Zur Homepage der Dissertation

Factors Influencing Sentence Intelligibility in Noise

Vom Institut für Physik an der
Fakultät für Mathematik und Naturwissenschaften
der Universität Oldenburg
zur Erlangung des Grades einer
Doktorin der Naturwissenschaften (Dr. rer. nat.)
angenommene Dissertation

Kirsten Carola Wagener
geb. am 19.8.1973
in Göttingen

Erstreferent: Prof. Dr. Dr. Birger Kollmeier
Korreferent: Prof. Dr. Volker Mellert
Tag der Disputation: 13. November 2003

ABSTRACT

The general goal of this thesis is to increase comparability, accuracy, and diagnostical benefit of sentence intelligibility tests within one language and to give suggestions for realizing comparable sentence intelligibility tests across languages.

Most sentence intelligibility tests are either composed of meaningful (everyday) sentences or syntactically fixed, but semantically unpredictable sentences. The second type tests were investigated in detail in this thesis that were based on the original Hagerman sentences (Hagerman, 1982) and the Oldenburg sentence test (Wagener *et al.*, 1999c; Wagener *et al.*, 1999a; Wagener *et al.*, 1999b). The development, optimization, and evaluation of the Danish DANTALE II test (Wagener *et al.*, 2003) is presented which closely resembles the Oldenburg sentence test. In order to test the comparability of this type of sentence intelligibility tests across languages, the average speech reception threshold (SRT), slope of the intelligibility function and spread across test lists was obtained and compared for these three tests. Only the Danish test yields a lower intelligibility function slope than both other tests, whereas a high comparability is maintained across test languages for most of the other parameters considered.

A large number of tests was conducted with normal listeners employing a systematic parameter variation to study further the influence of various test parameters on the expected outcome of the sentence test and its comparability even within one language (i. e. German). This includes the usage of test lists in quiet using sentence tests that were originally introduced and optimized for speech tests with interfering noise. In addition, measurement parameters like noise presentation level, type of interfering noise, and type of presentation were varied both with normal–hearing and hearing–impaired subjects. Fluctuating interfering noises were found to differentiate best between different degrees of hearing loss. Therefore, these noises were investigated in more detail.

In order to better understand the mechanisms of speech perception in fluctuating noise, speech intelligibility in such noises was predicted with different approaches. The most successful approach models speech perception in fluctuating noise by first computing the expected intelligibilities of sub–word units at the respective signal–to–noise ratio, considering the context effects of the sub–word units. Then the word intelligibilities (or error probabilities, respectively) are computed by multiplying the particular error probabilities for the sub–word units. Finally, the sentence intelligibility is computed by averaging across the words.

Taken together, the sentence tests and measurement procedures considered here both experimentally and by means of theoretical models appear to yield the highest practically achievable accuracy and comparability within and across languages. This might therefore help to harmonize speech audiometry across both laboratories, clinics, and languages.

Contents

Chapter 1

Introduction

Hearing is a dominant sense in our live. It helps us to recognize danger or friends, to communicate, to enjoy music. It is a very accurate sense. The healthy auditory system can differentiate between the roaring of the sea and traffic noise or between different melodies in a polyphonic symphony. We are able to understand our interlocutor within a noisy environment, even within a crowd of people which is talking with a higher level. These situations are most difficult for hearing–impaired listeners. Therefore, a hearing–impairment is often first recognized during communications within interfering noise (the so–called 'cocktail party situation').

Speech intelligibility tests in noise should be performed additionally to pure–tone audiograms during diagnosis and rehabilitation, because recognizing speech within interfering noise is more complex than perceiving pure tones and represents more the difficult everyday situation of the listener.

In contrast to speech recognition in quiet that shows large differences between normal–hearing and hearing–impaired listeners, speech recognition in noise only shows small differences between different degrees of hearing loss. Therefore, much effort was spent in the past on developing appropriate procedures to determine speech intelligibility in noise with high accuracy. Most approaches, however, led to a particular test set–up in one laboratory with results that could not directly be compared to those from other laboratories. The current thesis thus is concerned with improving the accuracy and applicability of speech tests. The main attention of this thesis is the comparability of speech test results and accuracy within and across languages.

As determining speech intelligibility is a statistical estimate, the accuracy of the intelligibility value is a function of the number of test items per measurement. Thus,

sentence tests are preferable for such purposes, because several words are tested within each sentence within a short time frame.

Most sentence tests can be divided into two different groups by the type of sentence material. First, high predictable everyday sentences as the German Göttingen sentence test (Kollmeier and Wesselkamp, 1997), and HSM test (Schmidt *et al.*, 1997), the Dutch Plomp and Mimpen sentences (Plomp and Mimpen, 1979), the sentences by Smoorenburg (1992), and Versfeld *et al.* (2000), and the American HINT test [hearing in noise test, Nilsson *et al.* (1994)]. The advantage of these tests is that there is no training effect when using the test lists only once. The disadvantage, however, is that the test lists usually cannot be used twice with the same subject, because the meaningful sentences can easily be memorized or words can be guessed from the context. This would generate an incorrect low SRT result. A repeated measurement with the same test list is not possible until a sufficient period of time has passed (i. e., half a year or even longer). As the amount of test lists is limited, these sentence tests are not suitable when many speech intelligibility measurements have to be performed, e.g. during hearing–aid or cochlear implant fitting, or in research. In order to overcome this problem, unpredictable sentence tests were developed. These tests consist of syntactically fixed, but semantically unpredictable (nonsense) sentences, i. e., sentences with a fixed grammatical structure but using words that do not necessarily make sense in their respective combinations. Hagerman (1982) first developed this test format for Swedish, and Wagener *et al.* (1999c) further adapted this format to the German Oldenburg sentence test and the Danish DANTALE II (Wagener *et al.*, 2003).

In order to establish comparable speech tests across different languages, this type of sentence test seems suitable and has been studied in detail in this thesis. In Chapter 2, the construction of one representative of such sentence tests is presented in detail. This Danish sentence test is an adaptation of the German Oldenburg sentence test. The chapter describes the design, optimization, and evaluation of the test in comparison to the German test [published in the International Journal of Audiology (Wagener *et al.*, 2003)].

To compare these both sentence intelligibility tests with the Swedish original test by Hagerman (1982) that differ in several details, Chapter 3 presents a re–analysis of Hagerman's original data and compares it to the respective data for the Oldenburg sentence test and DANTALE II test.

Although sentence tests often were intended and optimized to yield similar intelligibility in noise with the different test lists, these tests are sometimes used without any

interfering noise (Baumann, 2001; Brand and Kollmeier, 2002b). This differs from the usage of the tests in noise discussed so far, since speech intelligibility in quiet is mainly limited by the internal noise of the subject which also determines the individual hearing threshold for tones in quiet. The observed variability in speech intelligibility for a given subject is therefore determined by the variances of the external noise (measurements with interfering noise) or the variances of the internal noise (measurements in quiet). It is not clear if both types of variances influence the results of a given speech test in the same way and hence warrant that the test lists designed to be comparable in noise are also comparable in quiet. This problem is independent of language or speech test type. In Chapter 4, this problem was exemplarily investigated for two different German sentence tests. As one example of high predictable sentence intelligibility tests the German Göttingen sentence test and as one example of low predictable sentence intelligibility tests the German Oldenburg sentence test was used to investigate whether the homogeneity of test lists also holds for quiet test conditions.

Apart from the different languages and presentation in quiet or noise, there is a large number of language–independent methods to determine speech intelligibility in noise with one test. The measurement procedures are defined by different parameters. Hence, comparability of speech tests results is highly dependent on the influence of these parameters on the results. The aim of Chapter 5 is to investigate those parameters that typically vary between different test methods and different researchers. Critical factors that influence the test result, should be separated from non–critical factors. In this way, this study is intended to contribute to a harmonization in speech audiometry. The parameters presentation level, type of interfering noise, and different presentation modes (i. e. varying the speech or the noise level in an adaptive procedure and using interrupted or continuous interfering noise) were investigated. In order to exclude other influences, this extensive investigation was performed for one language (German) with the same test and subjects (normal–hearing and hearing–impaired) throughout the entire study.

As confirmed in Chapter 5 and by other authors (Versfeld and Dreschler, 2002; Hagerman, 2002; Eisenberg *et al.*, 1995; Festen and Plomp, 1990), fluctuating interfering noises are highly suitable for testing speech intelligibility in noise, as these noises differentiate best between different degrees of hearing loss. The ICRA group (International Collegium of Rehabilitative Audiology) presented a highly fluctuating noise as standard noise for investigating digital hearing aids and intelligibility in fluctuating noise (Dreschler *et al.*, 2001). This so–called *icra5* noise was also used as interfering

noise in Chapter 5. However, a disadvantage of this noise is that it includes silent intervals with durations up to 2 s. Hence, entire sentences of speech tests sometimes fall into these silent intervals. In order to eliminate this disadvantage, two noises were generated that limit the maximum silent interval duration of the *icra5* noise to 250 ms or 62.5 ms, respectively. In Chapter 6 the dependency of speech intelligibility in fluctuating noise on maximum pause duration is considered as a function of the individual hearing loss in hearing–impaired subjects. Since speech intelligibility in pause–limited fluctuating noise correlates both with speech intelligibility in quiet and in stationary noise, a prediction of speech intelligibility in fluctuating noise should be possible with these quantities.

Such a prediction is undertaken in the final chapter of this thesis: In order to understand the process of speech intelligibility in fluctuating interfering noise in more detail, Chapter 7 compares several approaches to model speech intelligibility in fluctuating noise. The most successful approaches base on Plomp's model about hearing loss for speech (Plomp, 1978). The input parameters of the speech perception model are the level histogram or level sequence of the fluctuating interfering noise and both the individual SRT in quiet and in stationary noise. First, the validity of Plomp's model for the results with the Oldenburg sentence test is evaluated. Further on, the predictions are compared to a large database of SRT measurements in fluctuating noise with different types and degrees of hearing loss and to predictions with the speech intelligibility index SII (Appendix A).

It will be shown in this thesis that it is possible to obtain comparable sentence intelligibility results even across languages and that there are measurement procedures that clearly represent the differences between hearing–impairments. Hence, the measurement procedures proposed and compared in this study should supply the audiologist with sophisticated tools to diagnose and rehabilitate the hearing–impaired listener.

Chapter 2

Design, Optimization, and Evaluation of a Danish Sentence Test in Noise

ABSTRACT

The Danish sentence test DANTALE II was developed in analogy to the Swedish sentence test by Hagerman and the German Oldenburg sentence test as a new Danish sentence test in noise to determine the speech reception threshold in noise (SRT, i. e. the signal–to–noise ratio (SNR) that yields 50% intelligibility). Each sentence is generated by a random combination of the alternatives of a base list. This base list consists of 10 sentences with the same syntactical structure (*name, verb, numeral, adjective, object*). The test sentences were recorded and segmented in such a way that the coarticulation effects were taken into account in order to achieve a high perceived sound quality of the resynthesized sentences: 100 sentences were recorded, therefore, each coarticulation between each word and the 10 possible following word alternatives were recorded, and the correct coarticulation was used to generate the test sentences. Word–specific speech recognition curves were measured for each recorded word to optimize the homogeneity of the speech material and the measurement accuracy. Level corrections of particular words and a careful selection of the test lists produced a noticeable reduction in the variation in the distribution of word–specific SRT (standard deviation 1.75 dB instead of 3.78 dB). Therefore, the slope of the total intelligibility function was expected to increase from 8.3 %/dB (raw test material)

to 13.2 %/dB (after modification). These theoretical expectations were evaluated by independent measurements with normal–hearing subjects, and, for the most part, confirmed. The reference data for the DANTALE II are: SRT = −8.43 dB SNR; slope at SRT, $s50 = 13.2\%$/dB. The training effect was 2.2 dB and could be reduced to less than 1 dB, if two training lists of 20 sentences were performed prior to data collection.

2.1 INTRODUCTION

Holding conversations in noisy environments represents a major problem for many hearing–impaired listeners. Therefore, recent speech audiometry assesses speech recognition in noise, with whole sentences as test material. The requirements for new audiological measurements have been formulated by the European project NATASHA (Network And Tools for the Assessment of Speech and Hearing Ability) in order to achieve similar audiological measurement procedures across Europe. These include requirements for new speech tests. One possible sentence test format that has the advantage in principle of being very similar across different languages is the use of syntactically fixed, but semantically unpredictable (nonsense), sentences, i. e. sentences with a fixed grammatical structure but using words that do not necessarily make sense in their respective combinations. Hagerman (1982) developed this test format for Swedish, and Wagener et al. (1999) further developed this format into the 'Oldenburger Satztest' (Oldenburg sentence test). One major advantage of these tests is their repeated usability with the same subject, as any sequence of these mostly nonsense sentences of the same general structure is unlikely to be memorized (in contrast to everyday sentences, as used in the Göttingen sentence test (Kollmeier and Wesselkamp, 1997), the Plomp sentences (Plomp and Mimpen, 1979), and the Hearing In Noise Test (Nilsson *et al.*, 1994)). Therefore, these tests are very useful for hearing aid evaluation. They are also useful for the fitting of cochlear implants, because the sentences are spoken relatively slowly, and the speech material consists of only 50 well–known words. These tests can be used for children, who are able to memorize five–word sentences.

The speech reception threshold, SRT (signal–to–noise ratio (SNR) that yields 50% intelligibility), in noise shows relatively small differences between normal–hearing and hearing–impaired subjects. In addition, the differences between different hearing aids or different hearing aid fittings are rather small. Therefore, a sentence test in noise should be capable of detecting even small changes in the 'effective' SNR by translating

them into large changes in intelligibility. This means that the intelligibility functions of the test lists have to be very steep. Also, the differences in intelligibility across different test lists have to be smaller than the effect to be measured, which calls for high comparability and low variability across test lists. Much effort has therefore to be invested to achieve high accuracy and good comparability of the different test lists.

In this article, a Danish version of this sentence test is presented (Dantale II). The DANTALE II was developed and produced as a Danish–German co–production to ensure maximum comparability with the Oldenburg sentence test. The word material and the speaker have been selected by the Danish authors (Josvassen and Ardenkjær). They also performed the cutting of the recorded material in Oldenburg, where the test was recorded in a way comparable to that used for the recordings of the Oldenburg sentence test. The Danish authors also performed the auditory quality check of the recorded and resynthesized speech material to ensure maximum subjective quality of the sentences for native Danish listeners. The German author (Wagener) generated the test sentences, produced the experimental design and performed all data analyses. All speech intelligibility measurements with normal–hearing subjects were performed at the Department of OtoRhinoLaryngology, Head and Neck Surgery, Copenhagen University Hospital, Rigshospitalet. The design, optimization and evaluation of this sentence test are presented here.

2.2 DESIGN

2.2.1 Construction

The test construction is based on the Swedish sentence test by Hagerman (Hagerman, 1982). The base test list consists of 10 different five–word sentences with an equal syntactical structure (*name, verb, numeral, adjective, object.*) (Table 2.1). The test sentences are generated by randomly choosing one of the 10 alternatives for each part of the sentence. Consequently, each test list consists of the same word material.

2.2.2 Selection of words for the sentence material

The words in the sentence material were chosen on the basis of an analysis of word frequency in the written language. This analysis includes the 5000 most frequently

Table 2.1: *Basic test list of the DANTALE II test. The bold and italic words are examples for two generated test sentences.*

Index	*Name*	*Verb*	*Numeral*	*Adjective*	*Object*
0	Anders	**ejer**	ti	gamle	jakker.
1	Birgit	havde	*fem*	røde	**kasser.**
2	Ingrid	ser	syv	pæne	ringe.
3	Ulla	købte	tre	*nye*	blomster.
4	Niels	vandt	seks	fine	skabe.
5	Kirsten	får	tolv	flotte	*masker.*
6	Henning	solgte	**otte**	smukke	biler.
7	Per	låner	fjorten	store	huse.
8	**Linda**	valgte	ni	**hvide**	gaver.
9	*Michael*	*finder*	tyve	sjove	planter.

used words in Danish. The latest count of the 5000 most frequently used words in Danish can be found in "Dansk frekvensordbog" (Dictionary of Word Frequency in Danish) by Bergenholtz (1992). The use of frequent words enables an equal difficulty level in the sentences to be achieved, placing the test subjects on an equal footing as far as possible with regard to familiarity with the sentence material words. As a result, the probability of achieving the same slope of the psychometric function for each word in the sentence material is increased.

Five of the sentences are represented in the present tense, and five are represented in the past tense. This was done to avoid having certain verb endings presented more often than others, which has an impact on the phonetic balance. An attempt has been made to avoid using words of an emotive and offending nature, as well as geographically, socially and professionally related words.

To avoid presenting certain speech sounds more often than others in the sentence material, and to maximize the validity of the test results across subjects, phonetic balancing of the sound material was carried out on the basis of Danish standard language. A statistical survey of the relative occurrences of speech sounds in the 5000 most frequent words in Danish (Bergenholtz, 1992) was done for this purpose by Peter Molbæk Hansen (a phonetician and teacher of linguistic science at the University of Copen-

Table 2.2: Basic test list of the DANTALE II test, English translation. The lines illustrate the way of recording the sentences for index 0 and four examples of following words. The same procedure was repeated for all following words (indicated by the dotted line) and all indices.

Index	Name	Verb	Numeral	Adjective	Object
0	Anders	owns	ten	old	jackets.
1	Birgit	had	five	red	boxes.
2	Ingrid	sees	seven	nice	rings.
3	Ulla	bought	three	new	flowers.
4	Niels	won	six	fine	cupboards.
5	Kirsten	gets	twelve	lovely	masks.
6	Henning	sold	eight	beautiful	cars.
7	Per	borrows	fourteen	big	houses.
8	Linda	chose	nine	white	presents.
9	Michael	finds	twenty	funny	plants.

hagen, Denmark). The results of this survey were compared to the relative occurrences of speech sounds in the sentence material (Figure 2.1).

The sound surveys are based on a modified version of the International Phonetic Alphabet (IPA). The consonant transcription level is rough, while the vowel transcription level is more detailed, which ensures a depiction of the vowel variants and length variations. A very accurate phonetic balance was achieved with a correlation of 0.97 using 'Pearson's r' coefficient.

A random sample of 10 semi–randomized lists was presented to 10 normal–hearing subjects (free field) in a pilot study. The purpose of the test was to prevent the sentence meanings from appearing puzzling to such a degree that there would be a reduction in test result validity and reliability. The pilot study resulted in a decision to change one of the nouns, as it deviated semantically to some degree from the other nouns in the sentence material. This change did not cause a loss of correlation with regard to the phonetic balance. The final sentence material is listed in the Table 2.1 (Ardenkjær-Madsen and Josvassen, 2001).

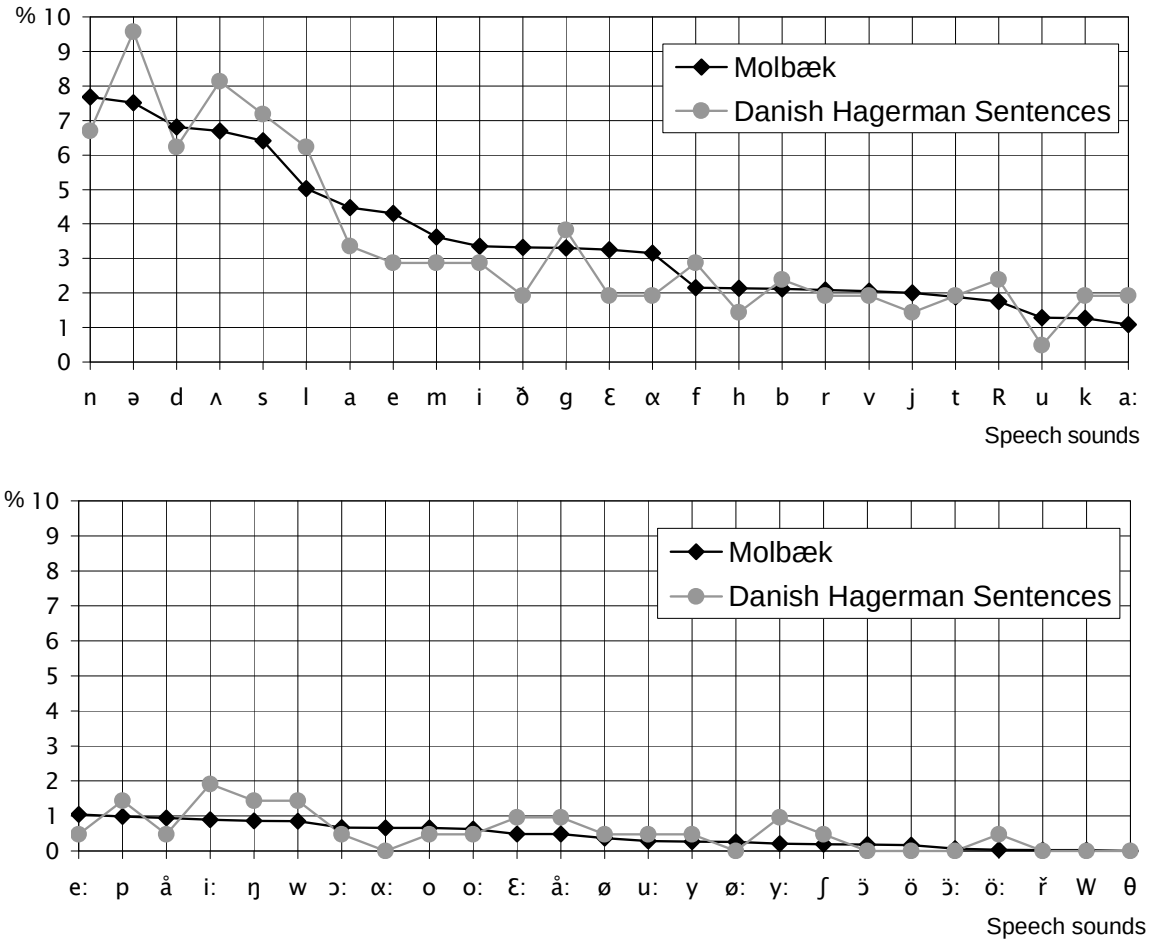

Figure 2.1: *A comparison expressed in percentages between the sound occurrences in the 5000 most frequently used words (Molbæk) and the sound occurrences in the final sentence material (Danish Hagerman sentences). For the sake of readability of the phonemes, the distribution is split into two figures.*

2.2.3 Recordings

The recordings were made in the radio studio of the University of Oldenburg, using an AKG C–1000S microphone and a DAT recorder (AIWA HHB1 Pro) with a sampling rate of 44100 Hz and a resolution of 16 bits. The sentences were spoken by Anne Bingen, a Danish speech and hearing therapist.

The main difference between the Hagerman and the Oldenburg sentence test is the method used to record and generate the sentences. For the Hagerman test, only the

sentences of the base list were recorded, without any transitions between the words, while for the Oldenburg sentence test, 100 sentences were recorded to take the coarticulation into account. For index 0, records were made as shown in Table 2.2. The same procedure was repeated for all following indices. In this way, all words in a given column were recorded in combination with all words in the following column. Since this approach yields a more natural sound of the test sentences, it was used for the Danish sentence test as well.

2.2.4 Generating test sentences

The test sentences were generated by combining the 10 alternatives for each word group at random (examples are indicated by bold and italic type in Table 2.1). The 100 recorded sentences were segmented into single words, very close to the beginning of the word, and including the part co–articulated to the following word at the end of the word. After some training, it became quite easy to identify the cutting point. We attempted to select the point in time for the cutting such that the following word would be perceived as 'naturally spoken' if it represented the first word of a new sentence. The cutting was performed with the CoolEdit program (by Syntrillium). The cutting points were identified by listening very carefully to the recorded material. In constructing sentences, a word in a given column was selected to produce the correct coarticulation for the following word, regardless of the previous word. As an example see Figure 2.2. The shortly ramped words (5–ms ramps) were strung together with a 5–ms overlap to generate a sentence. 25 test lists of 10 sentences were generated in this way. Each word of the base list occurs once in the resulting test list. Consequently, each test list consists of the same word material.

2.2.5 Interfering noise

The first step in achieving high accuracy in a speech test in noise (i. e., a steep intelligibility function) is to use stationary noise with the same long–term spectrum as the speech. This interfering noise yields optimal spectral masking. Each generated test sentence was strung with silence intervals in between the sentence repetitions to form a 2.5–min sequence. The lengths of the silence intervals were randomly chosen; for each particular sentence, the duration was fixed between 5 ms and 2 s. The starting points of the sentence repetitions also differed. These sequences were superimposed in order

Linda∼ ✂	ejer	...	
... ✂	**ejer∼**	✂ otte	...
	...	✂ **otte∼**	✂ hvide ...
		...	✂ **hvide kasser.**

Figure 2.2: *Taking the coarticulation effects into account to achieve a natural intonation. Only the utterances with the correct coarticulation to the following word in the final sentence are used, i. e. the words in bold type, to generate the sentence: Linda ejer otte hvide kasser. The coarticulation part is indicated by ∼, and the cutting place by scissors.*

to generate a speech–shaped interfering noise (Figure 2.3).

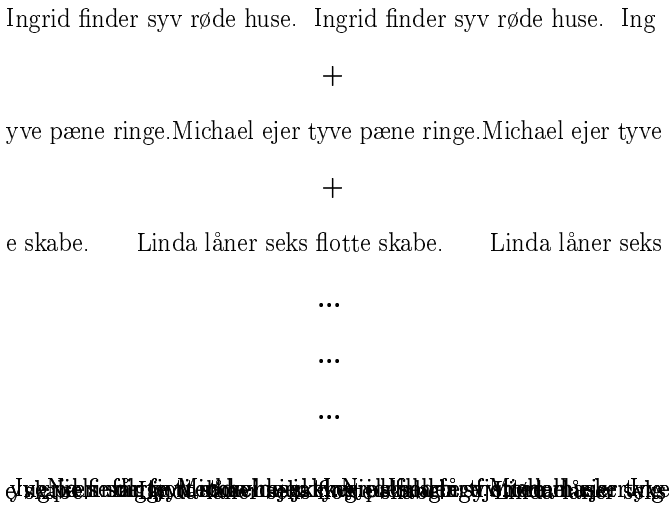

Figure 2.3: *Generation of the interfering noise. The speech material was superimposed using random silence duration and starting time.*

The superimposing was performed 30 times, to provide a more or less stationary noise without strong fluctuations. The long–term spectrum (Figure 2.4) of the resulting noise was comparable to the mean long–term spectrum of various languages, the LTASS spectrum (Byrne *et al.*, 1994).

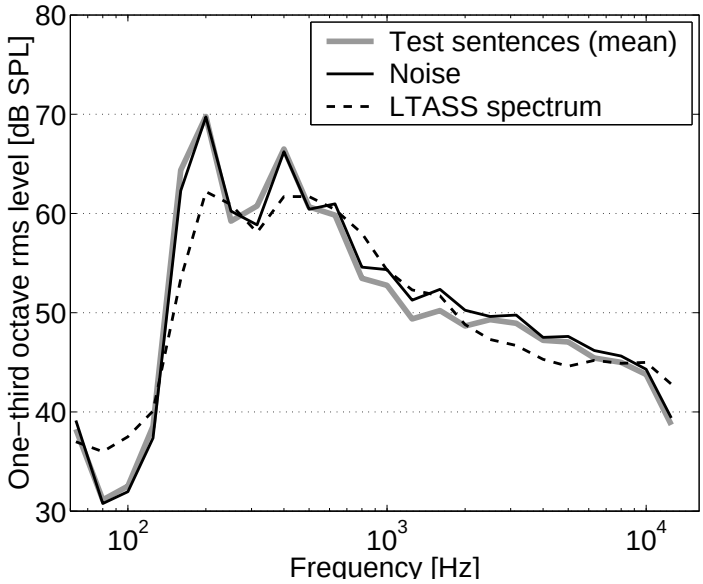

Figure 2.4: *Long–term spectrum of the interfering noise generated by superimposing the speech material of the DANTALE II test. The spectrum is given as root mean square (rms) levels in one third octave bands. The mean long–term spectrum of the test sentences and the LTASS spectrum (Byrne et al., 1994) are also given.*

2.3 OPTIMIZATION

The intelligibility function of a sentence test (i. e., dependency of speech intelligibility (SI) on sound pressure level or SNR) can be described by Equation 2.1.

$$SI(SNR) = \frac{1}{1 + e^{-4s50(SNR - SRT)}}, \quad s50\text{: slope at SRT} \tag{2.1}$$

After the probabilistic model of Kollmeier (1990), which is described in Wagener *et al.* (1999), the intelligibility function of a sentence test depends on the word–specific intelligibility functions as the convolution of the mean word–specific function and the distribution of the SRT values. Therefore, a steep slope $s50$ of the list–specific intelligibility function requires a small standard deviation σ_{SRT} of the word–specific SRT values and a steep slope s_{word} of the mean word–specific intelligibility function (Equation 2.2).

$$s50 \approx \frac{s_{word}}{\sqrt{1 + \frac{16 s_{word}^2 \sigma_{SRT}^2}{(ln(2e^{\frac{1}{2}} - 1 + 2e^{\frac{1}{4}}))^2}}} \tag{2.2}$$

The word–specific intelligibility functions of each generated test list were determined

using normal–hearing subjects in order to optimize the SRT distribution of the speech material by level adjustment of the single words and selection of the most homogenous lists.

2.3.1 Optimization measurements

Sixteen normal–hearing subjects (12 female, 4 male, age 20–37 years; median age 25 years; born and brought up in Northern Sealand or Copenhagen) participated in the measurements at the Rigshospital in Copenhagen. They had no otological problems, and their hearing thresholds did not exceed 20 dB HL at 0.5, 1, 2 and 4 kHz (a hearing threshold of 20 dB HL was allowed only once in this frequency range). The subjects were situated in a double–walled sound–insulated booth, fulfilling the requirements of ISO 8253–1. The sentences were presented monaurally via a Madsen type Midimate 622 audiometer, with Beyer Dynamic DT770 headphones. The subject's task was to repeat the words. The experimenter was situated outside the booth, and received the subject's answers via the audiometer intercom. The experimenter recorded the correct and incorrect responses (word scoring) for further analysis on a computer.

The 25 test lists of 10 sentences were combined to form 5 test lists of 30 sentences and 5 test lists of 20 sentences for the optimization measurements. All lists were presented at 10 different SNRs (−18 to 0 dB SNR, increments of 2 dB). The orders of the test lists and of the SNRs were chosen randomly.

The test sentences and the noise were mixed digitally at the particular SNRs. This mixed material was stored on Compact disks. The noise started 500 ms before the sentence started, and ended 500 ms after the sentence ended. Measurements with gated noise were chosen, because differences in SRT results were found for gated and continuous noise (Wagener *et al.*, 2000). In the case of the gated–noise condition, the delay between the noise onset and the speech onset was always fixed, while it depended on the response time of the subject in the case of the continuous–noise condition. Thus, in order to achieve a defined fixed onset–onset interval, we decided to use the gated–noise presentation for all experiments.

