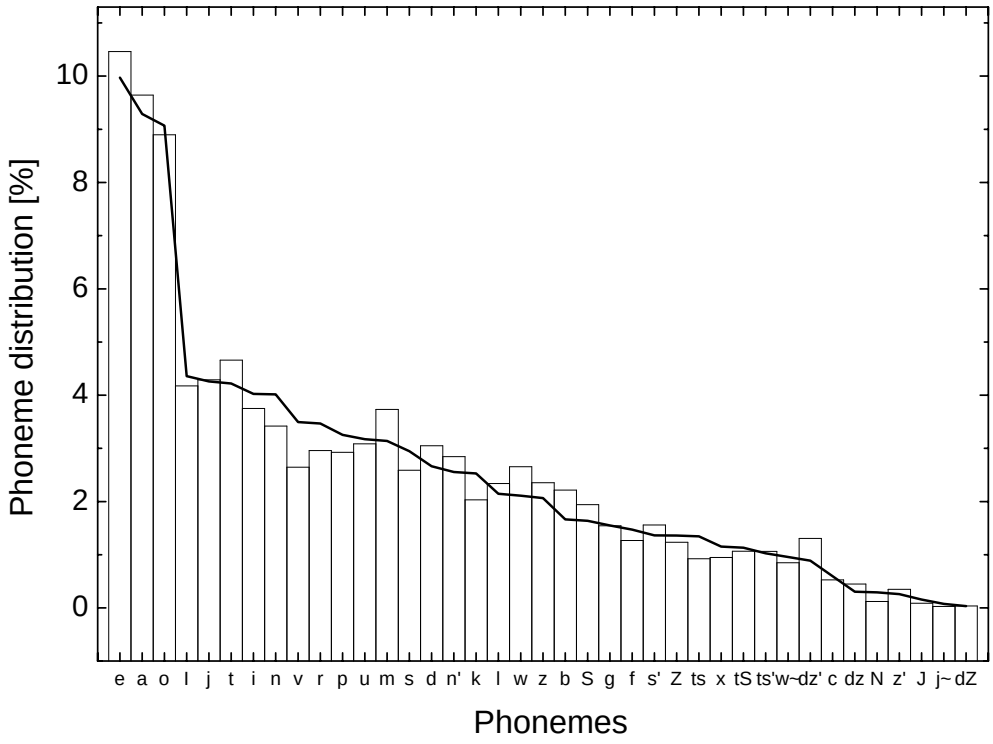

Fig. 3.1. *A comparison between the percent distributions of phonemes estimated for 1200 recorded sentences (bars) and average Polish speech (solid line).*

As seen from Fig. 3.1, these 1200 sentences can be regarded as phonemically balanced with respect to average Polish speech. The mean square error (MSE) between the phoneme distribution of the sentence set and average Polish speech is about 0.12 (corresponding to a Pearson correlation coefficient, *r,* of about 0.97). The 1200 sentences were read out in a natural intonation by a professional male speaker, keeping approximately the same loudness level over time. Recording was performed in a high quality radio studio using a Neumann U87 capacitor microphone.

The microphone output fed one of the input channels of a 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, with a resolution of 24 bits. It was digitally high-pass filtered at a cut-off frequency of 80 Hz. The signals were then sent via optical connection (ADAT-type) to a PC and stored on a computer hard disc using Samplitude Pro v.8.2 software. Special care was

taken to keep an approximately constant *rms* value during the recording session. A single recording session lasted no longer than 2 hours. To prepare the set of 1200 sentences 4 recording sessions were necessary.

3.1.2 Listening sessions

3.1.2.1 Apparatus, procedure and subjects

A computer controlled Tucker-Davis Technologies (TDT) System III with a 24-bit digital real-time signal processor RP2.1, and a headphone amplifier HB7 was used to play back the sentence material with the interfering noise at fixed SNRs. The level (in dB SPL) of the output signal was calibrated with B&K instruments (an artificial ear type 4153 connected with a microphone type 4134, preamplifier type 2669 and amplifier type 2610). Speech signals were presented monaurally via Sennheiser HD 580 headphones.

The test sentences were mixed digitally with a speech babble noise (masker) at a constant level of 70 dB SPL. The babble noise was generated by summing up all of the sentences and normalizing the *rms* value of a resultant wave. Waveforms representing single sentences were randomly shifted in the time domain with respect to each other and, additionally, half of them were reversed. In this way, the decrease in masker power resulting from a decline in the speaker's vocal effort during an utterance was reduced. As a result, a 15-s realisation of the speech babble noise was obtained. For this type of masker, average SNRs in the respective frequency bands (auditory filters) were kept constant for a given speaker.

The power spectral density of the babble noise, obtained by means of a Fast Fourier Transform (frequency resolution of 25 Hz, Hanning window with 50 % overlapping) is depicted in Fig. 3.2 (solid line).

The spectrum reveals an almost flat envelope between 5 and 10 kHz, reflecting the energy of consonants typical for the Polish language. As can be seen from the figure, the power spectrum density of the babble noise is slightly different from that used by Versfeld et al., (2000) and Wagener, (2003), which are also presented in the figure by means of dashed and dotted lines respectively. The difference between the presented spectra reflects the individual properties of German, Dutch and Polish as they were created in nearly the same way.

In the experiments, the sentence presentation level was changed to get different SNRs. The *rms-level* of the sentences was equalized (Versfeld et al., 2000). SNR was defined as the ratio of the sentence *rms* to the masking noise *rms*. The masking noise was a gated signal (20-ms ramps) and started 300 ms before the onset of the sentence and ended 300 ms after the end of the sentence. All sentences were presented at three main SNRs: -9, -5 and -1 dB SNR, randomly chosen during the listening sessions. However, the psychometric function fit to these data was not always found to be reliable (i.e. for some sentences the intelligibility scores did not optimally cover the

SRT point). Therefore, measurements for two additional SNRs falling in the range from –9 to –1 dB SNR have been performed. The lower additional SNR value was chosen in such a way that it was between –9 dB SNR and the value of the measured or expected SRT, while the second additional SNR was chosen between the measured or expected SRT and –1 dB SNR. In this way, the psychometric functions could be determined on the basis of five SNR values.

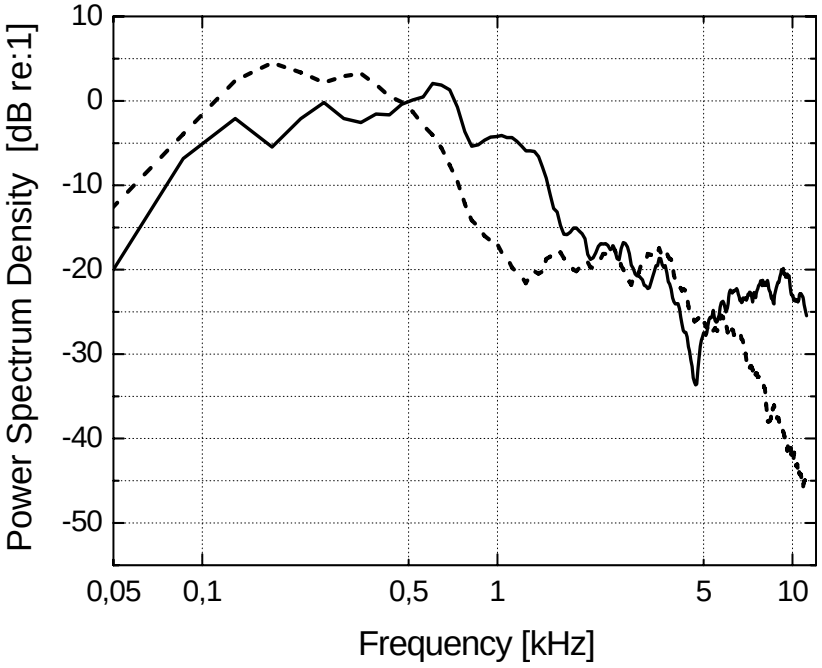

Fig. 3.2. *The long-term power spectrum of the Polish babble noise (solid line) used in the present study. For comparison the long-time power spectrum of the masker used in the study of Versfeld et al. (2000) is presented (dashed line). The average rms for both maskers was normalized to unity.*

Each sentence was presented to each subject only once. Therefore, each subject listened to 1200 sentences (three blocks of 400 sentences at each SNR value). Each sentence was presented to 21 subjects (7 subjects at each of the three SNR values). For simplicity, the set of 1200 sentences was divided into 40 subsets of 30 sentences each.

The listening session was self-paced and was controlled by the subject, who was placed in an acoustically isolated room. All instructions were displayed on an LCD screen. When the ENTER key was pressed for the first time, written instructions appeared on the monitor. Pressing this key for the second time caused the playback of one sentence via TDT III system as described above. The subject's task was to repeat this sentence as accurately as possible and his/her response was recorded on the computer hard disk. The recording of

the subjects' oral responses was made by means of a separate signal channel (a condenser E914 microphone, Yamaha MG10/2 mixer with a preamplifier and an external Sound Card of a PC). By pressing the ENTER key again, the recording of the subject's response was stopped, and the subject's task was to type in the sentence they heard on a keyboard. To listen to a new sentence the subject had to press the ENTER key again. It should be emphasised that the subjects were instructed to write an answer even when only one word of the sentence was understood. Both oral and 'typed' responses were collected in order to minimize potential ambiguities that might result from, for example, typing mistakes. If there were ambiguities or typing errors in the 'typed' response, the subject's recorded oral utterance was considered.

