



## UvA-DARE (Digital Academic Repository)

### Modeling the speech intelligibility in fluctuating noise

Rhebergen, K.S.

**Publication date**

2006

**Document Version**

Final published version

[Link to publication](#)

**Citation for published version (APA):**

Rhebergen, K. S. (2006). *Modeling the speech intelligibility in fluctuating noise*. [Thesis, fully internal, Universiteit van Amsterdam].

**General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

# **Modeling the speech intelligibility in fluctuating noise**

**ACADEMISCH PROEFSCHRIFT**

door

Koenraad Sjoerd Rhebergen

ISBN: 90-8559-202-X

Printed by: Optima Grafische Communicatie, Rotterdam, The Netherlands

Printing of this thesis was financially supported by Atze Spoor Fonds, Beltone Netherlands B.V., Beter Horen, GN Resound B.V., Oticon Nederland B.V., Schoonenberg Hoorcomfort, Siemens Audiologie Techniek B.V., SMA Hoorcomfort en Veenhuis Medical Audio B.V..

© K.S. Rhebergen, Leiden, 2006

All rights reserved. No part of this book may be reproduced in any form, by print, photocopying, microfilm, electronic data transmission, or otherwise, without prior permission in writing from the author.

# **Modeling the speech intelligibility in fluctuating noise**

**ACADEMISCH PROEFSCHRIFT**

ter verkrijging van de graad van doctor  
aan de Universiteit van Amsterdam  
op gezag van Rector Magnificus  
prof. mr. P.F. van de Heijden  
ten overstaan van een door het college voor promoties ingestelde  
commissie, in het openbaar te verdedigen in de Aula der Universiteit  
op dinsdag 26 september 2006, te 10.00 uur  
door Koenraad Sjoerd Rhebergen  
geboren te Amsterdam

## **Promotiecommissie:**

Promotor: Prof. dr. ir. W. A. Dreschler

Co-promotor: Dr. ir. N. J. Versfeld

Overige leden:

Prof. dr. E. de Boer

Prof. dr. A.W. Bronkhorst

Prof. dr. W.J. Fokkens

Prof. dr. ir. T. Houtgast

Prof. dr. ir. L.C.W. Pols

Prof. dr. J. Wouters

Faculteit der Geneeskunde

Voor mijn ouders



## Contents

	List of abbreviations	9
<b>Chapter 1:</b>	General Introduction.	11
<b>Chapter 2:</b>	A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. <i>J. Acoust. Soc. Am. (2005) 117 (4), 2181 –2192.</i>	21
<b>Chapter 3:</b>	Release from informational masking by time reversal of native and non-native interfering speech. <i>J. Acoust. Soc. Am. (2005) 118 (3), 1274 – 1277.</i>	53
<b>Chapter 4:</b>	Learning effect observed for the speech reception threshold in interrupted noise with normal-hearing listeners. <i>Submitted to J. Acoust. Soc. Am.</i>	63
<b>Chapter 5:</b>	Validation of the extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise for normal-hearing listeners, and suggestions for further improvement. <i>J. Acoust. Soc. Am. In press.</i>	71
<b>Chapter 6:</b>	Predicting the intelligibility for speech in real-life background noises. <i>Submitted to Ear &amp; Hearing</i>	99
<b>Chapter 7:</b>	The dynamic range of speech, compression, and its effect on the speech intelligibility in interrupted noise.	113
<b>Chapter 8:</b>	Predicting the Speech intelligibility in fluctuating noise in hearing impaired listeners	151

## *Contents*

<b>Chapter 9:</b>	General Discussion.	171
	Summary	185
	References	191
	Samenvatting	205
	Dankwoord	211
	Curriculum Vitae	215

## List of Abbreviations

AI	Articulation Index
CB	Critical Band
CR	Compression Ratio
CLD	Cumulative Level Distribution
CVC	Consonant Vowel Consonant
dB	decibel
dBA	decibel A-weighted
DC	Duty Cycle
DR	Dynamic Range
ER	Expansion Ratio
ERB	Equivalent Rectangular Bandwidth
ERD	Equivalent Rectangular Duration
ESII	Extended Speech Intelligibility Index
FIR	Finite Impulse Response
FFT	Fast Fourier Transform
FMF	Forward Masking Function
HI	Hearing-Impaired
HL	Hearing Level
Hz	Hertz
IIF	Intensity Importance Function
JFC	Just Following Conversation
L1	1st percentile level
L99	99th percentile level
LTASS	Long Term Average Speech Spectrum
NH	Normal-Hearing
PTA	Pure Tone Average
RMS	Root Mean Square
SII	Speech Intelligibility Index
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SRT	Speech Reception Threshold
SRTq	Speech Reception Threshold in quiet
STI	Speech Transmission Index
WDRC	Wide Dynamic Range Compression



## *Chapter 1*

### General Introduction

## **General Introduction**

In general, normal-hearing listeners have no serious problems to understand speech in different environments. Discussion partners intuitively adjust their vocal effort to the level of the surrounding noise such that they are able to understand each other at an acceptable level. The degree to which speech is intelligible to a listener depends on large number of factors. In most cases, the amount of speech information that is available to the listener is the most important variable. The more speech information is available, the easier it is to follow the speaker.

Attempts to model the speech intelligibility started many years ago. The first studies on intelligibility started at AT&T Bell Labs in the 1920s. The aim of these studies was to analyze speech quality, telephone quality, and speech transmission. The work remained largely classified and was first published internationally only shortly after World War II (French and Steinberg, 1947; Fletcher and Galt, 1950; Fletcher, 1953). This work formed the basis of the “articulation theory” (Allen, 1996). To that date, the articulation theory was used, amongst others, to evaluate pilot-navigator communications, to improve the signal-to-noise ratio (SNR) in the pilot’s ear, or to improve headphone damping (Allen, 1996). The principle of the calculation scheme was to calculate the remaining information present at the receiver (listener), and the outcome of the calculation scheme was a number between zero and unity, called the Articulation Index, or AI. An AI of zero meant that no information was present at the receiver’s ear, whereas an AI of unity meant that all information was transmitted in a proper manner. Unfortunately, the AI method of Fletcher and Galt (1950) was hardly used after its publication (Müsch, 2001), most likely due to the complex calculation scheme. To overcome this problem, Kryter (1962) introduced a calculation scheme that was based on the Fletcher and Galt method, but was much simpler to use. Eventually, his method led to the ANSI S3.5 (1969) standard AI calculation method. The AI model evolved and today, a revised version, called the Speech Intelligibility Index, or SII, is most commonly used (e.g., Noordhoek *et al.*, 2000; Hornsby and Ricketts, 2003; Kates and Arehart, 2005). The SII model has been designed to predict the speech intelligibility for communication channels, conditions with speech in noise, and hearing aid settings, adjustments, or evaluation of hearing aid benefit. The basic

## Chapter 1

principle of the SII model is as follows: For a given condition where speech is embedded in noise at a given SNR, the SII is calculated using the average noise spectrum, the average speech spectrum, and the listener's hearing threshold. By calculating the SNR in different frequency bands, and by assigning a weight to each band (the so-called Band Importance Function), the SII is finally determined by adding the amount of information for each band (e.g., Studebaker *et al.*, 1993). As with the AI, the SII is a number between zero and unity and can be interpreted as the proportion of the total speech information that is available to the listener. For example, the proportion of speech information required to understand about one half of a number of simple, everyday short sentences is about one-third. In terms of the model this means that  $SII=0.33$ . This is true for normal-hearing listeners, listening to their native language. For hearing-impaired listeners, the value of the SII is strongly influenced by their hearing loss and their ability to use their remaining auditory capacities. One often encounters hearing-impaired listeners who are unable to understand speech properly, even when it is presented, in comparison with a normal-hearing listener, sufficiently audible (e.g., with a hearing aid). Although speech is audible, the auditory system apparently is unable to process it properly (so-called supra-threshold deficits, Noordhoek, 2000). This is exactly the situation where hearing-impaired listeners complain when listening in noisy surroundings: "I can hear you, but I can't understand you". In general, these listeners require a higher SNR to reach a proper level of intelligibility.

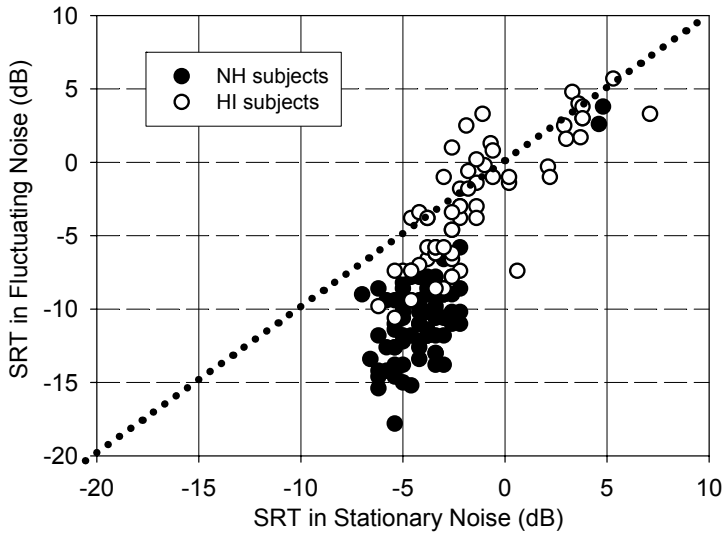
The performance of listeners with respect to speech intelligibility in different background noises can be assessed by means of the Speech Reception Threshold (SRT) test (Plomp and Mimpen, 1979). The SRT test is an adaptive test, set up to determine the SNR that is required to make half of a number of sentences completely intelligible. To that end, a list of 13 sentences, of course unknown to the listener, is monaurally presented via headphones. The listener's task is to repeat each sentence verbatim. Usually, the masking noise is presented at a fixed level, while the sentence level is varied adaptively. The SNRs of last 10 presentations are averaged, and this mean SNR is denoted as the Speech Reception Threshold, or SRT for that particular noise condition. To date, the SRT test of Plomp and Mimpen (1979) has been validated for a number of speech corpuses, not only for the Dutch language, but also for many other languages. Plomp conducted a series of studies on the SRT model to compare the effects of various factors on speech intelligibility with the aim to predict the

SRT in stationary noise. A review of these studies (Plomp, 1986) describes that the SRT in stationary noise for normal-hearing and hearing-impaired listeners can be predicted at various levels by two parameters. These two parameters are the SRT in quiet and the critical SNR in stationary noise at higher noise levels. The model prediction for SRT in stationary noise thus is not dependent on the presentation level but predicts a fixed SNR at various noise levels. Since no exact control over the presentation level is needed to measure a SRT in stationary noise, the SRT in stationary noise is thus very suitable for screening purposes, even by telephone (Smits, 2006).