The test sentences of the same test list had a different, randomly chosen, order at the different SNRs.

The SNR was adjusted in this way in order to avoid using the audiometer for adjusting the SNR. Therefore, any inaccuracy of the audiometer was avoided. The audiometer

was calibrated in such a way that the measurements were performed at a noise level of 65 dB SPL.

The model function (2.1) was fitted to each acoustical representation of the words by using a maximum likelihood procedure (more precisely, the negative logarithmic likelihood was minimized). In this way, the SRT and the slope at the SRT ($s50$) were determined for all different word representations of the speech material.

2.3.2 Level adjustment based on optimization results

Level adjustments of particular words and a test list selection were performed in order to minimize σ_{SRT} (Equation 2.2). This level adjustment was limited to ±4 dB maximum to preserve a natural intonation. This limit was chosen after a listening test of 10 Danish listeners, who determined the most natural sounding of the sentences out of six different limitations (maximum ±2-6 dB, increment 1 dB, and no limitation at all).

The SRT distribution before (grey line, 1250 words) and after (black line, 800 words) the modification are shown in Figure 2.5.

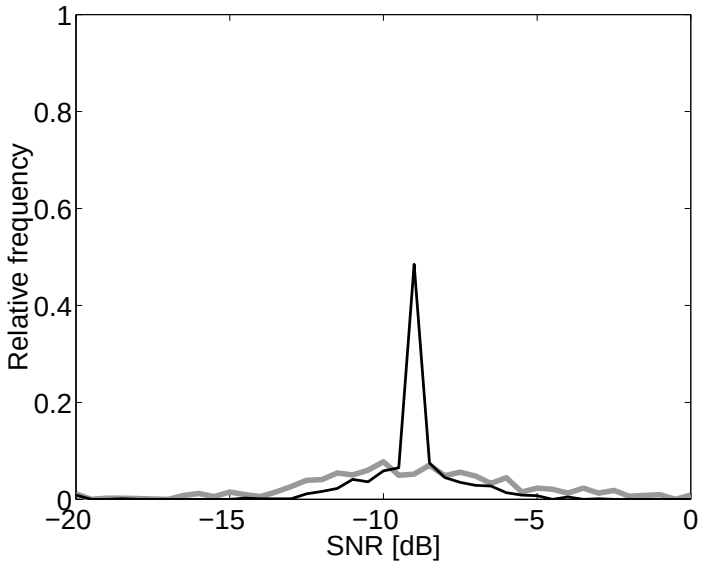

Figure 2.5: SRT distribution before (grey line, 1250 words) and after (black line, 800 words) the level adjustment and selection.

The expected slope $s50$ for the optimized test can be calculated using Equation 2.2. The mean word–specific slope s_{word} of the selected material equals 0.161 $dB^{-}1$, and the

standard deviation of the optimized SRT distribution equals 1.75 dB SNR. This yields an expected slope of $s50 = 0.132\,\mathrm{dB}^{-1}$ instead of $0.087\,\mathrm{dB}^{-1}$ before the optimization.

2.4 EVALUATION

The evaluation measurements were performed to evaluate the theoretical predictions by independent measurements with a large number of normal–hearing subjects. The list–specific intelligibility functions were calculated by determining the speech intelligibility at two different SNRs.

2.4.1 Evaluation measurements

Sixty normal–hearing subjects (41 female, 19 male; age 19–40 years; median age 27.5 years; born and brought up in Northern Sealand or Copenhagen) participated in the measurements at the Righospital in Copenhagen. They had no otological problems and their audiogram thresholds did not exceed 20 dB HL at 0.5, 1, 2 and 4 kHz.

For the evaluation measurements, test lists of 20 sentences were used. The experimental setup was the same as described in 'Optimization measurements' above. The subjects were divided into two groups. One group performed half of the lists at an SNR of −10 dB and the other lists at −6 dB SNR, the other group performed the particular lists at the respective other SNR. In this way, all subjects performed each test list just once in the evaluation measurements, and all lists were measured at two different SNRs. The order of the test lists was chosen randomly, and the two SNRs were presented alternately. The SNRs were chosen according to a corresponding intelligibility of above and below 50% (estimated from the optimization results). The noise was presented at a fixed level of 65 dB SPL. To achieve a similar training status for all subjects, all test lists were measured once with each subject before the evaluation measurements. These were adaptive measurements, determining the SRT. An adaptive procedure according to Brand (2002a) was used. The audiometer was used to adjust the SNRs during the training measurements. The SNRs for the evaluation measurements were adjusted by mixing the speech and noise signals digitally to avoid any inaccuracy of the audiometer.

As the model function (2.1) contains two parameters, and the speech intelligibility was determined at two different SNRs per test list, the model functions for each test list could be calculated using the measurement results.

2.4.2 Training effect

The SRT levels decreased with increasing number of lists performed per subject, due to familiarization with the measurement procedure and the word material. Therefore, eight test lists of 20 sentences were used for training purposes before the evaluation started. The training effect equals the SRT difference between the first and last performed training list. Figure 2.6 shows the SRT results by temporal order. The index on the x–axis indicates the temporal order of the measurements. The results of different test lists have been averaged for each index. The training effect of the training session equals 2.2 dB. The SRT given by the evaluation measurements (see below) hardly differs from that given by the last performed training list (difference: 0.3 dB).

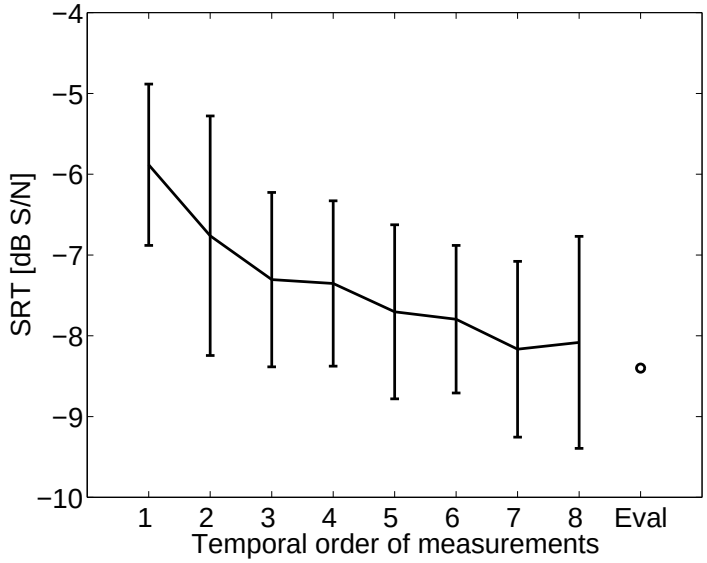

Figure 2.6: *SRT results during the training phase before the evaluation measurements were performed. The x–axis indicates the temporal order of the measurements using test lists of 20 sentences. The differences in SRT can be considered as training effects.*

2.4.3 Evaluation results

Figure 2.7 shows the results of the evaluation measurements. The intelligibility functions affiliated to the evaluation data are also shown.

The evaluation measurements result in a mean SRT of −8.38 dB SNR, with a standard deviation of 0.16 dB across test lists; the slope $s50$ equals 12.6%/dB±0.8. These results

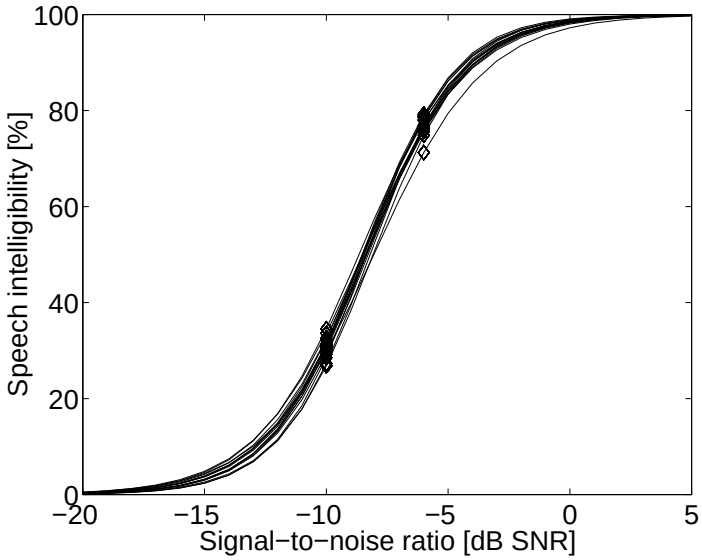

Figure 2.7: Speech intelligibility functions of the DANTALE II test. Evaluation data (diamonds) as well as the affiliated speech intelligibility functions (solid lines) are shown.

were derived by pooling the data of all subjects and calculating the mean values across the different test lists. It is not possible to determine threshold and slope for each subject and list individually, because each list was presented only once to each subject (two points are necessary for the calculation of SRT and $s50$ using Equation 2.1). In order to investigate the influence of the variability in threshold of the subjects on the resulting slope, data of all test lists can be pooled and the threshold and slope values for each individual subject can be determined. These values represent the individual intelligibility functions of the subjects. The mean SRT value across the subjects equals $-8.43\,\mathrm{dB\,SNR}$, with a standard deviation of $0.95\,\mathrm{dB}$ across subjects. The mean slope across the subjects equals 13.2%/dB, which exactly represents the slope that was expected after the optimization. The difference in the values for the slopes obtained for the pooled subjects and for the pooled test lists is due to the different thresholds for the subjects. In fact, the slope value for pooled subjects ($0.1262\,\mathrm{dB}^{-}1$) can be calculated by using the mean $s50$ value ($0.132\,\mathrm{dB}^{-}1$), the standard deviation of SRT across subjects ($0.95\,\mathrm{dB}$) with pooled test lists, and Equation 2.2: $\mathrm{s50} = 0.1261\,\mathrm{dB}^{-}1$.

No significant difference was found between the intelligibilities of the different test lists (single analysis of variance: F=0.80 at $-10\,\mathrm{dB\,SNR}$, and F=1.36 at $-6\,\mathrm{dB\,SNR}$).

2.5 DISCUSSION

A Danish sentence test based on the Swedish Hagerman and the German Oldenburg sentence test was introduced. The DANTALE II test was optimized to determine the SRT (SNR that yields 50% intelligibility) in noise. The differences between the results obtained by pooling the subjects (SRT=−8.38 dB SNR±0.16; $s50$=12.6%/dB±0.8) and pooling the test lists (SRT=−8.43 dB SNR±0.95; $s50$=13.2%/dB±1.9) show that the differences between the test lists are smaller than those between the normal–hearing subjects. Therefore, the reference data of the test are represented by the mean data across subjects (SRT=−8.43 dB SNR; $s50$=13.2%/dB). The mean slope is lower than the slope of the German test, which was realized in a similar way: $s50$ = 13.2%/dB instead of 17.1%/dB. This difference is due to a lower mean word–specific slope of the Danish test: 16.1%/dB, instead of 20%/dB for the Oldenburg word material. It is not clear, at present, whether this difference is due only to the different languages. This will be investigated further in the future.

All evaluation results confirm the expected values after the optimization (SRT = −9.1 dB SNR; $s50$ = 13.2%/dB). The slightly higher SRT of the evaluation can be attributed to better training of the subjects used for the optimization measurements. The DANTALE II consists of 16 test lists of 10 sentences, which can be combined to give 120 test lists of 20 sentences (test lists of 20 or 30 sentences give better reliability in adaptive threshold measurements (Brand and Kollmeier, 2002a)). There were no significant differences in the intelligibilities of the different test lists. The test provides a test–retest reproducibility of about 1 dB, if using 20 sentences for determining the threshold adaptively (if using 30 sentences, the reproducibility is less than 1 dB (Wagener *et al.*, 2000; Brand and Kollmeier, 2002a)). In these studies, training with 60 sentences before the first measurement and with 20 sentences before the retest measurement was used. There was a training effect of 1.4 dB during the first two test lists, making it essential to perform suitable training of the subject before doing speech intelligibility measurements. The overall training effect of 2.2 dB for 16 test lists of 10 sentences is similar to the training effect of the Swedish Hagerman sentences (Hagerman, personal communication and Hagerman and Kinnefors (1995). Hagerman found a training effect of 0.1 dB per list (10 sentences), and an additional 0.3 dB between the first–performed and second–performed lists. This yields an overall training effect of 1.9 dB when using 16 test lists. The advantage of the low–predictability sentences that are used in this test is that, after training, the test lists can be used repeatedly with the same

subject, because it is almost impossible to learn the lists by heart. Test lists that use highly predictable sentences, such as the German Göttingen sentence test (Kollmeier and Wesselkamp, 1997), the Dutch Plomp sentences (Plomp and Mimpen, 1979), or the English Hearing In Noise Test (Nilsson *et al.*, 1994), can only be used once per subject, as subjects tend to recognize the sentences if they are presented again.

2.6 CONCLUSIONS

The DANTALE II test is a sentence test with syntactically fixed, but semantically unpredictable (nonsense) sentences. The test was optimized for the determination of SRT in noise. The DANTALE II consists of 16 test lists of 10 sentences, which can be combined to give lists of 20 sentences. As the test lists can be used repeatedly with the same subject, the test is recommended for studies involving extensive measurements.

ACKNOWLEDGEMENTS

The present study was supported by GN Resound, Oticon, Widex, and DFG KO 942/13-3. We thank Anne Bingen for giving her voice, Line Bille Nugteren, and Annette Kristensen for collecting the data, Arne Nørby Rasmussen of the Rigshospitalet Copenhagen for his technical assistance in the measurements, Thomas Brand and Birger Kollmeier for their support. We also thank Andrew Oxenham for checking the manuscript and two anonymous reviewers for their helpful comments on the article.

Chapter 3

Sentence Intelligibility Tests with Syntactically Fixed, but Semantically Unpredictable Sentences: Comparison across Languages

ABSTRACT

Performance–intensity functions of three comparable sentence tests with syntactically fixed, but semantically unpredictable sentences (Swedish Hagerman sentences, German Oldenburg sentence test, Danish DANTALE II) were compared in order to test the perceptual similarities of these tests across languages. The same model function was employed to describe the intelligibility function in all three tests. The thus standardized reference data of all three sentence intelligibility tests were compared. They show similar reference values, except for the intelligibility function slope of the Danish test that was lower than of the other tests.

3.1 INTRODUCTION

Sentence intelligibility tests with syntactically fixed, but semantically unpredictable sentences are available in three different languages [Swedish: Hagerman (1982), German: Wagener *et al.* (1999c), Danish: Wagener *et al.* (2003)]. These tests can be described by their reference intelligibility function that was obtained by evaluation measurements with normal–hearing subjects. In order to compare the properties of such tests, these reference functions (sometimes denoted as performance–intensity functions) are compared. The reference intelligibility functions of the German Oldenburg sentence test (Wagener *et al.*, 1999b) and Danish DANTALE II test (Wagener *et al.*, 2003) could not be directly compared with the Swedish Hagerman sentences (Hagerman, 1982), as Hagerman originally used a different model function to describe the intelligibility function. In order to obtain comparable reference values, the raw data of Hagerman's evaluation measurements were analyzed with the logistic model function that was also used in the other tests (Equation 3.1):

$$p\ (L, SRT, s) = \frac{1}{1 + e^{4 \cdot s \cdot (SRT - L)}} \tag{3.1}$$

p denotes the mean probability that the words of a sentence are correctly repeated by the subject, if the sentence was presented with a signal–to–noise ratio L. The speech reception threshold SRT (signal–to–noise ratio that yields 50 % intelligibility) and the slope s of the intelligibility function at the SRT describe the entire function. Therefore, SRT and s describe the reference values of such speech tests and will be evaluated for the three tests under consideration.

3.2 HAGERMAN SENTENCES WITH LOGISTIC MODEL FUNCTION

The original raw data of Hagerman's evaluation measurements (Hagerman, 1982) consists of word–scored sentence scores of 10 normal–hearing subjects. Each subject performed five test lists (list no 6–10) at five different signal–to–noise ratios (SNRs). Three subjects performed these lists at -11, -9, -7, -5, -3 dB SNR, six subjects at -12, -10, -8, -6, -4 dB SNR, one subject at -10, -8, -6, -4, -2 dB SNR. The levels were chosen in ascending order for half of the subjects and in descending order for the other half.

The intelligibility function for each particular word was determined by averaging the data of all subjects and test lists for each single test word. For each particular word the logistic model function (Equation 3.1) was fitted to the SNR–intelligibility data by using a maximum likelihood procedure. When averaging across the intelligibility functions, the adjective 'hela' ('whole') was omitted, because not enough data were present at low SNRs. This would have resulted in an unrealistic slope of more than 400 %/dB.

The parameters of the mean word–specific intelligibility function equaled SRT = −8.3 dB SNR ± 1.7 dB and slope s_{word} = 23.5 %/dB (see Table 3.1).

In addition, the intelligibility function for each test list was determined by averaging the data of all subjects. For each test list the logistic model function (Equation 3.1) was fitted to the SNR–intelligibility data by using a maximum likelihood procedure. When averaging across the intelligibility functions of the lists, test list no 8 had to be omitted, because it was only performed at SNRs above -8 dB SNR. Therefore, too few data were below 50 % intelligibility (only one data point) and no reliable estimate of the slope at 50 % was possible.

The parameters of the reference intelligibility function equaled SRT = −8.1 dB SNR ± 0.3 and slope s = 16.0 %/dB ± 3.0 (see Table 3.1).

3.3 COMPARISON OF REFERENCE VALUES ACROSS LANGUAGES

The mean word–specific slopes s_{word}, the standard deviations of the word–specific SRT distribution σ_{SRT}, the experimentally determined mean list–specific slopes s, the mean SRTs, and the mean SRT standard deviations between test lists σ_{list} are given in Table 3.1 for each sentence intelligibility test, separately. The data of the Swedish test were determined in the present study. The data of the German test were determined in Wagener *et al.* (1999a) and Wagener *et al.* (1999b), the data of the Danish test in Wagener *et al.* (2003). In order to check the consistency of the data and the relation between word–specific and list–specific distribution parameters, the predicted mean list–specific slopes s_{pred} [predicted by the probabilistic model of Kollmeier *et al.* (1992)], is also included in Table 3.1. According to the model, the intelligibility function of a test list can be calculated by convoluting the mean intelligibility function of all words that build the list with the SRT distribution of these words.

Table 3.1: *Comparison of values that describe the perceptual properties across languages. Given are: mean word–specific slopes s_{word}, standard deviations of the word–specific SRT distribution σ_{SRT}, predicted mean list–specific slopes s_{pred}, experimentally determined mean list–specific slopes s, mean SRTs, and mean SRT standard deviations between test lists σ_{list}*

Language	s_{word} [%/dB]	σ_{SRT} [dB]	s_{pred} [%/dB]	s [%/dB]	SRT [dB SNR]	σ_{list} [dB]
Swedish	23.5	1.7	16.6	16.0	−8.1	0.30
German	20.0	1.1	17.2	17.1	−7.1	0.16
Danish	16.1	1.8	13.2	13.2	−8.4	0.16

SRT and s give the reference values of the particular test. The comparison between s_{pred} and s estimates the applicability of the probabilistic model. σ_{list} is a measure for the perceptual homogeneity of the test lists.

3.4 DISCUSSION

The results of the Swedish Hagerman sentences are very similar to the data of the German Oldenburg sentence test. The mean word–specific slope is slightly higher than of the German test (23.5 %/dB versus 20.0 %/dB) and clearly higher than of the Danish test (16.1 %/dB). The standard deviation of the word–specific SRT distribution is similar to the Danish test and slightly higher than the German test (Swedish: 1.7 dB, Danish: 1.8 dB, German: 1.1 dB). Therefore, the probabilistic model predicts a slightly lower slope of the mean list–specific intelligibility function of the Swedish test compared to the German test (16.6 %/dB versus 17.2 %/dB), both higher than the slope of the Danish test (13.2 %/dB). The probabilistic model predicts the mean list–specific intelligibility functions slope very accurately for all three sentence tests. The mean slope of the Swedish reference function is slightly lower than the German Oldenburg sentence test and higher than the Danish test (Swedish: 16 %/dB, German: 17.1 %/dB, Danish: 13.,2 %/dB), as it was predicted by the model.

The lower reference function slope of the Danish test is caused by the lower word–specific intelligibility function slope. This cannot be explained by eventually more fluctuations in the Danish interfering noise because both modulation spectra of the

German and Danish noises were similarly flat. The mean SRT of the Swedish Hagerman sentences were almost similar to the mean Danish DANTALE II SRT value and lower than of the German Oldenburg sentence test (Swedish: −8.1 dB SNR, Danish: −8.4 dB SNR, German: −7.1 dB SNR). The SRT standard deviation between test lists was slightly larger for the Swedish Hagerman sentences compared to the Danish and German test (Swedish: 0.30 dB, Danish and German: 0.16 dB). Nevertheless, the homogeneity of the test lists is high in all three sentence tests.

In conclusion, the mean expected SRT is highly comparable across similar tests in different languages. Using a 'calibration' offset of 1 dB, the reference data that describe the Swedish and German tests are highly comparable. However, the Danish test shows a shallower intelligibility function which is due to shallower word–specific intelligibility functions while the variability across test items (words, sentences, lists) is highly comparable across languages.

ACKNOWLEDGEMENTS

Dr. Björn Hagerman is gratefully acknowledged for entrusting the raw data of his evaluation measurements for this study. This study was supported by DFG KO 942/13-3.

Chapter 4

Test list homogeneity of high– and low–predictable sentence intelligibility tests in noise and in quiet

ABSTRACT

Sentence intelligibility tests are often developed and optimized for performing speech intelligibility measurements with interfering noise. The accuracy and homogeneity of such test lists is given by the external variance that is quantified by the masking of the particular test items by the noise. The accuracy and homogeneity in quiet is given by the internal variance that is quantified by the effect of hearing threshold on the items. It is a basic problem independent of the particular language that the external and internal variance cannot a priori be considered as being equal. In this study, this problem was exemplarily explored with two German sentence intelligibility tests that differ in predictability. The homogeneity of the different test lists with respect to speech intelligibility in noise was already shown for the German high–predictable Göttingen (Kollmeier and Wesselkamp, 1997) and the low–predictable Oldenburg sentence intelligibility test (Wagener *et al.*, 1999c; Wagener *et al.*, 1999a; Wagener *et al.*, 1999b). The equivalence of the respective test lists with respect to speech intelligibility in quiet was investigated with normal–hearing listeners. The standard deviations of the speech reception thresholds (SRT) across test lists were

smaller than the theoretical maximum accuracy of a single SRT determination. Therefore, both sentence tests can be used in quiet conditions without any changes of the test material. This result can be transferred to other speech intelligibility tests and languages that were developed with a similar optimization strategy.

4.1 INTRODUCTION

The accuracy of speech intelligibility tests and the homogeneity of different test lists is given by the perceptual variance of the particular test items. Therefore, speech intelligibility tests are optimized by minimizing this variance. Most sentence intelligibility tests were optimized with respect to intelligibility with interfering noise (Kalikow *et al.*, 1977; Plomp and Mimpen, 1979; Hagerman, 1982; Nilsson *et al.*, 1994; Kollmeier and Wesselkamp, 1997; Wagener *et al.*, 1999a; Wagener *et al.*, 2003), as this condition can be controlled best. The perceptual variance of such tests is given by the external variance that is quantified by the masking of the noise on the particular test items. It is not clear whether the homogeneity of such test lists can be transferred to speech intelligibility in quiet. The variance in quiet is given by the internal variance due to the effect of hearing threshold on the particular test items. It is questionable, whether the external and internal variance can be considered as being equal, since the frequency shape of hearing threshold is different from the interfering noise spectrum (mostly speech shaped). As sentence intelligibility tests are often applied both in noise and in quiet, especially when performed with cochlear implant users (Dorman *et al.*, 2002; Tyler *et al.*, 2002; Friesen *et al.*, 2001; Baumann, 2001; Brand and Kollmeier, 2002b), there is a need to explore the homogeneity of test lists in both conditions, noise and quiet. This language independent question was exemplarily investigated in this study using two German sentence intelligibility tests that vary in predictability of the respective words in each sentence.

The German Göttingen and Oldenburg sentence tests were developed, optimized, and evaluated for speech intelligibility in noise (Kollmeier and Wesselkamp, 1997; Wagener *et al.*, 1999c; Wagener *et al.*, 1999a; Wagener *et al.*, 1999b). The Göttingen high–predictable sentence test consists of 20 test list with 10 short meaningful sentences

each (similar speech material as the Dutch Plomp and Mimpen sentences (Plomp and Mimpen, 1979) and the American English HINT (Nilsson *et al.*, 1994)). As these sentences can easily be memorized by the subject, the test lists can only be used once with the same listener for a longer period of time. In order to overcome this disadvantage of meaningful test sentences, the Oldenburg sentence test was developed (Wagener *et al.*, 1999c), which is similar to the Swedish Hagerman sentences (Hagerman, 1982) and the Danish DANTALE II test (Wagener *et al.*, 2003). The Oldenburg low–predictable sentence test consists of 10 test lists with 10 syntactically fixed, but semantically non–predictable sentences of the form *name verb numeral adjective object* each. Since 10 alternative words exist for each position within each 5–word sentence, the speech material of this test consists of only 50 words that are used in a different permutation for each test list. Both sentence tests were recently analyzed with respect to efficient determination of speech intelligibility in noise (Brand and Kollmeier, 2002a) and context effects (Bronkhorst *et al.*, 2002).

In order to investigate whether the homogeneity of test lists also holds for quiet test conditions, both sentence tests were evaluated in quiet and compared to the respective results in noise.

4.2 METHOD

In order to determine the intelligibility functions of the particular test lists, the intelligibilities of all test lists were measured at two different presentation levels. The presentation levels should correspond to intelligibilities of about 20% and 80%, because these intelligibilities allow the most accurate concurrent estimates of speech reception threshold (SRT: speech level that corresponds to 50% intelligibility) and intelligibility function slope at SRT (Brand and Kollmeier, 2002a). An intelligibility function was fitted to the data, as follows:

$$p\left(L, SRT, s\right) = \frac{1}{1 + e^{4 \cdot s \cdot \left(SRT - L\right)}}, \tag{4.1}$$

where L= speech level; SRT= speech level that correspond to 50% intelligibility; s= slope at SRT.

In pilot measurements with five normal–hearing listeners, the SRT and slope values in quiet were determined adaptively for two test lists with 30 sentences of the Göttingen and Oldenburg tests. The mean results amount to SRT = 21.4 dB SPL,

s = 7.3 %/dB for the Göttingen sentence test and SRT = 21.4 dB SPL, s = 6.2 %/dB for the Oldenburg sentence test. Based on these results, the presentation levels for the evaluation of the Göttingen sentence test were chosen to 17 dB SPL (corresponding to 20 % intelligibility) and 26 dB SPL (corresponding to 80 % intelligibility). The presentation levels for the evaluation of the Oldenburg sentence test were chosen to 16 and 27 dB SPL.

Twenty normal–hearing subjects participated in the measurements (age: 21–49 years). They had hearing thresholds for pure tones better than 15 dB HL at the frequencies 0.125, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, and 6 kHz. They never participated in similar measurements before. The sentence tests were performed in their respective better ear.

The measurements were performed with a MS Windows–based program for conducting speech intelligibility measurements (developed and distributed by the national center of excellence HörTech, Oldenburg). The digital signals were D/A–converted by a 32–bit D/A–converter RME ADI 8–pro. The stimulus levels were adjusted by a computer–controlled custom–designed audiometer comprising attenuators, anti–aliasing filters, and headphone amplifiers. Signals were presented monaurally to the subjects via Sennheiser HDA 200 headphones. The headphones were free–field equalized according to international standard (ISO/DIS 389-8), using a FIR filter with 80 coefficients. The subjects were situated in a sound–insulated booth. Their task was to repeat each sentence or parts of the sentence presented over headphones as closely as possible. The instructor, also situated in the booth in front of the subject, marked each correctly repeated word. For this purpose, an Epson EHT 10S handheld touchscreen computer was used on which the target sentence was displayed. This handheld computer was connected to the personal computer via serial interface. All data were analyzed using word–scoring. The whole apparatus was calibrated to dB SPL with a B&K artificial ear 4153, a B&K 0.5 inch microphone 4143, a B&K preamplifier 2669, and a B&K measuring amplifier 2610.

In order to train the subjects for speech intelligibility measurements and to eliminate the training effect of the Oldenburg sentence test (Wagener *et al.*, 1999b), all subjects performed two test lists of the Oldenburg sentence test with 20 sentences at a presentation level of 35 and 25 dB SPL prior the evaluation measurements. The evaluation measurements were performed with presentation levels and test types in separate

blocks. The blocks were presented in random order. Prior to each block, the subject was told what sentence type should be repeated.

The sentences of the Göttingen test were combined to 10 test lists with 20 sentences each, the Oldenburg sentences were combined to five test lists with 20 sentences each for the evaluation measurements. The test lists were combined for the sake of measurement efficiency. It is time–saving to perform one test list consisting of 20 sentences instead of two test lists consisting of 10 sentences each.

Ten subjects performed one half of the Göttingen test lists at a presentation level of 17 dB SPL, and the other half of the lists at 26 dB SPL. The other ten subjects performed the particular Göttingen test lists at the respective other presentation level. In this way, each subject performed each test list of the meaningful Göttingen sentence test only once, and there was no influence by memorizing the sentences. Due to the structure of the Oldenburg sentence test, it is not possible to memorize the sentences. Therefore, the lists can be repeated. All 20 subjects performed each Oldenburg test list at both presentation levels (16 and 27 dB SPL).

4.3 RESULTS AND DISCUSSION

The data were analyzed separately for each 10–sentence–list. The data of all twenty subjects were pooled. These mean data were used to determine the speech intelligibility functions for each test list. The speech intelligibility functions of the Göttingen sentence test lists were determined in two different ways. First, the intelligibilities were calculated using the weighting factors for the particular words described by Kollmeier and Wesselkamp (1997). These factors were introduced in order to harmonize the intelligibilities of the words in noise. For comparison, the intelligibilities of the Göttingen sentence test were also calculated without using these weighting factors. The speech intelligibility functions of the Göttingen sentence test lists including the weighting factors are shown in Figure 4.1. The speech intelligibility functions of the Oldenburg sentence test lists are shown in Figure 4.2. As can be seen in the Figures, the speech presentation levels that were chosen according to the pilot measurements to obtain intelligibilities of 20 % and 80 %, respectively, did not exactly yield the expected intelligibilities. This was probably due to the limited accuracy of the SRT and intelligibility function slope estimates that was based on the pilot measurements.