Twenty-one native Polish subjects (aged from 18 to 25 years, 13 males and 8 females) participated in the experiment. The subjects had no otological problems and their pure-tone audiograms did not exceed 10 dB HL over the 6 octaves from 250 to 8000 Hz. Before the actual measurements were made, each subject was trained for 1-2 hours until stable answers were obtained. The sentences used in the training sessions were not used in the main experiment. However, they were spoken by the same speaker and the structure of those sentences was the same as that used in the main study. Collection of responses for all 1200 sentences for one subject took about 8 hours, but the duration of a single measurement session was limited to 2 hours. The subjects were allowed to have a rest whenever they wished.

3.1.3 Intelligibility functions

3.1.3.1 Intelligibility scoring

Intelligibility scores for each sentence were estimated in two ways. In the first type of scoring, the subject's response was treated as correct when all the words in a sentence were repeated correctly. In this case the score was 100 %. Otherwise the score was set to 0 % (Versfeld et al., 2000). This type of scoring is called sentence scoring.

The second way of scoring was based on a correct repetition of individual words in the sentence. The intelligibility score was estimated as a ratio of the number of correctly repeated words to the total number of words in the sentence, multiplied by 100 % (Kollmeier and Wesselkamp, 1997; Wagener,1999, 2003, 2005). In this way intermediate score values falling in the range between 0 and 100 % were also possible. This approach is called word-based scoring. The two methods give the same results if a given sentence is fully understood or totally misunderstood (this corresponds to intelligibility scores of 100 % or 0 %).

Sentence scoring may be less suited for testing profoundly hearing-impaired patients since such patients are often unable to repeat entire sentences correctly, especially under masking conditions. Consequently, derivation of intelligibility functions for these sentences would be impossible.

3.1.3.2 Determination of the intelligibility function parameters: SRT, SD and steepness (S_{50})

There are a number of methods for estimating the SRT. One of them is based on fitting a psychometric function to the data using a maximum-likelihood criterion (Versfeld et al., 2000, Brandt and Kollmeier, 2002). In this case all sentences of a given list are included in the SRT calculation. The psychometric function is usually approximated by the standardized cumulative normal distribution. It is generally assumed that the relationship between speech intelligibility and the SNR is expressed by the function $\varphi(SNR)$ (1):

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

The function $\varphi(SNR)$ has two parameters: SRT (the SNR ratio that produces 50 % correct responses) and σ (the standard deviation of the normal distribution). There is a direct relationship between the slope of the psychometric function at the SRT (denoted here as steepness S_{50}) and the standard deviation σ (denoted further as SD). The two parameters are easily convertible from the formula (2):

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

It is noteworthy that the steepness S_{50} of the psychometric function for a given sentence might be changed by a modification of intelligibility of particular words in this sentence (Kollmeier and Wesselkamp, 1997). For instance, the steepness increases and, consequently, the standard deviation decreases, if the intelligibility of the respective words is equalized. Hence, S_{50} indirectly describes the spread of intelligibilities of the individual words constituting the sentence.

It should be added that the relationship between the speech intelligibility and SNR can also be described by the logistic function (Kollmeier and Wesselkamp, 1997; Wagener et al., 1999, 2003):

$$\varphi(SNR) = \frac{100}{1 + e^{-4S_{50}(SNR-SRT)}} \qquad (3)$$

For each sentence, two psychometric functions were fitted to the averaged results across subjects: one for sentence scoring and one for word-based scoring. As a result, two groups of 1200 psychometric functions were obtained. Subsequently, two sets of 1200 SRT values and S_{50} values were determined.

3.2 Composition of the 20-sentence lists

Composition of the sentence lists aimed at minimizing the scatter of SRTs between lists. This is important taking into account possible applications of the lists in clinics. To do this it was decided that the SRT of any sentence to be included in the list should not exceed the range of ±1.5 dB SNR with respect to the mean SRT of all sentences. Moreover, to reduce the scatter resulting from inequality of the intelligibility of individual words, only sentences with steepness S_{50} not less than 15 %/dB were selected. As a result, 500 and 440 sentences fulfilling the above criteria for the sentence and word-based scoring methods were selected respectively. In this way it was possible to create 25 lists (20 sentences each) for sentence scoring and 22 lists (20 sentences each) for word-based scoring. Finally, the sentences need to be compiled into final lists, which should fulfil the following criteria:

- the lists should be equivalent, i.e. the average SRT and S_{50} of each list must not depend on the list number,
- the lists should contain a phonemically comparable content.

A special algorithm was prepared and implemented in Matlab 7.0 (*MathWorks*) which performed Monte Carlo simulations. The steps in the compilation were as follows:

- Random permutations of 500 and 440 sentences for the sentence and the word-based scoring were created respectively.

- The random series of sentences were grouped in 25 lists and 22 lists of 20 sentences each. The lists constituted a set of initial lists that were further analysed with respect to homogeneity of SRT and S_{50} steepness.

- For the 25 sentence scoring lists two separate one-way ANOVAs were performed with respect to the SRT and steepness S_{50} respectively. The lists were regarded as equivalent when there were no statistically significant differences ($p<0.05$) between the SRTs and S_{50}s across the lists. When the SRTs and S_{50}s across the lists were significantly different, a new random permutation was generated and this step was repeated. The same analysis was carried out with respect to the 22 word scoring lists.

- In the last step, phoneme analysis distribution was performed. A specially designed algorithm transformed the written sentences into SAMPA-broad[1] (Polish extension) phonemic code taking into account the co-articulation effects using phoneme distribution rules.

[1] SAMPA (Speech Assessment Methods Phonetic Alphabet) is a machine-readable phonetic alphabet, originally developed for Danish, Dutch, English, French, German, and Italian by an international group of phoneticians (ESPROT project) in the years 1987-1989. Under the BABEL project, SAMPA was extended to the Polish language (1996). http://www.phon.ucl.ac.uk/home/sampa/index.html

Subsequently, mean phonemic distributions for each list were determined and compared to the reference distribution for the Polish language (Jassem, 1973). The lists were regarded as phonemically balanced if for each phoneme and each list, the frequency of occurrence of any phoneme did not exceed the range ±2.5 percentage points with respect to the frequency of occurrence of that phoneme. This algorithm led to a set of equivalent lists, in which the phoneme balance accuracy was similar to that obtained by Kollmeier and Wesselkamp (1997).

Finally 25 statistically and phonemically equivalent lists containing 20 sentences each for sentence scoring (see Appendix A for an example list) and 22 statistically and phonemically independent lists containing 20 sentences each for word-based scoring (see Appendix B for an example list) were obtained. The main reason for composing two separate tests was the observation that there were differences between sentence-specific psychometric functions obtained for the sentence and word-based intelligibility data. In general, the psychometric functions for word-based scoring are slightly less steep and are characterized by lower SRT values than for sentence scoring. The two methods produce the same or very similar speech intelligibility scores at relatively high SNRs (higher than the SRT). However, at relatively low SNRs, the word-based intelligibility scores are higher than sentence scores.

The SRT_{mean} for a list of 20 sentences was calculated as the average SRT_k of the sentences in the lists:

$$SRT_{mean} = \frac{1}{N}\sum_{k=1}^{N} SRT_k \tag{4}$$

where k is the sentence index and N is the number of sentences in a list (N=20).

The so-called list-specific steepness S_{50list} was determined according to the probabilistic model described by Kollmeier (1990):

$$S_{50list} \approx \frac{S_{50mean}}{\sqrt{1+\frac{16 {S_{50mean}}^2 {\sigma_{SRT}}^2}{(\ln(2e^{1/2}-1+2e^{1/4}))^2}}} \tag{5}$$

where: S_{50mean} is the mean steepness of the sentences in a list, σ_{SRT} is the standard deviation of the sentence-specific SRT values. According to equation (5), a high list-specific steepness requires both a high mean sentence-specific steepness and a small spread of sentence-specific SRTs, i.e. a small standard deviation σ_{SRT}.

The effect of the subject was not statistically significant (according to the ANOVA). Thus, the intelligibility scores obtained for each SNR were averaged

across subjects. Independent averaging was performed for the data related to sentence scoring and to word-based scoring. The psychometric functions were fitted to the averaged intelligibility scores using the least squares (LS) method.

3.2.1 Sentence intelligibility based on sentence scoring

Fig. 3.3 presents a comparison between the distribution of phonemes estimated for averaged Polish speech and for the 25 lists. Fig. 3.4 presents examples of sentence intelligibility functions for two lists (No 2 and 6), fitted to the intelligibility data averaged across subjects (open circles). A juxtaposition of the psychometric functions for 25 lists is presented in Fig. 3.5.