With Dutch speech materials, consisting of simple, everyday short sentences, presented monaurally, normal-hearing listeners require an SNR in stationary noise with the long-term spectrum of the target speech on average of about -5 dB, meaning that the average speech level is 5 dB below the average noise level. The inability of some normal-hearing listeners to reach this score may be due to a number of factors. One is that the speech is distorted some way or another (Plomp, 1986). Another possibility is that ageing affects the nervous and cognitive system, making it more difficult to entangle the speech from the noise (Plomp and Mimpen, 1978; Duquesnoy, 1983, Marcell and Cohen, 1992). This group of normal-hearing subjects might have in some way or another difficulties with spectral and or temporal processing. Reduced temporal resolution results in difficulties in processing rapid changes in the speech signal, whereas reduced frequency resolution results in limited discrimination of spectral contrasts in the speech signal. Finally, it is well known that acquaintance with the language is a dominant factor in speech intelligibility. Non-native listeners need much more information to understand a message (van Wijngaarden, 2003). From the transmitter side, intelligibility is hampered if a non-native speaker is talking, or if the speaker pronounces or articulates in an unclear manner.

Traditionally, the SRT is determined for speech in stationary noise. Also, the SII has been validated only for speech in stationary noise. However, non-stationary or fluctuating noises are increasingly used, since they appear to be more sensitive to discriminate between conditions and listeners (Festen and Plomp, 1990; Middelweerd *et al.*, 1990; Peters *et al.*, 1998; Versfeld and Dreschler, 2002; Festen and Plomp, 2002). For example, Festen and Plomp (1990) introduced a fluctuating noise with speech-like properties, *viz.*, noise with a speech spectrum

and with speech modulations. They showed that differences between normal-hearing and hearing-impaired subjects were larger in fluctuating noise than in stationary noise. Existing differences between normal-hearing and hearing-impaired listeners were increased by the fact that normal-hearing listeners were able to make good use of the relatively silent period in the noise, resulting in a decrease in SRT, while hearing-impaired listeners only hardly benefited from the fluctuations, leaving their SRT unaltered.



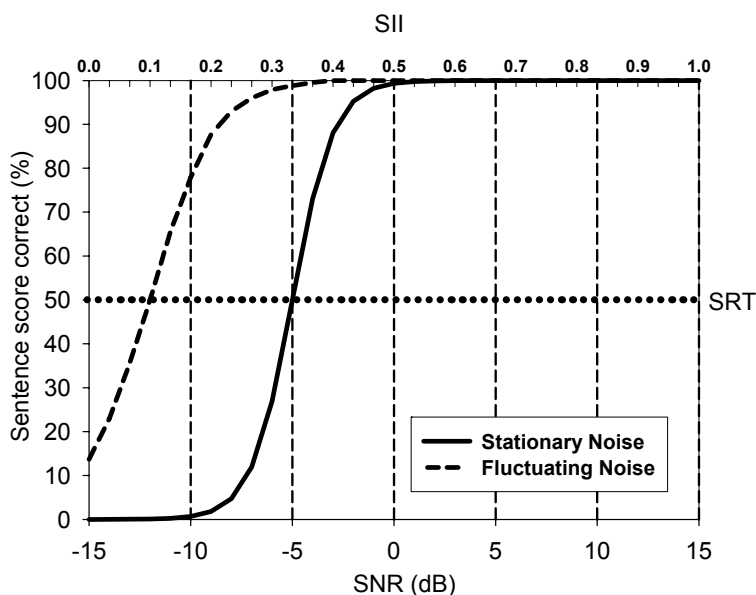
**Figure 1.1** The SRT in fluctuating noise as a function of the SRT in stationary noise. The dotted line indicates the points of equal SRT. Each circle denotes an individual normal-hearing (filled circle) or hearing-impaired (open circle) subject. Data taken from Versfeld and Dreschler (2002).

Figure 1.1 shows the relationship between the SRT measured in fluctuating speech shaped noise and stationary speech shaped noise for groups of young and elderly, normal-hearing or hearing-impaired listeners (data taken from Versfeld and Dreschler, 2002). Each data point denotes the result of an individual listener. Figure 1.1 shows that normal-hearing listeners (subjects

## General Introduction

with a normal pure-tone threshold) perform generally well in fluctuating noise (on average have SRTs of about -12 dB), whereas their SRTs in stationary noise are about -4.5 dB. Subjects who perform worse in fluctuating noise may still experience no decline in SRT in stationary noise. Consequently, a good performance in fluctuating noise implies a good performance in stationary noise, whereas the opposite is not true. The large improvement in SRT due to the fluctuations in the masking noise cannot be modeled by the SII model in its present form (ANSI S3.5-1997). Figure 1.2 displays the psychometric function where sentence intelligibility has been plotted as a function of SNR for normal-hearing subjects. The solid curve is obtained for speech in stationary speech shaped noise. It can be seen that this curve starts to deviate from zero at an SNR near -10 dB, and reaches 100 % at an SNR of about 0 dB. SII model calculations are given at the upper abscissa. As defined by the model, the SII increases linearly from zero at SNR=-15 dB to unity at SNR=15 dB. As mentioned above, near a score of 50 %, the SII is equal to 0.33, meaning that one-third of the speech information is required to obtain 50 % sentence intelligibility. The dashed curve in 1.2 denotes the psychometric function for the fluctuating masking noise of Festen and Plomp (1990). As discussed earlier, performance in fluctuating noise is much better, and 50 % intelligibility (the SRT) is at an SNR of about -12 dB. Given that the SII is calculated from the long-term average spectrum of noise and speech, it does not take into account any fluctuations in the masking noise. Accordingly, an SII value calculated for an SRT in fluctuating noise will be identical to that in stationary noise. Figure 1.2 shows that 50 % intelligibility corresponds to an SII of 0.09, which deviates severely from the existing model. At threshold, one would expect similar values of the SII, indicating that the amount of speech information at threshold (SRT) is equal irrespective of the masking noise. Many studies have shown convincingly that the SII model can adequately describe the SRT in different kinds of stationary noise. However, when fluctuating noise is involved, the model is clearly inadequate. Attempts have been made to modify the SII model in order to predict the speech intelligibility in interrupted noise (Ludvigsen, 1987; Dubno *et al.*, 2002). With interrupted noise, the signal can be partitioned into two portions: one part with masking noise present, and the other part with masking noise absent. For both separate parts, the conventional SII can be determined. The resulting SII then is the average of both parts. This technique is a special case of a model that considers the speech and noise stationary over a very short

time interval. Within each time interval the conventional SII can be determined, and where the resulting SII is the average across all time intervals. Indeed, Kates (1987) introduced the concept of the short-time articulation index to evaluate an adaptive noise-cancellation system by dividing the speech into segments, and plotting the change in the AI as a function of time to illustrate the adaptivity of the noise-cancellation system. Also, Houtgast *et al.* (1992) proposed that speech intelligibility in fluctuating noise might perhaps be predicted by calculating the instantaneous value of the AI. Unfortunately, although the general concept of an instantaneous AI or SII has been formulated about 20 years ago, it has never been properly implemented and used with the aim to predict the speech intelligibility in non-stationary noise. The present thesis describes the research that has been set up and conducted to gain more insight in the intelligibility of speech in fluctuating noise by means of listening experiments and by means of modeling.



**Figure 1.2** Percent sentence correct as function of SNR in stationary noise (solid line) and fluctuating speech shaped noise (dashed line) for normal-hearing listeners. The horizontal dotted line indicates the 50 % sentence correct score, or SRT. SRT is the SNR level in dB at which 50 % correct score is obtained. The upper axis indicates the SII (ANSI S3.5-1997) value for a given sentence correct score.

**Chapter 2** introduces an approach to model the Speech Reception Threshold (SRT) in fluctuating noise for normal-hearing listeners. The new method is an extension to the SII. Many SRT data available from the literature for a variety of noise types are used to evaluate the extension to the SII model. Furthermore, aspects that can influence the speech intelligibility in noise are discussed.

In **Chapter 3**, a method is presented with which the effect of informational masking on the speech intelligibility in interfering speech can be examined. The study shows that with speech in a language unknown to the subjects (and consequently unintelligible) as interferer, listeners suffer less from informational masking than with intelligible speech in their own language as interferer.

In **Chapter 4**, the learning effect is examined for the SRT in interrupted noise with normal-hearing listeners. It appears that with stationary noise as a masker, no learning effect is observed, whereas with interrupted noise there is a strong effect.

**Chapter 5** describes a study with normal-hearing listeners to validate the Extended SII (ESII) model introduced in **Chapter 2** of this thesis. The SRTs for a range of fluctuating masking noise conditions, critical to the model, have been measured with normal-hearing subjects and have been used to refine the model. A revision is proposed, which enables a better prediction for speech intelligibility in fluctuating noise, due to the introduction of forward masking.

**Chapter 6** is an evaluation study with normal-hearing listeners to examine the ESII introduced in **Chapter 5** of this thesis. The speech intelligibility is measured in a range of real-life background masking conditions. The results show that it is valid to use the ESII with real-life noises, i.e., sounds that show more complex spectro-temporal variations than do artificial masker signals used in most experiments.