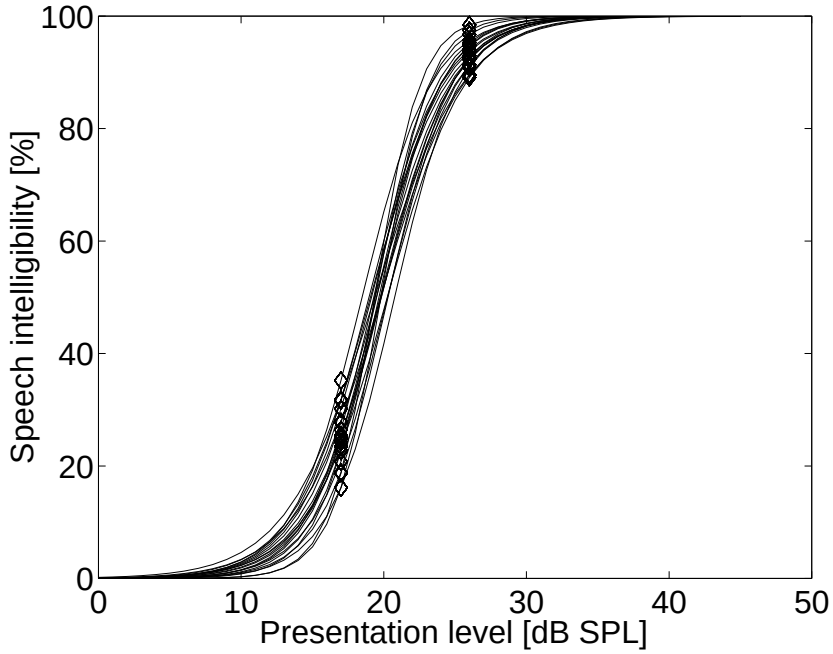

Figure 4.1: *Speech intelligibility functions of the Göttingen sentence test lists including the weighting factors (Kollmeier and Wesselkamp, 1997). The diamonds indicate the mean intelligibilities at 17 and 26 dB SPL.*

From a mathematical point of view, measuring two points of the intelligibility function and fitting the data with the model function (Equation 4.1) uniquely determines both parameters SRT and slope. However, an error in estimating the intelligibility influences both parameters. Therefore, it would be most suitable to determine a larger number of intelligibilities in order to determine the intelligibility function. For the sake of efficiency, only two points of the intelligibility function were determined. This simplification bases on the fact that at least in noise no mismatch between the model function and the data could be found: In the optimization measurements of the Oldenburg sentence test, the model function was fit to the data at eight different signal–to–noise ratios. The model function described the distribution of the data very accurately (Wagener *et al.*, 1999a). The two presentation levels in this study were chosen to statistically produce the best accuracy of concurrent SRT and slope determination according to Brand and Kollmeier (2002a). Each intelligibility estimate of a particular test list is based on at least 200 (Göttingen sentence test) or 800 (Oldenburg sentence test) statistically independent responses (number of subjects times sentences per list times number of statistically independent items per sentence

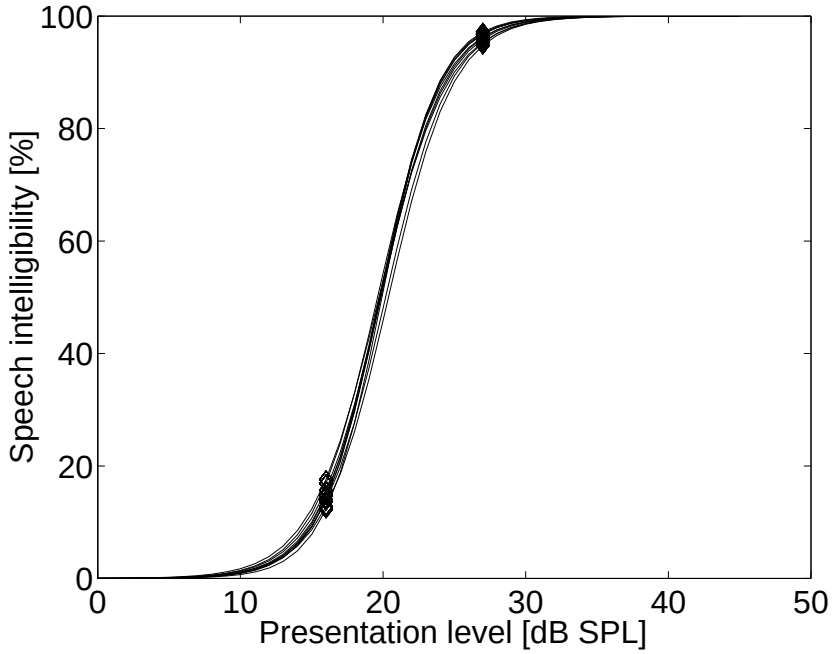

Figure 4.2: *Speech intelligibility functions of the Oldenburg sentence test lists. The diamonds indicate the mean intelligibilities measured at 16 and 27 dB SPL.*

j, see below). However, there is still one psychological problem if the SRT estimate is based on a rather low and a rather high intelligibility: The SRT estimate might be too large. It is possible that, at the lower presentation level, the subject had given up any effort since this is a very difficult situation. On the other hand, at the higher presentation level, the subject might not pay attention any more since this is a very easy situation. In this study, this effect should be reduced by deliberately maintaining the subject's motivation in both situations (the listeners were verbally motivated prior to each measurement to correctly repeat as much as possible).

The reference data in quiet were determined for both sentence tests by averaging SRT and slope values across subjects and test lists. The standard deviations across test lists give a measure for the homogeneity of the test lists in quiet measurement conditions. The mean results and standard deviations are given in Table 4.1. For the Göttingen sentence test, the results with and without weighting factors are given.

The theoretical minimum standard deviation σ_{SRT} of the SRT estimate from a mea-

Table 4.1: Mean SRT and slope values across subjects and test lists for the Göttingen and Oldenburg sentence test in quiet. The standard deviations of the mean data $\sigma_{SRT}, \sigma s$ are also given. For the Göttingen sentence test, the results with and without weighting factors are given.

	Göttingen sentence test		Oldenburg sentence test
	With weighting factors	Without weighting factors	
Mean SRT $\pm \sigma_{SRT}$	19.6 ± 0.5 dB SPL	19.3 ± 0.6 dB SPL	19.9 ± 0.2 dB SPL
Mean slope s $\pm \sigma_s$	10.7 ± 1.9 %/dB	10.8 ± 1.9 %/dB	11.3 ± 0.6 %/dB

surement with N sentences is given by:

$$\sigma_{SRT} = \sqrt{\frac{p\,(1-p)}{j \cdot N \cdot s^2}}, \tag{4.2}$$

where p = speech intelligibility, here $p = 0.5$; j = number of statistically independent items in each sentence; N = number of used sentences; s = slope of intelligibility function at SRT (Brand and Kollmeier, 2002a).

Assuming $j = 2$ (Kollmeier and Wesselkamp, 1997) and $s = 10.7\,\%/\mathrm{dB}$ for the Göttingen sentence test, the accuracy of determining the SRT with 10 sentences equals $\sigma_{SRT} = 1.0\,\mathrm{dB}$. The accuracy of the Oldenburg sentence test equals $\sigma_{SRT} = 0.7\,\mathrm{dB}$, due to a higher $j = 4$ (Wagener *et al.*, 1999b) and $s = 11.3\,\%/\mathrm{dB}$. Theoretically, the Göttingen sentence test needs twice as much measurement time to achieve the same accuracy as the Oldenburg sentence test. However, in adaptive SRT measurements, both sentence tests are equivalent with regard to test–retest accuracy (Brand and Kollmeier, 2002a). This shows that Equation 4.2 could not be directly transferred to adaptive procedures.

The SRT standard deviations across test lists of both sentence tests were smaller than the theoretical minimal SRT standard deviation of single SRT measurements, therefore no meaningful intelligibility differences in quiet were found between test lists.

All test lists of the Oldenburg sentence test consist of the same word material. The test lists differ only in the combination of the 50 words and in the usage of different recordings of the particular words from the same speaker (Wagener *et al.*, 1999c;

Wagener *et al.*, 2003). Each particular Göttingen test list consists of different word materials. Therefore, it is no surprise that the variations between the test lists of the Göttingen sentence test are larger than the variations between the Oldenburg sentence test lists. This holds also for both tests in noise [Göttingen sentence test: $\sigma_{SRT} = 0.27\,\mathrm{dB}$ (Kollmeier and Wesselkamp, 1997), Oldenburg sentence test: $\sigma_{SRT} = 0.16\,\mathrm{dB}$ (Wagener *et al.*, 1999b)]. It is also possible that the standard deviation across test lists of the Göttingen sentence test was slightly overestimated compared to the Oldenburg sentence test, because only 10 subjects were included in each data point (compared to 20 subjects in the Oldenburg sentence test). This might have resulted in a higher standard error of the particular intelligibilities and therefore in higher standard deviations across test lists.

The mean slopes of the intelligibility functions in quiet are smaller than the slopes of the intelligibility functions in noise for both sentence tests. The mean slopes in noise are as following: Göttingen sentence test $s = 20\,\%/\mathrm{dB}$ (Kollmeier and Wesselkamp, 1997); Oldenburg sentence test $s = 17\,\%/\mathrm{dB}$ (Wagener *et al.*, 1999b). There are two explanations for a lower slope in quiet compared to noisy conditions. First, the slope of a test list is highly determined by the intelligibility distribution of the particular test words. If the intelligibilities of the words within a list are very similar, the slope of the list is steep. Both the Göttingen and the Oldenburg sentence test were optimized in order to equalize the intelligibilities of the test words in noise, and therefore to obtain a steep slope in noise. Possibly, this equalization does not hold for quiet conditions. In other words, there might be no intelligibility differences across entire test lists in quiet, but there might still be intelligibility differences across the particular words within test lists in quiet. This may also explain why there was no difference when using the Göttingen sentence test with or without weighting factors. These factors were introduced in order to increase the homogeneity of intelligibility within the test words in noise. However, these factors do not increase the homogeneity in quiet.

The second explanation is a more general reflection: The long–term spectra of the interfering noises of both the Göttingen and Oldenburg sentence test were similar to the mean long–term spectra of the respective speech materials (Kollmeier and Wesselkamp, 1997; Wagener *et al.*, 1999c). Therefore, these noises generate an optimal spectral masking of the speech material. This leads to a well–defined limit between 'word not perceived' and 'word perceived'. In quiet conditions, this limit is less precise, because some frequency regions of the speech are audible whereas other frequency regions are inaudible. Therefore, the transition between 'word not perceived' and

'word perceived' is less distinct. This yields a lower intelligibility function slope.

The phenomenon of probably different external and internal variances of speech intelligibility tests in noise and in quiet is language independent. Therefore, it can be assumed that these results can be transferred to other sentence tests and languages that were developed using a similar optimization strategy (Plomp and Mimpen, 1979; Nilsson *et al.*, 1994; Hagerman, 1982; Wagener *et al.*, 2003) or to future tests of a similar type.

4.4 CONCLUSIONS

For both sentence tests (high–predictable and low–predictable) the mean SRT standard deviations across test lists were smaller than the accuracy of SRT determination from a single test list. This indicates that the homogeneity of the test lists is also given for both tests in quiet even though both tests were originally optimized to exhibit a high homogeneity in noise. Therefore, both sentence tests can be used in quiet measurement conditions without any changes of the test material. The Göttingen sentence test can be used in quiet with or without weighting factors for the particular words. In summary, achieving homogenous test lists with respect to intelligibility in noise seems to be an appropriate way to realize sentence intelligibility tests for noisy and quiet conditions.

ACKNOWLEDGEMENTS

We would like to thank Anita Gorges, Müge Kaya, and Dr. Birgitta Gabriel for performing the measurements. The valuable comments of Dr. Stefan Uppenkamp and three anonymous reviewers are gratefully acknowledged.

This study was supported by DFG KO 942/13-3.

Chapter 5

Influence of Measurement Procedure and Interfering Noise on Sentence Intelligibility in Noise

ABSTRACT

Clinical and research measurements of speech intelligibility are strongly dependent on several parameters of the measurement procedure. In order to obtain comparable results of different test procedures, it has to be investigated which parameters should be standardized and which parameters could be set freely. In this study, the influence of noise level, noise type, and presentation mode on speech reception thresholds (SRTs) and intelligibility function slopes in noise was therefore investigated for normal–hearing and hearing–impaired subjects using an adaptive sentence test [Wagener *et al.*, Zeitschrift für Audiologie 38, 4–15, 44–56, 86–95 (1999)]. The presentation level of the noise had no significant influence on either SRT or slope values, provided that the presentation level exceeded hearing threshold. Two stationary, speech shaped noises produced identical results. Speech–simulating fluctuating noise yielded about 14 dB lower SRTs for normal–hearing subjects and about 10 dB lower SRTs for some of the hearing–impaired subjects. Other hearing–impaired subjects did not benefit from the modulations and showed similar SRTs as for stationary noise. The slope values for fluctuating noise were significantly lower than for stationary noise. Using continuous noise yielded lower SRTs compared to gated noise that was only presented during the presentation of speech. However, the difference between continuous and gated noise

was not significant for the hearing–impaired subjects. As a proposal for comparable adaptive measurement procedures a presentation level of 65 dB SPL (normal–hearing subjects) or 80 dB SPL (hearing–impaired subjects) and a standard interfering noise with the LTASS–spectrum as long–term spectrum is suggested. A fluctuating, speech shaped noise is recommended to differentiate between subjects. The way of controlling signal–to–noise ratio (fixed noise or speech level) is non–critical in adaptive SRT–determination.

5.1 INTRODUCTION

A large number of methods have been used in the literature to determine speech intelligibility. They differ not only with respect to the use of different languages, but also with respect to a large number of procedural parameters such as, e. g., the presentation level, the type of interfering noise, and presentation modes such as varying the speech or the noise level in an adaptive procedure and using interrupted or continuous interfering noise. A comparison of studies investigating the same effect, but different speech tests is therefore difficult, because it remains unclear if differences are due to the effect under consideration or due to the respective test language, the respective group of subjects, or the details of the test procedure.

The aim of this study is therefore to quantify the influence of those parameters that typically vary between different test methods. Critical factors that systematically influence the test result, should be standardized whereas non–critical factors can be left open to the peculiarities of the respective speech test methods. Hence, the current study should contribute to harmonize speech audiometry across different research sites and languages - a process that has already begun within the NATASHA project (Droogendijk and Verschuure, 2000).

As the intelligibility differences in noise are rather small compared to intelligibility differences in quiet, the Oldenburg sentence test with less than 1 dB test–retest standard deviation (Wagener *et al.*, 1999c; Wagener *et al.*, 1999a; Wagener *et al.*, 1999b; Brand and Kollmeier, 2002a) was used to determine speech intelligibility functions adaptively for both normal–hearing and hearing–impaired subjects. The test lists of this sentence test can be repetitively used with the same subjects, therefore it was possible to use the same test and subjects throughout the whole study.

The first factor investigated here is the noise presentation level (in adaptive proce-

dures with fixed noise level). In literature, different opinions can be found about level dependency of speech intelligibility in noise. Some studies showed that the SRT is only dependent on the signal–to–noise ratio (SNR), when the noise level exceeds a threshold (Smoorenburg, 1992; Duquesnoy, 1983a; Plomp, 1978; Speaks *et al.*, 1967; Hirsh *et al.*, 1954; Hirsh and Bowman, 1953; Hawkins and Stevens, 1950). Other studies also showed SRT dependency on presentation level (Studebaker *et al.*, 1999; Beattie, 1989; Hagerman, 1982; Pickett and Pollack, 1958; Pollack and Pickett, 1958). While using adaptive procedures, it is important to investigate, whether there are any problems in determining SRT at high and low presentation levels.

The second factor is the interfering noise. Often, each speech test includes its own noise which was generated by superimposing the speech material to produce a noise with the same long–term spectrum as the speech material itself (Hagerman, 1982; Kollmeier and Wesselkamp, 1997; Wagener *et al.*, 1999c; Wagener *et al.*, 2003) or the noise was spectrally matched to the sentences (Nilsson *et al.*, 1994; Plomp and Mimpen, 1979). Since the long–term spectra of most languages are quite similar (Byrne *et al.*, 1994), it seems useful to standardize the interfering noise. In the present study it was proved whether an interfering noise with the mean speech spectrum over different languages like the LTASS–spectrum (Byrne *et al.*, 1994) is suitable as standard interfering noise for speech tests.

One important application of speech tests in noise is the evaluation of hearing aids (Green *et al.*, 1989; Arlinger, 1998; Humes, 1999). The time constants of the noise suppression method and dynamic–range compression are important parameters of modern hearing aid algorithms. In order to investigate the influence of such parameters on speech intelligibility in noise, these hearing aids should not only be tested by applying stationary but also fluctuating noise (van Toor and Verschuure, 2002). Different fluctuating versions of an LTASS–shaped noise exist (Dreschler *et al.*, 2001). The influence of these different fluctuating versions on speech intelligibility was also investigated in this study. Using fluctuating interfering noise is also interesting in clinical practise. From the diagnostic point of view, it is important to find a test configuration that differentiates significantly between different degrees of hearing impairment. A prerequisite for such a differentiation is a small intra–individual test–retest standard deviation of the testing procedure compared to the typical inter–individual standard deviation of SRT values. Large differences in SRT between normal–hearing and hearing–impaired subjects were found in the literature if fluctuating interfering noise has been used. Often, the SRT–benefit of hearing–impaired subjects was inves-

tigated when using fluctuating instead of stationary noise. A similar benefit due to modulations as in normal–hearing subjects was reported in hearing–impaired subjects (4–6 dB) by Festen and Plomp (1990). Little or no benefit was reported for hearing–impaired subjects in (Duquesnoy, 1983a; Hygge *et al.*, 1992; Eisenberg *et al.*, 1995; Gustafsson and Arlinger, 1994; Peters *et al.*, 1998). It should be investigated in this study, whether or not the intra–individual standard deviations of test–retest SRT measurements with fluctuating interfering noise is small enough to permit a better differentiation between subjects than using stationary interfering noise.

The third factor is the presentation mode of the speech and noise signals. The adjustment of the signal–to–noise ratio in adaptive measurements is performed differently by different authors or in clinical practise. Sometimes, the noise level was fixed and the speech level was adjusted corresponding to the subject's responses (Duquesnoy and Plomp, 1983; Nilsson *et al.*, 1994; Brand and Kollmeier, 2002a). Other authors (Dubno *et al.*, 1984; Gustafsson and Arlinger, 1994; Hagerman and Kinnefors, 1995) performed vice versa: the speech level was fixed and the noise level was adjusted. It was investigated in the present study, whether or not the results of these two different approaches differ significantly.

Another important aspect of the presentation mode is the continuity of the noise signal, i. e. the use of continuous noise instead of noise that is interrupted between the sentences. Hence the extent is investigated to which the results are influenced by such interruptions of the interfering noise.

The influence of these three factors (noise presentation level, type of noise, and presentation mode) on speech intelligibility of normal–hearing and hearing–impaired subjects was investigated in this study.

In this article, first the employed methods are described (both experimental and statistical). As the determination of speech intelligibility is a statistical estimate, the results are obtained with a limited accuracy. While investigating the parameter influence on speech intelligibility, it is important to investigate the parameter influence on the reliability of this measure. The reliability results of all parameter settings are presented prior to the SRT–differences within the parameter groups.

5.2 GENERAL METHOD

5.2.1 Oldenburg sentence test

The Oldenburg sentence test was employed throughout this study which has been constructed similar as the Swedish sentence test by Björn Hagerman (Hagerman, 1982; Hagerman, 1984) and equal to the Danish DANTALE II (Wagener *et al.*, 2003). It was realized using a base list consisting of ten sentences with five words each. The syntactic structure of all sentences is identical: *Name verb numeral adjective object.* (see Table 5.1). This base list approximates the mean phoneme distribution of the German language. The test lists of the Oldenburg sentence test were generated by choosing one of the ten alternatives for each word group in a pseudorandom way that used each word exactly once in each test list (Wagener *et al.*, 1999c). If only the ten sentences of the base list had been recorded, segmented, and synthesized in the desired pseudorandom order, a rather unnatural sounding speech pattern would have been resulted. For a more natural sound the co–articulation effects were taken into account. For this purpose, 100 sentences were recorded in such a way that all words in a given column (see Table 5.1) were recorded in combination with all words in the subsequent column. Thus, each co–articulation between a word and the ten possible subsequent word alternatives was recorded. In constructing the test sentences, the specific recording of a word in a given column was selected, which produces the correct co–articulation for the subsequent word, regardless of the previous word.

In order to achieve a steep speech intelligibility function, the words were selected and adjusted in level with respect to an optimized perceptual homogeneity. For this purpose, the speech intelligibility function of each word was determined with 12 normal–hearing subjects. In order to equalize these word specific speech intelligibility functions across all available words, some words were adjusted in level and the final 10 test lists with 10 sentences each were selected (Wagener *et al.*, 1999a).

These lists were evaluated in independent measurements with 20 normal–hearing subjects. It was verified that the different test lists are comparable with regard to intelligibility. The main advantage of the test is due to the semantically nonsense character of the sentences: the sentences cannot easily be memorized and there is no benefit available from sentence context. Therefore, the 10 final lists of 10 sentences can be used several times with the same subject and can be combined to 120 different test lists with 30 sentences each. One possible disadvantage of the test design is a strong

Table 5.1: Design of the Oldenburg sentence test (base list). The underlined or italic words show examples of newly generated sentences.

Name	**Verb**	**Numeral**	**Adjective**	**Object**
Peter	bekommt	drei	grosse	Blumen.
Kerstin	sieht	*neun*	kleine	Tassen.
Tanja	kauft	sieben	alte	Autos.
Ulrich	gibt	acht	*nasse*	Bilder.
Britta	schenkt	vier	schwere	Dosen.
Wolfgang	verleiht	fünf	grüne	*Sessel.*
Stefan	hat	zwei	teure	Messer.
Thomas	gewann	achtzehn	schöne	Schuhe.
Doris	nahm	zwölf	rote	Steine.
Nina	*malt*	elf	weisse	Ringe.

training effect during the first measurements with each new subject: comparing the first and second list of 30 sentences, there was an improvement in SRT of up to 2 dB (Wagener *et al.*, 1999b). This training effect is due to a general familiarization to the measurement procedure and to the limited number of words in the test material. Fortunately, there is no strong further training effect after this initial training (Wagener *et al.*, 1999b). Therefore the Oldenburg sentence test is highly suitable for studies with numerous repetitive speech intelligibility tests in noise, e.g. for research purposes or for fitting hearing aids or cochlear implants.

5.2.2 Apparatus

A computer–controlled audiometry workstation with a coprocessor board (Ariel DSP 32C) with 16–bit stereo AD–DA converters was used to control the complete experiment as well as stimulus presentation and storing of the subject's responses. This workstation was developed within a German joint research project on speech audiometry (Kollmeier *et al.*, 1992). The continuous and the fluctuating noise signals were played back by a Philips CD–Player (CD880), which was connected to the DSP board. The stimulus levels were adjusted by a computer–controlled custom–designed audiometer comprising attenuators, anti–aliasing filters, and headphone amplifiers. Signals were presented monaurally to the subjects via Sennheiser HDA 200 headphones. The headphones

were free–field equalized according to international standard (ISO/DIS 389–8), using a FIR filter with 80 coefficients. The subjects were situated in a sound–insulated booth. Their task was to repeat each sentence or parts of the sentence presented over headphones as closely as possible. The instructor, also situated in the booth in front of the subject, marked each incorrectly repeated word. For this purpose, an Epson EHT 10S handheld touchscreen computer was used on which the target sentence was displayed. This handheld computer was connected to the personal computer via serial interface. The whole apparatus was calibrated to dB SPL with a B&K artificial ear 4153, a B&K 0.5 inch microphone 4143, a B&K preamplifier 2669, and a B&K measuring amplifier 2610.

The subjects' responses were analyzed using word–scoring.

5.2.3 Test procedure

Test–retest measurements were performed for each parameter setting (described below) by determining the speech reception threshold (SRT, i. e. the signal to noise ratio that yields 50 % intelligibility) and the respective slope with 30 test sentences. Intelligibility is defined as the mean probability p that the words of a sentence are correctly repeated by the subject, if the sentence was presented with a signal–to–noise ratio L. The logistic function given by Equation 5.1 represents the intelligibility function. This function was used to determine the SRT and slope estimates (s).

$$p\ (L, SRT, s) = \frac{1}{1 + e^{4 \cdot s \cdot (SRT - L)}} \tag{5.1}$$

The adaptive procedure developed by Brand and Kollmeier (2002a) was used. This procedure determines SRT and slope concurrently. The SRT is estimated by this procedure with a mean test–retest standard deviation of 0.4 dB (normal–hearing subjects), and 0.6 dB (hearing–impaired subjects) (Brand and Kollmeier, 2002a). The slope is estimated with a mean standard deviation of 3.4 %/dB (normal–hearing subjects), and 4.1 %/dB (hearing–impaired subjects) (Brand and Kollmeier, 2002a).

Each subject performed four sessions: For the hearing–impaired subjects, a categorical loudness measurement using the *olnoise* signal (see Section 5.2.5) was performed (Brand and Hohmann, 2002). Each subject performed two practice lists with 30 sentences in the first session and one practice list at the beginning of each subsequent session. The results were discarded. Subsequently, up to five tracks were performed using the

different parameter settings in random order. The test and retest measurements of one condition were performed on two different days.

5.2.4 Statistical methods

All parameter settings were analyzed together with regard to test–retest reliability (Section 5.3) and separately for each experiment with regard to effects of different parameter settings. The test–retest reliability was determined by calculating the test–retest differences and both the inter–individual and intra–individual standard deviations. The inter–individual and intra–individual standard deviations of both SRT and slope values were determined to quantify the potential for differentiating between subjects with various degrees of hearing impairment. Equation 5.2 was used to calculate the mean intra–individual standard deviation of two data points per subject (calculating the 2nd moment without calculating the 1st moment).

$$\sigma_{intra} = \frac{1}{\sqrt{2}} \cdot \sum_{n} \frac{1}{n} \cdot (x_{n,1} - x_{n,2})^2 \tag{5.2}$$

In order to determine the effect of different parameter settings, the data of the 10 normal–hearing subjects were corrected by the overall training effect (Section 5.3) and pooled within the particular experiments. The mean SRT and slope values were determined within the experiments. The differences between different parameter settings were investigated with a one way ANOVA ($p<0.01$). In order to test whether the variances of different parameter settings were identical and the ANOVA could be used, the Hartley test was used. It was assumed that the normal–hearing data were normal distributed. When the ANOVA detected significant differences, the modified LSD test (least significant difference test) according to Hayter was performed as post–hoc test in order to investigate the differences in more detail.

The data of all 10 hearing–impaired subjects were tested using the Bartlett test (a combined test of normality of distribution and equality of group variance), if an ANOVA could be performed. If so, the data were analyzed analogously to the normal–hearing data.

5.2.5 Test conditions

Three different types of parameters were tested: 1) noise presentation level, 2) type of noise, and 3) presentation mode (see below). The standard measurement procedure

was as follows: A stationary speech–simulating noise - the *olnoise* - generated by randomly superimposing the speech material of the Oldenburg sentence test, was used as interfering noise. The interfering noise was presented synchronous to the sentences (that is with interruptions between sentences) with the noise starting 500 ms before and ending 500 ms after each sentence. The noise level was held constant at 65 dB SPL in the normal–hearing subjects and at the individual medium–loudness level in the hearing–impaired subjects. The noise level used in hearing–impaired subjects was limited to maximum 85 dB SPL, therefore in the standard configuration, one subject used a noise level of 78 dB SPL, and nine subjects 85 dB SPL. The sentence level was adjusted according to the subject's responses using the adaptive procedure. As this standard parameter setting represents one setting of each tested parameter, it is referred to differently in the three experiments: In the first experiment the parameter set is referred to as *25 cu*, in the second as *olnoise*, and in the third as *synch*, respectively.

Each parameter type will be presented as one experiment in the following, although the measurements were performed in an interleaved way, using the same subjects over four days with a random order of the parameter settings.

5.2.5.1 Experiment I: Noise presentation level

The normal–hearing subjects performed measurements with noise presentation levels of 45, 55, 65, 75, and 80 dB SPL. These values approximate the different categorical loudness judgments in normal–hearing subjects for broadband noise ['soft' (15 cu), 'between soft and medium' (20 cu), 'medium' (25 cu), 'between medium and loud' (30 cu), 'loud' (35 cu), compare (Brand and Hohmann, 2001)]. The level that would yield 'loud' is about 85 dB SPL. 80 dB SPL was used as maximum level instead of 85 dB SPL to avoid any discomfort.

The hearing–impaired subjects performed measurements at a noise level of 65 dB SPL and at levels corresponding to their individual categorical loudness judgments, determined in a prior adaptive categorical loudness scaling measurement. 85 dB SPL was the maximum noise level used in the measurements, therefore the loudness categories 30 and 35 cu were not represented at all, and 20 and 25 cu were not represented for some subjects.

5.2.5.2 Experiment II: Type of noise

Two stationary speech–simulating noises were tested. The *olnoise* was generated from the speech material of the Oldenburg sentence test and provides the same long–term frequency spectrum as the sum of all test sentences (Wagener *et al.*, 1999c). This noise type was used in the standard parameter setting. A second, stationary interfering noise, the *icra1* noise (Dreschler *et al.*, 2001) was used that was generated for the International Collegium of Rehabilitative Audiology by the HACTES work group (Hearing Aid Clinical Test Environment Standardization). The *icra1* noise is a random Gaussian noise with a male–weighted idealized speech spectrum according to the ANSI S3.5 that is consistent with the LTASS spectrum (Byrne *et al.*, 1994) throughout most of the frequency range (Dreschler *et al.*, 2001). In addition, two fluctuating speech simulating noises were tested, the *icra5* and *icra7* noises (Dreschler *et al.*, 2001). The *icra5* noise is a three–band–speech–fluctuating noise with a male weighted idealized speech spectrum, which represents the modulations of one male speaker. The *icra7* noise represents a six persons babble: one female, one male speaker and two female and male speaker with a softer effort of −6 dB each. [1] The names of the *icra* noises in this study were given according to the track numbers of the ICRA compact disk (International Collegium of Rehabilitative Audiology, 1997). Root mean–square levels were identical for all four noises. The normal–hearing subjects performed all measurements at a noise presentation level of 65 dB SPL. The hearing–impaired subjects performed all measurements at a noise presentation level chosen according to their individual medium–loudness level. The maximum level was 85 dB SPL, this maximum generated a softer category than 'medium' for some hearing–impaired subjects.

5.2.5.3 Experiment III: Presentation mode

Measurements were performed with three different presentation modes.