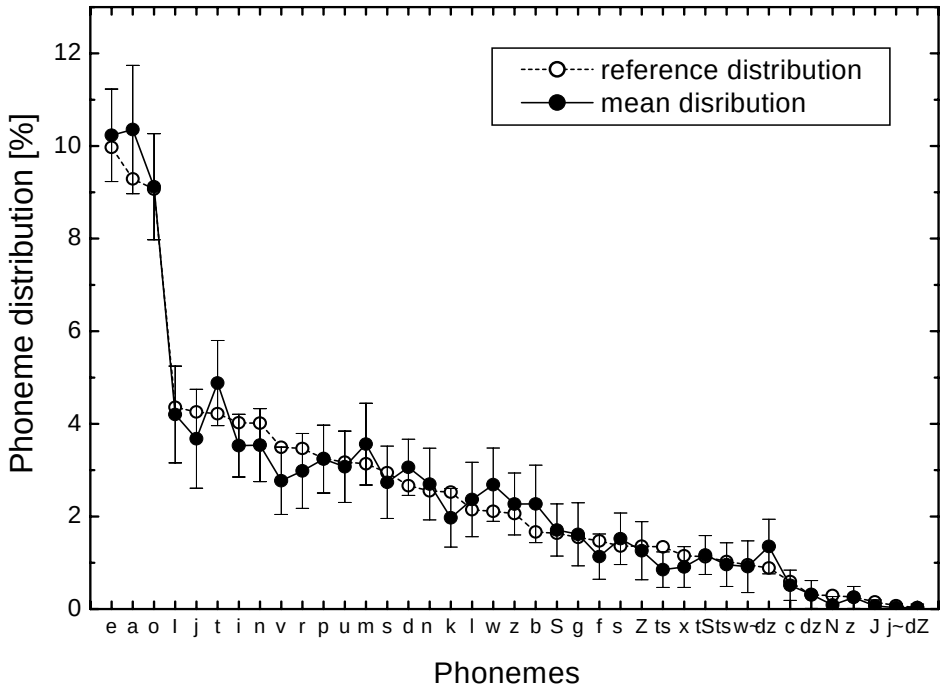

Fig. 3.3. *A comparison between the percent distribution of phonemes estimated for a large database of Polish speech (open circles) and the mean percent distribution for the 25 test lists (filled circles). Vertical bars show the standard deviation values across the lists.*

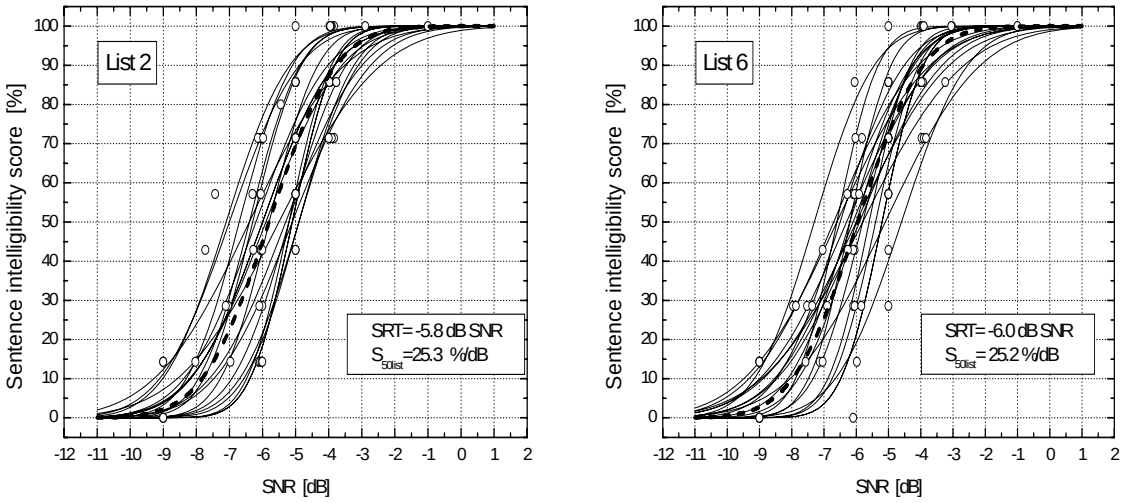

Fig. 3.4. *Examples of the intelligibility functions versus SNR for two lists for sentence scoring. Symbols denote experimental results averaged across subjects and the thin solid lines present intelligibility functions fitted by formula (2). The broken lines show the list-specific intelligibility functions.*

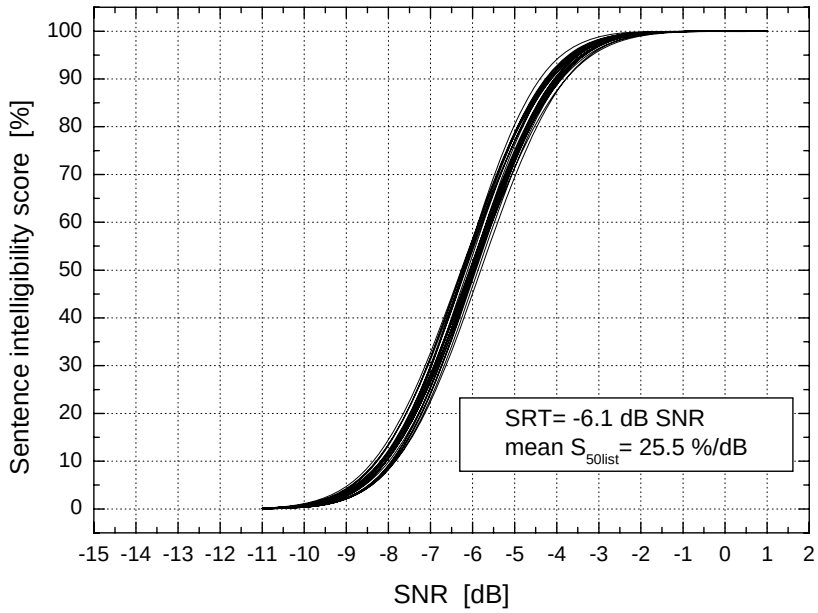

Fig. 3.5. *List-specific psychometric functions for 25 lists (sentence scoring).*

Fig.3.6 presents the average SRT values and the corresponding standard deviations for each list. ANOVA analyses showed that neither the SRTs {F(24,499)=0.54, p=0.96} nor the S_{50} {F(24,499)=0.55, p=0.96} varied significantly across the lists, meaning the lists are statistically equivalent.

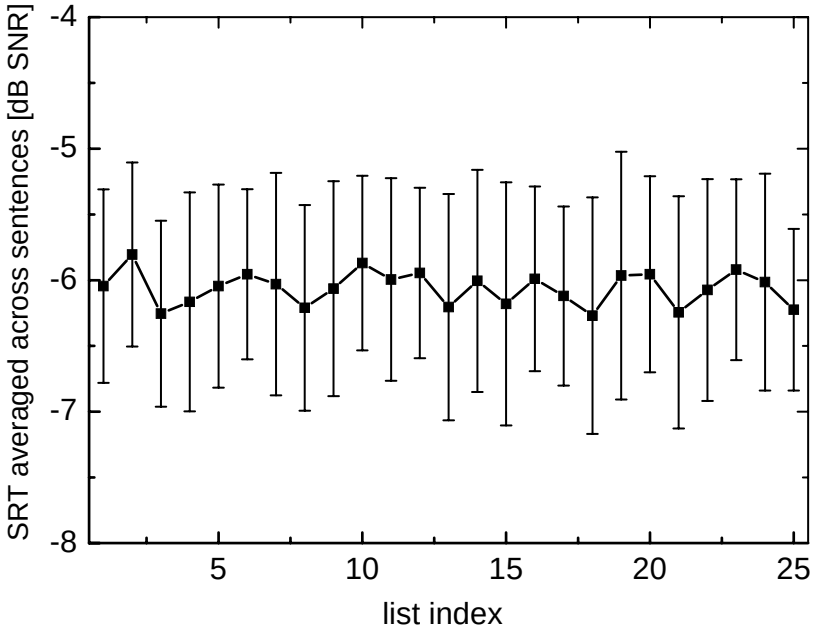

Fig. 3.6. *SRT averaged across sentences for each list and the corresponding standard deviations (sentence scoring).*

As can be seen in the figures, the average SRT values per list range from –6.3 dB SNR to –5.8 dB SNR (mean SRT=-6.1 dB SNR). The average list-specific steepness S_{50list} equals to 25.5 %/dB.

3.2.2 Sentence intelligibility based on word scoring

The similar procedure as in paragraph 3.2.1 was applied to the sentence intelligibility data based on word scoring. In this case, the average SRT values were between -7.6 dB SNR and -7.2 dB SNR (mean SRT=-7.5 dB SNR). The ANOVA showed that neither the SRT nor the S_{50} varied significantly across lists ({F(21,439)=0.29, p=0.99} and {F(21,439)=0.40, p=0.99} respectively). The average list-specific S_{50list} equaled to 22.4 %/dB.