**Chapter 7** describes the effect of the dynamic range of speech on the intelligibility in stationary and interrupted noise for normal-hearing listeners. In addition, a Wide Dynamic Range Compression (WDRC) scheme is used to measure the speech intelligibility for compressed speech-in-stationary, and speech-in-interrupted noise. The observed SRTs are used to refine and extend the model.

In **Chapter 8**, the ESII model is used to predict the speech intelligibility for hearing-impaired listeners in stationary and interrupted noise. Moreover, a

## *Chapter 1*

method is suggested to model the effect of cochlear compression for normal-hearing and hearing-impaired listeners for speech intelligibility in noise.

Finally, **Chapter 9** describes the general conclusions of this thesis, considers some limitations for the prediction of speech intelligibility in noise, and gives suggestions for further study opportunities to examine the speech intelligibility for hearing-impaired and aided hearing-impaired listeners.

This thesis is composed of seven papers (chapters 2 - 8) published or (to be) submitted as research paper. Therefore, chapter 2 to 8 can be read separately. Consequently, there may be some overlap in the introduction and method sections among these chapters.

## *General Introduction*

## *Chapter 2*

# A Speech Intelligibility Index-Based Approach to Predict the Speech Reception Threshold for Sentences in Fluctuating Noise for Normal-Hearing Listeners

*Koenraad S. Rhebergen and Niek J. Versfeld*

*Journal of the Acoustical Society of America* (2005) 117 (4), 2181 –2192.

## **Abstract**

The SII model in its present form (ANSI S3.5-1997) can accurately describe intelligibility for speech in stationary noise but fails to do so for non-stationary noise maskers. Here, an extension to the SII model is proposed with the aim to predict the speech intelligibility in both stationary and fluctuating noise. The basic principle of the present approach is that both speech and noise signal are partitioned into small time frames. Within each time frame, the conventional SII is determined, yielding the speech information available to the listener at that time frame. Next, the SII values of these time frames are averaged, resulting in the SII for that particular condition. Using Speech Reception Threshold (SRT) data from the literature, the extension to the present SII model can give a good account for SRTs in stationary noise, fluctuating speech noise, interrupted noise, and multiple-talker noise. The predictions for sinusoidally-intensity modulated (SIM) noise and real speech or speech-like maskers are better than with the original SII model, but are still not accurate. For the latter type of maskers, informational masking may play a role.

## I. Introduction

In daily life, speech is not always equally intelligible due to the presence of background noise. This noise may mask part of the speech signal such that not all speech information is available to the listener. In order to be able to predict the speech intelligibility under such masking conditions, French and Steinberg (1947), Fletcher and Galt (1950), and later Kryter (1962) initiated a calculation scheme, known as the Articulation Index (AI), which at present still is used by a number of investigators (Rankovic, 1998, 2002; Hogan and Turner, 1998; Müsch and Buus, 2002; Brungart, 2001; Turner and Henry, 2002; Dubno *et al.*, 2002, 2003). In 1984, Pavlovic and others (Dirks *et al.*, 1986; Kamm *et al.*, 1985; Pavlovic, 1984, 1987; Pavlovic and Studebaker, 1984; Pavlovic *et al.*, 1986; Studebaker *et al.*, 1987, 1994) started to re-examine the AI calculation scheme, which has led to a new method accepted as the ANSI S3.5-1997 (1997). Since its revision in 1997, the method is named the Speech Intelligibility Index (SII).

For a given speech-in-noise condition, the SII is calculated from the speech spectrum, the noise spectrum and the listener's hearing threshold. Both speech and noise signal are filtered into frequency bands. Within each frequency band the factor Audibility is derived from the Signal-to-Noise Ratio (SNR) in that band indicating the degree to which the speech is audible. Since not all frequency bands contain an equal amount of speech information (i.e., are not equally important for intelligibility), bands are weighted by the so-called Band-Importance function. The Band-Importance function indicates to which degree each frequency band contributes to intelligibility. It depends on the type of speech material involved (e.g., single words or sentences), and other factors. Finally, the SII is determined by accumulation of the Audibility across the different frequency bands, weighted by the Band-Importance function. The resulting SII is a number between zero and unity. The SII can be seen as the proportion of the total speech information available to the listener. An SII of zero indicates that no speech information is available to the listener; an SII of unity indicates that all speech information is available. Model parameters have been chosen such that the SII is highly correlated to intelligibility. The SII model has been developed to predict the *average* speech intelligibility for a given speech-in-noise condition; it does not attempt to predict the intelligibility of the individual utterances (phonemes or words) of a speech fragment. Also, speech

redundancy or contextual effects, which are inherent to meaningful speech, are captured in the SII model by choice of the model parameters. Higher speech redundancy simply results in less information (i.e., a lower value for the SII) required for understanding the speech message. Within the context of the present paper, an important observation is that the existing SII model does not take into account any fluctuation in the masking noise, since the SII is computed from the long term speech and noise spectrum. Therefore, the SII is independent of the amount of fluctuations in the noise signal.

Numerous papers have reported on experiments dealing with speech intelligibility in fluctuating noise. In almost all cases, normal-hearing listeners perform better in conditions with fluctuating noise compared to those with stationary noise of the same RMS level (Miller, 1947; Miller and Licklider, 1950; Licklider and Guttman, 1957; de Laat and Plomp, 1983; Duquesnoy, 1983; Festen, 1987, 1993; Festen and Plomp, 1990; Gustafsson and Arlinger, 1994; Bacon, *et al.*, 1998; Peters *et al.*, 1998; Brungart, 2001; Versfeld and Dreschler, 2002; Dubno, 2002; Nelson *et al.*, 2003). In many cases, this finding has been phenomenologically explained by stating that the listener is “able to catch glimpses of the speech during the short silent periods of the masking noise” (Howard-Jones and Rosen, 1992, 1993; Festen, 1993; Peters *et al.*, 1998). Recently, Oxenham and co-workers (Oxenham and Plack, 1997; Plack and Oxenham, 1998; Oxenham *et al.*, 2004) proposed that the nonlinear behavior of the basilar membrane enables increased gain during the silent periods, allowing increased audibility. In hearing-impaired subjects, this non-linear behavior is less or even absent, which results in decreased audibility during absence of masking noise. So far, the SII model has been validated only for stationary masking noises, for which it works well. However, it fails to predict speech intelligibility accurately in case of fluctuating noise maskers (Festen and Plomp, 1990; Houtgast *et al.*, 1992; Versfeld and Dreschler, 2002). Other methods, such as the Speech Transmission Index (STI, Steeneken and Houtgast, 1980), or even the speech-based STI (van Wijngaarden, 2002) also fail at this point. To our knowledge, there is still no method that can predict the speech intelligibility in fluctuating noise accurately. Yet, since most real-life noises do exhibit strong variations over time, there is a great interest in a procedure that is able to predict speech intelligibility in fluctuating noises adequately.

In the present paper, an extension to the SII model is proposed in order to be able to predict the speech intelligibility not only in stationary noise, but also in fluctuating noise. The extension consists of an approach where, for a given condition, both speech and noise signal are partitioned into small time frames. Within each time frame, the conventional SII is determined, yielding the speech information available to the listener at that time frame. Next, the SII values of these time frames are averaged, resulting in the SII for that particular noise type. It is hypothesized that this averaged SII is closely related to the speech intelligibility for that condition.

In the next section, an outline of the existing SII model is given. It is followed by a detailed description of the extensions to the existing model, which is introduced to allow predictions of the speech intelligibility in fluctuating noises as well. In extending the SII model, attention has been given to stay as close as possible to the original SII model, thus making as few adaptations as possible. In the choice of the model parameters, this paper concentrates on experiments where speech intelligibility has been assessed with the method of the so-called Speech Reception Threshold (SRT), as described by Plomp and Mimpen (1979). With this method, short everyday sentences are used as speech materials. In Section II-C the SRT method is described in some detail. Next, (in section III) data from the literature are used to evaluate the extended SII model. Finally, in Section IV, predictions and limitations of the extended SII model will be discussed.

## **II. Model description**

### **A. The SII model**

A detailed description of the SII model is given in ANSI S3.5-1997 (1997). Here, a brief overview is given, so that in the next section the extensions to the existing model are easier to follow.

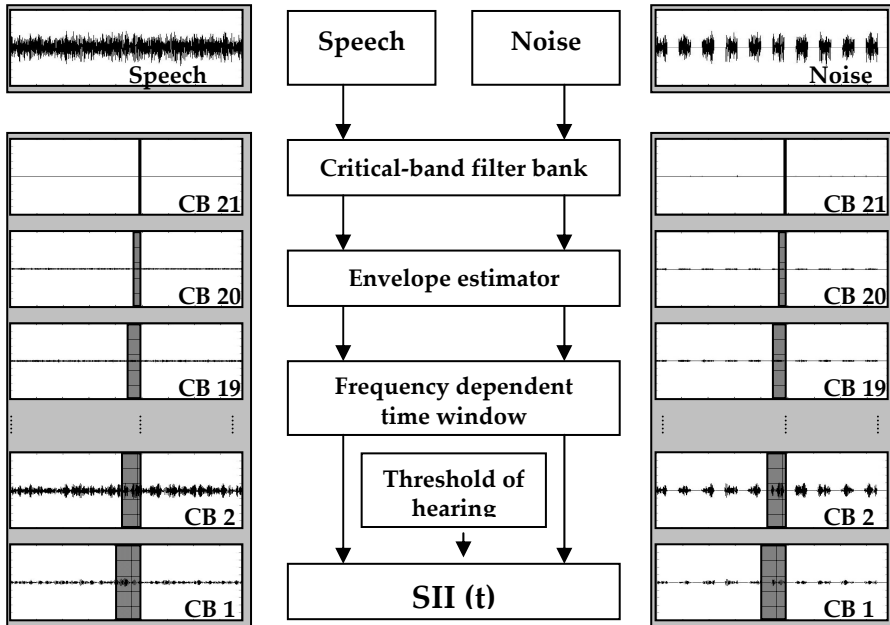
The SII model basically calculates the average amount of speech information available to a listener. To that extent, the model uses the long-term averaged speech spectrum and the long-term averaged noise spectrum as input. Both speech and noise spectrum are defined as the spectrum level (in dB/Hz) at the