[1] These noises were generated as follows: A spoken text about arithmetical notation was split up into three different frequency bands. In each frequency band the sign of each sample was randomly either reversed or kept with a probability of 50 % (Schroeder, 1968), therefore the resulting signals showed the same modulation properties as the originals, but had a flat, white spectrum. These signals were again filtered into the three frequency bands and in each band the RMS–values were equalized. The bands were added together to get one signal with a white spectrum but the original modulations. In order to obtain the correct long–term spectrum, the signal was filtered with a male or female speech shaped filter according to the LTASS spectrum (Byrne *et al.*, 1994). To avoid an unpleasant scratchy sound, the phase of the signal was randomized.

1) During the synchronous presentation mode (referred to as *synch*), the noise was interrupted between the sentences and the noise level was fixed while the speech level was adaptively adjusted according to the subject's response. This presentation mode was the standard parameter setting.

2) The presentation mode with continuous noise (not interrupted between sentences) is referred to as *cont*. The noise level was fixed and the speech level varied, too.

3) In contrast to the *synch* and *cont* presentation mode, the noise level was adjusted adaptively according to the subject's response while the speech level was held constant in the *inv* presentation mode (for inverse). Similar to the *synch* presentation, the noise was interrupted between sentences in this presentation mode.

5.2.6 Subjects

5.2.6.1 Normal–hearing subjects

Ten normal–hearing subjects (five females, five males; aged 22–40 years; median age: 26.5) participated in the measurements. They had hearing thresholds for pure tones better than 15 dB HL at the frequencies 0.125, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, and 6 kHz. Thresholds of 15 dB HL were permitted at no more than 2 frequencies. The sentence tests were presented to each listeners' better ear.

5.2.6.2 Hearing–impaired subjects

Ten sensorineural hearing–impaired subjects (three females, seven males; aged 59–79 years; median age: 70) participated in the measurements. They showed different types and degrees of sensorineural hearing loss with no conductive hearing loss. Pure–tone hearing thresholds ranged from 10 dB HL up to more than 100 dB HL. The types of hearing loss were: three broadband hearing losses, two pure high–frequency hearing losses, one combined low– and high–frequency hearing loss, and four sloping hearing losses. The audiogram data of each ear that was tested in this study are shown in Figure 5.1.

All subjects were paid for their participation on an hourly basis.

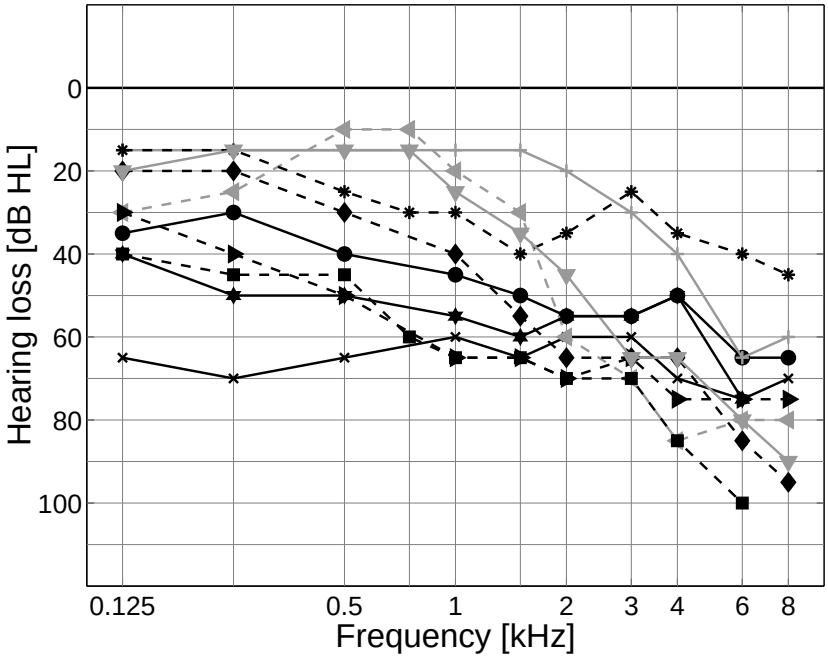

Figure 5.1: *Audiogram data of the respective tested ear of each hearing–impaired subject (broadband hearing loss: black solid line; pure high–frequency hearing loss: gray solid line; combined low– and high–frequency hearing loss: dashed gray line; sloping hearing loss: black dashed line).*

5.3 TEST ACCURACY

5.3.0.3 Test–retest differences

One measure of reliability (and hence accuracy of the speech test for different parameter configurations) is the test–retest difference of results. Table 5.2 shows the median SRT test–retest differences for the different parameter settings averaged across all normal–hearing and hearing–impaired subjects.

The test–retest differences between SRT values were positive for 84 % of the measurements with normal–hearing subjects, this indicated a training effect of 0.67 dB between the test and retest measurements (median test–retest difference across all settings). As training occurred in most retest measurements with normal–hearing subjects, the retest data were corrected by this 'overall training effect' of 0.67 dB. All analysis in the experiments was performed with these corrected normal–hearing data. The test–retest differences between SRT values were positive for 61 % of the measurements with

Table 5.2: Median test–retest differences between SRT values for the respective parameter settings averaged across all normal–hearing (NH) and hearing–impaired subjects (HI) in dB. The medians are given for each parameter setting and across all settings.

	15 cu	20 cu	25 cu	30 cu	35 cu	65 dB	85 dB
NH:	0.9	0.4	0.8	0.9	0.7		
HI:	0.2	0.3	0.0			0.4	0.0
	icra1	icra5	icra7	cont	inv	median	
NH:	0.5	2.0	0.4	0.5	0.1	0.67 dB	
HI:	0.4	-0.6	0.9	0.8	0.1	0.23 dB	

hearing–impaired subjects, the median training effect was 0.2 dB (75 % and 25 % percentile: 1.1 and -0.3 dB). The hearing–impaired retest data were not corrected, because the overall training effect was negligible.

The largest SRT test–retest differences occurred in the strongly fluctuating noise *icra5*. The slightly fluctuating noise *icra7* and the stationary noises *olnoise* and *icra1* result in comparable SRT test–retest differences within the normal–hearing and the hearing–impaired group.

5.3.0.4 Intra– and inter–individual standard deviations

Another measure of test reliability is the intra–individual standard deviation of test– and retest results. Speech intelligibility measurements should differentiate between subjects with different degrees of hearing loss. Therefore, the results should show small intra–individual and large inter–individual standard deviations. Figure 5.2, upper left panel, shows the inter–individual and intra–individual SRT standard deviations of the normal–hearing subjects as well as the intra–individual standard deviations corrected by the median training effect. The inter–individual and intra–individual slope standard deviations of the normal–hearing subjects are shown in the same figure, upper right panel. The lower panels in Figure 5.2 show the same values of the hearing–impaired subjects (for hearing–impaired subjects, no training effect correction was necessary). As the inter–individual standard deviation should exceed twice the intra–individual standard deviation in order to differentiate significantly between the subjects ('2σ–criterion'), the 2σ–criterion limits are given as a line in Figure 5.2.

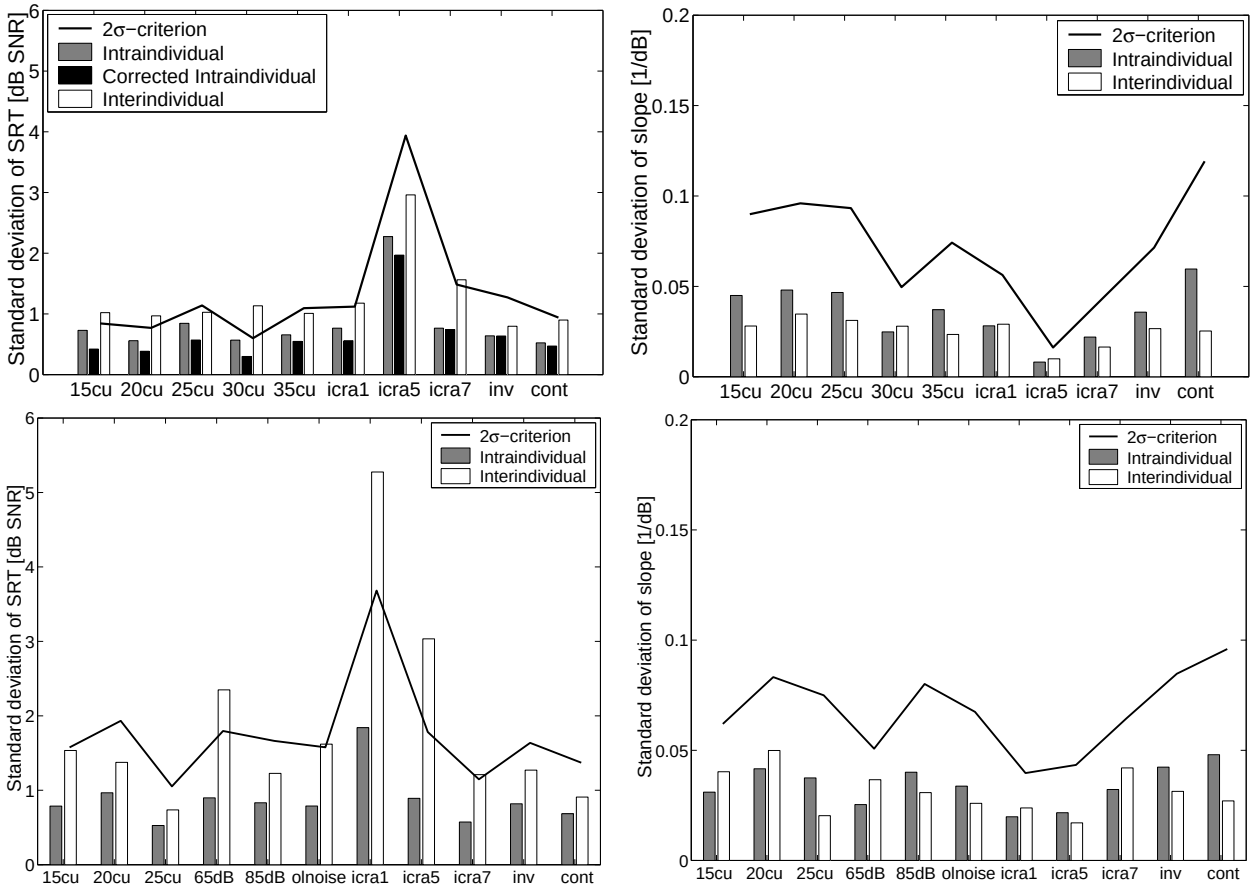

Figure 5.2: *Inter–individual (white bars) and intra–individual standard deviations (gray bars) of SRT (left panels) and slope (right panels) for the different parameter settings. Upper panels: normal–hearing subjects, lower panels: hearing–impaired subjects. The corrected intra–individual standard deviation (black bars) of SRT are also given for the normal–hearing subjects (the retest results were corrected by the median SRT test–retest difference across all parameter settings). The 2σ–criterion is indicated as a line (the criterion for normal–hearing subjects is based on the corrected intra–individual standard deviations of SRT).*

The inter–individual and corrected intra–individual of normal–hearing subjects were rather similar in all configurations using stationary interfering noise (mean inter–individual: 1 dB, mean intra–individual: 0.5 dB). The same holds for hearing–impaired subjects (mean inter–individual: 1.2 dB, mean intra–individual: 0.7 dB), with an exception at a presentation level of 65 dB (mean inter–individual: 2.4 dB, mean intra–individual: 0.9 dB). The inter–individual standard deviation of hearing–impaired sub-

jects at 65 dB was more than twice the intra–individual standard deviation. [2] The intra–individual SRT standard deviation of the stationary and slightly fluctuating noises were similar within the normal–hearing and hearing–impaired subjects. The intra–individual SRT standard deviation of the strongly fluctuating noise *icra5* was larger than of the other noise types for both groups of subjects. The inter–individual SRT standard deviation of both fluctuating noises *icra5* and *icra7* exceeded twice the intra–individual SRT standard deviations for the hearing–impaired subjects.

The intra– and inter–individual standard deviations of the slope data were comparable in all settings for both groups of subjects. Therefore, it was not possible to use the slope data to differentiate significantly between the subjects. Both intra–individual and inter–individual slope standard deviations were smaller for fluctuating interfering noises than for stationary noises for both groups of subjects, because the slope results were rather low (all standard deviations were similar when given relatively to the respective slope: $0.2 - 0.3 \cdot \text{slope}$).

5.4 EXPERIMENT I: NOISE PRESENTATION LEVEL

5.4.1 Results

Table 5.3 shows the mean SRT and slope values for various noise levels for normal–hearing subjects (upper part) and hearing–impaired subjects (lower part). In the latter case, the mean SRT and slope values for the noise levels belonging to the individual loudness categories *15, 20* and *25 cu* as well as for the noise levels *65* and *85 dB SPL* are reported.

Neither the SRTs nor the mean slopes differed significantly for the normal–hearing or the hearing–impaired subjects across noise levels (one way ANOVA).

The mean slope values of the hearing–impaired subjects were slightly, but not significantly lower than the normal–hearing data, which is consistent with the expectation.

[2] The intra– and inter–individual SRT standard deviations of hearing–impaired subjects at 25 cu were smaller than at the other levels, because only six subjects were included in these values. For the missing subjects, the levels corresponding to 25 cu exceeded 85 dB SPL. The parameter setting *olnoise* in the lower panels of Figure 5.2 shows the standard deviations of all 10 hearing–impaired subjects (including six results at 25 cu, three at 20 cu, and one at 15 cu).

Table 5.3: Mean SRT and slope values of different noise levels for normal–hearing (NH) and hearing–impaired subjects (HI).

		15 cu	20 cu	25 cu	30 cu	35 cu
NH:	SRT [dB SNR]	-6.2	-6.5	-6.2	-6.2	-5.9
	slope [1/dB]	0.19	0.21	0.18	0.18	0.17
		15 cu	20 cu	25 cu	65 dB	85 dB
HI:	SRT [dB SNR]	-3.0	-3.7	-3.9	-2.0	-3.3
	slope [1/dB]	0.16	0.17	0.17	0.16	0.16

5.4.2 Discussion

The dependency of presentation level on speech intelligibility has often been investigated with different results. Some studies showed an effect of level (Studebaker *et al.*, 1999; Goshorn and Studebaker, 1994; Beattie, 1989; Hagerman, 1982; Pickett and Pollack, 1958): speech intelligibility became worse with increasing speech level. Other studies did not show any effect of level (Smoorenburg, 1992; Duquesnoy, 1983a; Plomp, 1978; Speaks *et al.*, 1967; Hirsh *et al.*, 1954; Hirsh and Bowman, 1953; Hawkins and Stevens, 1950), when the noise level exceeded a certain threshold (40–50 dB SPL). Some of the former articles only report a level effect in hearing–impaired subjects when audibility is considered (Studebaker *et al.*, 1999). In this study, we investigated the influence of level corresponding to the individual different loudness categories, therefore overall audibility was considered (as no hearing loss compensation was applied, this does not hold implicitly for all frequencies). In this study, there was no statistically significant level effect. This is remarkable when comparing to the results for normal–hearing subjects by Hagerman (1982), who found a level effect using a similar sentence test. The only intelligibility–intensity relation similar to Hagerman's results could be found regarding the mean normal–hearing values of this study: best performance at 20 cu (55 dB SPL), worst performance at 35 cu (80 dB SPL). However, the differences between the presentation levels were smaller than the intra–individual SRT standard deviations and smaller than the differences reported by Hagerman.

One possible explanation for the different results in former literature is that the main level effects can only be seen at significantly high presentation levels where distortion effects may become significant both in the subject's auditory system and in some outdated audiometric equipment. The level range used in this study was motivated by

the levels mostly used in diagnosis, and was limited to 80 dB SPL (normal–hearing subjects) and 85 dB SPL (hearing–impaired subjects), in order to avoid any inconvenience. No statistically significant level effect could be found within this range. Unfortunately, the literature data were often only presented descriptively and no statistical test was presented about the level effect. Using the given normal–hearing inter–individual standard deviations as a measure of test–retest reliability, no significant level effect of the signal–to–noise ratios according to 50 rau [rationalized arcsine units (Studebaker, 1985)] was assumed within a speech level range of 64–79 dB SPL in (Studebaker *et al.*, 1999). Other studies (Pickett and Pollack, 1958; Pollack and Pickett, 1958) used 75 or 80 dB SPL as lowest noise presentation level. Beattie (1989) compared SRT results for 65 and 85 dB SPL presentation level with normal–hearing subjects. Even though the difference of 1.9 dB was statistically significant, the results were nevertheless combined in further analysis, because the standard deviations indicated substantial overlap between both distributions.

5.5 EXPERIMENT II: TYPE OF NOISE

5.5.1 Results

Table 5.4 shows the mean SRT and slope values for the noise types *olnoise, icra1, icra5* and *icra7* for normal–hearing subjects (upper part) and hearing–impaired subjects (lower part). While the results for the two stationary noises *olnoise* and *icra1* are very similar, the result for *icra5* shows a large difference and the less fluctuating noise *icra7* shows similar results as the two stationary noises. These findings are also represented by statistical tests:

Table 5.4: Mean SRT and slope values of different types of noise for normal–hearing (NH) and hearing–impaired subjects (HI).

		olnoise	icra1	icra5	icra7
NH:	SRT [dB SNR]	-6.2	-7.4	-21.6	-9.9
	slope [1/dB]	0.18	0.19	0.05	0.11
HI:	SRT [dB SNR]	-3.3	-3.0	-8.0	-2.1
	slope [1/dB]	0.16	0.16	0.06	0.11

The Hartley test on equality of variances (a prerequisite for the ANOVA) rejected the equality in variance if testing all types of noise for the normal–hearing subjects. Therefore, an analysis of variance was not justified. By omitting *icra5*, which obviously shows different variances and results, the equality in variance of the remaining types could not be rejected.

According to the one way ANOVA (SRT: $F = 144.7$; slope: $F = 78.7$; $p < 0.01$), there was a significant effect of type of noise on both normal–hearing SRT and slope values. The post–hoc LSD test showed the following results for both normal–hearing SRT and slope values: There were no differences in the results for the stationary noises *olnoise* and *icra1*. The stationary values were significantly higher than the results for the *icra7* noise which were significantly higher than the *icra5* results.

The Bartlett test (a prerequisite for the ANOVA) showed normality of distribution and equality of variances for the hearing–impaired slope data. The group variances for SRT (hearing–impaired subjects) did not differ significantly, when the stationary *olnoise, icra1* and the fluctuating noises *icra5, icra7* were analyzed separately.

The SRTs of the hearing–impaired subjects with the stationary noises *olnoise* and *icra1* did not differ significantly (one way ANOVA). There was a significant effect of the type of the fluctuating noise on the hearing–impaired SRTs ($F = 9.3$, $p < 0.01$) as well as for the type of noise on the slope values ($F = 32.3$, $p < 0.01$). The slope data were further investigated with the post–hoc LSD test. There was no difference in slope between the stationary noises *olnoise* and *icra1*. The stationary slopes were significantly higher than the slopes for the *icra7* noise, which were significantly higher than the *icra5* results.

5.5.2 Discussion

Regarding both the normal–hearing and hearing–impaired subjects, the stationary noises *olnoise* and *icra1* generated the largest SRT and slope values. These noises mask the test sentences most efficiently because both long–term spectra correspond to the speech material (Wagener *et al.*, 1999c; Dreschler *et al.*, 2001). Since no differences between the results of the stationary noises could be found, it can be recommended to use the *icra1* noise as standard masking noise for the current speech test. This recommendation also holds for speech tests in other languages that show a long–term spectrum comparable to the mean international long–term spectrum (Byrne *et al.*, 1994). No remarkable changes in the reference values will be expected compared to

using a speech–shaped noise generated by superimposing the speech test's word material.

For the normal–hearing subjects, the fluctuating noises *icra5* and *icra7* generated lower SRT values and shallower slopes than the stationary noises, because these subjects benefited from the valleys in the envelope of the masker (mean benefit for *icra5*: 14 dB). The mean SRT (*icra5:* −21.6 dB SNR) and slope values (*icra5:* 5.2 %/dB) are comparable to the results of Hagerman (Hagerman, 1997), who also used fully fluctuating noise with normal–hearing subjects. That noise was generated by using the original noise of the Hagerman sentences (Hagerman, 1982) and fully modulating it. The sentence material Hagerman used with the fluctuating noise was an optimized version of the original sentences, adapted for the use in that frozen fluctuating noise, i. e. the modulation was phase–locked to the individual sentence and not randomized, as in our case. The mean speech reception threshold equaled about -23.5 dB SNR and the respective slope equaled 9.0 %/dB (Hagerman, 1997).

Different results were found for the hearing–impaired subjects. The fluctuating noise *icra7* (fluctuating like 3 female and 3 male speakers) generated a similar mean SRT as the stationary noises. However, the inter–individual standard deviation (*icra7*: $\sigma_{inter} = 3.0$ dB) was twice that for the stationary noises (*olnoise*: $\sigma_{inter} = 1.3$ and *icra1*: $\sigma_{inter} = 1.6$ dB), while the intra–individual standard deviations were comparable (*icra7*: $\sigma_{intra} = 0.9$ dB, *olnoise, icra1*: $\sigma_{intra} = 0.8$ dB). This indicates a lower SRT for *icra7* for some hearing–impaired subjects than for the stationary noises (similar to normal–hearing subjects), while another group of hearing–impaired subjects exhibits even higher SRT values than for the stationary noises. Averaged across all hearing–impaired subjects, the strongly fluctuating noise *icra5* generated lower SRTs than the less fluctuating noises. However, the subjects who could not benefit from the slight modulations of the *icra7* noise, could also not benefit from the strong modulations of the *icra5* noise. Other hearing–impaired subjects showed a benefit up to 10 dB (compared to stationary noise). Such a spread in hearing–impaired SRT results with fluctuating interfering noise can also be found in Versfeld and Dreschler (2002), though the maximum benefit was smaller in that study (7 dB). Similar to Versfeld and Dreschler (2002), subjects that perform well in fluctuating noise also perform well in stationary noise, but not vice versa. This means that for some hearing–impaired subjects, the valleys of the fluctuating noise became smeared, and therefore act as stationary noise. This division across different hearing–impaired subjects (which also appeared less distinctly using *icra7*) should be further investigated in a further study.

Hagerman also investigated SRTs in fully fluctuating noise with hearing–impaired subjects (Hagerman, 2002). He found a mean benefit of 3.6 dB by the fully fluctuating interfering noise compared to only slightly fluctuating noise. The mean SRT values were comparable to the respective SRTs of this study. Hagerman: slightly fluctuating noise: −3.9 dB SNR, fully fluctuating noise: −7.5 dB SNR. This study: *icra1*: −3.0 dB SNR, *icra5*: −8.0 dB SNR.

The hearing–impaired slope values for the different types of noise were comparable to the normal–hearing data. This means that the slope was influenced more strongly by the type of noise than by the hearing ability, indicating that the external variability introduced by the noise maskers dominates the overall variability across test items (which is represented in the slope).

5.6 EXPERIMENT III: PRESENTATION MODE

5.6.1 Results

The mean SRT and slope values for the presentation modes *synch*, *cont* and *inv*, using normal–hearing (NH) and hearing–impaired subjects (HI), are given in Table 5.5. The slope values did not differ significantly for the different presentation modes for both normal–hearing and hearing–impaired subjects. There was a significant effect of the presentation mode on the normal–hearing SRTs ($F = 7.7$, $p < 0.01$) amounting to 1.4 dB. For the hearing–impaired subjects, no significant SRT difference could be found. According to the post–hoc modified LSD test of the normal–hearing data, there was no difference between *inverse* and *synchronous* presentation mode. These presentation modes generated significant higher SRT values than the *continuous* presentation.

Table 5.5: Mean SRT and slope values of different types of presentation for normal–hearing (NH) and hearing–impaired subjects (HI).

		synchronous	continuous	inverse
NH:	SRT [dB SNR]	-6.2	-7.6	-6.2
	slope [1/dB]	0.18	0.21	0.18
HI:	SRT [dB SNR]	-3.3	-4.5	-3.2
	slope [1/dB]	0.16	0.18	0.18

5.6.2 Discussion

While the SRT results did not differ when either the speech level or the noise level was held constant in the adaptive procedure with gated noise, there was a difference between gated and continuous noise presentation. One hypothesis for the deviation between the *continuous* and the gated (or synchronous) conditions is that the first word (the name) was additionally masked by the abrupt beginning of the noise in the non–continuous conditions. In order to test this hypothesis, the SRT of each word group (name, verb, numeral, adjective, object) was calculated separately for the *synchronous* and *continuous* presentation. The mean SRTs, using normal–hearing (NH) and hearing–impaired subjects (HI), are given in Table 5.6.

Table 5.6: *Mean SRT values for each word group using the synchronous and continuous presentation (normal–hearing subjects: NH, hearing–impaired subjects: HI) in dB SNR.*

		synchronous	continuous
NH:	Name	-7.1	-8.2
	Verb	-6.6	-7.7
	Numeral	-7.0	-8.2
	Adjective	-6.1	-7.6
	Object	-6.3	-7.9
HI:	Name	-4.2	-5.2
	Verb	-2.7	-4.1
	Numeral	-3.1	-4.3
	Adjective	-3.1	-4.7
	Object	-3.7	-4.3

If the starting of the noise masked the name more than the other words, the difference between the SRT values of the name (SRT_{name}) and the other words (SRT_{others}) should be higher for the *synchronous* mode than for the *continuous* mode. The difference amounted to -0.6 dB (normal–hearing) and -1.0 dB (hearing–impaired), respectively, for the synchronous presentation (normal–hearing: SRT_{name} = -7.1 dB SNR, SRT_{others} = -6.5 dB SNR, hearing–impaired: SRT_{name} = -4.2 dB SNR, SRT_{others} = -3.2 dB SNR). For the continuous presentation, on the other hand, the difference amounted to -0.3 dB (normal–hearing) and -0.8 dB (hearing–impaired) (normal–hearing subjects: SRT_{name} = -8.2 dB SNR, SRT_{others} = -7.9 dB SNR, hearing–impaired subjects:

SRT_{name} = -5.2 dB SNR, SRT_{others} = -4.4 dB SNR). Hence, for hearing–impaired subjects, a similar difference occurred for continuous and synchronous noise. For normal–hearing subjects, the difference for synchronous presentation exceeded the one for continuous presentation by only 0.3 dB. This cannot explain the difference of 1.4 dB between the mean overall speech reception thresholds for synchronous and continuous presentation. Table 5.6 indicates that each word group exhibited a SRT difference of about 1 dB between continuous and synchronous representation. This shows that there is a main difference in the effect of synchronous and continuous presentation mode for normal–hearing subjects which distributes evenly across the whole sentence.

5.7 GENERAL DISCUSSION

In this study, the parameters that significantly influence the results of adaptive speech intelligibility measurements in noise should be identified. The presentation levels up to 85 dB SPL, adaptive procedures with gated noise (constant speech or noise level), and different stationary speech shaped noises did not significantly influence SRT results of normal–hearing and hearing–impaired subjects. Using fluctuating noises strongly influence the results, especially the differentiation between subjects is better with this types of noise compared to stationary noise. Using continuous noise instead of gated (synchronous) noise yielded lower SRTs. In the following, the test–retest accuracy and the results for the different parameter conditions are discussed.

Test–retest accuracy

The normal–hearing subjects showed a training effect of 0.7 dB between the test and retest measurements (with 30 sentences each). This is remarkable because the retest measurements were performed on another day and the subjects performed two practice lists (60 sentences) prior to the first measurement. In order to compensate for this training, the retest data of the normal–hearing subjects were corrected by the overall training effect prior to further analysis. The training effect for hearing–impaired subjects was negligible: 0.2 dB was smaller than the measurement accuracy, but there were also some hearing–impaired subjects with large differences between particular test and retest measurement. Hagerman found a similar training effect of 0.1 dB per list with 10 sentences for normal–hearing subjects (Hagerman and Kinnefors, 1995), and

0.3 dB per list for hearing–impaired subjects, using the test–specific, slightly fluctuating interfering noise (Hagerman, 2002).

This noticeable learning effect between test and retest measurements with normal–hearing subjects is certainly due to the structure of the sentence test (cf. Section 5.2.1): All test lists consist of the same word material. Therefore the set size of the test is limited, it is more or less a 'semi–open' test if the subject knows more or less the base word material. [3] The subject benefits from the reduced set size in contrast to a real open test by an elevated chance of guessing a word if only a part of the word is perceived (Bronkhorst *et al.*, 2002). Therefore, it may be necessary to perform even more training before measuring with normal–hearing subjects than used in this study. It has to be proved whether training using 80 sentences instead of 60 sentences before the first measurement will exclude any remaining training effect. It has also to be proved by a long–term study if this extended training has to be performed only once - i. e. if the subjects are thereby 'pre–trained for their entire life' and can later on be retrained by a just few training sentences. Even though this training effect clearly limits the usage of the current speech test for clinical and research purposes, there is no alternative test procedure available: Sentence tests consisting of highly predictable sentences [such as the Göttingen sentence test (Kollmeier and Wesselkamp, 1997) or the Hearing In Noise Test (Nilsson *et al.*, 1994)] do not show any training effect (Brand and Kollmeier, 2002a), but usually cannot be used twice with the same subject, because the meaningful sentences can easily be memorized or words can be guessed by the context. This would generate an incorrect low SRT result. A repeated measurement with the same test list is not possible until a sufficient period of time has passed (i. e., half a year or even longer). Subjects intimately familiar with these meaningful sentence tests will always show lower SRTs than naive subjects with comparable hearing ability. These differences do not occur to the same extent when syntactically fixed, but semantically unpredictable sentences are used as in this study.

The largest test–retest difference between SRT results of the normal–hearing subjects were found using the highly fluctuating noise *icra5* (2 dB). One reason for the large variability in SRT results using this type of noise is the fact that intelligibility strongly varies if the words were presented synchronously to a maximum or minimum of the noise amplitude. Some of the hearing–impaired subjects also showed the largest individual

[3] 'Open' means, the subject can choose the response out of an infinite number of alternatives, i. e. the size of vocabulary; in contrast to a 'closed' test, when a finite number of response alternatives is given to the subject.

test–retest difference for the *icra5* noise, but most of them showed a negative difference - i. e. the retest SRTs were higher than the first measured values. These results were comparable to (Hagerman, 1997; Hagerman, 2002), who used frozen fluctuating noise.

As the training effect is a disadvantage of speech tests with such speech materials (Hagerman, 1982; Wagener *et al.*, 1999b; Wagener *et al.*, 2003), it should be investigated in the future wheter the individual training status could be estimated by analyzing the performed measurement in order to calculate a correction term that compensates for training.