3.2.3 Verification of the reliability of the sentence tests

In the final stage of this study, a verification (retest) experiment was carried out to investigate the reliability of the Polish sentence materials when presented in an interfering babble noise. Both sets of lists (sentence and word-based scoring) were evaluated. Five randomly chosen lists from each set of the two lists were presented to a new group of 10 normal-hearing subjects (5 male and 5 female). For 5 subjects, the intelligibility score was estimated using sentence scoring and for the other 5 subjects it was estimated using the word-based scoring. Each list was evaluated at different SNRs (constant stimuli paradigm method, five SNRs: –9, -7.5, -6, -4.5 and –3 dB).

The mean SRT and S_{50list} obtained in the verification experiment were –6.0 dB SNR and 24.3 %/dB, respectively. These correspond well with the values of SRT=-6.1 dB SNR and S_{50list}=25.5 %/dB obtained in the first experiment.

The method of verification of the sentence material for word-based scoring was similar to that for the sentence scoring. The mean SRT and S_{50} for five subjects are -7.4 dB SNR and 20.8 %/dB respectively. These values are similar to the expected values of - 7.5 dB SNR and 22.4 %/dB obtained in the second experiment.

3.3 Composition of lists with the optimal number of sentences

The tests with lists of 20 sentences each, give very stable and reliable SRT estimates. As the lists are statistically equivalent the use of any of them yields similar SRT and S_{50}. However, the 20-sentence lists have one important disadvantage, namely the time that is necessary to determine the SRT reaches approximately 5 minutes. This is rather a long time especially for speech intelligibility measurements in the clinic. Diagnostic tools should be as fast as possible without compromising on its accuracy. As shown by Wouters at al. (1999) it is possible to design a shorter speech intelligibility test (13-sentence test) keeping similar SRT, SD and S_{50} to those determined using the 20-sentence list as described above. In this way, the overall time required for a single SRT determination is reduced by 30 % relative to that for the 20-sentence test.

3.3.1 Composition of the 13-sentence lists

A 13-sentence test was composed using the same sound material that constituted the previously developed lists. However, in this case only the sentence scoring method was used. The lists were intended to fulfill two basic criteria: they should be statistically equivalent, i.e. should produce very similar psychometric functions, and the lists should contain a phonemically comparable content (phonemic balance).

The way in which the 13-sentence lists were composed out of the 500 sentences was different than in the case of the 20-sentence lists. The main steps of this process were as follows. A subgroup containing n=13 sentences was randomly selected until the mean SRT and the mean S_{50} for the subgroup fell into the ranges of ±1.5 dB SNR and ±1 %/dB, respectively, around the mean SRT and S_{50}. Then, the phoneme distribution was analysed: if the mean distribution of each phoneme for the subgroup fell into a range of ±1.25 percent point with respect to the reference phoneme distribution for Polish language, then the subgroup was accepted as a final list. From the remaining sentences, using the same rules as described above, a next subgroup of 13 sentences was randomly chosen. Each subgroup of 13 sentences chosen as a final list decreased the number of the remaining sentences. In the above described way, 37 lists, each containing 13 sentences, were selected (see

Appendix C for an exemplary list). For the 19 remaining sentences none of 13-sentence subgroups fulfilled the above mentioned requirements for SRT, S_{50} and phoneme distribution. Therefore these sentences were not included in the final lists.

The above presented method allowed the creation of 37 lists with 13 sentences each for sentence scoring (see Appendix C for an exemplary list). The SRT_{mean} of each list was calculated from equation (4) and the standard deviation of a single list (SD_f) from equation (5). A comparison of the phoneme distribution estimated for averaged Polish speech (open circles) and for all the 13-sentence lists (filled circles) is presented in Fig. 3.7.

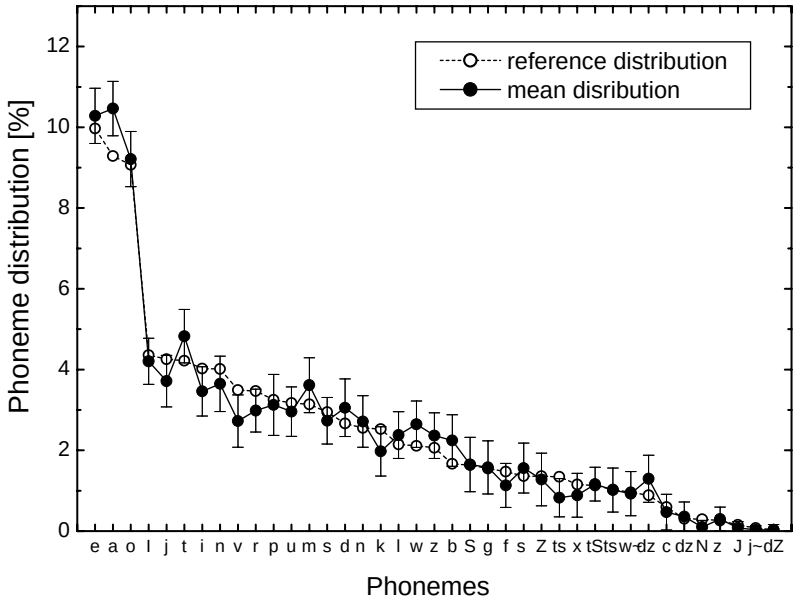

Fig. 3.7. *A comparison between the percent distribution of phonemes estimated for a large database of Polish speech (open circles) and the mean percent distribution for the 25 test lists (filled circles). Vertical bars show the standard deviation values across the lists.*

3.3.2 Sentence intelligibility for 13-sentence list using the sentence scoring method

Examples of sentence-specific intelligibility functions for two 13-sentence lists (No. 1 and 21), fitted to the intelligibility data averaged across subjects (open circles), are presented in Fig. 3.8. A juxtaposition of the average psychometric functions for 37 lists is presented in Fig. 3.9. Fig. 3.10 presents the average SRT values and the standard deviations for each list. ANOVA analyses showed that neither the SRTs {F(36,480)=0.4, p=0.99} nor the steepness S_{50} {F(36,480)=0.34, p=0.99} varied significantly across the lists.

As can be seen in the figures, the average SRT of the 13-sentence lists ranges from –7.1 dB SNR to –4.9 dB SNR (mean SRT=-6.2 dB SNR). The mean list specific S_{50list} is 25.6 %/dB. The average SRTand S_{50list} obtained for the 13-

sentence list are very close to the analogous values obtained for the 20-sentence lists. This means that the newly created lists may be regarded as reliable sound material for much faster speech intelligibility test.

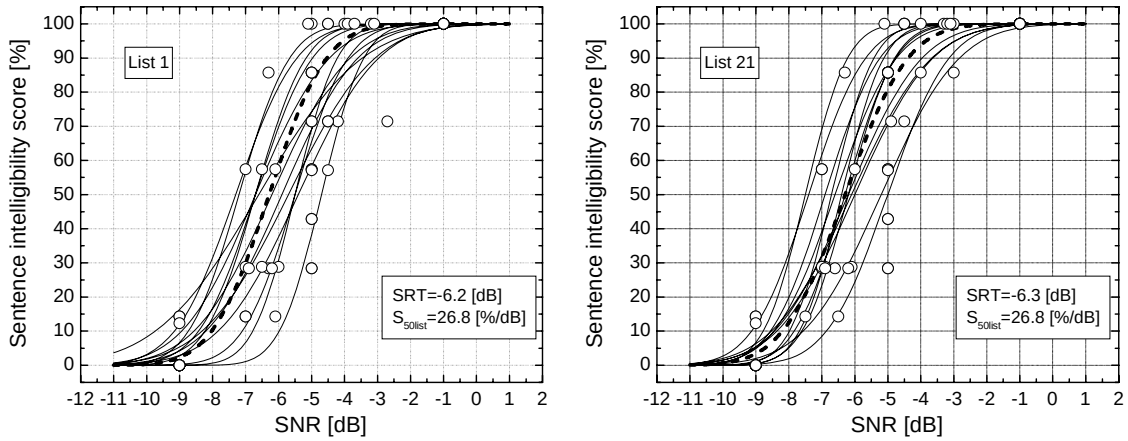

Fig. 3.8. *Examples of the intelligibility functions versus SNR for two lists. Symbols denote experimental results averaged across subjects and thin solid lines present intelligibility functions fitted by formula (2).*