eardrum of the listener. Within the model, an option exists to partition the speech and noise spectrum into octave bands, one-third-octave bands, or critical bands. In this paper, spectra are partitioned into critical bands (given in Table I of the ANSI S3.5-1997 standard) although the other two options are equally valid. Within each critical band, the spectrum level is separately determined for both speech and noise. Next, correction factors are taken into account for effects such as upward spread of masking for both speech and noise, inaudibility due to the auditory threshold for pure tones, and distortion due to excessive high speech or noise levels. Then, within each frequency band, the difference between the speech and noise level (signal-to-noise ratio or SNR) is calculated and this value is multiplied with the so-called Band-Importance function, which results in the proportion of information in that band that is available to the listener. The Band-Importance function may depend on the type of speech materials (e.g., sentences or words), or level. Finally, these values are added, yielding the Speech Intelligibility Index (SII), or the amount of speech information available to the listener. For normal-hearing listeners, the SII has proven to be closely related to the average intelligibility in a given condition where speech is masked by a stationary noise masker (Pavlovic, 1987).

## **B. Extension to the SII model**

Since the SII model uses the long-term averaged speech and noise spectrum as input, all temporal characteristics of these signals are lost. As mentioned in the Introduction, large differences in intelligibility exist between masking noises that differ from each other solely with respect to temporal fluctuations (e.g., steady versus fluctuating noise). In this section, an extension is presented that does take the temporal characteristics of the masking noise into account. In essence, the SII model is adapted such that the SII is calculated within small time frames, after which the average SII is calculated. A block diagram of the calculation scheme is presented in Figure 2.1. Both speech and noise are analyzed separately for the SII calculation. Although, in principle, regular speech could be used as the speech input signal, speech-shaped noise (i.e., stationary gaussian noise with the long-term average spectrum of speech) was used. The main reason for this is that in combination with stationary noise as a noise masker, all SII values are identical to those obtained with the existing SII

model. This prerequisite is not easily fulfilled when normal speech signals would be used.



**Figure 2.1** Schematic overview of the calculation scheme for the extended SII model. A detailed description is given in the main text. The input speech signal (stationary gaussian noise with the long-term average spectrum of speech) and input noise (in this example interrupted noise with the long-term average spectrum of speech) are separately filtered by a 21 Critical-Band (CB) filter bank. The envelope of the input speech and noise are estimated in every CB (1-21); the instantaneous intensity is estimated in a frequency dependent time window, as indicated by the shaded bars (CB1=35ms to CB21=9.4ms). Every 9.4 ms an SII is calculated as described by ANSI S3.5-1997. For each of the approximately 200 steps (of 9.4ms), the instantaneous SII(t) is determined (sentence of about 2 seconds). Lastly, the SII for that speech-in-noise condition is determined by averaging across all instantaneous SII(t) values.

## *Extended Speech Intelligibility Index*

The SII is in principle designed to predict the average intelligibility of speech in noise and not the intelligibility of individual words or phonemes. In any case, the SII is badly defined in case of silent periods occurring within the normal speech signal because, regardless of the masking noise, the SII will always be zero. Thus, even when a speech signal is presented at a clear level without any masking noise, the SII based on regular speech never will reach unity, due to the inherent silent periods in the speech signal. Moreover, problems will occur if one considers the silent periods between sentences. It is clear that large differences in SII may occur when the silent periods between sentences vary, whereas the actual intelligibility should not be different.

The most straightforward approach to determine the SII within small time frames is to window the speech and noise signal at a given point in time, to calculate the frequency spectrum (by means of a Fast Fourier Transform, FFT), and to derive an SII from the resulting speech and noise spectrum and the threshold of hearing. However, in order to be able to track the perceptually relevant fluctuations over time, the window length should be small enough. This means that the time window should have a duration of several milliseconds, which is the temporal resolution for normal-hearing listeners based on gap detection thresholds in the higher frequency bands (Plomp, 1964; Shailer and Moore, 1983, 1987; Glasberg and Moore, 1992; Eddins *et al.*, 1992; Oxenham and Moore, 1994, 1997; Moore *et al.*, 1996; Plack and Oxenham, 1998; Moore, 1997). Unfortunately, such a short time window leads to the signal-analytical problem that the level in the lower frequency bands is not estimated accurately. On the other hand, a longer time window leads to a poorer grasp of the temporal variations of the signal.

It is known that the temporal resolution of the auditory system is frequency dependent (Shailer and Moore, 1983, 1987). Time constants (i.e., integration times) for the lower frequency bands are larger than those for the higher bands. To overcome the analysis problems on the one hand, and to stay close to the characteristics of the auditory system with respect to temporal resolution on the other hand, the signal was first filtered into 21 critical bands, and the window length was chosen to be relatively short in the higher bands and relatively long in the lower bands. Since in the original SII calculations the frequency bands are essentially non-overlapping (after all, the intensity within each filter band was

derived from the frequency spectrum), a FIR filter bank of order 200 (600 dB/oct.) was used to filter the entire speech and noise signal into the separate bands. Within each band, the temporal envelope was determined by means of a Hilbert Transform. At a given time frame, rectangular windows were used with window lengths ranging from 35 ms at the lowest band (150 Hz), to 9.4 ms at the highest band (8000 Hz). These window lengths were taken from Moore (1997, Chapter 4) for gap detection and have been multiplied by 2.5. The factor 2.5 was chosen to provide a good fit to the present data-set, as will be discussed below. The windows were aligned such that they ended simultaneously. Within each time frame the intensity was determined, and these, together with the absolute threshold for hearing were used as input to calculate the instantaneous SII, for that given time frame. To calculate the SII, the so-called Speech Perception In Noise (SPIN) weighting function (ANSI S3.5-1997, 1997, Table B.1) was used. This choice seems to be valid, since the speech materials of Plomp and Mimpen (1979) are closely related to the SPIN materials with respect to sentence length and redundancy. Last, the SII for the speech-in-noise condition under consideration was determined by averaging across all instantaneous SII values.

### **C. Speech Reception Threshold**

In the present paper, the proposed extension to the SII model was evaluated using existing data from the literature. The data differ from each other with respect to a number of variables that all can have an effect on intelligibility, hence on the parameter settings of the SII model. For example, it is known that the type of speech material (monosyllables, words, sentences, etc.), open or closed response set, and native or non-native language acquisition can have a large effect on intelligibility (Bosman and Smoorenburg, 1995; Drullman and Bronkhorst, 2000; van Wijngaarden, 2003). Next, similarity between masker and target, e.g., in the case where both target and masker consist of a male voice (Bronkhorst and Plomp, 1992; Bronkhorst, 2000), has a detrimental effect on the actual threshold (i.e., the signal-to-noise ratio that results in just-intelligible speech). Also, the experimental paradigm influences threshold to a large extent. The adaptive SRT-procedure according to Plomp and Mimpen (1979), and the Just to Follow Conversation (Hygge *et al.*, 1992; Larsby and Arlinger, 1994) result in different threshold levels for the same speech material. Additionally,

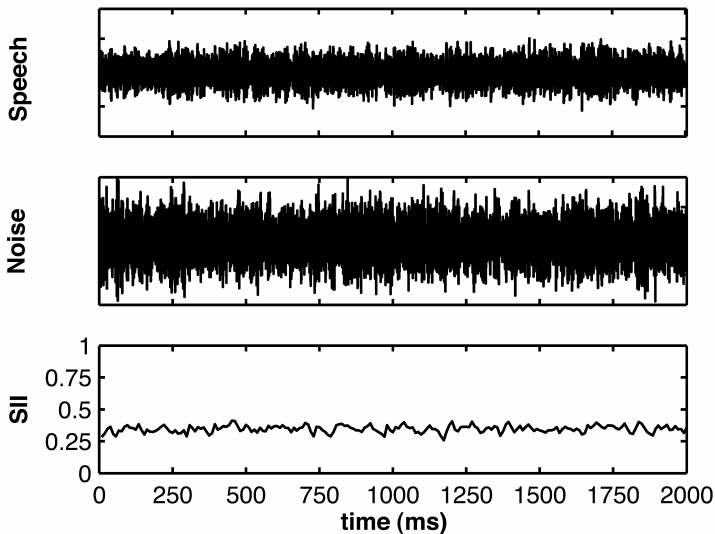
differences in data acquisition (e.g., strictness of sentence scoring) may have an effect on threshold level. Furthermore, different presentation methods (through headphones, loudspeakers, monaural, binaural, diotic, or dichotic presentation) evidently affect threshold level. If one considers masking noises bearing silent periods, it is likely that, even within a group of normal-hearing subjects, differences in hearing level may affect audibility, and thus intelligibility. Finally, when dealing with spectral differences between masker and target, the method used for calibrating signal levels (e.g., RMS, dBA) may have a clear effect.