The particular slope values of the hearing–impaired subjects hardly differed from the normal–hearing data (median slope of all parameter settings: hearing–impaired: 14.9 %/dB, normal–hearing: 17.3 %/dB). These values are comparable to other sentence tests in noise: Plomp sentences yield slopes of about 17–20 %/dB (Duquesnoy, 1983a; Duquesnoy, 1983b; Festen and Plomp, 1990) for both normal–hearing and hearing–impaired subjects. Therefore, the SRT values could be determined with a comparable accuracy for normal–hearing and hearing–impaired subjects because the standard deviation of SRT estimate is inversely proportional to the slope. On the other hand it was shown that slope values do not act as a good diagnostic indicator of hearing loss.

The intra–individual standard deviation of slope estimates with normal–hearing subjects was mostly higher than the inter–individual standard deviation. That means that it was not possible to differentiate between different (normal–hearing) subjects based on slope results.

However, the intra–individual standard deviation of SRT estimates with normal–hearing subjects was always smaller than the inter–individual standard deviation, therefore it was possible to differentiate between the different (normal–hearing) subjects based on SRT. In comparison with other measurements using the same sentence test and adaptive procedure (Brand and Kollmeier, 2002a), the intra–individual standard deviations were higher in the present study (0.8 dB instead of 0.4 dB). One main difference between the studies was that some test lists of the Göttingen sentence test (Kollmeier and Wesselkamp, 1997) were performed between the different test lists of the Oldenburg sentence test in (Brand and Kollmeier, 2002a). When the retest data of the present study were corrected by the overall training effect of 0.7 dB, also an intra–individual standard deviation of 0.4 dB was obtained. The inter–individual standard deviations of the fluctuating noises (*icra5*: 3.0 dB, *icra7*: 1.6 dB) were higher than the standard deviations of the other parameter settings. In other words, there were larger

differences in intelligibility between different normal–hearing subjects using fluctuating noise. Using the highly fluctuating noise *icra5*, the intra–individual standard deviation was also significantly higher than the other intra–individual standard deviations. The noise was fully fluctuating, therefore the valleys and peaks in amplitude were very conspicuous. Since no frozen noise was used, the test words could be presented synchronously to a noise valley or a peak. Therefore the intelligibility function of a speech test in highly fluctuating noise was much flatter than the function in a stationary noise. This yielded a lower accuracy for SRT estimation.

The inter–individual standard deviation should be larger than twice the intra–individual standard deviation ('2σ–criterion'), in order to discriminate significantly between the subjects. It was possible to discriminate significantly between the SRTs of the hearing–impaired subjects for the parameter settings *65 dB SPL, icra5, icra7, icra1* and *inv*. The largest differences between the hearing–impaired subjects (i. e., the highest inter–individual standard deviations) were found for a constant noise level of 65 dB SPL and for the fluctuating noises. The differences using a noise presentation level of 65 dB SPL resulted from the audibility of the noise. For some hearing–impaired subjects, a noise at 65 dB SPL is not audible at all frequencies, therefore they performed the test partly 'in quiet'. For other subjects it was a test in noise. Both the intra–individual and the inter–individual standard deviations of stationary noise and strongly fluctuating noise for hearing–impaired subjects were comparable to results by Hagerman (2002). Hagerman, slightly fluctuating noise: $\sigma_{intra} \approx 0.8$, $\sigma_{inter} = 2$; fully fluctuating noise: $\sigma_{intra} \approx 1.9$, $\sigma_{inter} \approx 3$. *icra1*: $\sigma_{intra} \approx 0.8$, $\sigma_{inter} \approx 1.6$; *icra5*: $\sigma_{intra} \approx 1.8$, $\sigma_{inter} \approx 5.3$.

Experiments I–III

Regarding the mean SRT and slope values there was no influence of the noise presentation levels either for normal–hearing or for hearing–impaired subjects. This means that the intelligibility in stationary noise was mainly determined by the signal–to–noise ratio if the noise exceeded hearing threshold. This is consistent with Plomp (1978), Duquesnoy (1983b), and Smoorenburg (1992).

The stationary noises *olnoise* and *icra1* did not show any differences in SRTs and slopes for both normal–hearing and hearing–impaired subjects. The normal–hearing subjects showed a much lower SRT using fluctuating noise. The hearing–impaired subjects appeared to form two groups when using the fluctuating interfering noises:

there were several subjects who benefited from the modulations, and there were some subjects who showed similar SRT values as for stationary noise. These benefits were higher compared to data in literature: Festen and Plomp (1990): 4–6 dB, Peters *et al.* (1998): 6 dB normal–hearing, 1 dB hearing–impaired. The benefit strongly depends on the fraction and duration of sub–threshold amplitudes in the noise waveform. As the *icra5* also simulates speech pauses, it includes silence intervals of up to 2 s duration. The *icra7* does not include such pauses, its benefit is therefore comparable to the cited data (3.7 dB for normal–hearing, 1.2 dB for hearing–impaired subjects).

The stationary noises generated the largest slope for both normal–hearing and hearing–impaired subjects. These slopes were significantly higher than the slope using the *icra7* noise, which was significantly higher than the slope using the *icra5* noise (normal–hearing and hearing–impaired subjects).

Using continuous noise, the mean normal–hearing SRT was significantly lower than using synchronous noise. The data of the hearing–impaired showed the same trend, but was not statistically significant. The difference was not due to an additional masking by the onset of the noise in the synchronous presentation. It seems that the human ear performs a kind of noise reduction mechanism, which performs the best in continuous noise.

5.8 CONCLUSIONS

In order to obtain comparable results for different speech intelligibility tests in noise, the following points should be considered:

1.) While using speech intelligibility tests with unpredictable sentences, but a limited word material, it is necessary to perform appropriate training prior the measurements. After this training, these speech materials can be repeated arbitrarily without any further training effect. For this reason, such speech tests (Hagerman, 1982; Wagener *et al.*, 1999c; Wagener *et al.*, 2003) are recommended for extensive studies. It should be investigated in the future, if it will be possible to compensate for training by adequately analyzing the measurement.

2.) In general, there is no difference in adaptive SRT and slope estimates if using different noise presentation levels up to 85 dB SPL. It is only necessary that the noise is audible at most frequencies to hearing–impaired subjects. Using presentation levels above 85 dB SPL would result in different values according to Studebaker *et al.* (1999),

Pickett and Pollack (1958), Pollack and Pickett (1958). However, these high noise levels may not be suitable in diagnosis or rehabilitation of hearing impairment.

3.) No differences were found in adaptive SRT and slope estimates between the two stationary, speech–shaped noises (*olnoise, icra1*) as interfering noise. As the speech spectra of different languages are rather similar (Byrne *et al.*, 1994), the use of standardized noises like the *icra1* noise that represents the LTASS–spectrum, seems to be adequate even for different speech materials and possibly for different languages. This indicates that it is not necessary to generate a particular interfering noise for a given speech intelligibility test.

4.) The highly fluctuating icra–noise differentiates best between subjects with different degrees of hearing impairment. Unfortunately, it contains sub–threshold intervals of rather long durations up to 2 s. These long pauses may coincide with major parts of speech (like whole sentences), thus yielding high intra–individual variability both for SRT and slope values. Modified highly fluctuating noises with briefer pauses should be considered in the future.

5.) No differences were found between an adaptive procedure with fixed noise level and a similar adaptive procedure with fixed speech level. This indicates that experimenters are free to choose the procedure that fits best to their demands, because the results of both procedures seem to be perfectly comparable.

6.) Continuous interfering noise yielded slightly lower SRTs than noise interrupted between the test sentences. However, this effect is not statistically significant for hearing–impaired subjects. This indicates that studies using continuous and gated noise can only be compared with care.

ACKNOWLEDGEMENTS

We would like to thank Anita Gorges, Müge Kaya, and Dr. Birgitta Gabriel for performing the measurements. The comments on the manuscript by Dr. John Culling are gratefully acknowledged. This study was supported by DFG KO 942/13-3.

Chapter 6

Factors influencing Sentence Intelligibility for Hearing–impaired Subjects in Fluctuating Noise

ABSTRACT

Fluctuating interfering noises seem to be highly suitable for speech audiometry, because they provide a larger inter–individual variability in intelligibility results across subjects compared to stationary noises [Versfeld and Dreschler (2002) and Chapter 5]. However, the underlying mechanisms of this phenomenon and its consequences for speech audiology is not yet completely clear. Therefore, this study explores some of the mechanisms underlying sentence intelligibility in fluctuating noise. Among other factors that describe the individual hearing ability, the influence of speech intelligibility in quiet and in stationary noise on intelligibility in fluctuating noise with different maximum pause durations was investigated in this study. For this purpose, three versions of speech simulating fluctuating interfering noises based on the *icra* noises [International Collegium of Rehabilitative Audiology, (Dreschler *et al.*, 2001)] were used: The original *icra5* noise which simulates one interfering speaker and contains pause durations up to 2 s as well as two modified versions with pause durations limited to 250 ms and 62.5 ms, respectively (*icra5–250* and *icra5–62.5*). Additionally, a speech–spectrum matched stationary noise (*icra1*) was used. The main measurement properties, such as test–retest reliability as well as SRT and speech intelligibility function slope were determined for all interfering noises with 10 hearing–impaired subjects. The low–

predictable Oldenburg sentence test (Wagener *et al.*, 1999c; Wagener *et al.*, 1999a; Wagener *et al.*, 1999b) was used. All three highly fluctuating noises (*icra5, icra5–250, icra5–62.5*) differentiate very well between subjects. Partial rank correlation analysis showed that SRT for fluctuating noise with longest maximum pause durations (*icra5*) mostly depended on SRT in quiet, while SRT for the other fluctuating noises with smaller maximum pause durations (*icra5–250, icra5–62.5*) correlated both with SRT in quiet and in stationary noise.

6.1 INTRODUCTION

Speech intelligibility measurements using highly fluctuating noise [Versfeld and Dreschler (2002) and Chapter 5] showed, that the benefit over stationary interfering noise by exploiting the passages containing soft noise levels or even pauses of the noise varies a lot across hearing–impaired subjects. The subjects in Chapter 5, for example, could be divided into one group who did benefit from the modulation valleys of the noise, i. e., their speech reception threshold (SRT, i. e. signal–to–noise ratio SNR, that yields 50% intelligibility) was lower compared to stationary noise with the same long–term spectrum (about 10 dB difference). The other group, however, did not benefit from the modulations, since they showed equal SRTs for both fluctuating and stationary noise. The benefit a subject may obtain from fluctuating noise strongly depends on the fraction and duration of soft intervals in the noise (Nelson *et al.*, 2003; Dubno *et al.*, 2002; Eisenberg *et al.*, 1995; Duquesnoy, 1983a; Carhart *et al.*, 1969). The strongly fluctuating noise used in Chapter 5, referred to as *icra5*, simulates the long–term spectrum and the modulation properties of one male speaker. However, this noise includes silent intervals of up to 2 s duration. This means, that the short sentences most often used in sentence intelligibility tests could fall completely into such a pause. The SRT differences of the hearing–impaired subjects could be due to a poorer audibility of speech in the noise pauses.

The current chapter evaluates the influence of these pause durations on speech intelligibility for hearing–impaired subjects. Therefore, the test–retest reliability, SRT and intelligibility function slope results, and correlations with SRT in quiet and stationary noise as well as with other factors that describe the individual hearing ability were compared between fluctuating noises with different maximum pause durations. Mostly, the benefit from the modulations is expected to depend on hearing ability in the modulation valleys (Dubno *et al.*, 2002; Eisenberg *et al.*, 1995; Duquesnoy, 1983a;

Carhart *et al.*, 1969). Therefore, individual speech intelligibility in quiet seems to be important for the modulation benefit. The noises with different pause durations were applied in the speech intelligibility measurements in order to investigate if the differences in modulation benefit across hearing–impaired subjects disappear if the silent interval duration is limited. If so, it can be concluded that the modulation benefit is mainly caused by audibility in quiet.

6.2 METHODS

6.2.1 Fluctuating noise with limited pause durations

Two modified versions of the original *icra5* noise were generated. The modification denotes a limitation of the maximum silence interval duration. The version with a maximum pause length of 250 ms is referred to as *icra5–250*. Only 2–3 phonemes could fall into the silent intervals of this noise. The version with a maximum pause length of 62.5 ms is referred to as *icra5–62.5*. Only parts of a phoneme could fall into the silent intervals of this noise. The *icra5–250* noise sounds like simulating continuous speech without any speech pauses. The *icra5–62.5* noise sounds like 'breathless continuous speech'. Fig. 6.1 shows the respective modulation spectra of the original *icra5* noise and the modified noises (left panels). The modulation spectra were calculated within octave frequency bands similar to (Dreschler *et al.*, 2001).[1] The level histograms of the noises are also given in Fig. 6.1 (right panels). The RMS levels were calculated with a time resolution of 1024 samples (approx. 23 ms) and a level resolution of 1 dB. As can be seen in the modulation spectra, a maximum pause length of 250 ms lowers the occurrence of modulation frequencies below 2 Hz considerably, and changes the occurrence of speech like modulations of about 3–4 Hz barely. A maximum pause length of 62.5 ms lowers the occurrence of modulation frequencies up to 4 Hz considerably.

[1]The Hilbert envelopes of the noise signals (2–min tracks) were calculated, exempted from the dc component, and normalized. The envelopes were down–sampled to 100 Hz, a spectrum was calculated by using the FFT. The values were summed within octaves (range from 0.5 to 32 Hz). The modulation indices were calculated by determining the octave band RMS level.

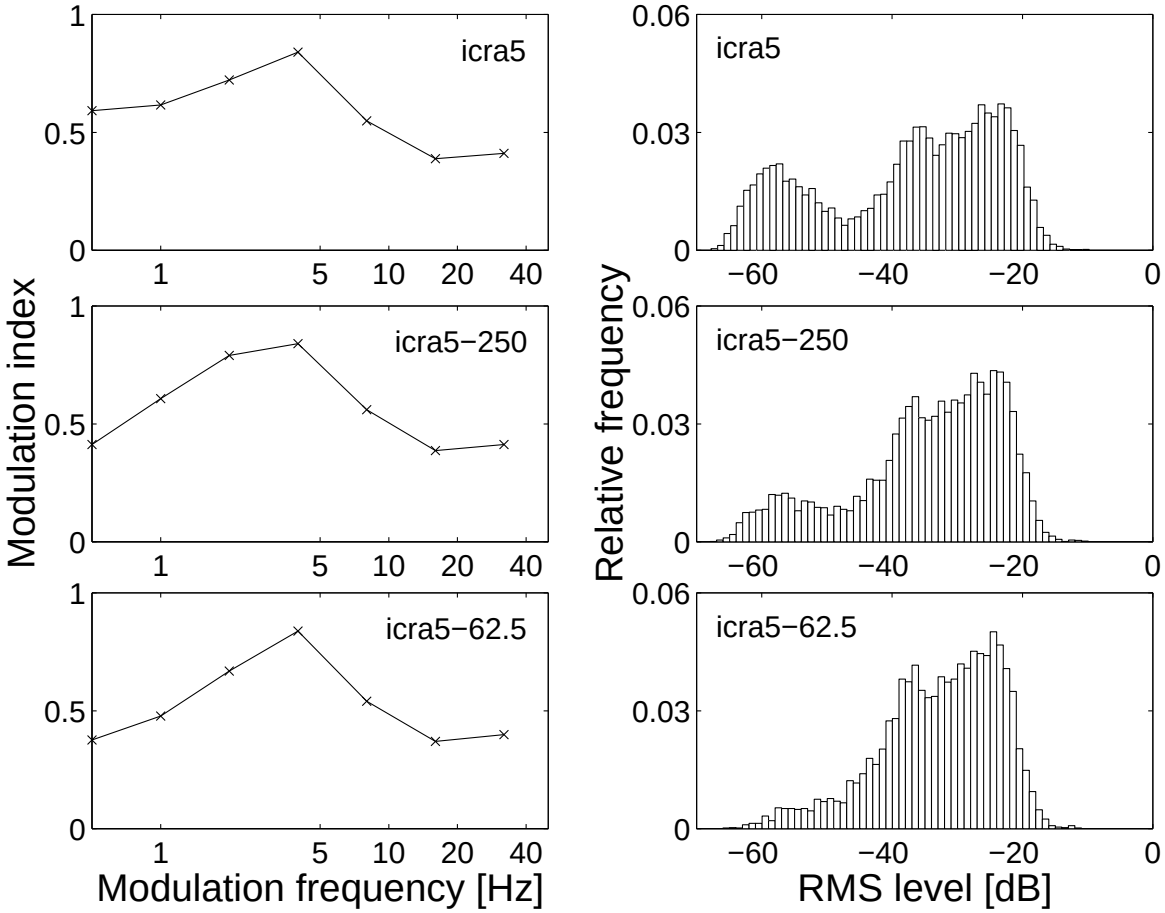

Figure 6.1: *Modulation spectra (left panels) and affiliated level histograms (right panels) of the original icra5 (upper panel), the icra5–250 (medium panel), and the icra5–62.5 noise (lower panel). The modulation indices were calculated within octave frequency bands. The RMS levels were calculated with a time resolution of approx. 23 ms and a level resolution of 1 dB.*

6.2.2 Measurements

SRT and slope of the intelligibility function were determined with both modified *icra5* noises and in quiet using the same procedures and subjects as described in Chapter 5 [Oldenburg sentence test (Wagener *et al.*, 1999c; Wagener *et al.*, 1999a; Wagener *et al.*, 1999b) with 30 test sentences per measurement, adaptive procedure (Brand and Kollmeier, 2002a)].

SRT and slope values were determined as test and retest measurements on two different days (with an interception of about a week). Although the same subjects as in Chapter 5 were used to obtain high comparability of the data, some training was

performed, because about nine months passed between both studies. On the first day, each subject performed two practice lists with 30 sentences; on the second day, only one practice list with 30 sentences was performed.

All measurements were performed in random order, except for the adaptive measurement of SRT in quiet that was performed first on the first session. Both (test and retest) sessions started with determining the individual hearing threshold for the speech shaped stationary noise *olnoise* [2] two times prior the practice measurements.

The pause–limited fluctuating noises were analyzed with respect to test–retest reliability: The comparison of inter– and intra–individual standard deviations quantifies the possibility to differentiate between the subjects. Thus, a small intra and a large inter–individual standard deviation indicate good differentiation.

The SRT and intelligibility function slope results in the pause–limited fluctuating noises were compared with the results in the original fluctuating *icra5* noise and in stationary speech–shaped *icra1* noise that were determined with the same subjects in Chapter 5.

In order to investigate the relation between sentence intelligibility in fluctuating interfering noise and sentence intelligibility in quiet or other factors describing the individual hearing loss, a rank correlation according to Spearman was calculated. The individual SRT and slope data of the noises *icra1, icra5, icra7, icra5–250,* and *icra5–62.5* were correlated with the individual SRT and slope in quiet and *icra1*, medium–loudness level (level that corresponds to loudness 'medium' in a loudness scaling measurement) of speech–shaped stationary *olnoise*, lower slope of the loudness function determined with *olnoise* as stimuli that describes recruitment (Brand and Hohmann, 2002)[3], hearing threshold of *olnoise*, age, hearing loss for pure tones at 500 Hz, and hearing loss for pure tones at 4000 Hz. All values reported here were determined either in this study or were taken from Chapter 5.

[2] *olnoise* denotes the speech–shaped noise that was generated by randomly superimposing the speech material of the Oldenburg sentence test. Therefore, the long–term spectrum of this noise is similar to the mean long–term spectrum of the speech material.

[3] All subjects performed a categorical loudness scaling measurement with the *olnoise* prior the speech intelligibility measurements in Chapter 5. The individual loudness growth function was described by two straight lines that were connected at the medium–loudness level and smoothed in the transition area. The lower slope of the loudness function, i. e. the slope below the medium–loudness level, characterizes recruitment. Loudness functions that represent recruitment show a higher slope than those without recruitment.

6.2.3 Apparatus

The same apparatus and test setup was used for the adaptive speech tests as described in Chapter 5.

A different apparatus (in the same sound insulated booth) was used for determining the individual hearing threshold for the *olnoise*. The same freefield–equalized Sennheiser HDA 200 headphones, connected to a Tucker–Davis system 2 headphone–attenuator, were used to present the *olnoise* played back by a Kenwood DP–5090 compact disc player. The attenuator was controlled manually in order to determine the hearing threshold similar to a pure–tone audiometer.

Both apparatus were calibrated to dB SPL with a B&K artificial ear 4153, a B&K 0.5 inch microphone 4143, a B&K preamplifier 2669, and a B&K measuring amplifier 2610.

6.2.4 Subjects

The same ten sensorineural hearing–impaired subjects (three females, seven males; aged 59–79 years; median age: 70) as in Chapter 5 participated in the measurements. They showed different types and degrees of sensorineural hearing loss with no conductive hearing loss. Pure–tone hearing thresholds ranged from 10 dB HL up to more than 100 dB HL. The types of hearing loss were: three broadband hearing losses, two pure high–frequency hearing losses, one combined low– and high–frequency hearing loss, and four sloping hearing losses. The audiogram data of each ear that was tested in this study are shown in Figure 6.2. All subjects were paid for their participation on an hourly basis.

6.3 RESULTS

6.3.1 Test–retest reliability

In order to investigate the suitability of the modified fluctuating noises *icra5–250* and *icra5–62.5* for sentence intelligibility tests, the test–retest differences for each condition as well as the inter–individual and intra–individual standard deviations of SRT and slope were determined for each test condition employed here.

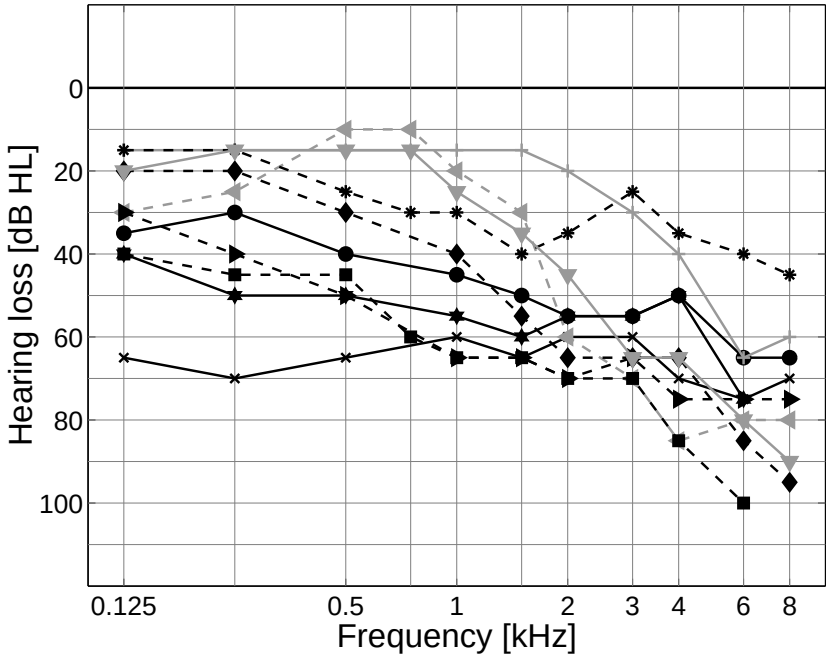

Figure 6.2: *Audiogram data of the respective tested ear of each hearing–impaired subject (broadband hearing loss: black solid line; pure high–frequency hearing loss: gray solid line; combined low– and high–frequency hearing loss: dashed gray line; sloping hearing loss: black dashed line).*

Table 6.1 shows the median SRT test–retest differences as well as the inter– and intra–individual standard deviations of SRT and intelligibility function slope for each test condition of all hearing–impaired subjects. The slope standard deviations are given relatively to the respective slope. For this purpose, the particular mean slope data of all subjects were used for each condition.

The median test–retest differences of the *icra1*, the *icra5* and the *icra5–250* noise were smaller than the test–retest differences of the *icra5–62.5* noise or the quiet condition. While 60 % of the differences using *icra5* or *icra5–250* were negative, 80 % were positive in *icra5–62.5* or quiet (70 % in *icra1*). Some of the individual test and retest differences amounted to 3–4 dB.

The intra–individual SRT standard deviations were similar in all three conditions. The inter–individual SRT standard deviations were at least three times the intra–individual values for the respective conditions. The inter–individual SRT standard deviation in quiet exceeded three times the largest inter–individual SRT standard deviation in

Table 6.1: Median test–retest differences between SRT values (Δ SRT) as well as the intra– (σ_{intra}) and inter–individual standard deviations (σ_{inter}) of SRT and intelligibility function slope for each test condition of all hearing–impaired subjects. The slope standard deviations are given relatively to the respective mean slope.

	quiet	icra5	icra5–250	icra5–62.5	icra1
Δ SRT [dB]	1.4	-0.6	-0.5	1.3	0.4
σ_{intra}, SRT [dB]	1.3	1.8	1.5	1.3	0.8
σ_{inter}, SRT [dB]	17.0	5.3	4.9	5.4	1.6
σ_{intra}, slope	0.5	0.3	0.2	0.2	0.2
σ_{inter}, slope	0.5	0.4	0.2	0.3	0.2

fluctuating noise. This confirms that the differences between different degrees of hearing impairment are larger in quiet than in fluctuating noise. The 2σ–criterion was fulfilled in all conditions: The inter–individual standard deviation should exceed twice the intra–individual standard deviation in order to discriminate significantly between the subjects.

6.3.2 SRT and Intelligibility Function Slopes

The SRT and slope results for the pause–limited *icra5–250* and *icra5–62.5* noise were compared to results with the original fluctuating *icra5* and the stationary *icra1* noise, obtained in Chapter 5, nine months before with the same subjects. Fig. 6.3 shows the individual SRT data for all different types of noise. The individual test and retest data are connected by a line. The noises on the x–axis are ordered from long to short (none) maximum pause durations.

The benefit of using those fluctuating noise instead of stationary was different between subjects. There was a division of the subjects in those who benefited from the modulations and those who did not benefit at all (Chapter 5). These division could also be observed in the pause–limited fluctuating noises *icra5–250, icra5–62.5* of this study. The subjects with the lowest SRT values for the *icra5* noise showed a slightly higher SRT if the pauses were limited to 250 or 62.5 ms. The SRTs of the subjects without any benefit were not influenced by the pause limitations.

The Bartlett test (a prerequisite for the ANOVA) showed normality of distribution

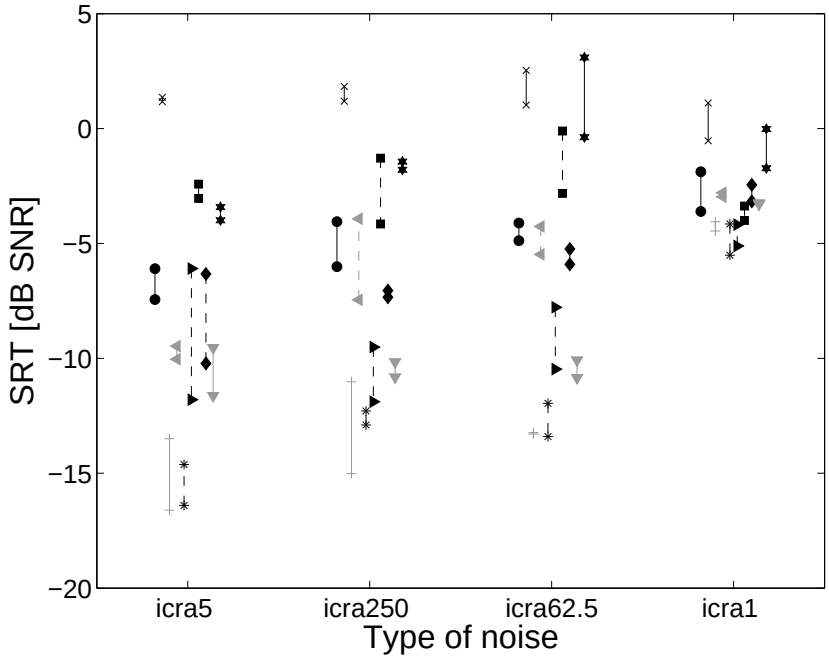

Figure 6.3: *Individual SRTs for different types of noises. The test and retest measurements of each subject are indicated by a particular symbol, they are connected by a line. The same symbols and line styles are used for the particular subjects as in the audiogram data (Fig. 6.2).*

and equality of group variances for all SRTs and slopes in fluctuating noise obtained in Chapter 5 and this study, respectively. On this basis, the data of all subjects were pooled and further analyzed with a one way ANOVA. No significant dependency of SRT or slope on the factor 'type of fluctuating noise' could be found with an error probability of 1 %. This indicates that limiting the maximum pause duration of the fluctuating noises did not significantly affect the SRT and slope results determined in these noises.

The intelligibility functions slope results in quiet of this study (mean slope: 12.3 %/dB) was comparable to the reference value of the Oldenburg sentence test (11.3 %/dB, compare Chapter4). This indicates, that the slope results of hearing–impaired subjects in quiet are similar to those of normal–hearing subjects.

6.3.3 Factors influencing sentence intelligibility in fluctuating noise

Table 6.2 gives the rank correlation coefficients between SRT and slope data in fluctuating / stationary noise on one hand and SRT and slope in quiet and stationary noise, medium–loudness level, lower slope of loudness function, hearing threshold of *olnoise*, age, hearing loss for pure tones at 500 Hz and at 4000 Hz on the other hand. Only significant correlations (error probability of 5% or less) are given. Correlation coefficients with an error probability of less than 1 % are marked by an asterisk.

Table 6.2: *Significant rank correlation coefficients (error probability of 5% or less). SRT and slope data for the noises icra5, icra5–250, icra5–62.5, and icra1 were correlated with the individual SRT and slope in quiet, SRT and slope in icra1, medium–loudness level of speech simulating stationary olnoise (L_{cut}), lower slope of the olnoise loudness function (Brand and Hohmann, 2002) (s_{low}), hearing threshold of olnoise (ht_{ol}), age, hearing loss for pure tones at 500 Hz (hl_{500}), and hearing loss for pure tones at 4000 Hz (hl_{4000}). Significant coefficients on the basis of an error probability of 1% are annotated by an asterisk.*

	icra5		icra5–250		icra5–62.5		icra1	
	SRT	slope	SRT	slope	SRT	slope	SRT	slope
SRT_{quiet}	0.9*	0.7	0.8		0.8*			
$slope_{quiet}$								
SRT_{icra1}	0.7		0.8*		0.8		·/.	
$slope_{icra1}$								·/.
L_{cut}								
s_{low}	0.7		0.8		0.8			
ht_{ol}	0.9*	0.7	0.8*	0.7	0.8*			
Age			0.7		0.7			
hl_{500}	0.9*		0.7		0.7			
hl_{4000}				0.7				

Often, several significant correlations could be found, because only one factor determined all correlations. The subject's age could be such a factor (Dubno *et al.*, 2002). In order to separate the influence of the subject's age, the partial rank correlation of

the same data was calculated. The partial rank coefficients are given in Table 6.3. Similar to Table 6.2, only significant correlations (error probability of 5% or less) are given. Correlation coefficients with an error probability of less than 1 % are marked by an asterisk.