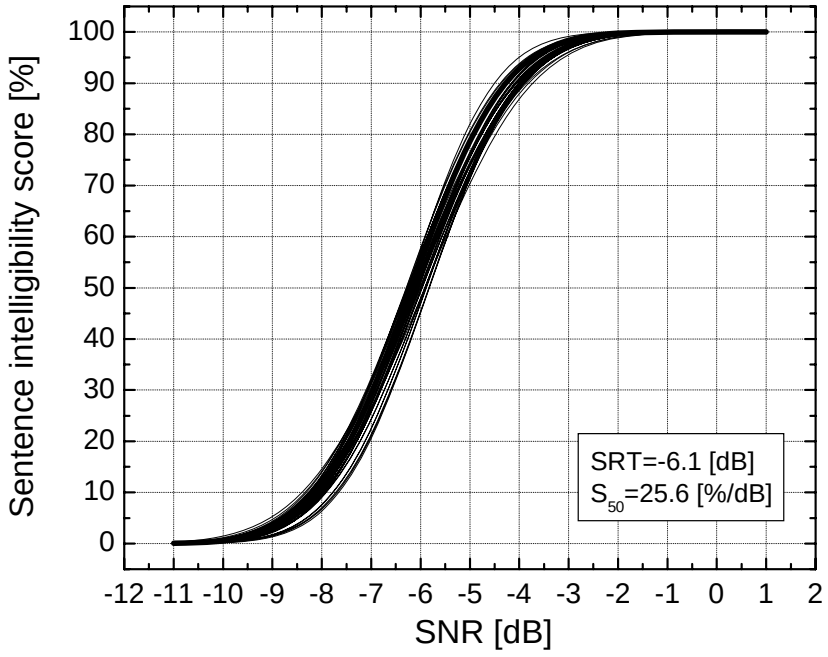

Fig. 3.9. *List-specific psychometric functions for 37 sentence lists (sentence scoring).*

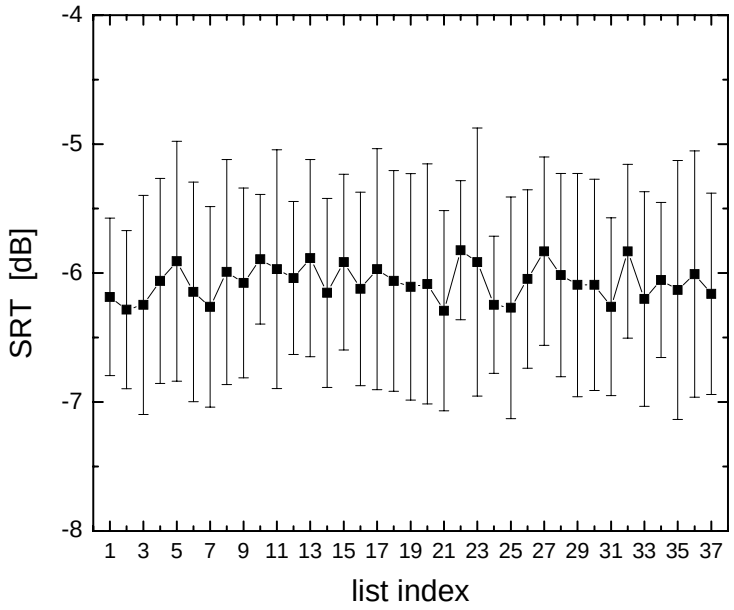

Fig. 3.10. *SRT averaged across sentences for each list and the corresponding standard deviations (sentence scoring).*

In the final stage of this part of the study, a verification (retest) experiment was carried out with respect to the 13-sentence lists. The purpose of this was to check the reliability of the 13-sentence material. A new group of 10 normal-hearing subjects was presented with five randomly chosen lists and the intelligibility score was estimated using the sentence scoring method.

The retest study was carried out using an adaptive method. The measurement session started with the presentation of the first sentence in a given list with a moderately high SNR (-2 or –3 dB). If the subject's answer was correct then the SNR was decreased by an SNR step or was increased by the same step in the case of an incorrect answer determination of the SRT only. Therefore, the steepness S_{50} was not determined in this part of the study. The initial size of the SNR step was set to 2 dB and after presentation of the fifth sentence it was reduced to 1 dB. The SRT was calculated by averaging the 8 last SNRs that appeared in the measuring session. It is worth adding that while calculating the SRT a 'virtual' SNR was also taken into account (Smits *at al*., 2004). The virtual SNR was the signal-to-noise ratio that would have occurred if the test had consisted of 14 sentences.

The mean SRT obtained in the verification experiment was –6.6 dB SNR, while its mean standard deviation of SRT estimation was 0.8 dB SNR. This corresponds very well with the value of SRT=-6.1 dB SNR obtained in the basic experiment. A *within subject* ANOVA revealed that the 'list factor' was statistically insignificant {F(36,369)=1.23, p=0.5}. Thus, the results of the retest measurements indicated that the 37 lists developed consisting of 13 different sentences could be regarded as statistically balanced as regards SRT.

4 The French sentence test

It appears that different speech tests are used for the French language, but little is known about standardised sentence tests. Audiological centres have custom-developed tests, and many evaluations are done with live voice. Recently, the Hearing In Noise Test was developed for Canadian Francophone listeners (Vaillancourt et al., 2005). Because of the large differences between Canadian and European French, the French Intelligibility Sentence Test (FIST) was developed for adaptive testing of speech intelligibility in noise in Europe. It is a Plomp-type test based on the sentence material collected by Wable (2001). Reference data were obtained for normal-hearing listeners and the performance of Francophone listeners from France and Belgium was compared.

4.1 Preparation and recording of the sentences

The initial test material was developed by Wable (2001). Sentences with various length, syntactic structures and word content were created using a database with typical French words (Content et al., 1990). The written sentences were judged by an informal jury according to difficulty, level of abstraction, naturalness and word content. A total of 582 sentences were selected. These were read by a male speaker, who was instructed to pronounce all sentences with the same clarity, level and articulation. Recordings were done in a sound-proof room. A Beyer Dynamic microphone (M69 TG) was placed at 30 cm in front of the speaker. The signal was monitored and stored on a Tascam DAT (DA-P1) and then digitized with a 16 bit A/D converter at a 44.1 kHz sampling rate. The sentences were edited using Cool Edit Pro software. The sentences were rescaled to equalize their *rms* level.

Within this project, a stationary speech-weighted noise with the long-term average speech spectrum of the sentences was generated. The power spectral density of the speech-weighted noise, obtained by means of a Fast Fourier Transform (frequency resolution of 25 Hz, Hanning window with 50% overlap) is depicted in Fig. 4.1

4.2 Equalization of sentence difficulty

The sentence intelligibility was equalized without modifying the sentence level, but by selecting a subset of sentences with equal intelligibility scores. In two phases, Wable (2001) reduced the number of sentences to 160, based on the speech recognition scores in noise of hearing-impaired subjects at different SNRs. In total 433 subjects participated, with Pure Tone Averages ranging from 20 to 83 dB HL. For this final selection of sentences the number of syllables per sentence ranged from 6 to 15, with an average of 10.

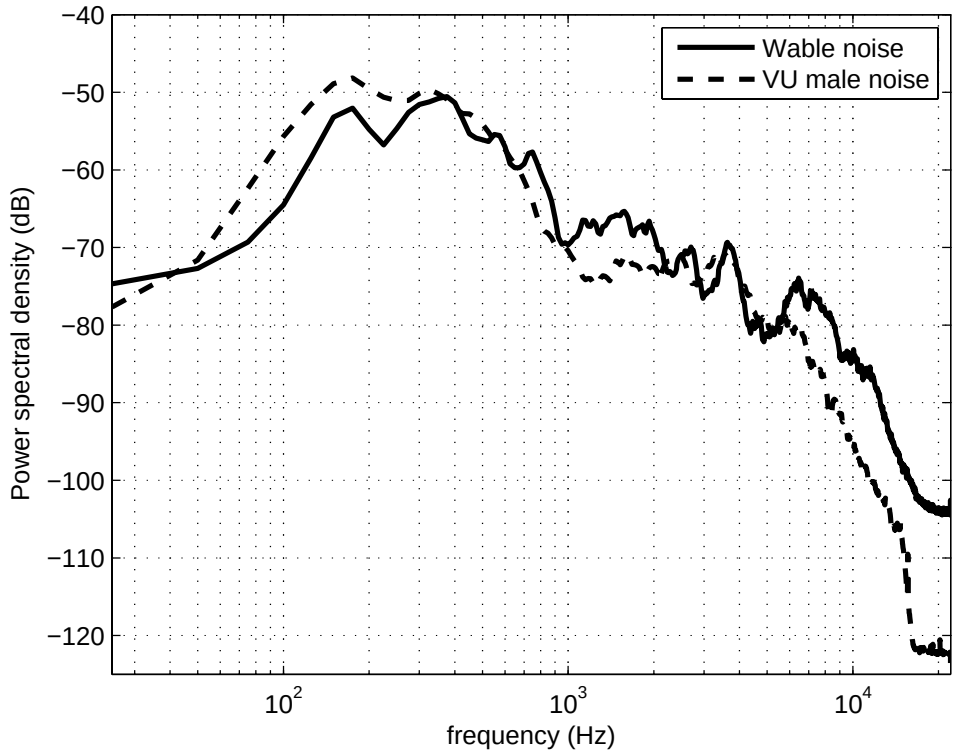

Fig. 4.1 *The long-term power spectrum of the speech-weighted noise used for the French sentence test (solid line). For comparison, the spectrum of Versfeld's masker (dashed line) is shown (Versfeld* et al.*, 2000).*

4.3 Formation of sentence lists

The remaining sentences provided by Wable (2001) were, within this project, distributed into lists in order to obtain the most consistent mean intelligibility scores. Therefore, the sentences were evaluated at SNR levels of 0, -5 and -10 dB in a speech-weighted noise presented at 65 dB SPL. These SNR levels were chosen based on the results of a pilot study. Twelve normal-hearing Francophone Belgian subjects, with hearing thresholds of 15 dBHL or better for all octave frequencies between 125 and 8000 Hz, participated. The sentences were evaluated using sentence scoring. Based on these results, sentences were combined into 10-sentence lists, with the aim to minimize the scatter of average SRTs and slopes between lists. In this way, 14 test lists of 10 sentences and one training list of 20 sentences were created (see appendix D for an example). Creation of lists was thus based on intelligibility equalization rather than phoneme equalization.