To enable a comparison between data obtained in different studies, in the present study only thresholds are used that were obtained with the so-called Speech Reception Threshold (SRT) method for sentences, as described by Plomp and Mimpen (1979). Speech materials consist of simple everyday sentences, having a length of 8 to 9 syllables (Plomp and Mimpen, 1979; Nilsson *et al.*, 1994; Versfeld *et al.*, 2000). The SRT is defined as the signal-to-noise ratio (SNR) needed for 50 % sentence intelligibility. The SRT is estimated as described by Plomp and Mimpen (1979): A list of 13 sentences, unknown to the listener, is monaurally presented via headphones. The masking noise is presented at a fixed level, whereas the sentence level is varied adaptively. The first sentence starts at a very unfavorable SNR, and is repeated each time at a 4-dB higher level until the listener is able to repeat every word of this sentence exactly. The SNR of the twelve remaining sentences is varied adaptively with a step size of 2 dB using a one-up, one-down procedure. The SNR of the next sentence is increased by 2 dB after an incorrect response and decreased by 2 dB after a correct response. The average adjusted SNR of sentence 5 through 13 plus the estimated SNR that would have been used for the 14<sup>th</sup> sentence is adopted as the SRT for that particular noise condition. With the speech material of Plomp and Mimpen (1979), normal-hearing listeners require an SNR in stationary speech-shaped noise of -5 to -4 dB, which corresponds to an SII between 0.3 and 0.4 (Steeneken, 1992; Bronkhorst, 2000; Noordhoek, 2000; Versfeld and Dreschler, 2002; van Wijngaarden, 2002, 2003). This means that roughly one-third of the speech information is required to the normal-hearing listener (i.e., the SII is between 0.3 and 0.4) to reach the SRT for these sentences

### III. Model predictions

#### A. Steady-state speech noise

Speech intelligibility in stationary speech-shaped noise can be well predicted by the existing SII model. There are numerous papers dealing with the SRT in stationary speech noise, and all report for normal-hearing listeners at a fixed noise level between 60 and 80 dBA an SRT for sentences of approximately  $-4.5$  dB (de Laat and Plomp, 1983; Middelweerd *et al.*, 1990; Festen, 1987; Festen and Plomp, 1990; ter Keurs *et al.*, 1993; Versfeld and Dreschler, 2002; Neijenhuis, 2002). For speech in stationary speech noise, an SRT of  $-4.5$  dB results for the existing SII model in an SII value of 0.35.

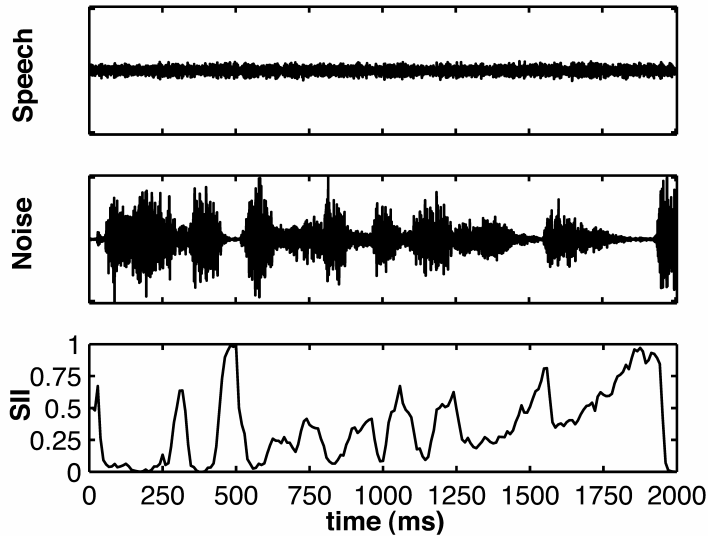


**Figure 2.2** Representation of the SII with the extended SII model for a speech-in-noise sample of 2 seconds. The upper panel represents a speech signal of a female speaker. The middle panel represents a stationary speech-shaped masking speech noise. The noise has been scaled to 60 dBA. The target has been scaled to 55.5 dBA, which results in an SNR of  $-4.5$  dB. The lower panel displays the resulting instantaneous SII as a function of time. The SII averaged across time is equal to 0.35.

Figure 2.2 displays the results of a calculation with the extended SII model for speech in stationary speech noise. The upper panel in Figure 2.2 displays the waveform of a speech signal representation (that is— a stationary speech-shaped noise signal instead of an actual speech signal, as discussed in the previous section) with a duration of two seconds, presented at a level of 55.5 dBA. Here, speech noise was taken from Versfeld *et al.* (2000) for the female speaker. The middle panel shows a 2-s sample of the stationary speech-shaped noise masker derived from the same female speaker, at a level of 60 dBA. The lower panel in Figure 2.2 shows the resulting instantaneous SII, where the SII has been determined every 9.4 ms. Due to the fact that speech and noise signal are uncorrelated (different noise samples), small fluctuations in the instantaneous SII occur. It is easy to see that the SII, averaged across the 2-s sample is between 0.3 and 0.4. In fact, the average is 0.35, which is identical to the value obtained by the existing SII model. Many conditions with speech in stationary noise have been studied, and all calculations show that neither speech type nor noise type result in differences between the existing SII model and the present extended SII model. In conclusion, the extended SII model yields exactly the same results as the existing SII model, as long as a stationary masking noise is used.

## **B. Speech noise with a speech-like modulation spectrum**

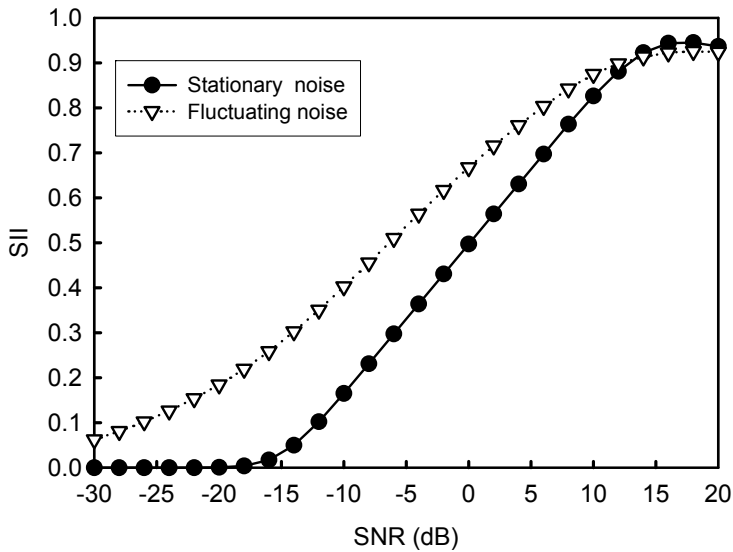
As discussed above, the existing SII model is not able to correctly predict intelligibility for speech in modulated noise. This section deals with speech intelligibility for speech in noise with a speech-like spectrum and a single-speaker modulation spectrum. The generation of this type of noise is described by Festen and Plomp (1990). With normal-hearing subjects, several papers report for this condition an SRT around  $-12$  dB (Festen and Plomp, 1990; ter Keurs *et al.*, 1993; Versfeld and Dreschler, 2002; Neijenhuis, 2002), when the noise level is between 60 and 80 dBA. Computations with the existing SII model yield a score of 0.089, which is far too low.



**Figure 2.3** Representation of the SII with the extended SII model for a speech-in-noise sample of 2 seconds. The upper panel represents a speech signal of a female speaker. The middle panel represents a fluctuating speech-shaped masking speech noise, as used by Festen and Plomp (1990). The noise has been scaled to 60 dBA. The target has been scaled to 48 dBA, which results in an SNR of  $-12$  dB. The lower panel displays the resulting instantaneous SII as a function of time. The SII averaged across time is equal to 0.35.

Figure 2.3 displays the results of the calculations with the extended SII model, similar to the previous section. The upper panel displays the waveform of a speech signal (again, taken as a stationary speech-shaped noise signal) with a duration of two seconds, presented at a level of 48 dBA. The middle panel shows a 2-s sample of the modulated speech noise masker, at a level of 60 dBA. The lower panel in Figure 2.3 shows the resulting instantaneous SII, where, in contrast to the findings in Figure 2.2, the SII value greatly varies over time. It ranges from values close to zero (at points in time where the speech is entirely masked by the masking noise) to values near unity (at points where the masking noise is momentarily absent). The lower panel thus denotes the amount of speech information available to the listener as a function of time. Averaging across time results in an SII score of 0.35. Because large fluctuations

exist over time, a suitably long period has to be chosen to average across. The time interval required to reach stable values for the SII depends on the periodicity, or alternatively, randomness, of the signal as well as on the modulation frequencies in the masking signal. With the present type of masking noise, where the modulations are most prominent near 4 Hz, a period of 2 seconds appears to be long enough to reach a between samples standard deviation for the SII of 0.0056. Increasing the period to 4 seconds decreases the standard deviation of the SII to 0.0030.



*Figure 2.4 SII as a function of SNR as calculated with the Extended SII model. Filled symbols denote calculations with a stationary noise masker with the long-term spectrum of the female target speaker. Open symbols denote calculations with a fluctuating noise masker with the long-term spectrum of the female target speaker and a speech-like modulation spectrum. The level of the noises was set to 60 dBA.*

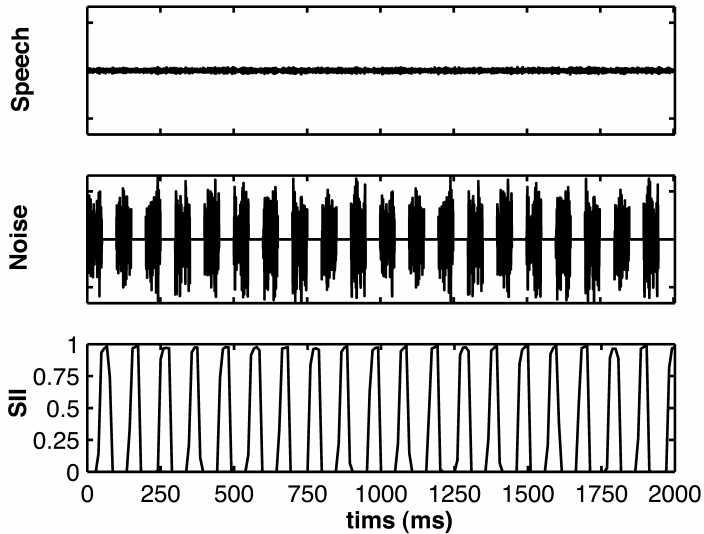
Figure 2.4 displays the SII as a function of the SNR. Here, the masking noise has been kept fixed at 60 dBA, and the level of the speech has been varied between

30 and 80 dBA (thus between SNRs of  $-30$  and  $+20$  dB). With stationary speech noise (denoted as filled symbols in Figure 2.4) the SII starts to deviate from zero as the SNR reaches a value of  $-15$  dB and increases almost linearly with the SNR up to a value of  $+15$  dB. At this value, the speech level is about 75 dBA, and the distortion factor in the SII model prevents the SII from reaching unity. The behavior of the SII as a function of SNR with stationary noise is identical for the existing and the extended SII model. Differences between the two models arise when fluctuating noise is used as a masker. Since the existing SII model does not take the amplitude modulations in the noise masker into account, the SII as calculated with the existing SII model will be identical to that calculated for stationary noise. The SII as a function of SNR for fluctuating noise predicted by the extended SII model is given with open symbols in Figure 2.4. Even at very low signal-to-noise ratios, there is still some speech information available to the listener and the SII exceeds zero. Increasing the SNR causes the SII to increase, but the slope of the function is not as steep as that calculated for speech in stationary noise. Again, at higher speech levels, the distortion factor of the SII model causes the function to level off, such that the SII does not reach unity. An important observation seen in Figure 2.4 is that a constant SII value of 0.35 (the information required to reach threshold) results in an SRT of  $-4.5$  dB for stationary masking noise and  $-12$  dB for fluctuating masking noise.