Table 6.3: *Significant partial rank correlation coefficients (the influence of age was neutralized, error probability of 5% or less). SRT and slope data for the noises icra5, icra5–250, icra5–62.5, and icra1 were correlated with the same values as in Table 6.2. Significant coefficients on the basis of an error probability of 1% are annotated by an asterisk.*

	icra5		icra5–250		icra5–62.5		icra1	
	SRT	slope	SRT	slope	SRT	slope	SRT	slope
SRT_{quiet}	0.9*		0.8		0.8*			
$slope_{quiet}$		0.7						
SRT_{icra1}			0.7		0.7		·/.	
$slope_{icra1}$								·/.
L_{cut}								
s_{low}			0.7		0.7			
ht_{ol}	0.9*		0.8	0.7	0.8*			
hl_{500}	0.9*		0.7		0.8*			
hl_{4000}								

Only few correlations disappear when partializing out the influence of age (i.è. SRT of *icra5* with lower slope of loudness function, SRT of *icra5* with SRT of *icra1*, slope of *icra5* with hearing threshold of *olnoise*, and slope of *icra5–250* with hearing loss at 4000 Hz). All other correlations were only minimally affected by partializing out the factor age, indicating that these correlations are not due to a common factor with subject's age. The SRTs in all types of fluctuating noise were significantly correlated with SRT in quiet. The SRTs in pause–limited fluctuating noises were also significantly correlated with SRT in stationary *icra1* noise. SRTs in all fluctuating noises were significantly correlated with hearing threshold of speech–shaped stationary noise and pure–tone threshold at 500 Hz. The SRTs of the pause–limited noises used in this study (*icra5–250* and *icra5–62.5*) also correlated with the individual recruitment, given by the lower slope of the loudness function with speech–shaped stationary noise as stimuli. SRTs in stationary *icra1* noise were not significantly correlated with any other factor

investigated in this study.

6.4 DISCUSSION

In this study, the influence of the maximum pause length in fluctuating noises on speech intelligibility was investigated. Before exploring the detailed results of the pause–limited fluctuating noises it was necessary to investigate the test accuracy when using those noises. The training effect, i. e. the difference in SRT between test and retest, was similar to that reported in Chapter 5 even though the subjects were trained 9 months before data collection on a different task. This indicates that a break of about nine months neutralizes the previous training with the same test. It was shown that the test–retest reliability of all fluctuating noises were similar. The individual test–retest variability was much smaller than the inter–subject variability. Thus, the measurement accuracy and the potential to differentiate between the subjects was not influenced by the maximum length of silence intervals in the noises used in this study. This advantage of the *icra5* noise is not lost by limiting the maximum pause durations.

No significant differences between the mean SRT data could be found for different fluctuating *icra* noises. The separation into two groups of subjects that received benefit from fluctuating noise versus who did not was found in the pause–limited noises *icra5–250, icra5–62.5* as well as in the original *icra5*. Only a few subjects who showed a large benefit in the *icra5* noise, showed a slightly smaller benefit in the pause–limited noises. Those subjects without any benefit did not show any difference if the pause duration was limited. This indicates that the benefit from the modulations is determined by the length of the silence intervals only if a substantial benefit exists for the respective subject. A silence interval length of 62.5 ms is long enough to produce significant benefit compared to stationary interfering noise. Therefore, these pause–limited noises are suitable and should be recommended as interfering noise in speech audiometry in order to preserve the advantages of the speech simulating fluctuating noise *icra5* (simulation of a competitive speaker, high differentiation) while eliminating the disadvantage of that noise (i. e. a coincidental presentation of entire sentences in quiet).

All SRT results in the different fluctuating noises correlated significantly with the SRT in quiet, the hearing threshold of speech shaped stationary noise, and the hearing loss at 500 Hz. All three factors were linked, as the SRT in quiet depends on the hearing threshold of the sentence material (that was simulated by the stationary noise) and

SRT in quiet directly correlates with hearing loss at low frequencies around 500 Hz (Peters *et al.*, 1998; Smoorenburg, 1992). As a striking fact, the SRT in stationary noise did not significantly correlate with those factors. Therefore, speech audiometry with interfering noise provides additional information to that already given by speech intelligibility in quiet and pure–tone thresholds.

Similar to Versfeld and Dreschler (2002), speech intelligibility measurements with fluctuating interfering noises provided a different information than that obtained from stationary interfering noise, since subjects showed small differences in stationary but large differences in fluctuating noise. The pause–limited fluctuating noises *icra5–250* and *icra5–62.5* correlated significantly with both SRT in quiet and SRT in stationary noise. Different to this, the SRT for the original fluctuating noise *icra5* only correlated significantly with SRT in quiet. This indicates that the intelligibility in *icra5* noise was mostly determined by speech intelligibility in quiet. The SRTs of the other fluctuating noises were influenced both by speech intelligibility in quiet as well as in noise. This finding is consistent with the shape of the level distributions, as relatively more silence intervals occur in the *icra5* noise compared to *icra5–250* or *icra5–62.5* (Figure 6.1). Thus, the pause–limited fluctuating noises preserve the advantage of the *icra5* noise (good differentiation between subjects) without showing the disadvantage of extraordinary long pauses.

6.5 CONCLUSIONS

Limiting the maximal duration of sub–threshold intervals in the highly fluctuating *icra5* noise only minimally affects test–retest reliability and the potential of the intelligibility test to differentiate across subjects. The *icra5–250* noise is recommended as interfering noise for future investigation in order to simulate an every–day conversation situation and to achieve good differentiation between subjects. This noise includes the mean modulation frequencies of speech at 3-4 Hz, but does not contain lower frequencies due to longer speech pauses. Sentence intelligibility in fluctuating noise with maximum pause durations up to 250 ms was significantly correlated to both intelligibility in quiet and in stationary noise.

Modeling the individual SRT results in fluctuating noise in more detail is necessary. Thus, it should be further investigated if speech intelligibility in fluctuating noise can be predicted by individual intelligibility in quiet and in stationary noise. This should

be investigated on the basis of a large database with different normal–hearing and hearing–impaired subjects.

ACKNOWLEDGEMENTS

We would like to thank Anita Gorges, Müge Kaya, and Dr. Birgitta Gabriel for performing the measurements. This study was supported by DFG KO 942/13-3.

Chapter 7

Prediction of Sentence Intelligibility in Fluctuating Noise

ABSTRACT

Speech intelligibility measurements using fluctuating interfering noise simulate everyday situations more appropriate than measurements in stationary noises and yield much better differentiation between hearing–impaired subjects with different degrees of hearing loss compared to stationary noises [Versfeld and Dreschler (2002) and Chapter 5]. However, the prediction of speech intelligibility in fluctuating interfering noise is still an unsolved problem: Speech intelligibility predictions using the pure tone audiogram–based Speech Intelligibility Index (SII) or Articulation Index (AI) approach could only partly predict observed data with fluctuating interfering noises (Brand and Kollmeier, 2002b; Dubno *et al.*, 2002; Versfeld and Dreschler, 2002; Peters *et al.*, 1998). Since speech intelligibility in fluctuating noise is highly correlated with both speech intelligibility in quiet and in noise, this study describes two approaches, how the short–time level distribution of the interfering fluctuating noise can be used to predict speech reception threshold (SRT) data in such noises by combining stationary short–time speech intelligibilities. The first approach (one–stage approach) uses the level histogram of the fluctuating noise as description of the interferer and averages across short–time SRTs. The second approach (two–stage model) uses the complete level sequence of the fluctuating noise. Within this approach, sentence intelligibility in fluctuating noise is simulated as a two–stage process. In the first stage, the intelligibility of a particular word is simulated by combining the intelligibilities of

sub–word units. The second stage calculates the intelligibility of the entire sentence as average across word intelligibilities. A context concept similar to the k factor approach by Boothroyd and Nittrouer (1988) is integrated in the first step. In this study, the particular stationary short–time speech intelligibilities were calculated using two different methods: first using the Plomp model about the effect of hearing loss on speech perception (Plomp, 1978) and second using the pure–tone audiogram based SII (ANSI S3.5, 1997). In the Plomp model, the individual SRT in quiet and in stationary noise were used as input to predict the individual SRT at different noise presentation levels. All predictions were compared with speech intelligibility data from 131 ears, including various types of hearing loss. Good predictions with the one–stage approach were obtained when applying an additional linear transformation to the predictions. Using the two–stage approach yields more precise predictions. Compared to the SII predictions with the one–stage approach, a higher prediction accuracy was obtained using the Plomp–based prediction approaches.

7.1 INTRODUCTION

The determination of the individual patients speech intelligibility in noise is relevant for the diagnosis of hearing impairment and the evaluation of rehabilitation with hearing aids. Using fluctuating interfering noises in speech audiometry provides several advantages: Fluctuating noises simulate everyday situations more appropriately than stationary noises because most environmental noises are fluctuating. Furthermore, stationary interfering noises have the disadvantage that the intelligibility differences between different hearing impaired subjects are only small. To overcome this disadvantage, one could perform speech intelligibility measurements in quiet, since the speech reception threshold in quiet (SRT, level that yields 50% speech intelligibility) differentiates very well between different hearing impairments. On the other hand, performing speech intelligibility measurements in quiet gives no additional information, since speech intelligibility can already be very precisely predicted for all listeners with different types of hearing disorders by using input parameters like the individual pure tone audiogram data (Brand and Kollmeier, 2002b). Fluctuating noises yield a better differentiation between hearing–impaired subjects with different degrees of hearing loss compared to stationary noises [Versfeld and Dreschler (2002) and Chapter 5] and can not easily be predicted by audiogram–based approaches like the SII.

Nevertheless, it is worth to investigate more appropriate prediction models for fluctu-

ating interfering noises, since a prediction model is beneficial not only to quantify and test our current understanding of the effects of hearing loss on speech intelligibility, but also to quantify the influence of additional factors that are not included in the model. Such additional factors seem to play a role, if the model fails to predict individual results even though average speech intelligibility data can be reproduced quite well. This study can be regarded as feasibility study of SRT prediction in fluctuating noise. Different approaches to realize such predictions will be presented that vary in complexity of the underlying theory.

A typical situation where audiogram–based models of speech perception like the SII fail is the prediction of speech intelligibility of hearing–impaired listeners in stationary noise: Even though average data can be predicted quite accurately (Brand and Kollmeier, 2002b; Dubno *et al.*, 2002), certain differences between predictions and actual performance still remain. They can most probably be attributed to suprathreshold processing deficits and to an increased 'internal' noise of the individual hearing–impaired listener.

The prediction of speech intelligibility in fluctuating noise is even more complicated since this listening situation includes both speech perception in quiet and in noisy parts of the interferer and the appropriate combination of both situations. Hence, the prediction can be separated in predicting individual performance in quiet and in stationary noise (so–called stationary predictions) and – in a second step – in the prediction of the combined performance in fluctuating noise based on the correct individual performance in quiet and in noise (so–called fluctuating predictions). For the stationary predictions two standard methods are used. That is the Plomp model about speech hearing loss in quiet and stationary noise (Plomp, 1978) and the SII. The validity of Plomp's model for the sentence material used here was evaluated in this study. The validity of the SII was evaluated by Brand and Kollmeier (2002b).

For the fluctuating predictions two approaches were used: a one–stage and a two–stage approach. Both approaches are motivated by the fact that SRTs in fluctuating noises correlate both with SRTs in quiet and in stationary speech shaped noise (Chapter 6). Both approaches have in common that the interfering fluctuating noise was assumed as being stationary within short–time intervals. For each of these short–time levels of the noise an individual short–time intelligibility was calculated. These intelligibilities were combined to obtain the individual effective SRT result for the entire noise. The difference between one–stage and two–stage approach concerns the combination of the short–time SRT results. The one–stage approach simply averages the short–time SRTs

that correspond to the respective short–time levels of the fluctuating noise to obtain the effective SRT result for the entire fluctuating noise. The second–stage approach tries to simulate a possible sentence perception process in fluctuating noise: In the first stage, the speech intelligibility functions of particular sub–word units are statistically combined to obtain the intelligibility function of one word. In the second stage, the word–based intelligibility functions are averaged to obtain the effective SRT for the entire fluctuating noise.

The respective predictions are compared with observations and SII predictions (compare Appendix A) using a large database of speech intelligibility measurements with the Oldenburg sentence test (Brand and Kollmeier, 2002b).

7.2 STATIONARY PREDICTIONS

In order to compute the short–time speech intelligibilities the respective short–time sequence of the fluctuating noise is assumed as being stationary. This means that speech intelligibility has to be predicted in stationary quiet or noisy conditions. First, Plomp's model of speech hearing loss is presented and evaluated. Second, an SII approach is presented.

7.2.1 Plomp's model of speech hearing loss

Plomp's model is based on the fact, that the SRT in noise of normal–hearing listeners depends mostly upon the signal–to–noise ratio SNR (Hawkins and Stevens, 1950). This behavior can be described by Equation 7.1.

$$SRT = 10\ log\,[10^{L_0/10\,dB} + 10^{(L_n - \Delta L_{SN})/10\,dB}]\ \text{in dB} \tag{7.1}$$

L_0 denotes the SRT in quiet [dB(A)], L_n denotes the sound pressure level of interfering noise [dB(A) SPL], and ΔL_{SN} denotes the SRT relative to L_n [dB(A)] (= SNR at threshold with opposite sign).

The first term describes the effect of the ear's internal noise (i. e. hearing threshold), the second term the effect of the external noise. A hearing impairment was separated by Plomp into an attenuation (denoted by 'class A') and a distortion part (denoted by 'class D'). The attenuation part affects only speech intelligibility in quiet, therefore it

changes only the first term of Equation 7.1. The distortion part affects speech intelligibility in quiet as well as in noise, therefore it changes both terms. The resulting SRT of hearing–impaired listeners was given by Plomp as described by Equation 7.2.

$$SRT = 10\ log\,[10^{(L_0+A+D)/10\,dB} + 10^{(L_n-\Delta L_{SN}+D)/10\,dB}]\ \text{in dB} \tag{7.2}$$

A + D denotes the individual speech hearing loss in quiet [dB] and D denotes the individual speech hearing loss in noise [dB].

Equation 7.2 indicates, that SRT values are constant for lower noise levels and depend on the SNR for higher noise levels. Figure 7.1 illustrates the dependency of SRT from the interfering noise level. The reference normal–hearing SRT in quiet for the Oldenburg sentence test ($L_0 = 20$ dB SPL, determined in Chapter 4) and in noise [$\Delta L_{SN} = 7.1$ dB SNR, determined in (Wagener *et al.*, 1999b)] were used to derive the figure. In addition to the normal–hearing behavior of SRT, the influence of a pure attenuation speech hearing loss A = 30 dB, the influence of a pure distortion speech hearing loss D = 10 dB, and the influence of the combination of both is illustrated.

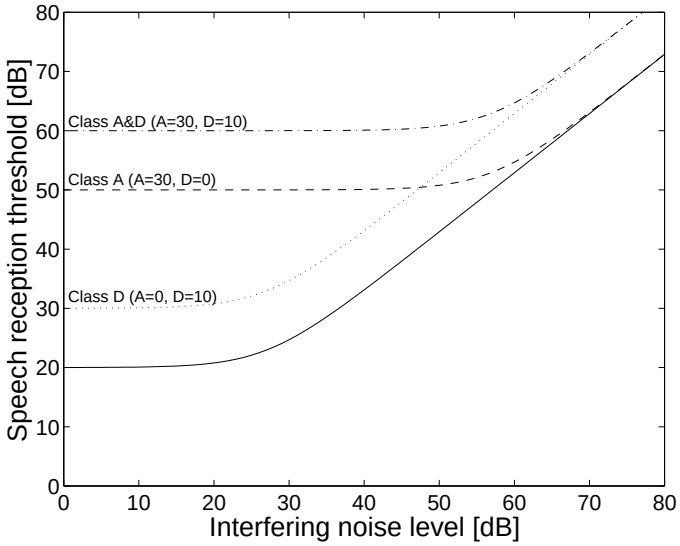

Figure 7.1: *Speech reception thresholds (SRT) as a function of interfering noise level. The solid line illustrates normal–hearing behavior, given by Equation 7.1. The dashed line illustrates a theoretical class A speech hearing loss, the dotted line a class D speech hearing loss, and the dash dotted line a combination of both. Adopted from Plomp (1978), using the reference data of the Oldenburg sentence test.*

Plomp's model denotes that the speech hearing loss can be determined by measuring two speech reception thresholds, namely the SRTs in quiet and at a noise level

sufficiently higher than hearing threshold. This approach was verified by Duquesnoy (1983a) using the Plomp and Mimpen sentences (1979). Plomp's model gave an accurate description of any hearing loss for speech used in Duquesnoy (1983a). In those data for elderly listeners, every 7 dB hearing loss in quiet (A + D) was accompanied by 1 dB hearing loss in noise (D). Plomp estimated the ratio of $D \sim \frac{1}{3}(A + D)$ by analyzing different SRT data of single words as well as sentences (Plomp, 1978).

Note, however, that Plomp's A–component does not take any frequency–dependence of the audibility into account, but attributes any frequency shaping to the D–component. Hence, this 'Distortion'–component includes both the individual frequency shape of the audiogram and any suprathreshold processing deficits. This makes a straight–forward interpretation of the model difficult and is a main reason why the model has not yet been widely accepted in clinical audiology.

7.2.2 Evaluation of Plomp's model with the Oldenburg sentence test

In order to test the applicability of Plomp's model for word–scored sentence tests as the Oldenburg sentence test, the data of Chapter 5 and 6 were analyzed according to Duquesnoy (1983a). Duquesnoy verified Plomp's model experimentally, measuring SRT with the Plomp and Mimpen sentence test based on sentence–scoring (Plomp and Mimpen, 1979) monaurally over headphones in quiet and at four noise levels (28, 43, 58, and 73 dBA) using normal–hearing and hearing–impaired subjects.

In Chapter 5, the SRT data of 10 hearing–impaired subjects at different noise presentation levels were presented for stationary noise. In Chapter 6, the SRT data in quiet were presented for the same listeners. The reference SRT data for normal–hearing subjects using the Oldenburg sentence test in quiet and in noise [Chapter 4 and Wagener *et al.* (1999b)] were used as reference data for $L_0 = 20\,\text{dB SPL}$ and $\Delta L_{SN} = 7.1\,\text{dB SNR}$. The validity of Plomp's model was tested using these subjects with different hearing losses (Figure 7.2).

According to Duquesnoy, Equation 7.2 was fitted to these experimental data individually for each subject. As error criterion the sum of the quadratic differences was used. The initial values of A and D were chosen similarly to Duquesnoy as $A_0 + D_0 = SRT_0 - L_0$ and $D_0 = 0.3 \cdot (SRT_0 - L_0)$, with SRT_0: individual threshold in quiet. The iteration process was stopped when the changes of both A and D were

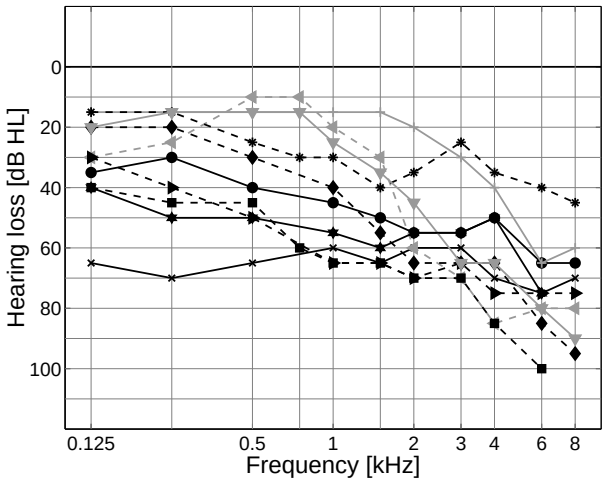

Figure 7.2: Audiogram data of the respective ear of each hearing–impaired subject (broadband hearing loss: black solid line, pure high–frequency hearing loss: gray solid line, combined low– and high–frequency hearing loss: dashed gray line, sloping hearing loss: black dashed line) used for the validation of Plomp's model.

smaller than 0.002 dB. Differently to Duquesnoy, who used the steepest descent method as minimizing procedure, the simplex method (performed by the function 'fmins' of MATLAB R11) was used in this study.

The mean standard deviations of all data points from the fitted curve were determined individually as a measure of goodness of fit.

Figure 7.3 shows the fitted model functions as well the experimental data for each subject.

The median of the standard deviations across subjects equaled 0.47 dB. This was even smaller than the median standard deviation of the normal–hearing group (0.84 dB) in Duquesnoy (1983a).

Duquesnoy stated that the individual hearing loss for speech can be specified by measuring only two thresholds, SRT in quiet and SRT at a high noise level (> 50 dBA). As an additional validation test the model function was fitted to the experimental data using only the data in quiet and at a noise presentation level of 85 dB. The median difference across subjects of the model parameters A (-0.20 dB) and D (-0.24 dB) between the two–points fit and the fit to the complete experimental data set were smaller than the measurement accuracy. These absolute difference values were comparable to those given in Duquesnoy (1983a), although with opposite sign.

The small standard deviation and difference between multi–data based model parameters and parameters based on two measurement results extends Plomp's model to the word–score based Oldenburg sentence test. The model obviously gives a very accurate description of the hearing loss for speech as a function of noise level, using a variety of different hearing losses. The average ratio between A + D and D of 7:1 found by Duquesnoy could not be confirmed, as the average ratio amounts 8:1 with a high standard deviation of 6.6 between all 10 listeners.

7.2.3 Speech Intelligibility Index SII

The audiogram–based SII could predict SRTs in quiet very accurately, whereas the predictions of SRTs in stationary interfering noise are less accurate (Brand and Kollmeier, 2002b). The SII considers the individual hearing threshold in the particular signal–to–noise ratio calculation of the respective frequency band (details see Appendix A). The individual hearing threshold is interpreted as internal masking noise. There are different ways to combine the internal masking noise and the interfering noise of the speech test. Different to a maximum norm that is included in the standard [paragraph 4.5., (ANSI S3.5, 1997)], in this study the amplitudes of both noises were added, assuming that both noises were correlated. This should model that speech perception is most difficult near threshold.

7.3 PREDICTING SENTENCE INTELLIGIBILITY IN FLUCTUATING NOISE

Fluctuating noise can be considered as a sequence of small time intervals of stationary noise with a fixed level that varies across the particular time intervals. Therefore it should be possible to predict the individual SRT in fluctuating noise by calculating a combined speech reception threshold based on the individual stationary short–time intelligibilities in these time intervals. Two different approaches to calculate this 'effective' SRT are presented in the following. One approach considers context effects between the time intervals, the other does not consider any context effects. The predictions were compared to a database of SRT measurements with fluctuating interfering noise.

7.3.1 Measurement database

In order to verify the validity of speech intelligibility predictions with the SII, a large database from audiological measurement results was constructed in cooperation with the center of excellence 'HörTech' (Brand and Kollmeier, 2002b). The database includes data of 131 different ears, including normal–hearing subjects as well as subjects with various types and degrees of hearing loss. All subjects performed the following measurements monaurally via headphones: pure–tone audiogram, categorical loudness scaling with speech–shaped noise (Brand and Hohmann, 2002), adaptive SRT and speech intelligibility function's slope determination (Brand and Kollmeier, 2002a) in quiet as well as in noise employing stationary speech–shaped noise *icra1* [Dreschler *et al.* (2001) and Chapter 5] and fluctuating noise that simulates one interfering speaker with a maximum silence interval length of 250 ms, referred to as *icra5–250* (Chapter 6). The speech intelligibility measurements in noise were performed with a fixed noise presentation level that corresponded to the individual loudness perception category 'medium' of the speech–shaped noise. The individual data in quiet and in *icra1* noise were used to determine two components A and D of the individual hearing loss for speech according to Plomp that are fitted to the individual relation of SRT and noise level (Equation 7.2 and Figure 7.1). These data were used to calculate the short–time intelligibilities in the fluctuating *icra5–250* noise. The effective predictions of the entire noise were compared with the individual measurement results using *icra5–250* noise as interfering noise.

7.3.2 Intelligibility prediction by a one–stage model (weighted sum)

Figure 7.4 shows the level histogram of the fluctuating *icra5–250* noise used in the measurements. The noise was analyzed within time intervals of 11 ms (512 samples). Within each interval, the digital root–mean–square (RMS) level was determined in 1–dB–steps. Level resolutions from 0.2 dB to 6 dB, and time resolutions from 128 to 4096 samples were also tested, but did not yield largely deviating results.

The noise presentation level in the speech intelligibility measurements was calibrated to the overall RMS level of the fluctuating noise. The level histogram was used to calculate the individual short–time presentation levels of the noise. The individual short–time SRT for each particular short–time noise presentation level was calculated both by using the relation of SRT and noise level by Plomp's model (Equation 7.2)

based on the individually fitted A and D values and by the SII [1]. As a first approach, these individual short–time SRT values were averaged across levels in order to obtain an 'effective' SRT for the entire fluctuating *icra5–250* noise. This averaging was realized by calculating the weighted sum across the individual short–time SRT values. The weighting was given by the relative occurrence of the particular short time noise levels in the level histogram. This one–stage approach is a very basic and technical approach without any consideration of a possible speech perception process.

7.3.3 Intelligibility prediction by a two–stage speech perception model

The second prediction approach tries to model a possible speech perception process by using the knowledge about speech perception in stationary noise and quiet concerning context effects and statistics. The perception process is assumed as a two–stage model. The first stage describes the perception of a single word by combining the information of sub–word units. Context effects can be modeled by the assumption that it is sufficient to perceive only a certain fraction of sub–word intervals in order to perceive an entire word. This means that the probability p of correctly perceiving an entire word can be modeled by Equation 7.3.

$$1 - p = \prod_i (1 - p_i) \tag{7.3}$$

p_i denotes the particular probability of correctly perceiving the sub–word unit. This is a generalization of the k factor concept of Boothroyd and Nittrouer (1988).

The second stage describes the overall intelligibility of a sentence. As word–scoring was used in the measurements, this overall intelligibility was modeled by averaging the intelligibilities across all words.

In order to calculate the probabilities of correctly perceiving the sub–word units p_i, the level sequence of the *icra5–250* noise was calculated by determining the short–time presentation levels of the noise in time steps according to the length of the sub–word units. For each short–time noise presentation level, the probability p_i of correctly perceiving a sub–word unit at a given SNR can be derived from the measured intelligibility results

[1] For each short–time noise level an individual short–time SII was calculated. These short–time SIIs were averaged across noise level to obtain the effective individual SII for the entire fluctuating noise. In order to compare the predictions with measured data, the resulting individual SIIs were transformed to SRT values.

of entire words p using Equation 7.4.

$$1 - p_i = (1 - p)^{\frac{1}{k'}} \tag{7.4}$$

with k' denoting the number of sub–word units per word. In this backwards calculation from p to p_i it is assumed for simplicity that all p_i have the same intelligibility function. When combining the short–time intelligibilities of the particular sub–word units to compute the word intelligibility, different intelligibility functions were applied for the respective unit, depending on the short–time noise presentation level. The speech intelligibility p of the entire words is given by the logistic model function Equation 7.5.

$$p\,(SNR, SRT, s) = \frac{1}{1 + e^{4 \cdot s \cdot (SRT - SNR)}} \tag{7.5}$$

As a first attempt the SRT in Equation 7.5 was individually calculated with the fitted A and D values from Plomp's model and the short–time noise presentation level. The slope s of the intelligibility functions was set to $s = 12.3\,\%/\mathrm{dB}$ for intelligibility in quiet (mean value for hearing–impaired subjects, determined in Chapter 6), and $s = 16\,\%/\mathrm{dB}$ for intelligibility in stationary noise (mean value for hearing–impaired subjects, determined in Chapter 5). The slope for noise was used if the current SRT exceeded the individual SRT in quiet by more than 5 dB. Otherwise, the slope for quiet was used.

k' and effective word length

The k' for the words used in this study is an unknown parameter of the model. Preliminary, it was assumed that it is sufficient to correctly perceive two phonemes out of the particular word (that contains on average 6 phonemes) in order to correctly perceive the entire word. This implies a high context between the sub–word units, given by an assumed k' values of three. The relatively large number of phonemes per word is due to the fact that not the real number of words per sentences (five) has to be used here, but the number of statistically independent words per sentence. This number is expressed by the j factor according to Boothroyd and Nittrouer (1988). The mean number of statistically independent perceived parts per sentence j in the Oldenburg test sentences is $\mathrm{j} \approx 3$ to 4 depending on the SNR (Wagener *et al.*, 1999b). j also determines the effective word length wherein the sub–word units were combined. The mean word duration of the Oldenburg sentence test is 400 ms. This indicates a mean length of the statistically independent perceived parts of approximately 600 ms (correspond to

a j of 3.3). Therefore, approximately 6 phonemes are included in one 'effective' word and the error probabilities for the sub–word units were multiplied within time intervals of 600 ms in order to calculate the intelligibility for the particular words. In order to model the word–scoring data analysis, the intelligibilities of these 600 ms intervals were averaged to calculate the 'effective' SRT for the entire fluctuating noise.

7.4 RESULTS

All SRT predictions were compared with the respective observed SRT data for the fluctuating interfering noise.

As a baseline, the one–stage approach of averaging the predicted short–time SRTs was applied to the audiogram–based SII concept. Figure 7.5 shows the prediction results of the one–stage SII approach. 63 % of the predictions with fluctuating noise fell within the estimated 95 %–confidence interval of the observed data, the correlation coefficient equaled r = 0.78, the slope of the regression line equaled 2.5.

Figure 7.6 shows the predictions of the one–stage approach of averaging the predicted short–time SRTs based on Plomp's model across short–time noise levels compared to the observed data. 31 % of the prediction data fell within the estimated 95 %–confidence interval, the correlation equaled r = 0.87. It can be seen in Figure 7.6, that the slope of the data distribution is too high (2.2) since an 'ideal' prediction of the data would result in a regression line with slope = 1. A linear transformation of the predicted data was applied in order to obtain a lower slope (Equation 7.6).

$$SRT_{transf} = 2.5 \cdot SRT_{pred} + 20 \tag{7.6}$$

Figure 7.7 shows the comparison of the linear transformed predictions and the observed data. The transformation function was fitted by eye, in order to result in smallest deviations for both normal–hearing and hearing–impaired subjects. 75 % of the transformed prediction data fell within the estimated 95 %–confidence interval, the correlation equaled r = 0.87, slope = 0.9. The linear relation within the prediction data and therefore the correlation did not change, because a linear transformation was used. The transformation yielded better prediction of the data than predictions with the simple one–stage approach (based on Plomp's model or SII). As only a linear transformation was applied, and therefore no additional information was included in the predictions, it seems that sufficient information for good predictions was included in the input data

used for the predictions. Apparently, this information was not adequately utilized in the simple one–stage approach using the weighted sum to adequately predict the 'effective' SRT in fluctuating noise. As the transformation includes both a shift and a factor, the predictions with the simple one–stage approach includes both a bias and a compression compared with the data.