4.4 Development of norms and reliability, and estimation of the psychometric function

The 10-sentence lists were evaluated by a new series of listening tests. Speech reception thresholds (SRTs) were obtained in 20 Francophone normal-hearing adults. Ten subjects were French, the other ten were Belgian. Their

hearing thresholds were equal to or better than 25 dBHL for all octave frequencies between 125 and 8000 Hz. The sentence material was presented through headphones (TDH-39) with an adaptive test procedure, in a stationary speech-weighted noise with a fixed level of 65 dBSPL. Starting at 51 dBSPL the level of the first sentence of each list was increased in steps of 2 dB until the sentence was identified correctly. Subsequently, the intensity level of the sentences within the list was adjusted adaptively in steps of 2 dB, with a 1-down, 1-up procedure to target the 50% intercept.

Average SRTs are shown in Tab. 4.1. The SRT is calculated as the average of the presentation levels of the following sentence for the 5^{th} to the 10^{th} sentence in the adaptive procedure. It is expressed as the SNR required for 50% speech intelligibility. Standard deviations (SD) across subjects are shown. There is a small, but significant difference in SRT of 0.7 dB SNR between both groups of listeners (independent-samples *t*-test, p=0.027). The SRT of the psychometric function, as well as the slope S_{50}, are also calculated based on fitted data. The precision is expressed as the quadratically averaged standard errors (SE) of the fit to the data of each subject.

The psychometric function is estimated based on the individual scores at the different presentation levels within the adaptive procedure. The SRTs and slopes at 50% scores are calculated based on non-linear regression fits to a logistic function of the psychometric function of each subject (see Tab. 4.1. and Fig. 4.2). The average SRTs calculated from the adaptive tests or deduced from the fitted data are similar (paired-samples *t*-test, p=0.178). For the fitted SRT values the difference between both groups is also significant (independent-samples *t*-test, p=0.047). The slope S_{50} of the psychometric function is not significantly different between both subject groups (independent-samples *t*-test, p=0.224).

Tab. 4.1 *SRT and slope of the psychometric function of the French sentence test.*

		Average SRT (dB SNR)	SD		
Adaptive procedure	France	-7.8	0.6		
	Belgium	-7.1	0.7		
	Total	-7.4	0.7		
		Average SRT (dB SNR)	**SD**	**Average slope (%/dB)**	**SE**
Fitted,	France	-7.9	0.2	21.2	2.5
based on adaptive data	Belgium	-7.3	0.2	19.2	2.7
	Total	-7.6	0.2	20.2	2.6

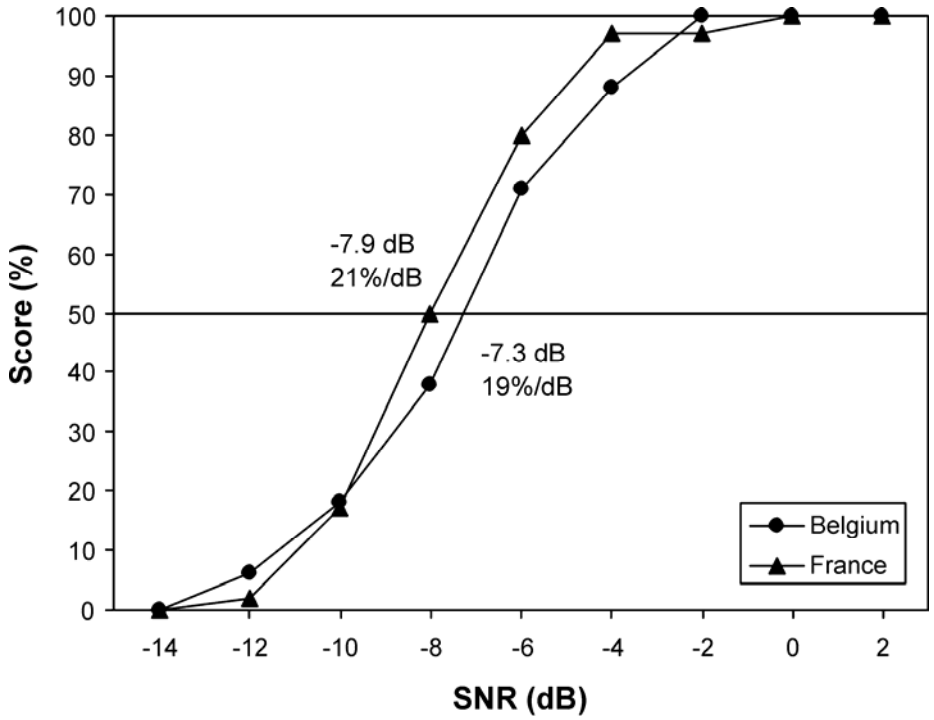

Fig. 4.2 *The psychometric functions of the sentence test for French and Belgian listeners.*

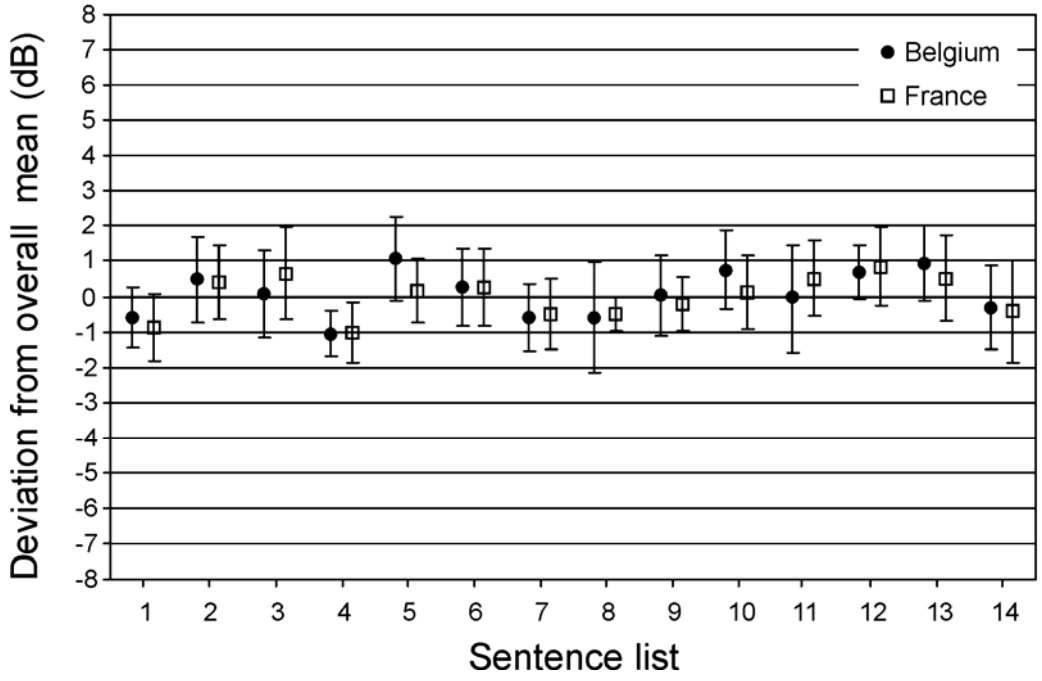

Fig. 4.3 *Differences between means for list-specific SRTs and the mean SRT across all lists for the French (n=10) and Belgian (n=10) listeners. Error bars show ± 1 standard deviation across listeners for each mean.*

4.5 List equivalency

Fig. 4.3 illustrates the variation in SRT of the 14 lists determined with the adaptive method (not the fitted values). The values are plotted in terms of a deviation score from the overall mean, together with their respective standard deviations. For the Belgian as well as for the French subjects, all list means deviate maximum 1.1 dB from the overall mean. This is comparable to the English HINT (Nilsson et al., 1994) and to the Dutch LIST (van Wieringen & Wouters, 2007).

4.6 Reliability

The reliability of the SRT measurements was determined by considering the quadratically averaged within-subject standard deviation of repeated measurements (Plomp and Mimpen, 1979). The within-subjects standard deviation is 1.0 dB and 1.1 dB for French and Belgian subjects respectively. These values for a list of ten sentences are in the same order of magnitude as within-subject standard deviations of other speech materials: 1.2 dB for 10 LIST sentences per list (van Wieringen & Wouters, 2007), 1.1 dB for 13 VU sentences per list (Versfeld et al., 2000), 1.1 dB for 12 HINT sentences per list (Nilsson et al., 1994) & 1.1 dB for 20 Canadian French sentences per list (Vaillancourt et al., 2005).