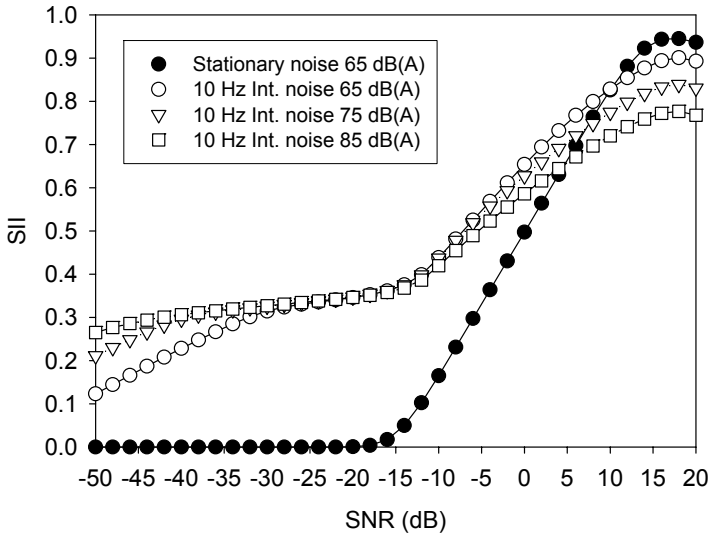
### C. Interrupted speech noise

De Laat and Plomp (1983) measured SRTs for sentences in interrupted (gated) speech noise with a duty cycle of 50 %. Modulation frequency was 10 Hz. Masking noise was presented at 65, 75, or 85 dBA. Figure 2.5 displays the calculations with the extended SII model, similar to Figures 2.2 and 2.3. The upper and middle panel show the speech signal and masking noise signal, respectively. Signal and noise level are 42 dBA and 65 dBA, respectively. The SNR thus is  $-23$  dB. The lower panel shows the SII as a function of time. As seen earlier, the SII is close to zero when the masking noise is present, and is close to unity when the masking noise is absent. Due to the longer integration times in the lower frequency bands, the SII does not change as rapidly as the interrupted noise, but rather smears out over time. Again, the SII averaged across time is equal to 0.35.

## Extended Speech Intelligibility Index



*Figure 2.5* Representation of the SII with the extended SII model for a speech-in-noise sample of 2 seconds. The upper panel represents a speech signal of a female speaker. The middle panel represents an interrupted speech-shaped masking speech noise, as used by de Laat and Plomp (1983). The noise has been scaled to 65 dBA. The target has been scaled to 42 dBA, which results in an SNR of  $-23$  dB. The lower panel displays the resulting instantaneous SII as a function of time. The SII averaged across time is equal to 0.35.

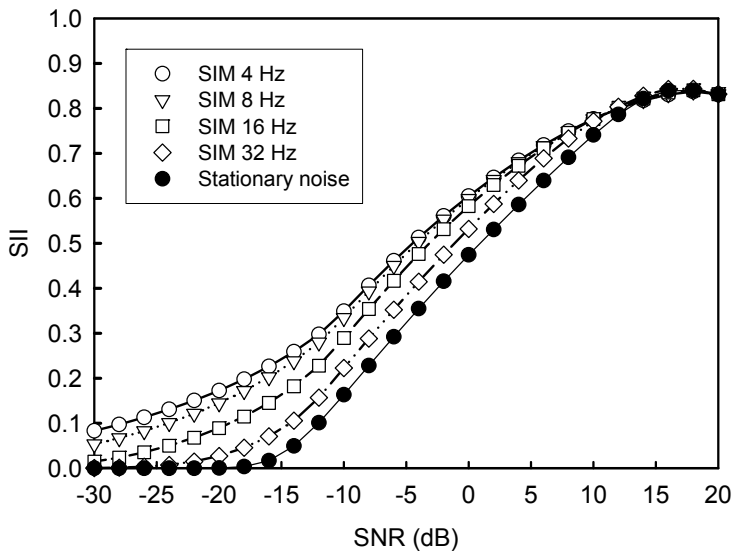


**Figure 2.6** SII as a function of SNR as calculated with the Extended SII model. Filled symbols denote calculations with a stationary noise masker with the long-term spectrum of the female target speaker at a level of 60 dBA. Open squares, circles, and triangles denote calculations with the interrupted noise masker with the long-term spectrum of the female target speaker where the level of the noise was set to 65, 75, and 85 dBA, respectively.

Figure 2.6 displays the SII as a function of SNR for stationary speech noise (filled symbols), and for the three conditions with 10 Hz interrupted noise used in de Laat and Plomp (1983, open symbols; noise at 65, 75 and 85 dBA). At low SNRs (between  $-15$  and  $-35$  dB), speech is entirely masked at moments when the masking noise is present, and it is audible in the gaps. Due to the gaps in the masking noise, values for the SII are relatively independent of SNR and are still quite large, in the order of 0.3. At even lower SNRs (below  $-35$  dB), SII eventually decreases to zero, due to the fact that the speech signal will fall below the absolute threshold. Absolute threshold here has been taken equal to 0 dBHL. At an SNR of  $-15$  and larger, portions of the speech signal start to exceed the noise signal, and SII increases. Again, at high speech levels, distortion occurs which causes the function to level off. De Laat and Plomp (1983) found

## Extended Speech Intelligibility Index

an SRT of  $-23$  dB,  $-26$  dB, and  $-29$  dB at a presentation level of the noise of 65 dBA, 75 dBA, and 85 dBA, respectively. Figure 2.6 shows that for these conditions a large variation in the SNR causes only a slight variation in the SII. At time frames where the noise signal is present, no speech information is available; but at time frames where the noise masker is absent, the amount of speech information available is determined by the degree of temporal resolution (i.e., forward and backward masking) as well as by the absolute threshold of hearing. Nevertheless, while computations with the existing SII model give an SII of zero, the extended SII model results in values near 0.35.



*Figure 2.7* SII as a function of SNR as calculated with the Extended SII model. Filled symbols denote calculations with a stationary noise masker with the long-term spectrum of the female target speaker at a level of 75 dBA. Open squares, circles, diamonds, and triangles denote calculations with SIM noise as a masker at a level of 75 dBA, and a modulation frequency of 4, 8, 16, and 32 Hz, respectively.

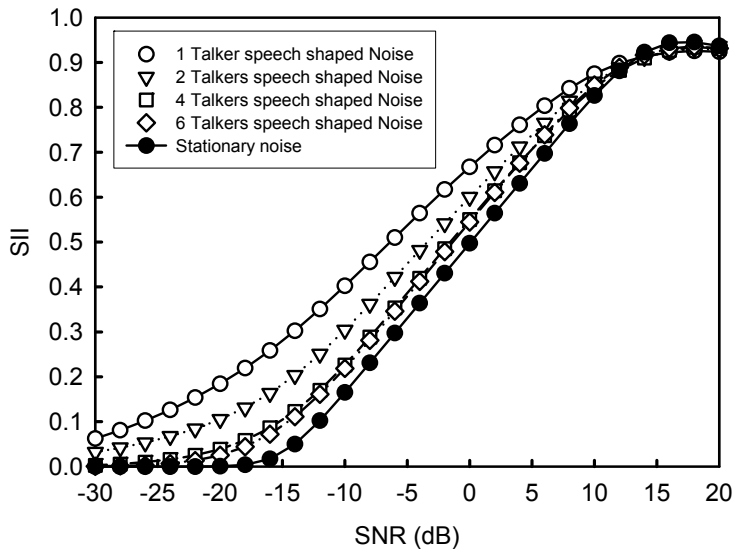
## D. Sinusoidally Intensity-Modulated speech noise

Festen (1987) measured the SRT for sentences in 100 % Sinusoidally Intensity-Modulated (SIM) speech noise. At a presentation level of the noise of 75 dBA he found SRTs of -7.5 dB, -9 dB, -10 dB, -10.2 dB, and -4 dB for modulation frequencies of 4, 8, 16, 32 and "infinity" Hz (steady state), respectively. Figure 2.7 displays the SII as a function of SNR for stationary speech noise (filled symbols), and for four conditions with SIM noise used in the study of Festen (1987, open symbols). Computations with the extended SII model, given an SII of 0.35, result in SRTs of -10, -9, -8, -6.3, and -4 dB for the above-mentioned conditions. The predicted SRT in a 4 Hz SIM noise with the extended SII model seems to be lower compared to SRT values obtained by Festen (1987). Furthermore, the predicted SRT in a 16 Hz or a 32 Hz SIM noise with the extended SII model seems to be higher compared to SRT values obtained by Festen (1987). Although the SRT values obtained with the extended SII model indicate an improvement over the existing model (which predicts an SRT of -4 dB for all conditions), there are still some deviations. So far, no explanation can be given for this result.