The results of the two–stage model with $k' = 3$ is shown in Figure 7.8. Only 15 % of the prediction data fell into the 95 %–confidence interval (namely the normal–hearing data), the correlation equaled r = 0.86. The slope of the regression line equaled 2.4.

The two–stage model was then modified to include different context values for normal–hearing and hearing–impaired subjects. It was assumed that hearing–impaired listeners need to perceive all sub–word units in order to perceive the entire word. This implied a k' of 1. In order to simulate the gradual 'context loss' between normal–hearing and hearing–impaired listeners, a k' of 2 was applied at the limits between normal and impaired hearing. The individual SRT in quiet was used to define the limits. The most accurate predictions were obtained when using $k = 3$ if the individual SRT in quiet was less than 20 dB SPL; $k = 2$ if the individual SRT in quiet was less than 30 dB SPL; and $k = 1$ if the individual STR in quiet exceeded 30 dB SPL. Figure 7.9 shows the predictions of the two–stage model with different context values. 80 % of these predictions fell within the estimated 95 %–confidence interval of the observed data, the correlation equaled r = 0.89, the slope of the regression line equaled 1.0.

The parameters describing the goodness of the predictions for each approach are given in Table 7.1. This includes the percentage of predictions that fell within the estimated 95 %–confidence interval of the observed data (this interval was given by ± twice the mean estimated intra–individual standard deviation of the data = ± 4 dB [2]). The correlation coefficient between predictions and observed data as well as the slope of the regression line are given in Table 7.1.

7.5 DISCUSSION

Plomp's model about hearing loss for speech (Plomp, 1978) is valid for speech intelligibility determined with the word–scoring Oldenburg sentence test. It is sufficient to use

[2] The mean intra–individual standard deviation of the SRT data was conservatively estimated by 2 dB. This value exceeded the mean intra–individual SRT standard deviation of 1.5 dB that was determined with a smaller subject collective that included less severe hearing impairments (Chapter 6).

Table 7.1: Comparison of SRT predictions with observed SRT data in fluctuating interfering noise. Percentage of predictions N_{95} that fell within the estimated 95 %–confidence interval of the observed data, correlation coefficient r between predictions and observed data, and slope of the regression line s_{reg} are given. Results for the one–stage approach based on SII (denoted as '1 stage, SII'); based on Plomp's model (denoted as '1 stage, Plomp'); the linear transformed one–stage approach based on Plomp's model (denoted as '1 stage, Plomp, linear transformed'); the two–stage model with $k' = 3$ (denoted as '2 stage, $k' = 3$'); and the two–stage model with three different k' (denoted as '2 stage, $k' = 1 - 3$') are shown.

	N_{95}	r	s_{reg}
1 stage, SII	63 %	0.78	2.5
1 stage, Plomp	31 %	0.87	2.2
1 stage, Plomp, linear transformed	75 %	0.87	0.9
2 stage, $k' = 3$	15 %	0.86	2.4
2 stage, $k' = 1 - 3$	80 %	0.89	1.0

only one individual SRT measurement in quiet and one in stationary noise to determine the SRT relation to noise level with a high accuracy. It is possible to determine the sum of the individual A– and D–component (A+D) of Plomp's model with a high accuracy when using the individual audiogram data in an SII prediction, because A+D simply describes the SRT difference in quiet between the individual subject and the normal–hearing reference value. According to Brand and Kollmeier (2002b), individual SRTs in quiet can be predicted very accurately with the SII. Since it is not possible to predict SRTs in stationary noise with sufficient accuracy using the SII, it is not possible to determine the individual D–component of Plomp's model based only on the audiogram data and SII calculation. This shows that the D–component includes both the individual audiogram and a suprathreshold processing deficit that is not modeled by the SII.

The basic approach of averaging the short–time SRTs in order to predict the effective SRT in fluctuating noise (one–stage approach based on Plomp's model) gave unsatisfying results. However, it was possible to improve the predictions considerably by applying a linear transformation to the predictions. This clearly shows that the desired information is somehow included in the predictions. However, the need of the linear transformation shows that the model is not yet adequate. It was not possible to

find such a common linear transformation for the one–stage approach based on the SII that improved those predictions, since there was no linear relationship within the SII predictions.

In order to model the speech perception process in fluctuating noise by a two–stage approach, a context concept was integrated in the first stage. The concept was formulated following the k factor approach of Boothroyd and Nittrouer (1988). However, the meaning of k' that was used in the present study is different to the k factor: Boothroyd and Nittrouer (1988) defines k as the amount of context calculated between a word without context and a word with context. The k' used here can be regarded as context within a word caused by the fluctuations of the interferer. This is connected to the benefit in speech intelligibility between stationary and fluctuating interfering noise. A high benefit means high context. This can motivate the necessity of using a k' which depends on the hearing impairment: Normal–hearing listeners show a larger benefit than hearing–impaired listeners when using fluctuating interfering noise instead of stationary noise. This means that normal–hearing listeners can correctly guess missing sub–word units of a word when just one or two sub–word units per word were correctly perceived, since they are able to listen into the noise level valleys. Therefore, the context between the sub–word units is high for normal–hearing listeners and it is only necessary to correctly perceive some parts of a word (for the speech material of this study equivalent to 2 phonemes: $k' = 3$). In hearing–impaired listeners, the context effect is small: the subjects takes less advantage of the 'listening into the gaps' of the noise because these gaps are masked by an internal noise (i. e. the raised hearing threshold level). Because of the decreased probability of perceiving sub–word units, the probability of guessing not perceived sub–word units decreases as well. Consequently, hearing–impaired listeners show less context and have to perceive the entire word correctly ($k = 1$).

Although the underlying context concept is different, the differences between the $k' = 3$ for fluctuating noise and the $k = 1.3$ for stationary noise (Boothroyd and Nittrouer, 1988) using single words in normal–hearing listeners possibly imply that the context within words is higher in fluctuating noise than in stationary noise. However, this effect is overlapping with other factors that influence the context within words. Boothroyd and Nittrouer (1988) used one–syllable words. 70 % of the word material in this study consists of two–syllable words. In addition, the context of the Oldenburg word material is very high because of the limited word material: the chance of correctly guessing a sub–word unit is higher when only 50 different words are used.

The main disadvantage of the model in its current form is that the very rapid decrease of k (from 3 to 1) for SRT values in quiet of approximately 10 dB above normal can hardly be motivated. Therefore, these context effects should be further investigated regarding more elaborate context models. An advanced context model was developed by Bronkhorst *et al.* (1993). That model was successfully applied to sentence intelligibility data determined with the Oldenburg sentence test (Bronkhorst *et al.*, 2002). In the future this model should be applied to the present two–stage model of speech perception in fluctuating noise. It is very likely that a complete model of speech intelligibility in fluctuating noise (if it exists at all) has to include not only a very elaborate context model but also the effects of time resolution and forward– and backward–masking. Such a model would be beyond the scope of this thesis.

It is difficult to compare the predictions with other SII data in literature, because mostly speech intelligibility predictions were used in order to quantify the hearing deficits that were due to masking in noise and to separate these deficits from other factors of hearing loss (Peters *et al.*, 1998). Peters *et al.* (1998) applied the Articulation Index AI without integrating the interfering noise in the calculation of AI. In another study, the absolute values of the SII predictions were accounted to be irrelevant, therefore only the relative relationship between measured SRT and predicted SII was confirmed (Versfeld and Dreschler, 2002). In that particular study, the SII was calculated assuming the fluctuating noise was stationary. Dubno *et al.* (2002) directly compared two types of AI predictions with observed data in fluctuating interfering noise. The first approach used the corrections described in Sec. 5.1.2 (Masking by Non–Steady–State Noise) of the AI standard (ANSI S3.5, 1969). These predictions gave nearly always substantially higher scores than observed. According to Dubno *et al.* (2002), the best results were obtained with the second approach described by Houtgast *et al.* (1992). This approach was similar to the one–stage approach based on short–time SIIs in the present study as the AI was calculated by averaging the noise level over time. As the calculations included in the AI are linear within a 30 dB range, this should result in similar results as calculating short–time SIIs and averaging these values over time (one–stage, SII). Dubno *et al.* (2002) could predict the observed data of younger subjects quiet well [nearly 70 % of the predictions fell into the estimated 95 %–confidence interval of the measurements, compare Figure 8 in Dubno *et al.* (2002)]. The results of nearly all older subjects were poorer than predicted [only approx. 13 % of the predictions fell into the estimated 95 %–confidence interval of the measurements, compare Figure 8 in Dubno *et al.* (2002)]. The SII predictions of the

present study were comparable to the AI predictions of Dubno *et al.* (2002), Houtgast approach, with younger subjects.

The two–stage speech perception model with integrated context concept introduced in the present study could predict the speech intelligibility results in fluctuating interfering noise more precisely than the one–stage model based on SII and SII / AI predictions in the literature. It seems that the two–stage approach reflects the 'effective' behavior of the sentence perception process in fluctuating noise in a more appropriate way. However, the necessity to use different k' values for different hearing losses can be motivated but can not yet be confirmed by literature data. Therefore, it should be further investigated whether such an context loss in hearing–impaired subjects can also be described by stronger forward masking compared to normal–hearing subjects (Nelson and Freyman, 1987; Glasberg *et al.*, 1987). Such an approach would use an 'effective' level course of the noise, considering forward masking of soft noise intervals by previous loud intervals when determining the short–time noise levels.

7.6 CONCLUSIONS

1) Plomp's model about hearing loss for speech (Plomp, 1978) is valid for speech intelligibility determined with the word–scoring Oldenburg sentence test.

2) It is possible to predict speech intelligibility in fluctuating noise of normal–hearing and hearing–impaired subjects by using individual SRTs in quiet and in stationary noise as input parameters.

3) The perception of sentences in fluctuating noise can be described by a two–stage process that integrates a context concept. The perception of sub–word units considering the respective context yields the intelligibility of the particular words. Word–scoring implies that the intelligibility of the particular words are averaged to obtain the total intelligibility in fluctuating noise.

4) Assuming different context effects for different degrees of hearing loss yields very accurate predictions by the two–stage speech perception model.

5) It should be further investigated whether the different characteristics of normal–hearing and hearing–impaired subjects can also be modeled by considering forward masking in the two–stage speech perception model.

6) The two–stage speech perception model with integrated context concept should be

adapted to the SII concept.

7) A measure of suprathreshold processing deficits should be included in the SII concept for fluctuating noise.

ACKNOWLEDGEMENTS

This study was supported by DFG KO 942/13-3/4.

The measurement database was built in cooperation with the National Center of Excellence 'HörTech' (BMBF).

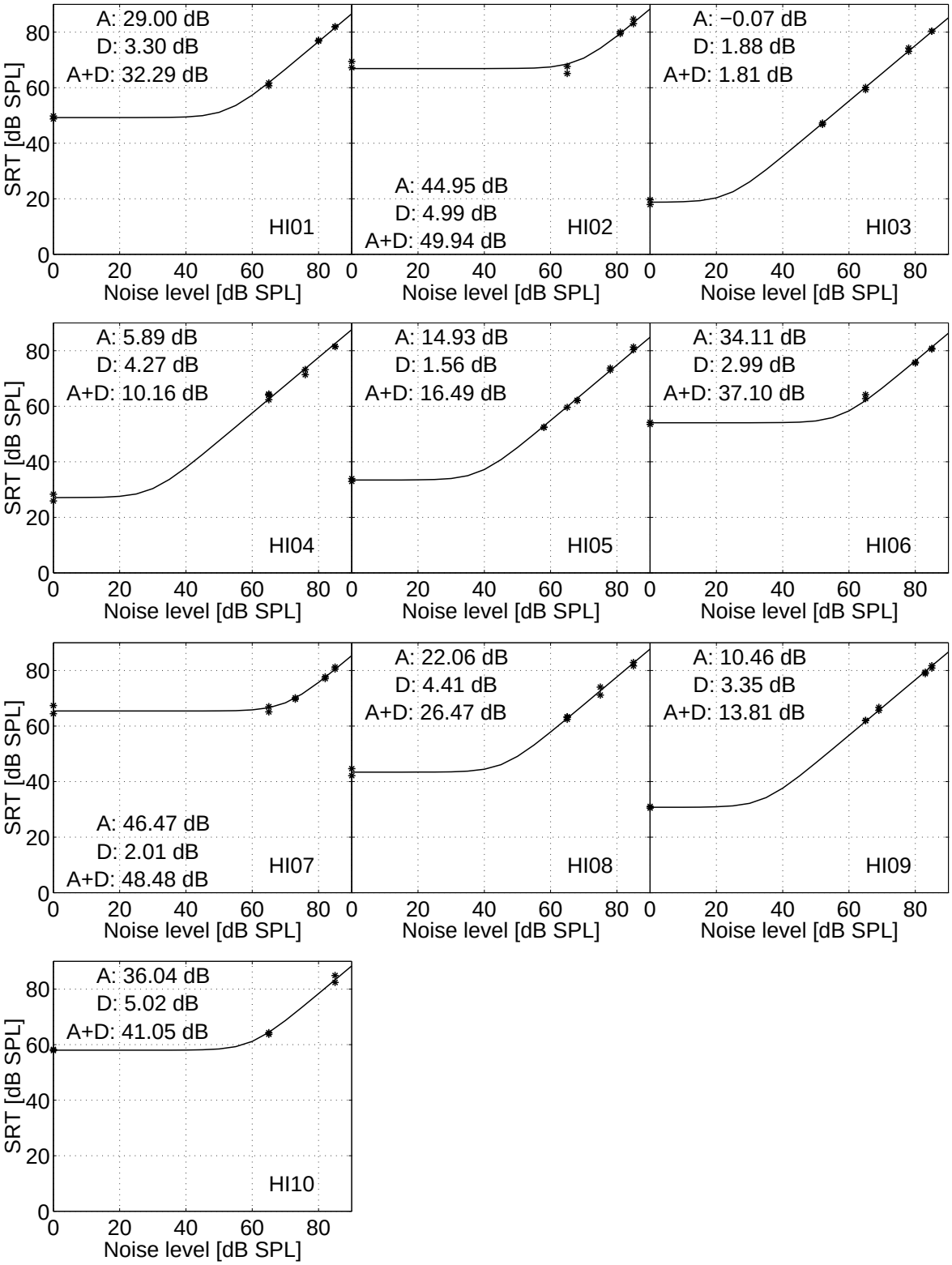

Figure 7.3: *Plomp's model function (Equation 7.2) fitted to the Oldenburg sentence test results of 10 hearing–impaired subjects with different types of hearing loss.*

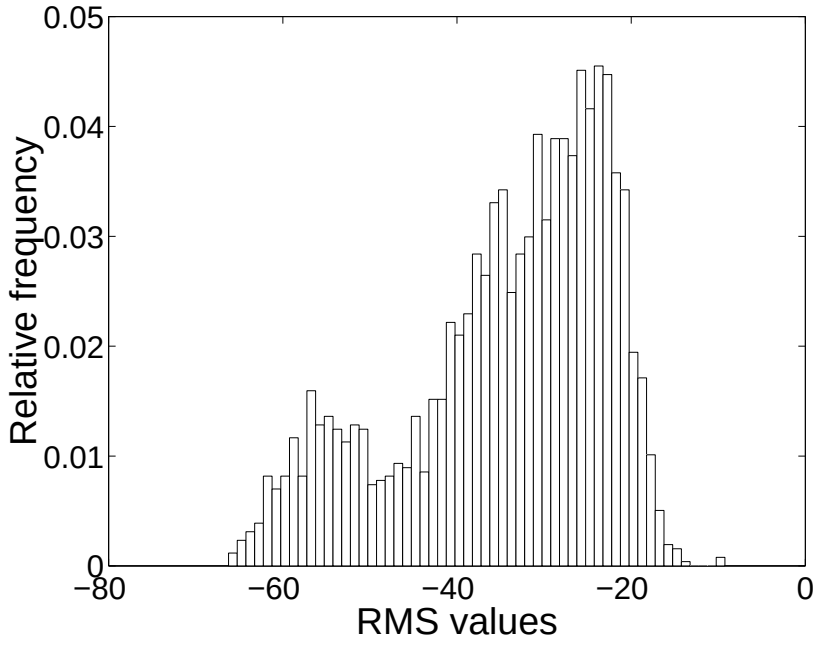

Figure 7.4: *Level histogram of fluctuating noise icra5–250. The noise was analyzed within time intervals of 11 ms. Within each interval, the digital root–mean–square (RMS) level was determined in 1–dB–steps.*

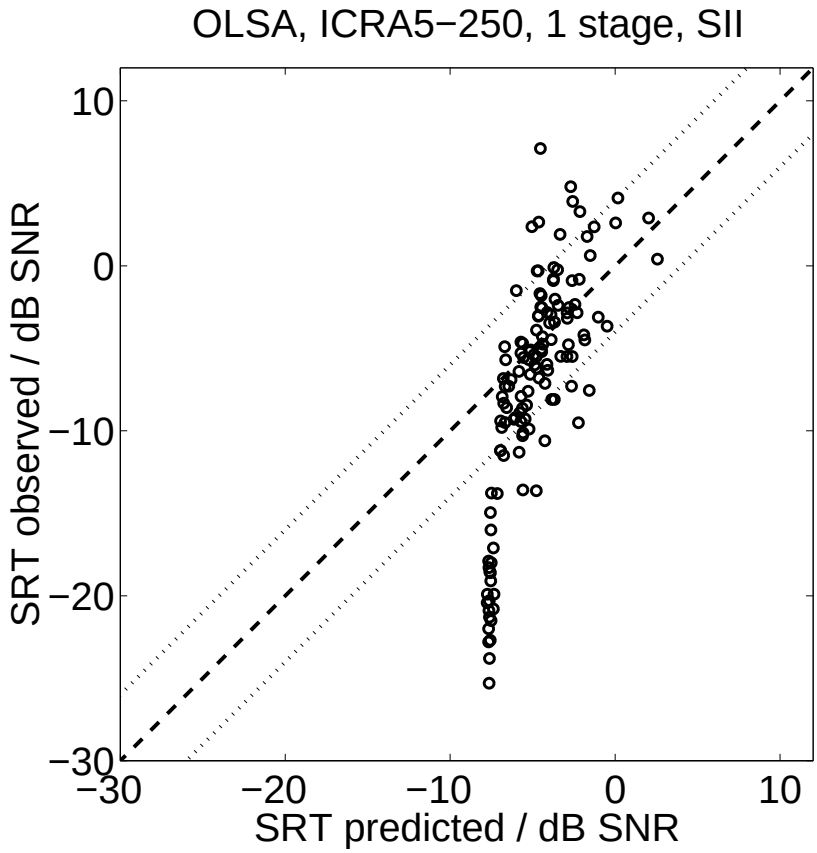

Figure 7.5: Predictions of SRT in icra5–250 noise with the one–stage approach based on SII compared to observed data determined with the Oldenburg sentence test. The dotted lines indicate the estimated 95 % confidence interval of the measurement results.

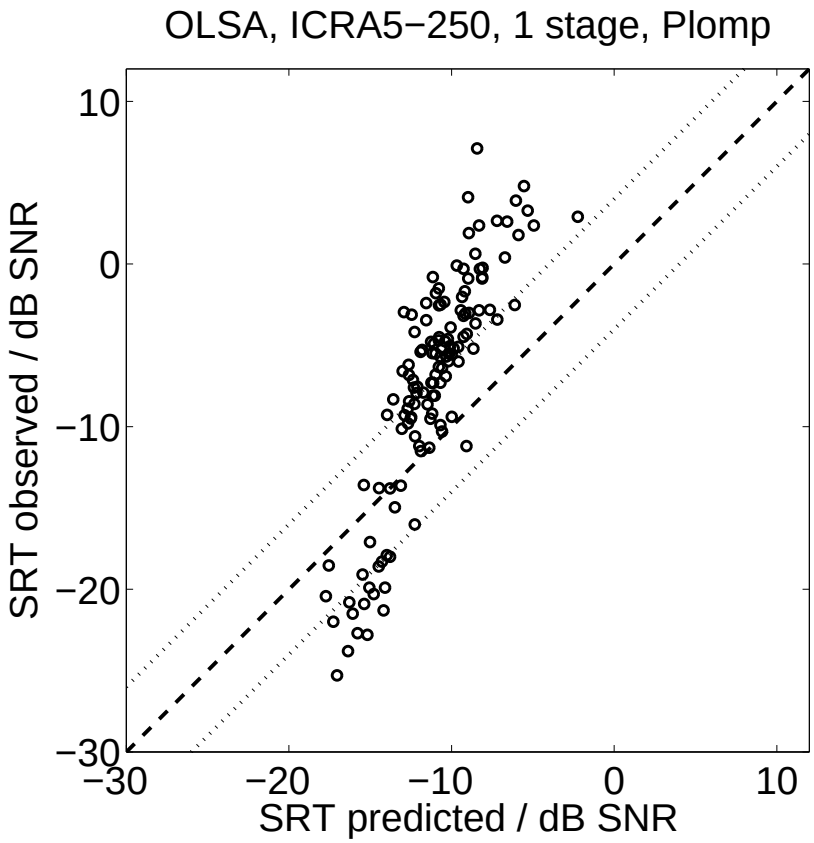

Figure 7.6: *Intelligibility predictions for icra5–250 noise by the one–stage approach based on Plomp's model. Each circle indicates one pair of observed/predicted SRT. The dashed lines indicate the estimated 95 %–confidence interval of the observed data.*

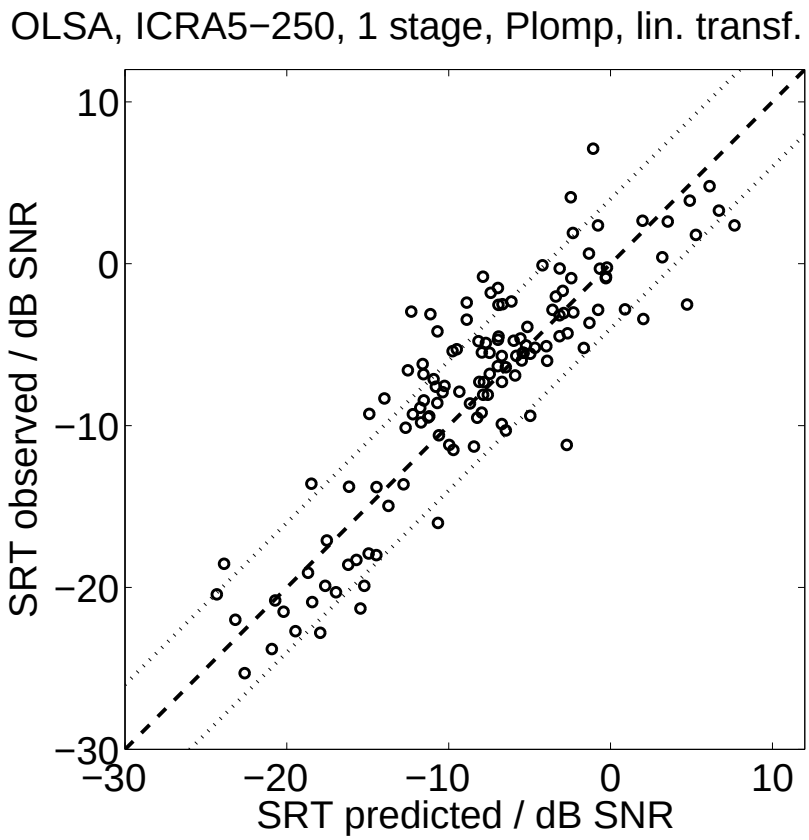

Figure 7.7: *Intelligibility predictions for icra5–250 noise by the one–stage approach based on Plomp's model and linear transformation. Each circle indicates one pair of observed/predicted SRT. The dashed lines indicate the estimated 95 %–confidence interval of the observed data.*

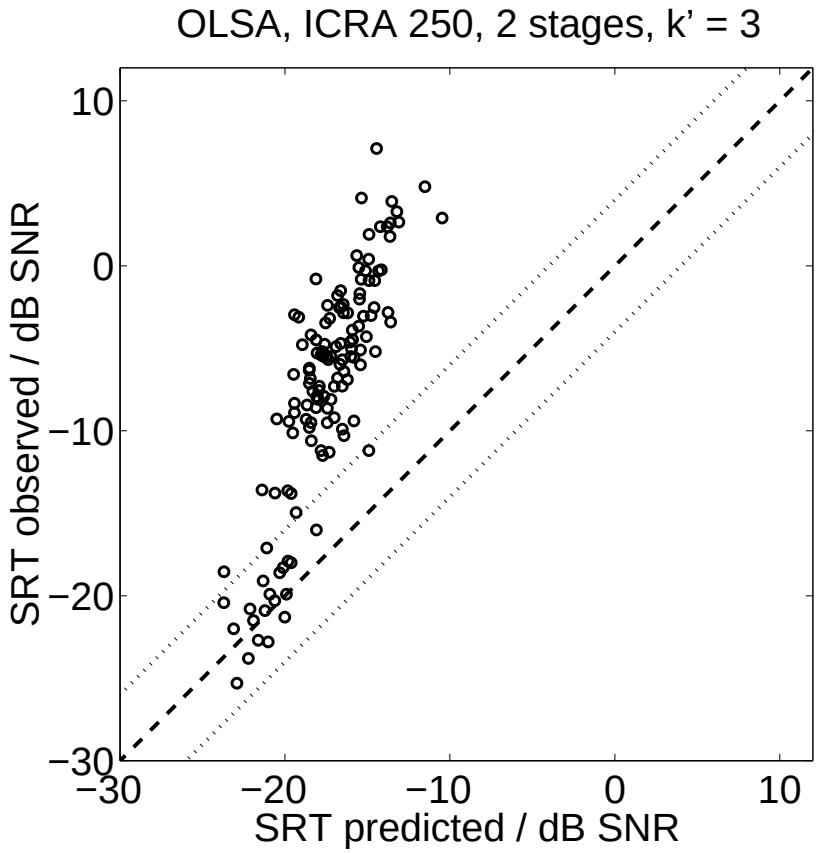

Figure 7.8: *Intelligibility predictions for icra5–250 noise by two–stage speech perception model, $k' = 3$. Each circle indicates one pair of observed/predicted SRT. The dashed lines indicate the estimated 95 %–confidence interval of the observed data.*

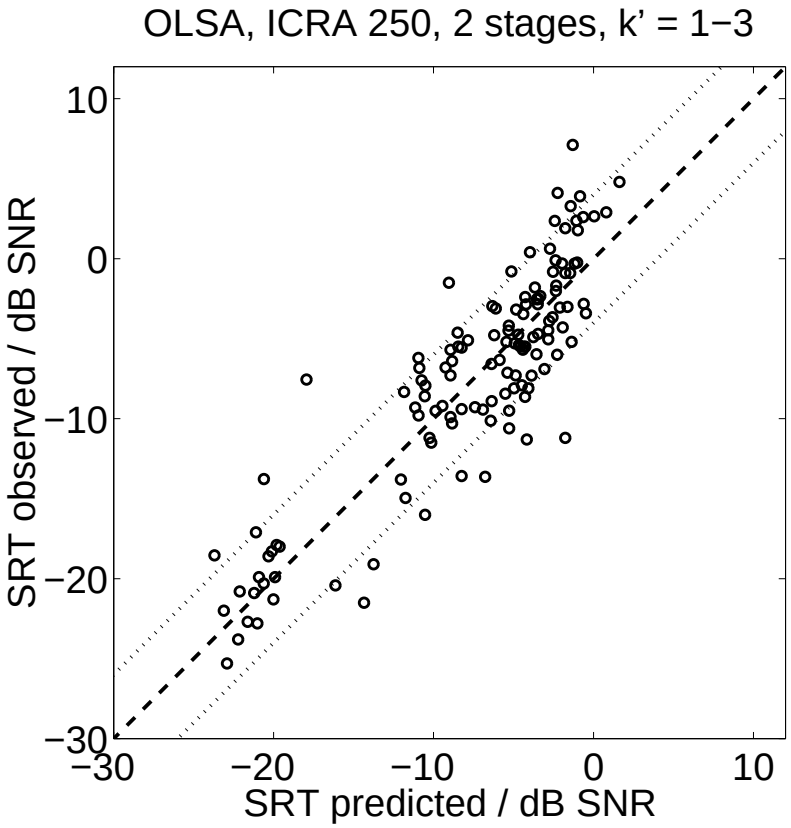

Figure 7.9: *Intelligibility predictions for icra5–250 noise by two–stage speech perception model with integrated context concept and three different k' values. Each circle indicates one pair of observed/predicted SRT. The dashed lines indicate the estimated 95 %–confidence interval of the observed data.*

Chapter 8

Summary and Conclusion

The aim of this thesis was to evaluate and model speech intelligibility with regard to comparability of sentence tests both within one language and across languages. Comparability was explored with one particular test type (syntactically fixed, but semantically unpredictable sentences). In addition to comparability across languages, comparability of test lists both in noise and in quiet was investigated and comparability of measurement procedures (different presentation levels, noises, and presentation types) was explored. Using fluctuating interfering noises gave the largest diagnostical benefit. Therefore, speech intelligibility in fluctuating noise was explored in more detail.

The second chapter presented the Danish version of a sentence test with syntactically fixed, but semantically unpredictable sentences. This test was developed similar to the German Oldenburg sentence test (Wagener *et al.*, 1999c; Wagener *et al.*, 1999a; Wagener *et al.*, 1999b) that was adopted from the Swedish Hagerman sentences (Hagerman, 1982). The design, the optimization, and the evaluation of the Danish test were presented.