5 Discussion

Tab. 5.1 compares the properties of various sentence intelligibility tests obtained by different authors. As can be seen, there are significant differences in the masking signal, scoring method, predictability and psychometric function parameters. SRT values vary across the tests from –8.4 dB SNR (Wagener, 2003) to –2.9 dB SNR (Nilsson et al., 1994). SRTs are typically lower for semantically unpredictable sentences and word scoring (Hagerman, 1982; Wagener et al., 1999, 2003), while SRT are highest for Plomp-type tests and sentence scoring (Plomp, 1979; Smoorenburg, 1992; Versfeld et al., 2000). This difference is both the consequence of using a limited word material like 50 words (Hagerman, 1982; Wagener et al., 1999, 2003) that reduces SRT and the scoring method (sentence scoring increases SRT).

The Polish sentence test, using the 20-sentence lists and sentence scoring, results in a mean SRT of –6.1 dB SNR and a mean S_{50list} of 25.5 %/dB. With the word scoring method the mean SRT and mean S_{50list} were –7.4 dB SNR and 22.4 %/dB, respectively. As expected, SRT and S_{50list} values resulting from the sentence scoring are higher than those following from the word-based scoring. For the 13-sentence test the average SRT and S_{50list} were equal to -6.1 dB SNR and 25.6 %/dB, respectively. Test-retest experiments have shown that the Polish sentence test is highly reliable and brings accurate and repeatable speech intelligibility results.

The difference between the parameters of the Polish materials and previously designed tests for other languages is the steepness of the psychometric function. Although, S_{50} obtained with word scoring is comparable to S_{50} of a very similar German Göttingen test (Kollmeier and Wesselkamp, 1997), the value of this parameter is higher than the mean S_{50} of the Dutch tests (Plomp, 1979; Smoorenburg, 1992; Versfeld et al., 2000). It seems likely that the difference in steepness is affected by many factors such as the type of masking noise, the method of intelligibility calculation and the linguistic structure of the speech material.

The influence of the masker type, and particularly its temporal structure, on the steepness of the psychometric function is now under detailed investigation (Ozimek et al., 2007). Preliminary data have shown that for the same speech material and sentence scoring, the babble noise masker leads to a mean steepness S_{50list} of about 25.6 %/dB and the speech-shaped white noise to S_{50list}=22.5 %/dB (i.e. comparable value to other languages). If the same type of masking signal (speech-shaped noise) and scoring method (sentence scoring) were used (see Tab.5.1, rows 13 and 15) in the intelligibility experiment, the slopes of psychometric functions obtained for the Polish (S_{50list}=22.5 %/dB) and French tests (S_{50}=21.2 %/dB) were similar. Close similarity can also be obtained comparing Polish and German sentence tests for word scoring method and the same maskers (Polish: S_{50list}=22.4%/dB (row 12); German: S_{50list}=19.2%/dB (row 4)).

Tab. 5.1 *Properties of sentence intelligibility obtained by different authors*

No.	*Language*	*Masker*	*Scoring method*	*SRT [dB SNR]*	*Mean steepness [%/dB]*	*Speaker*	*Remarks*
1	Dutch (Plomp and Mimpen, 1979)	speech-shaped stationary noise	sentence scoring	-4.5	15.9*	female	semantically predictable sentences (Plomp-type test)
2	Dutch (Smoorenburg, 1992)	speech-shaped stationary noise	sentence scoring	-3.7	17.7*	male	Plomp-type test
3	Dutch (Versfeld et al., 2000)	individually shaped white noise	sentence scoring	-4.1	16.3*	male, female	Plomp-type test, signals presented via loudspeaker
4	German (Kollmeier and Wesselkamp, 1997)	superposition of monosyllabic words	word scoring	-6.2	19.2**	male	Plomp-type test ('Götingen test')
5	German (Wagener, 1999)	babble noise	word scoring	-7.1	17.1**	male	semantically unpredictable sentences ('Oldenburg Satztest, OLSA), fixed grammatical structure
6	Danish (Wagener, 2003)	babble noise	word scoring	-8.4	13.2**	female	semantically unpredictable sentences ('DANTALE II' test), fixed grammatical structure
7	Swedish (Hagerman, 1982)	babble noise	word scoring	-8.1	16.0*	female	semantically unpredictable sentences, fixed grammatical structure
8	American English (Nilsson, 1994)	speech-shaped stationary noise	sentence scoring	-2.9	-	male	Plomp-type test, adaptive procedure
9	Dutch (for Flanders) (Wieringen and Wouters, 2007)	speech-shaped stationary noise	key-words scoring	-7.8	17.0*	female	Plomp-type test, a constant, stimuli method, adapative procedure
10	Polish	babble noise	sentence scoring	-6.1	25.6**	male	Plomp-type test (37 lists of 13 sentences)
11	Polish	babble noise	sentence scoring	-6.1	25.5**	male	Plomp-type test (25 lists of 20 sentences)
12	Polish	babble noise	word scoring	-7.5	22.4**	male	Plomp-type test (25 lists of 20 sentences)
13	Polish	speech-shaped stationary noise	sentence scoring	–6.7	22.5**	male	Plomp-type test (25 lists of 20 sentences)
14	Polish	speech-shaped stationary noise	word scoring	–7.8	18.6**	male	Plomp-type test (25 lists of 20 sentences)
15	French (France)	speech-shaped stationary noise	sentence scoring	–7.8	21.2*	female	Plomp-type test
16	French (Belgium)	speech-shaped stationary noise	sentence scoring	–7.1	19.2*	female	Plomp-type test

* mean steepness S_{50mean},** list-specific steepness S_{50list} determined according to formula (5)

Although the babble noise and the speech-shaped stationary noise have the same power spectra, the parameters of the psychometric functions determined for the Polish test and the above mentioned maskers are different, i.e. S_{50list} is 25.5 %/dB for the babble noise, while for the speech-shaped noise it is 22.5 %/dB. This implies that some temporal factors might have influenced speech intelligibility. The babble noise revealed larger temporal envelope fluctuations than the speech-shaped noise. These fluctuations may have masked more effectively some phonemes at relatively low SNRs. Consequently, these phonemes were not perceived correctly, which brought about a decrease in the intelligibility score (sentence scoring method). This mechanism does not seem to influence the speech intelligibility score at a relatively high SNR (i.e. –1 dB), when the speech is clearly understood regardless to the masker fluctuation. Therefore, these envelope fluctuations in the masker lead to an increase in S_{50} and SRT values.

It was additionally assumed that some differences in perceptual quantities characterizing both maskers might also have affected the speech perception in noise. Accordingly, parameters as fluctuation strength, loudness, roughness and sharpness were analyzed using the Artemis Software. The obtained values of these parameters turned out to be very similar for both the maskers, except for roughness. For the babble noise the determined roughness was larger (r=2.55 asper) than for the speech-shaped noise (r=2.19 asper). This fact was also supported by subjective assessments of the maskers. It seems that higher roughness of the babble noise might worsen the listening comfort and decrease the speech intelligibility score at low SNRs. This led to an increase in steepness of the psychometric function for the babble noise. However, this is not the case at higher SNRs since speech is well perceived then.

The last factor that might differentiate the steepness obtained by the respective authors is the way of determining its value. As can be seen from the table, some authors tend to characterize their materials in terms of an average steepness S_{50}, while the Polish and German sentence tests are characterized by list specific steepness S_{50list} computed according to formula (5) (see notes given in bottom of Tab.5.1).

For the French sentence test, the number of sentence lists (N=14) is rather limited. However, the psychometric function of the test material has a relatively steep slope and can provide accurate speech intelligibility scores in noise. There is a small but significant difference in SRT between the French and Belgian listeners. This has two possible explanations. The speech material is uttered by a French male voice, which could be slightly more difficult for the Belgian listeners. However, within France the accent can also differ substantially. A second explanation could be the fact that six out of ten Belgian subjects were bilingual, and they might have more difficulties with speech recognition in noise than monolingual listeners (Grosjean, 1989).

The suitability of the French test for severely hearing-impaired listeners or cochlear implant users is limited, especially because of the demand to repeat

the entire sentence. Moreover, some of the sentences are rather long, up to 15 syllables per sentence.

The test battery for the assessment of speech intelligibility in noise will soon be completed with a closed-set sentence test for French (MATRIX-F) and a digit triplet test suited for telephone screening for French and Polish. These tests are under development within WP1 of the HEARCOM-project.

6 Dissemination and Exploitation

The speech materials and the measurement procedures described in this report will be mainly of interest to audiologists and ENT physicians to test intelligibility of hearing-impaired people. The tests developed will be made available for use via the Internet (HearCom portal). The way in which they will be exploited via the Internet is under preparation. Several possible methods of exploitation are considered.

Results will be presented at conferences and in international journal publications.

7 Conclusions

Summarizing, the main purpose of this study, i.e., to prepare Polish and French sentence materials for accurate intelligibility measurements under noisy conditions, has been met. Both test materials have steep psychometric functions, which increases the accuracy and reliability of the test results. In order to accurately determine an individual SRT value in a real clinical situation, it is sufficient to collect the intelligibility scores for several randomly selected lists presented at different SNRs or preferably to perform one adaptive test.