## E. Multiple-talker noise

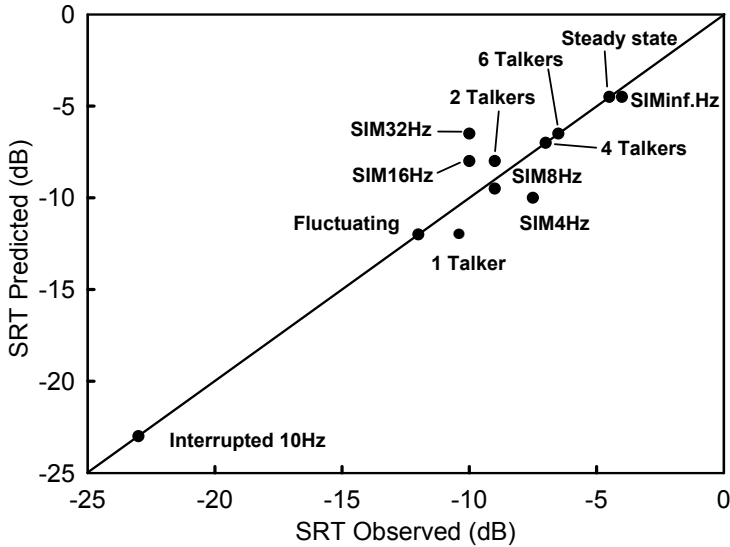
There are numerous papers dealing with the SRT for speech in the presence of one or more competing talkers (e.g., Festen and Plomp, 1990; Bronkhorst and Plomp, 1992; Bronkhorst, 2000; Drullman and Bronkhorst, 2000; Brungart, 2001; Brungart *et al.*, 2001, 2002). It is generally observed that the SRT becomes worse as the number of competing voices increases (Miller, 1947; Carhart *et al.*, 1969; Bronkhorst and Plomp, 1992), eventually resulting in the SRT for stationary speech noise. Bronkhorst and Plomp (1992) measured the SRT for sentences masked by speech-shaped noise modulated by the envelope derived from one, two, four, or six interfering speakers. Observed SRTs were -9.7, -9.9, -7.2, and -6.4 dB, respectively. The stimuli, i.e. speech and fluctuating speech noise, were recorded with a KEMAR manikin and presented monaurally to the subjects.

## Extended Speech Intelligibility Index



*Figure 2.8* SII as a function of SNR as calculated with the extended SII model. Filled symbols denote calculations with a stationary noise masker with the long-term spectrum of the female target speaker. Open squares, circles, diamonds, and triangles denote calculations with noise derived from a single, two, four, and six speakers speech shaped noise. The level of the noises was set to 65 dBA.

Figure 2.8 displays for the four conditions of Bronkhorst (1992) calculations of the extended SII model as a function of the signal-to-noise ratio where it was attempted to simulate Bronkhorst and Plomp (1992) speech shaped noises. It shows that at an SII value fixed at 0.35, the SRT increases from  $-12$  dB (for a single interfering speech shaped noise) to  $-6$  dB (for six interfering speech shaped noises). Although the masking noises were regenerated, since the original masking noises of Bronkhorst and Plomp (1992) were not available, the trend is similar to that reported in the original study.



*Figure 2.9* For a number of different masking noises, the SRT (dB) predicted with the Extended SII model is plotted as a function of the observed SRT (dB). Conditions are denoted in short in the figure.

#### IV. Discussion

Figure 2.9 displays the relationship between the observed SRT (i.e. as measured in actual experiments) and the SRT as predicted by the extended SII model for all conditions described in the previous section, as well as some other conditions that will be discussed below. SRTs were calculated by taking the hearing loss fixed at 0 dB(HL) at all audiometric frequencies, and by setting the threshold value of the SII to 0.35. Different SRTs were obtained by taking the associated sample of the masking noise. The diagonal indicates the points where the observed and predicted SRT are equal. Points under the diagonal indicate an overestimation (with respect to performance) of the predicted SRT; points above the diagonal indicate that listeners generally perform better than predicted by the extended SII model. All predicted SRT values are within a few decibels of the diagonal, or even lie on the diagonal, indicating that the model

does well with the present set of data. The extended SII model yields a substantial improvement over the existing model. Since the latter is insensitive to modulations in the masking noise, it thus predicts for practically all conditions an SRT of  $-4.5$  dB. The most important finding of this paper is that average speech intelligibility in fluctuating noise can be modeled by averaging the amount of speech information across time.

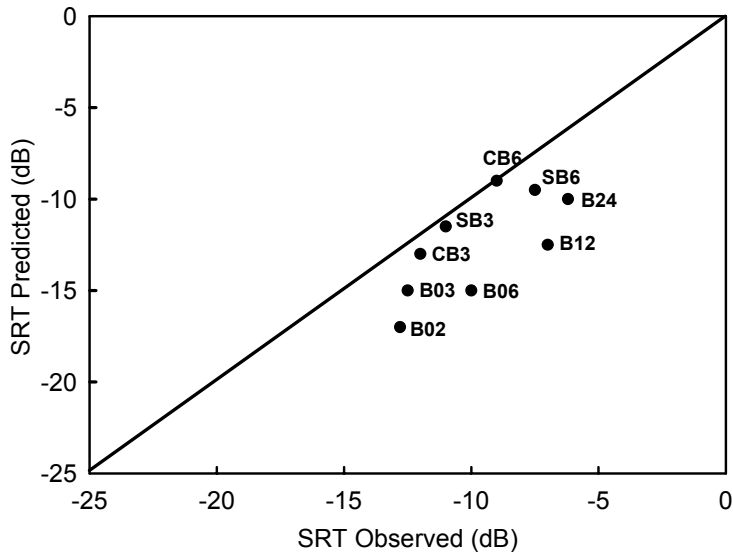
If the data in Figure 2.9 are considered in detail, some of the results obtained with the SIM-noises of Festen (1987) seem to deviate to some degree from the diagonal. Festen (1987) found lowest SRTs for modulation frequencies of 16 and 32 Hz. His finding is in contrast with most data from the literature that indicate maximum performance at 10 Hz (Miller and Licklider, 1950; Licklider and Guttman, 1957; Gustafsson and Arlinger, 1994; Trine, 1995; Bronkhorst, 2000; Nelson *et al.*, 2003). The difference in the position of the minimum may be attributable to differences in stimulus type (gated noise versus SIM noise) and speech materials (word versus sentence scoring). There appears to be a large difference in the SRT results (about 16 dB) found by de Laat and Plomp (1983) and Festen (1987) obtained with about the same modulation frequencies [modulation frequency: 10 Hz; SRT:  $-26$  dB for de Laat and Plomp (1983), compared to modulation frequency: 8 Hz: SRT  $-10$  dB for Festen (1987)]. Festen (1987) suggested that this discrepancy can be due to the relatively broad and deep minimum in the interrupted noise compared to that in the SIM-noise (Figure 2 from Festen, 1987).

The SRT values, obtained with 16-Hz and 32-Hz SIM noise are very similar, *viz.*,  $-10$  dB, and are 2 to 3 dB better than predicted by the extended SII model. As for now, we have no explanation for this part of Festen's (1987) data. Increasing the modulation frequency of the SIM noise results in gaps that are sufficiently small such that they start to fall within the time window of the extended SII model (i.e., smaller than 35 ms). This results in a decrease in performance, and finally performance will approach that of stationary noise. This condition is indicated by "SIMinf.Hz" in Figure 2. 9, and is close to the diagonal. Decreasing the modulation frequency to 8 Hz also results in a point close to the diagonal. However, a further decrease of the modulation frequency to 4 Hz again results in a deviation from the diagonal. The overestimation of the 4-Hz SIM noise may be accounted for by the fact that with these slow modulation rates, masking of complete words in a sentence can occur. This phenomenon has already been

observed by Miller and Licklider (1950), who found optimal performance around modulation rates of 10 Hz. The mere fact that complete words are masked implies that the SRT procedure –where every word of the sentence needs to be repeated correctly– is unsuitable for these low modulation frequencies. Indeed, Trine (1995) shows that in the so-called Just-to-Follow-Conversation (JFC) procedure, the signal-to-noise ratio keeps on decreasing below modulation rates of 8 Hz. In this procedure, the subject is asked to adjust the level of speech in a fixed given noise masker such that he or she is able to “just follow” the speech. This procedure does not require the intelligibility of individual syllables, words, or even sentences. Therefore, the optimum performance for 8 Hz is a procedural artifact. Hence to validate the extended SII model for masking noises comprising modulation rates of, say, 8 Hz and below, procedures other than the SRT-procedure of Plomp and Mimpen (1979) should be utilized.

### **A. Effect of informational masking**

The extended SII model may not be able to predict SRTs accurately in conditions where speech and masking noise interfere at a higher level. One example of such interference is when both target speech and masking noise are derived from the same speaker. In that condition, the listener is confused since he or she does not know which signal represents the target and which components of the signal represents the masker. Festen and Plomp (1990) describe a number of conditions where speech is masked by a single speaker or by multiple speakers. Indeed, performance for speech intelligibility in time-reversed masking speech is better than for forward masking speech. This additional masking, on top of energetic masking, is called informational masking (Bronkhorst, 2000; Brungart, 2001; Brungart *et al.*, 2001): The spoken message of real interfering speech accounts for a rise in SRT.



**Figure 2.10** The SRT (dB) predicted with the Extended SII model is plotted as a function of the observed SRT (dB) for the noise maskers used in Festen (1993). Conditions are denoted by abbreviations in the figure. In conditions B02 through B24, conditions consisted of speech fragments that were manipulated by shifting individual frequency bands of the noise masker independently over time. In conditions CB3, CB6, SB3, and SB6, half of the speech masker was replaced by stationary speech noise. For further details the reader is referred to the main text.