From the speech audiometric point of view, these tests are characterized by their reference intelligibility functions determined with normal–hearing subjects (sometimes denoted as performance–intensity functions). To compare sentence intelligibility tests across different languages, these intelligibility functions were compared across Swedish, German, and Danish (third chapter). A logistic model function with two parameters was applied to describe the dependency of speech intelligibility and signal–to–noise ratio. The parameters were speech reception threshold SRT (signal–to–noise ratio that yields 50 % intelligibility) and slope of the intelligibility function at SRT. The main dif-

ference between the Swedish, German, and Danish test apart from language and speaker (Swedish and Danish: female speaker, German: male speaker) was the consideration of coarticulation effects. In the Swedish test, no coarticulation was preserved between the particular words of the test sentences leading to an unnatural speech flow at very high signal–to–noise ratios. The German and Danish tests preserved the natural coarticulation between words. The comparability of the Swedish and German test is high both with respect to the SRT and slope values (Hagerman sentences: SRT = -8.1 dB SNR, slope = 16 %/dB; Oldenburg sentence test: SRT = -7.1 dB SNR, slope = 17.1 %/dB). Although the Danish test was realized very similar to the German test, the different language and speaker resulted in lower comparability of both tests (SRT = -8.4 dB SNR, slope = 13.2 %dB). The lower slope of the Danish test is due to a lower slope of the mean word–specific intelligibility function [16.1 %/dB versus 20 %/dB (Oldenburg sentence test) and 23.5 %/dB (Hagerman sentences)]. This shows that there are some difficulties in establishing comparable speech tests across different languages even though most parameters have been set to be the same. In comparison with the sentence tests considered so far, however, the reference intelligibility functions of sentence tests with everyday sentences vary more: The slopes of these tests (Kalikow *et al.*, 1977; Plomp and Mimpen, 1979; Nilsson *et al.*, 1995; Smoorenburg, 1992; Kollmeier and Wesselkamp, 1997; Versfeld *et al.*, 2000) cover a range from 9 %/dB (Nilsson *et al.*, 1995) to 19.2 %/dB (Kollmeier and Wesselkamp, 1997).

Speech tests were often optimized with respect to test list homogeneity in interfering noise (Kalikow *et al.*, 1977; Plomp and Mimpen, 1979; Hagerman, 1982; Nilsson *et al.*, 1994; Kollmeier and Wesselkamp, 1997; Wagener *et al.*, 1999a; Wagener *et al.*, 2003). It was questionable whether the so achieved homogeneity also holds for intelligibility in quiet. Therefore, it was investigated in chapter four, whether sentence tests that were optimized for performing speech tests in noise are also suitable for speech tests in quiet. The German Göttingen [meaningful sentences, (Kollmeier and Wesselkamp, 1997)] and the Oldenburg sentence test [syntactically fixed, but semantically unpredictable sentences, (Wagener *et al.*, 1999c)] were evaluated with normal–hearing subjects. It was confirmed that test list homogeneity in noise (i. e., small standard deviations of SRT and intelligibility functions slope values across test lists) is also valid in quiet. It can be assumed that this holds also for other tests in different languages that were optimized in an similar way.

In the fifth chapter, comparability within one language was explored by investigating the influence of the factors presentation level, type of interfering noise, and presentation

mode on sentence intelligibility with normal–hearing and hearing–impaired subjects. Critical factors that influence the comparability of speech test results should be separated from non–critical factors that can be chosen arbitrary. The presentation level of the interfering noise was found to be non–critical within the range that is mostly used in diagnostics and rehabilitation (up to 85 dB SPL) as long as the noise presentation level exceeds the individual hearing threshold. SRT results were only depending on the signal–to–noise ratio and not on the presentation level within this range as long as the noise presentation level exceeded hearing threshold. Therefore, it is recommended to use a presentation level of 65 dB SPL for normal–hearing and 80 dB SPL for hearing–impaired subjects. When using this recommendation, however, it has to be assured that the hearing–impaired subjects perceive the noise, otherwise the level has to be increased. Furthermore, the presentation level should not approximate the individual uncomfortable level. It is also non–critical whether the noise level or the speech level is held constant in an adaptive procedure for determining SRT. Using continuous noise instead of noise that is interrupted between the particular test sentences yielded a slightly lower SRT. This was only statistically significant for normal–hearing subjects. This indicates that the healthy human auditory system may adapt to a noise with a certain time constant that yields a certain noise reduction in continuous noise (similar to some hearing aids with noise suppression algorithms). Since hearing–impaired subjects may also show an impairment of this mechanism, the difference between continuous and synchronous noise presentation may be less prominent for these subjects. In addition, no difference was found when using the stationary speech–shaped *icra1* noise or the *olnoise*. Therefore, it is recommended to use the stationary *icra1* noise as interfering noise in speech tests when the respective language shows a long–term spectrum similar to the LTASS–spectrum (Byrne *et al.*, 1994).

Using fluctuating noises gives a better differentiation between subjects with different degrees of hearing loss. Therefore, these noises are highly recommended for speech audiometry. The highly fluctuating *icra5* noise has the disadvantage that silent intervals with durations of up to 2 s are included. As entire sentences could fall into such silence intervals, the original *icra5* noise was modified by limiting the maximum silence interval duration of the original *icra5* noise to 250 ms (*icra5–250*) and 62.5 ms (*icra5–62.5*). Chapter six explores speech intelligibility in fluctuating noise depending on the maximum pause duration. By comparing the speech intelligibility results of the pause limited noises to the *icra5* data (determined with the same hearing–impaired subjects) it was shown that the advantage of fluctuating noises (good differentiation between

subjects) was preserved. The relation of speech intelligibility in fluctuating noise with other factors that describe the hearing ability was investigated. Speech intelligibility in the original *icra5* noise was correlated to speech intelligibility in quiet, but not to speech intelligibility in noise. Speech intelligibility in the pause limited fluctuating noises was highly correlated to both intelligibility in quiet and in stationary noise. This is consistent with the idea that the benefit in fluctuating noise over stationary noise is due to the ability to 'listen in the valleys' (Dubno *et al.*, 2002; Eisenberg *et al.*, 1995; Duquesnoy, 1983a; Carhart *et al.*, 1969). As the rhythm perception and modulation content of the *icra5–250* noise resembles a simulated interfering speaker more than the *icra5–62.5* noise, the *icra5–250* noise is recommended for speech tests in fluctuating noise.

In chapter seven, a feasibility study of predicting speech intelligibility in fluctuating noise was presented. All presented approaches were motivated by the fact that speech intelligibility in fluctuating noise highly correlates with both speech intelligibility in quiet and in stationary noise. The fluctuating noise was subdivided into short–time intervals wherein the noise presentation level was assumed to be stationary. Within these intervals, the intelligibility was predicted using two different SRT prediction concepts for stationary conditions. The Plomp model describes the relationship between SRT and noise presentation level relationship using both one SRT measurement in quiet and one SRT measurement in stationary noise as individual input data. The SII approach predicts the intelligibility function based on the individual pure–tone audiogram. These predicted stationary short–time intelligibilities were used to compute the SRT prediction for the entire fluctuating noise. Different approaches of combining the short–time values were presented that vary in complexity of the underlying theory. A simple one–stage approach of averaging the SRT values that correspond to the particular noise levels together with a linear transformation gave fairly good predictions. A more sophisticated approach simulated the perception process in more detail. This so–called two–stage speech perception model describes the perception of a test sentence by combining independent sub–word intervals in order to understand a word within the first stage. The word intelligibilities were averaged to compute the total intelligibility within the second stage in order to simulate word–scoring. The intelligibility functions of the sub–word intervals were derived from the intelligibility functions of the complete sentences using a context concept similar to the k factor proposed by Boothroyd and Nittrouer (1988). The two–stage model with integrated context concept predicts the SRTs in fluctuating interfering noise very accurately. However, this agreement between

predictions and measurements could so far only be achieved by assuming a strong dependency of context values (k') and the degree of hearing impairment. The high k' values of 3 in normal–hearing subjects can be motivated by a large context effect due to the modulation valleys of the noise level and the limited set size of the word material. The rapid decrease of k' (from 3 to 1) for SRT values in quiet of approximately 10 dB above normal, however, is hard to motivate. Therefore, further research is needed to investigate these context effects in more detail.

Conclusions

Using syntactically fixed, but semantically unpredictable sentences for speech tests in noise allows to perform extensive measurements with the same subjects, because the test lists can be used several times with the same subject. The comparison of all three existing speech tests with syntactically fixed, but semantically unpredictable sentences showed that there are still differences between these tests due to the different languages and speakers. These differences are small compared to the differences across speech tests with everyday sentences in different languages. Therefore, using syntactically fixed, but semantically unpredictable sentences is recommended in order to establish comparable speech tests across different languages.

It was shown that test lists that were originally optimized with respect to homogeneity in noise are also perceptually homogeneous in quiet conditions. This holds for both syntactically fixed, but semantically unpredictable sentences and everyday sentences.

The comparability of speech test results within one language and one sentence test was explored by separating measurement parameters that influence the results from those that can be arbitrarily chosen. As standard parameter setting for determining SRT adaptively, the stationary *icra1* noise should be used with a presentation level of 65 dB SPL (normal–hearing) and 80 dB SPL (hearing–impaired subjects). It should be documented, whether continuous or interrupted noise was used. In order to investigate the SRT benefit in fluctuating noise, the speech simulating modified *icra5–250* noise should be used that was generated by limiting the maximum silence interval durations of the original *icra5* noise to 250 ms.

In order to improve understanding of speech perception in fluctuating noise, SRT predictions were performed. Intelligibility in fluctuating noise can be fairly well predicted by a two–stage speech perception model that includes a context concept with different

amount of context for normal–hearing and hearing–impaired listeners.

In summary, the methods and recommendations given by this thesis appear to yield high comparability of speech intelligibility measures both within and across languages.

Appendix A

Speech Intelligibility Index for fluctuating noise

The Speech Intelligibility Index (SII) (ANSI S3.5, 1997) is a common method to predict speech intelligibility. It was developed for predictions in stationary noise. The basis of this measure is that both signal and noise properties are known separately. The SII is defined as the weighted sum of signal–to–noise ratios measured in different frequency bands [Equation A.1, paragraph 4.7.1 in ANSI S3.5 (1997)].

$$\mathrm{SII} = \frac{1}{30} \sum_{i=1\ldots N} w_i (SNR_i + 15) \qquad \text{(A.1)}$$
$$SNR_i \; \epsilon \; [-15; 15]$$

SNR_i in Eq. A.1 denotes the signal–to–noise ratio SNR in frequency band i between speech signal and noise signal. In this study, the real long–term spectrum of the sentence material was used. The calculated SNR_i considers effects of upward–/downward– and self–masking. The weighting factors w_i for each frequency band i are given for different types of speech material and characterize the importance of the particular frequency bands for understanding the respective speech material. In this study, N = 18 frequency bands and the weighting factors for 'short passages of easy reading material' were used [Tab. B.2., ANSI S3.5 (1997)]. These factors show a maximum at 500 and 4000 Hz. The SII computed in this way takes values between 0 and 1. In order to transform the SII to speech intelligibility S, the transformation function for sentence intelligibility of Fig. 7 / Tab. III, curve I (Flechter and Galt, 1950) was used.

In order to determine SRT values (speech reception threshold, i. e. signal–to–noise

ratio that yields 50 % intelligibility), the intelligibilities S were calculated within a 1–dB level grid. The SRT was calculated by linearly interpolating between two adjacent level values that yield intelligibilities below and above 50 %, respectively.

References

ANSI S3.5 (**1969**). *Methods for the calculation of the articulation index.* American National Standards Institute, New York.

ANSI S3.5 (**1997**). *Methods for Calculation of the Speech Intelligibility Index.* American National Standards Institute, New York.

Ardenkjær-Madsen, R. and J.L. Josvassen (**2001**). *Forslag til en ny dansk sætningsbaseret audiovisuel taleaudiometrisk test i støj [Suggestion for a New Danish Sentence Based Audiovisual Speech Audiometric Test in Noise]*, (M.A. Thesis in Speech and Hearing Science). University of Copenhagen.

Arlinger, S. (**1998**). Clinical Assessment of Modern Hearing Aids. *Scand Audiol*, **27**(Suppl. 49):50–53.

Baumann, U. (**2001**). Sprachverständnis mit Cochlea Implantat und Richtmikrofon–Hörgerät. *Fortschritte der Akustik - DAGA 2001.*

Beattie, R.C. (**1989**). Word recognition functions for the CID W–22 test in multitalker noise for normally hearing and hearing–imapaired subjects. *J. Speech Hear. Disord.*, **54**:20–32.

Bergenholtz, H. (Ed.) (**1992**). *Dansk frekvensordbog baseret på danske romaner, ugeblade og aviser 1987 - 1990 [Dictionary of Word Frequency in Danish based on danish novels, magazines and papers 1987 - 1990].* Copenhagen, Gad.

Boothroyd, A. and S. Nittrouer (**1988**). Mathematical treatment of context effects in phoneme and word recognition. *J. Acoust. Soc. Am.*, **84**(1):101–114.

Brand, T. and V. Hohmann (**2001**). Effect of Hearing Loss, Center Frequency and Bandwidth on the Shape of Loudness Functions in Categorical Loudness Scaling. *Audiology*, **40**(2):92–103.

Brand, T. and V. Hohmann (**2002**). An adaptive procedure for categorical loudness scaling. *J. Acoust. Soc. Am.*, **112**(4):1597–1604.

Brand, T. and B. Kollmeier (**2002a**). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *J. Acoust. Soc. Am.*, **111**(6):2801–2810.

Brand, T. and B. Kollmeier (**2002b**). Vorhersage der Sprachverständlichkeit in Ruhe und im Störgeräusch aufgrund des Reintonaudiogramms. *Zeitschrift für Audiologie*, **Suppl. V**.

Bronkhorst, A. W., T. Brand and K. Wagener (**2002**). Evaluation of context effects in sentence recognition. *J. Acoust. Soc. Am.*, **111**(6):2874–2886.

Bronkhorst, A.W., A.J. Bosman and G.F. Smoorenburg (**1993**). A model for context effects in speech recognition. *J. Acoust. Soc. Am.*, **93**(1):499–509.

Byrne, B., H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hagerman, R. Hetu, J. Kei, C. Lui, J Kießling, M. N. Kothby, N. H. A. Nasser, W. A. H. El Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. Stephens, R. Meredith, T. Sirimanna, G. Tavartkiladze, G. I. Frolenkov, S. Westerman and C. Ludvigsen (**1994**). An international comparison of long–term average speech spectra. *J. Acoust. Soc. Am.*, **96**(4):2108–2120.

Carhart, R., T.W. Tillman and E.S. Greetis (**1969**). Perceptual Masking in Multiple Sound Backgrounds. *J. Acoust. Soc. Am.*, **45**(3):694–703.

Dorman, M.F., P.C. Loizou, A.J. Spahr and E. Maloff (**2002**). A comparison of the speech understanding provided by acoustic models of fixed–channel and channel–picking signal processors for cochlear implants. *J. Speech Lang. Hear. Res.*, **45**(4):783–788.

Draft international standard ISO/DIS 389–8 (**2001**). *Acoustics - Reference zero for the calibration of audiometric equipment - Part 8: Reference equivalent threshold sound pressure levels for pure tones and circumaural earphones.* Annex C.

Dreschler, WA., H. Verschuure, C. Ludvigsen and S. Westermann (**2001**). ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. International Collegium for Rehabilitative Audiology. *Audiology*, **40**(3):148–157.

Droogendijk, M. and H. Verschuure (**2000**). Inventory and usage of audiological procedures in Europe. *Zeitschrift für Audiologie*, **Supplementum III**:121–124.

Dubno, J.R., D.D. Dirks and D.E. Morgan (**1984**). Effects of age and mild hearing loss on speech recognition in noise. *J. Acoust. Soc. Am.*, **76**(1):87–96.

Dubno, J.R., A.R. Horwitz and J.B. Ahlstrom (**2002**). Benefit of modulated maskers for speech recognition by younger and older adults with normal hearing. *J. Acoust. Soc. Am.*, **111**(6):2897–2907.

Duquesnoy, A.J. (**1983a**). Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons. *J. Acoust. Soc. Am.*, **74**(3).

Duquesnoy, A.J. (**1983b**). The intelligbility of sentences in quiet and in noise in aged listeners. *J. Acoust. Soc. Am.*, **74**(4).

Duquesnoy, A.J. and R. Plomp (**1983**). The effect of a hearing aid on the speech–reception threshold of hearing–impaired listeners in quiet and in noise. *J. Acoust. Soc. Am.*, **73**(6).

Eisenberg, L.S., D.D. Dirks and T.S. Bell (**1995**). Speech Recognition in Amplitude–Modulated Noise of Listeners With Normal and Listeners With Impaired Hearing. *J. Speech Hear. Res.*, **38**.

Festen, J.M. and R. Plomp (**1990**). Effect of fluctuating noise and interfering speech on the speech–reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am.*, **88**(4).

Flechter, H. and R.H. Galt (**1950**). The Perception of Speech and Its Relation to Telephony. *J. Acoust. Soc. Am.*, **22**(2).

Friesen, L.M., R.V. Shannon, D. Baskent and X. Wang (**2001**). Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. *J. Acoust. Soc. Am.*, **110**(2):1150–1163.

Glasberg, B.R., B.C.J. Moore and S.P. Bacon (**1987**). Gap detection and masking in hearing–impaired and normal–hearing subjects. *J. Acoust. Soc. Am.*, **81**(5):1546–1556.

Goshorn, E.L. and G.A. Studebaker (**1994**). Effects of intensity on speech recognition in high– and low–frequency bands. *Ear Hear.*, **15**:454–460.

Green, R., S. Day and J. Bamford (**1989**). A comparative evaluation of four hearing aid selection procedures. I–Speech discrimination measures of benefit. *British Journal of Audiology*, **23**:185–199.

Gustafsson, H.Å. and S. Arlinger (**1994**). Masking of speech by amplitude–modulated noise. *J. Acoust. Soc. Am.*, **95**(1):518–529.

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

Hagerman, B. (**1984**). *Some Aspects of Methodology in Speech Audiometry*, (Dissertation). Karolinska Institute Stockholm, Schweden.

Hagerman, B. (**1997**). Attempts to Develop an Efficient Speech Test in Fully Modulated Noise. *Scand Audiol*, **26**:93–98.

Hagerman, B. (**2002**). Speech recognition threshold in slightly and fully modulated noise for hearing–impaired subjects. *International Journal of Audiology*, **41**:321–329.

Hagerman, B. and C. Kinnefors (**1995**). Efficient Adaptive Methods for Measuring Speech Reception Threshold in Quiet and in Noise. *Scand Audiol*, **24**:71–77.

Hawkins, J.E.Jr. and S.S. Stevens (**1950**). The Masking of Pure Tones and of Speech by White Noise. *J. Acoust. Soc. Am.*, **22**(1).

Hirsh, I.J. and W.D. Bowman (**1953**). Masking of speech by bands of noise. *J. Acoust. Soc. Am.*, **25**:1175–1180.

Hirsh, I.J., E.G. Reynolds and M. Joseph (**1954**). The intelligibility of different speech materials. *J. Acoust. Soc. Am.*, **26**:530–538.

Houtgast, T., H.J.M. Steeneken and A.W. Bronkhorst (**1992**). Speech communication in noise with strong variations in the spectral or the temporal domain. *Proceedings of the 14th International Congress on Acoustics*, **3**:H2–6.

Humes, L.E. (**1999**). Dimensions of Hearing Aid Outcome. *Journal of the American Academy of Audiology*, **10**:26–39.

Hygge, S., J. Ronnberg, B. Larsby and S. Arlinger (**1992**). Normal–hearing and hearing–impaired subjects' ability to just follow conversation in competing speech,

reversed speech, and noise backgrounds. *Journal of Speech & Hearing Research*, **35**(1):208–215.

International Collegium of Rehabilitative Audiology (**1997**). *ICRA, noise signals, Ver. 0.3*.

Kalikow, D. N., K. N. Stevens and L. L. Elliot (**1977**). Development of a test of speech intelligibility in noise using sentences with controlled word predictability. *J. Acoust. Soc. Am.*, **61**:1337–1351.

Kollmeier, B. (**1990**). *Messmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache*, (Habilitation). Georg–August–Universität Göttingen.

Kollmeier, B., C. Müller, M. Wesselkamp and K. Kliem (**1992**). Weiterentwicklung des Reimtests nach Sotscheck. In B. Kollmeier (Ed.), *Moderne Verfahren der Sprachaudiometrie*, pp. 216–237. Median Verlag, Heidelberg.

Kollmeier, B. and M. Wesselkamp (**1997**). Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *J. Acoust. Soc. Am.*, **102**(4):2412–2421.

Nelson, D.A. and R.L. Freyman (**1987**). Temporal resolution in sensorineural hearing–impaired listeners. *J. Acoust. Soc. Am.*, **81**(3):709–720.

Nelson, P.B., S.-H. Jin, A.E. Carney and D.A. Nelson (**2003**). Understanding speech in modulated interference: Cochlear implant users and normal–hearing listeners. *J. Acoust. Soc. Am.*, **113**(2):961–968.

Nilsson, M., S.D. Soli and J.A. Sullivan (**1994**). Development of hearing in noise test for measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.*, **95**:1085–1099.

Nilsson, M., S.D. Soli and A. Sumida (**1995**). Development of Norms and Percent Intelligibility Functions for the HINT. *House Ear Institute*, pp. 1–9.

Peters, R.P, B.C.J. Moore and T. Baer (**1998**). Speech reception thresholds in noise with and without spectral and temporal dips for hearing–impaired and normally hearing people. *J. Acoust. Soc. Am.*, **103**(1):577–587.

Pickett, J.M. and I. Pollack (**1958**). Prediction of speech intelligibility at high noise levels. *J. Acoust. Soc. Am.*, **30**:955–963.

Plomp, R. (**1978**). Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J. Acoust. Soc. Am.*, **63**(2).

Plomp, R. and A. M. Mimpen (**1979**). Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, **18**:43–52.

Pollack, I. and J.M. Pickett (**1958**). Masking of Speech by Noise at High Sound Levels. *J. Acoust. Soc. Am.*, **30**:127–130.

Schmidt, M., I. Hochmair-Desoyer, E. Schulz and L. Moser (**1997**). Der HSM–Satztest. *Fortschritte der Akustik - DAGA '97*, pp. 93–94.

Schroeder, M. R. (**1968**). Reference Signal for Signal Quality Studies. *J. Acoust. Soc. Am.*, **44**:1735–1736.

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. *J. Acoust. Soc. Am.*, **91**:421–437.

Speaks, C., J.L. Karmen and L. Benitez (**1967**). Effect of a competing message on synthetic sentence identification. *J. Speech Hear. Res.*, **10**:390–396.

Studebaker, G.A. (**1985**). A rationalized arcsine transform. *J. Speech Hear. Res.*, **28**:455–462.

Studebaker, G.A., R.L. Shernecoe, D.M. McDaniel and C.A. Gwaltney (**1999**). Monosyllabic word recognition at higher–than–normal speech and noise levels. *J. Acoust. Soc. Am.*, **105**(4):2431–2443.

Tyler, R.S., A.J. Parkinson, B.S. Wilson, S. Witt and J.P. Preece (**2002**). Patients utilizing a hearing aid and a cochlear implant: speech perception and localization. *Ear & Hearing*, **23**(2):98–105.

van Toor, T. and H. Verschuure (**2002**). Effects of high–frequency emphasis and compression time constants on speech intelligibility in noise. *International Journal of Audiology*, **41**:379–394.

Versfeld, N.J., L. Daalder, J.M. Festen and T. Houtgast (**2000**). Method for the selection of sentence material for efficient measurement of the speech reception threshold. *J. Acoust. Soc. Am.*, **107**:1671–1684.

Versfeld, N.J. and W.A. Dreschler (**2002**). The relationship between the intelligibility of time–compressed speech and speech in noise in young and elderly listeners. *J. Acoust. Soc. Am.*, **111**:401–408.

Wagener, K., T. Brand and B. Kollmeier (**1999a**). Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests (Development and evaluation of a German sentence test II: Optimization of the Oldenburg sentence test). *Zeitschrift für Audiologie*, **38**(2):44–56.

Wagener, K., T. Brand and B. Kollmeier (**1999b**). Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests (Development and evaluation of a German sentence test III: Evaluation of the Oldenburg sentence test). *Zeitschrift für Audiologie*, **38**(3):86–95.

Wagener, K., T. Brand and B. Kollmeier (**2000**). Einfluss verschiedener Parameter auf die Sprachverständlichkeit im Störgeräusch. *Fortschritte der Akustik - DAGA 2000*, pp. 258–259.

Wagener, K., J.L. Josvassen and R. Ardenkjær (**2003**). Design, Optimization, and Evaluation of a Danish Sentence Test in Noise. *Journal of International Audiology*, **42**(1):10–17.

Wagener, K., V. Kühnel and B. Kollmeier (**1999c**). Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests (Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test). *Zeitschrift für Audiologie*, **38**(1):4–15.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

Oldenburg, den 12. September 2003 Kirsten Wagener

Danksagung

Wenn man anfängt, die Danksagung zu schreiben, dann wird es ernst. Das Gefühl im Bauch ähnelt der freudigen Aufregung, die einen befällt, wenn der Dirigent den Taktstock zum Auftakt des Konzerts hebt. Man möchte alle umarmen und ihnen danken, dass man mit ihnen zusammen das alles erarbeiten und nun aufführen darf.

Ich bedanke mich ganz besonders bei Prof. Dr. Dr. Birger Kollmeier, der in konstruktiven Gesprächen immer neue Denkanstösse und Perspektiven eröffnet und ideale Arbeitsbedingungen schafft. Er hat wesentlich dazu beigetragen, dass der OlSa in andere Sprachen umgesetzt wird. Ich bekam schon manchmal etwas Angst, dass er von jeder Tagung mit jeweils zehn Basis–Sätzen in den jeweiligen Muttersprachen aller Tagungsteilnehmer wiederkäme.

Prof. Dr. Volker Mellert möchte ich danken, dass er so freundlich das Korreferat dieser Arbeit übernommen hat.

Dr. Thomas Brand möchte ich ganz herzlich für die ausdauernde Betreuung dieser Arbeit danken. Er zeichnet sich durch geradezu triathletische Fähigkeiten aus: motivieren, diskutieren, korrigieren. Das alles auch noch zu jeder Zeit, an fast jedem Ort. Das ihm nicht irgendwann die Korrekturseiten dieser Arbeit wieder aus den Augen rausgetropft sind, ist mir ein Rätsel.

Manfred Mauermann danke ich für die netteste Büro–WG, die ich kenne. Nur wir beiden wissen das Geheimnis, wie man sowohl ein $6\,\mathrm{m}^2$ als auch ein $25\,\mathrm{m}^2$ großes Büro mit dem gleichen Inhalt absolut ausfüllen kann.

Dr. Oliver Fobel möchte ich für die Hotline–Hilfen danken, die ich bei Computerfragen immer gerne in Anspruch genommen habe. Ich denke gerne an unsere Wetten und unsere Erweckungsversuche der kirk.

Dr. Martin Hansen möchte ich besonders für die gute Zusammenarbeit bei der Realisierung des dänischen OlSa danken. Es ist mir immer unangenehm, dass wir seinen Namen in der Danksagung des Artikels vergessen haben.

Ein besonderer Dank geht an Jane Josvassen, Regitze Ardenkjær, Line Bille Nugteren und Arne Nørby Rasmussen. Es hat großen Spaß gemacht, mit Euch zusammen den dänischen Satztest auf die Beine zu stellen und zu Papier zu bringen. Auch wenn mein dänischer Wortschatz immer noch nicht über 50 Wörter hinaus geht, war das eine im wahrsten Sinne des Wortes vorbildliche und produktive Zusammenarbeit.

Allen, die durch interessante und motivierende Diskussionen meine Arbeit bereichert haben, möchte ich danken, besonders Prof. Dr. Björn Hagerman, Prof. Dr. Tammo Houtgast, Dr. Adelbert Bronkhorst und Prof. Dr. Wouter Dreschler.

Prof. Dr. Torsten Dau, Dr. Stefan Uppenkamp, Dr. Andrew Oxenham und Dr. John Culling haben netterweise Teile dieser Arbeit Korrektur gelesen – vielen Dank!

Anita Gorges, Müge Kaya und Dr. Birgitta Gabriel möchte ich ganz besonders für die Logistik und Durchführung der zahlreichen Messungen danken, ohne die diese Arbeit nur aus leeren Blättern bestehen würde. Daran schließt sich direkt mein Dank an alle ProbandInnen an, die mit ihren Ohren und dem "dazwischen" wesentlich zu dieser Arbeit beigetragen haben.

Dr. Volker Hohmann kennt sich hervorragend in Signalverarbeitung aus. Da ist es ein Glücksfall, dass man ihn kennt. Denn er beantwortet geduldig alle Fragen, die man dazu hat – Danke!

Natürlich sollen hier nicht die Leute vergessen werden, die durch ihre programmatischen Fähigkeiten dafür gesorgt haben und sorgen, dass Sprachverständlichkeitsmessungen in den unterschiedlichsten Konfigurationen nicht zusammen mit den noch diese Arbeit bestimmenden DOS–Messprogrammen und Ariel–Karten ins Jenseits gelangt sind, sondern bunt und frisch vielerorts verfügbar sein können: Ein Hoch auf Dr. Daniel Berg, Dr. Thomas Wittkop und Dr. Jens Appell!

Aber was sind schon ein paar einzelne Musiker, die immer nur etwas "für sich" spielen? Eigentlich gibt es erst den richtigen Kick, wenn es genügend andere Mitstreiter gibt, die sich gegenseitig motivieren, helfen, die Einsätze zeigen, mitzählen und zusammenhalten. Deshalb möchte ich ein dickes Dankeschön an die Arbeitsgruppe Medizinische Physik loswerden. In Euren Reihen macht das Arbeiten Spaß!

Bevor dies noch das längste Kapitel der Arbeit wird ...

Mille grazie, tom!

Diese Arbeit wurde gefördert von der DFG, BMBF, EG, GN Resound, Oticon und Widex.

Lebenslauf

Am 19. August 1973 wurde ich, Kirsten Carola Wagener, in Göttingen als dritte Tochter von Anneliese Wagener, geb. Junghänel und Bernd Wagener geboren. Von 1979 bis 1983 besuchte ich die Hermann–Ehlers–Grundschule in Oldenburg, von 1983 bis 1985 die Orientierungsstufe Marschweg und wechselte 1985 auf das Herbartgymnasium (bis 1988 Hindenburgschule), das ich im Mai 1992 mit dem Abitur abschloss. Im Oktober 1992 nahm ich das Studium der Physik an der Carl–von–Ossietzky Universität Oldenburg auf. Im Dezember 1994 legte ich dort die Diplomvorprüfung ab. Im Juni 1997 begann ich in der Arbeitsgruppe Medizinische Physik unter der Anleitung von Prof. Dr. Dr. Birger Kollmeier mit der Anfertigung meiner Diplomarbeit mit dem Thema: "Entwicklung und Evaluation eines Satztests für die deutsche Sprache". Mein Studium schloss ich am 14. September 1998 mit der Diplomprüfung ab. Von Oktober 1998 bis März 2001 war ich Stipendiatin im Graduiertenkolleg Psychoakustik der Carl–von–Ossietzky Universität Oldenburg. Seit April 2001 arbeite ich als wissenschaftliche Mitarbeiterin in der Arbeitsgruppe Medizinische Physik an der Carl–von–Ossietzky Universität Oldenburg. Hier fertigte ich unter Anleitung von Prof. Dr. Dr. Birger Kollmeier die vorliegende Dissertation an.