In view of the required harmonisation across European countries, the Polish and French sentence tests are a welcome addition to the existing sentence tests for Danish, Dutch, Swedish, (British) English and German.

8 References

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

Brand, T. and Kollmeier, B. (**2002**). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," Journal of Acoustical Society of America 111, 2801-2810.

Content, A., Mousty, P., and Radeau, M. (**1990**). "Brulex. Une base de données lexicales informatisée pour le Français écrit et parlé," L'Année Psychologique 90, 551-566.

Grosjean, F. (**1989**). "Neurolinguists, beware! The bilingual is not two monolinguals in one person," Brain & Language 36, 3-15.

Hagerman, B. (**1982**). "Sentences for testing speech intelligibility in noise," Scand. Audiol. 11, 79-87.

Jassem, W. (**1973**). "Podstawy akustyki fonetycznej," (PWN, Warszawa).

Kalikow, D. N., Stevens, K. N., and 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, 1337-1351.

Kohlrausch, A., Fassel, R., van der Heijden, M., Kortekaas, R., van de Par, S., and Oxenham, A. (**1997**). "Detection of tones in low-noise noise: Further evidence for the role of envelope fluctuations," Acustica united with Acta-Acustica 83, 659-669.

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

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

Martin, M. (**1997**). *Speech Audiometry*, Whurr Publishers Ltd, Berlin.

Nilsson, M., Soli, S. D., and 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 Acoustical Society of America 95, 1085-1099.

Ozimek, E., Kutzner, D., Sęk, A., and Wicher, A. (**2007**). "An influence of masker type on sentence intelligibility for Polish language," (in preparation).

Plomp, R. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Audiology 18, 43-52.

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

Plomp, R., and Mimpen, A. M. (**1979b**). "Speech-reception threshold for sentences as a function of age and noise level," Journal of Acoustical Society of America 66, 1333-1342.

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

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

Runge, C. A., and Hosford-Dunn, H. (**1985**). "Word Recognition Performance with Modified CID W-22 Word Lists," J. Speech Hear. Res. 28, 355-362.

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

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

Studebaker, G. A., Sherbecoe, R. L., McDaniel, D. M., and Gwaltney, C. A. (**1999**). "Monosyllabic word recognition at higher-than-normal speech and noise levels," Journal of Acoustical Society of America, 2431-2444.

Vaillancourt, V., Laroche, C., Mayer, C., Basque, C., Nali, M., Eriks-Brophy, A., Soli, S.D., and Giguere, C. (**2005**). "Adaptation of the HINT (hearing in noise test) for adult Canadian Francophone populations," International Journal of Audiology 44, 358-69.

van Wieringen, A., Wouters, J (**2007**). "LIST and LINT: sentences and numbers for qualifying speech understanding in severely impaired listeners for Flanders and the Netherlands", International Journal of Audiology, in press.

Versfeld, N. J., Daalder, L., Festen, J. M., and 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.

Wable, J. (**2001**). "Normalisation d'un test de reconnaissance de la parole dans le bruit chez le sujet déficient auditif," Cahiers de l'Audition 14, 29-38.

Wagener, K., Brand, T., and Kollmeier, B. (**1999**). "Development and evaluation of a German sentence test I-III: Design, Optimalization, and Evaluation of the Oldenburg sentence tests (in German)," Z.Audiol 38, 4-15, 44-56, 86-95.

Wagener, K.C. and Brand, T. (**2005**). "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters," International Journal of Audiology 44, 144-157.

Wagener, K.C., Brand, T., and Kollmeier, B. (**2006a**). " The role of silent intervals for sentence intelligibility in fluctuating noise in hearing-impaired listeners," International Journal of Audiology 45, 26-33.

Wagener, K.C., Eeenboom, F., Brand, T., and Kollmeier, B. (**2006b**). "Ziffern-Tripel-Test: Sprachverstandlichkeitstest über das Telefon," DGA. 8. Jahrestangung Göttingen 2005. TagungsCD, ISBN 3-9809869-4-2

Wagener, K.C., Bräcker, T., Brand, T., and Kollmeier, B. (**2007**). "Evaluation des Ziffern-Tripel-Tests über Kopfhörer und Telefon," DGA. 9. Jahrestangung Köln 2006. TagungsCD, ISBN 3-9809869-5-0

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

Wouters, J., Literie L., van Wieringen A. (**1999**). "Speech Intelligibility in Noisy Environments with One- and Two-microphone hearing aids," Audiology 38, 91-98.

Appendix A. Example of a 20-sentence list of the Polish test for the sentence scoring

1) Nie pomogli nam przy budowie (They didn't help us with building)
2) Zapiął jej starannie kurtkę (He buttoned up carefully her jacket)
3) Starała się pomagać ojcu (She tried to help her father)
4) Teraz nie mają na to czasu (They have no time now)
5) Mimo to doszło do konfliktu (It came to a conflict, however)
6) Dlatego proszę cię o to (So, I am asking you to do it)
7) Listonosz wszedł do werandy (A postman came up to the veranda)
8) To jest różnica między nami (This is the difference between us)
9) Wziął latarkę i tam poszedł (He took a torch and went there)
10) Czaiło się jakieś zwierzę (There was any animal lurking)
11) Nikomu to się nie podoba (Nobody likes it)
12) Przysunął do siebie krzesło (He moved the chair towards himself)
13) Na czele grupy stoi wódz (A group is led by a commander)
14) Można też kupować na raty (One may pay in instalments)
15) Te sprzed roku były słabe (These from the last year were poor quality)
16) Doktor roześmiał się wesoło (The doctor laughed joyfully)
17) Wskazał palcem na budynek (He pointed at a building)
18) Zawsze chcieli nam dorównać (They have always wanted to catch up with us)
19) Rzadko chodzą do teatru (They rarely go to a theatre)
20) Nie sposób ich nie zauważyć (It is impossible to miss them)

Appendix B. Example of a 20-sentence list of the Polish test for the word based scoring

1) Znowu ta winda nie działa (This lift doesn't work again)
2) Najpierw zwabiło go światło (At first he was lured by the light)
3) Wracam późno do hotelu (I come back late to the hotel)
4) Taśma przesuwa się ciszej (The tape moves silent)
5) Nie znamy wielu powodów (We don't know many reasons)
6) To były nasze pomysły (These were our ideas)
7) Dla ciebie była zawsze dobra (She has always been good for you)
8) Na czele grupy stoi wódz (A group is led by a commander)
9) Ale potrzebny był następny (But next one was needed)
10) Były pewnie zbyt głęboko (They were probably too deep)
11) Wskazał palcem na budynek (He pointed at a building)
12) Podbiegli na boso do okna (They ran barefoot to the window)
13) Jeszcze zjadłem ten kawałek (I ate this piece as well)
14) Spogląda bezmyślnie w okno (She/he looks thoughtlessly trough a window)
15) Musimy czytać ten artykuł (We have to read this article)
16) Przeważnie jednak ludzie milczą (Usually people keep quiet)
17) Ta drukarka już nie działa (This printer does not work any more)
18) Teraz ważny jest ten obraz (This painting is important now)
19) Trwało to ułamek sekundy (It lasted fraction of a second)
20) Wiele dzieci tam się bawi (There are many children playing there)

Appendix C. Example of a 13-sentence list of the Polish test for the sentence scoring

1) Naprawdę jestem pod wrażeniem (I am really impressed)
2) Cała rodzina była w lesie (The whole family was in a forest)
3) Na próbie byli zmęczeni (They were tired during the rehearsal)
4) Pod palcami czuł chłodny dotyk (He felt cold touch under fingers)
5) Sam dyrektor rękę podniósł (The boss himself raised his hand)
6) Wzięli z domu dwie latarki (They took two torches from home)
7) Miała dzisiaj znowu ten sen (She had the same dream today)
8) Mogą podnieść swoje książki (They can lift their books)
9) Można też kupować na raty (One may pay in instalments)
10) Samochód podjechał znienacka (A car approached suddenly)
11) Czaiło się jakieś zwierzę (There was an animal lurking)
12) Trwało to ułamek sekundy (It lasted fraction of a second)
13) To ty zabrałeś te koce (You have taken these blankets)

Appendix D. Example of a 10-sentence list of the French test

1) Ce pantalon est trop cher pour moi (These trousers are too expensive for me)

2) Son garage a été complètement inondé (His garage has been completely flooded)

3) Il ne trouve pas de pantalon à sa taille (He can not find these trousers in his own size)

4) Elle se regarde dans le miroir (She is looking at herself in the mirror)

5) Elle a appris le karaté pour savoir se défendre (She has learned karate to be able to defend herself)

6) Ce geste m'a fait plaisir de sa part (The favour from his end pleased me)

7) Chaque matin je me rends à la gare (Every morning I’m going to the train station)

8) Il est content de son nouveau travail (He is happy of his new job)

9) Le berger rassemble son troupeau (The sheppard gathers his sheep)

10) Ce film d'horreur donne des cauchemars (This horror film leaves me with nightmares)