In another experiment, Festen (1993) measured SRTs in other speech-like maskers. The target speech was uttered by a female speaker (of Plomp and Mimpen, 1979). The interfering speech consisted of comparable sentences from a male voice (Smootenburg, 1992). In the reference condition, the interfering speech signal consisted of a concatenation of sentences, with no pauses between the sentences. Five other conditions were derived from this reference condition by first dividing the masking speech stream into 2, 3, 6, 12, or 24 separate frequency bands, that next were independently shifted in time. One may see this masker as an addition of 2, 3, 4, 6, 12, or 24 speakers where the speech of

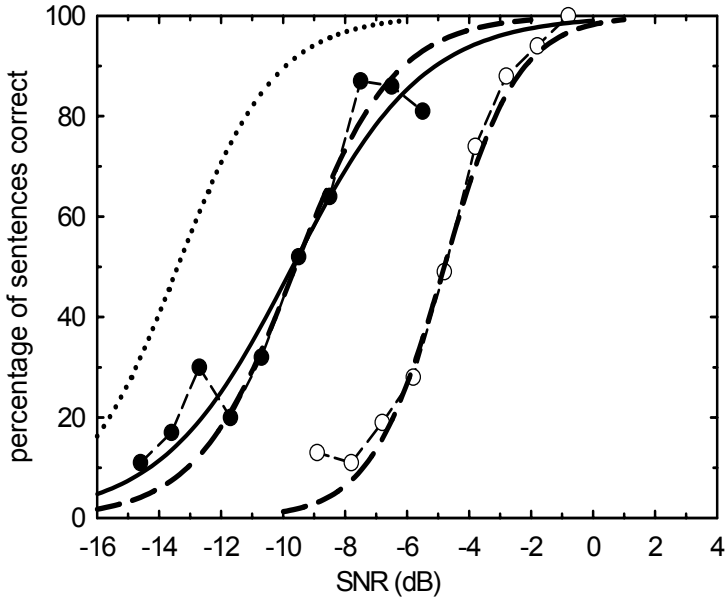
the individual speakers do not overlap in frequency. The result is a masker that sounds very speech-like. The measured SRTs as well as the SRTs calculated with the extended SII model are displayed in Figure 2.10. Different conditions are denoted as B02, B03, B06, B12, and B24, where the number denotes the number of frequency bands. The extended SII model appears to overestimate the observed SRT values of all conditions by 4 to 5 dB. Although speech and noise masker were well discernable, informational masking may have played a role, since the maskers still resembled running speech.

In addition to these conditions, Festen (1993) generated other maskers, where the upper 1/3 octave of each frequency band in the 3- and 6-band speech masker was replaced by noise of the same level as the time average of the original masker. Maskers therefore consisted half of stationary speech-shaped noise. The modulated part was either synchronous in time (labeled in Figure 2.10 as “CB” for “constant bands”) or shifted in time (labeled in Figure 10 as “SB” for “shifted bands”). As can be seen in Figure 2.10, the extended SII model is able to predict the SRT of all these noise conditions (CB3, CB6, SB3, and SB6) reasonably well, probably due to the fact that the masker is less speech like.

In summary, when speech like maskers are used, it is expected that the obtained thresholds are worse than predicted by the extended SII model due to additional (i.e., informational) masking.

## **B. Steepness of the psychometric function**

Festen and Plomp (1990) measured entire psychometric functions for speech in stationary and fluctuating noise. Given the larger dynamic range of fluctuating noise, one would expect a larger range in SNR in which the speech is audible, hence a shallower slope for the fluctuating noise masker. Indeed, with normal-hearing subjects, at the level for which a score of 50 % is obtained, Festen and Plomp (1990) found a slope of 21.0 %/dB and 11.9 %/dB for stationary noise and fluctuating noise, respectively. The present Figure 2.4, too, shows a shallower slope for fluctuating noise. With the extended SII model, it is possible to predict the slope of the curve obtained with fluctuating noise from that obtained with stationary noise. To that end, it first should be noted that for SNRs from -9 to -1 dB the psychometric curve with stationary noise in Figure 6 of Festen and Plomp (1990) ranges from 0 % to 100 %.



**Figure 2.11** Percentage of sentences correct as a function of signal-to-noise ratio (dB), for a stationary noise masker (open symbols) and fluctuating noise masker (filled symbols) (replotted from Festen and Plomp, 1990). The two solid curves represent Festen and Plomp’s (1990) fit to the data. The dotted curve is predicted by the extended SII model, based on the curve given by Festen and Plomp (1990) for stationary noise. The dashed curve (without symbols) is identical to the dotted curve, except for a shift of 3.8 dB to the right.

Figure 2.4 shows that this SNR range corresponds to a range for the SII of 0.2 to 0.5. An important observation hence is that within the range of 0.2 to 0.5 of the SII, sentence intelligibility changes from 0 % to 100 %. Within that range for the SII, both curves in Figure 2.4 can be well approximated by a linear function. The curve for stationary noise is given by

$$SII_S = (15 + SNR_S) / 30, \quad (1)$$

the curve for fluctuating noise is given by

$$SII_F = (27 + SNR_F) / 40. \quad (2)$$

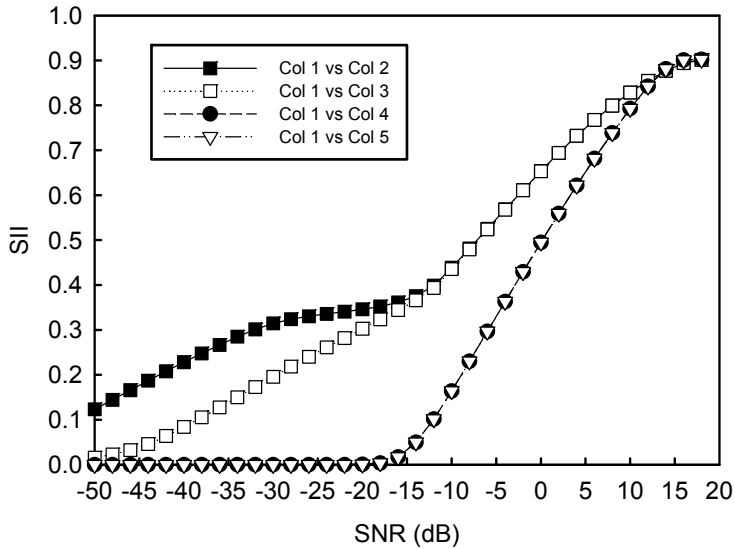
Festen and Plomp (1990) describe their curves with a logistic function

$$p(SNR) = \frac{1}{1 + e^{(M-SNR)/S}}, \quad (3)$$

where  $M$  is the SNR for which the probability on a correct response  $p(SNR)$  is equal to 0.5, and  $S$  is the steepness of the function at  $p(SNR)=0.5$ . For the stationary noise curve in Figure 6 of Festen and Plomp (1990),  $M=-4.7$  dB and  $S=1.19$  dB (corresponding to 21.0 %/dB as given by Festen and Plomp, 1990). For the fluctuating noise curve,  $M=-9.7$  dB and  $S=2.10$  dB (corresponding to 11.9 %/dB). The data of Figure 6 of Festen and Plomp (1990) are replotted in Figure 2.11, together with the two functions given by Festen and Plomp (1990), given as solid curves. When  $SII_S = SII_F$ , Equations (1) and (2) give the relation between  $SNR_S$  and  $SNR_F$ :

$$SNR_S = (21 + 3SNR_F) / 4. \quad (4)$$

By insertion of Equation (4) into Equation (3), the shape of the function for fluctuating noise is obtained. This curve is plotted as a dotted line in Figure 2.11. The predicted curve for fluctuating noise has a slope of 15.6 %/dB and a value for  $M$  of  $-13.3$  dB. The curve is about 3.8 dB to the left of the data of Festen and Plomp (1990), but has a slope that fits very well to the data of Festen and Plomp (1990), as can be seen when the curve is shifted 3.8 dB to the right, as has been done in Figure 2.11 (dashed curve). The slope fits their data even better than their calculated slope of 11.9 %/dB. The fact that the calculated curve does not fall on top of the data of Festen and Plomp (1990) is due to the fact that Festen and Plomp (1990) shifted their data to the average results.



**Figure 2.12** SII as a function of SNR as calculated with the extended SII model. Filled symbols denote calculations with the absolute threshold set to 0 dB(HL). Open symbols denote calculations with the threshold set to 15 dB(HL). Circles and triangles indicate calculations with a stationary noise masker and squares indicate calculations with interrupted noise masker, respectively, both with the long-term spectrum of the female target speaker. The level of the noises was set to 65 dBA

### C. Effect of absolute threshold

With the calculation of the SII, it was assumed that all subjects had normal-hearing, that is – thresholds for all frequencies were taken equal to 0 dB(HL). In real life, thresholds deviate to some degree from this value, but with the normal-hearing group it is generally assumed (ANSI S3.6-1996, 1996) that the hearing level is equal to or less than 15 dB(HL). Given the dynamic range of speech (30 dB) and the presentation level of the masking noise one can calculate the effect of an elevated threshold. With stationary speech noise as a masker, audibility of average conversational speech starts to play a role only at losses of 50 dB(HL) and larger, as can be calculated with the existing SII model. In contrast, with fluctuating noise and interrupted noise, effects become already

noticeable at thresholds of 30 dB(HL) or 15 dB(HL), respectively. The effect of hearing loss on the SII is depicted in Figure 2.12 for both a stationary noise masker and an interrupted noise masker. As can be seen in this Figure, elevating the threshold from 0 to 15 dB(HL) has no effect on the SII with stationary noise, but has a clear effect with interrupted noise. The two curves with interrupted noise start to overlap near an SNR of  $-15$  dB. For the calculations with the extended SII model, little differences in prediction of the SRT in stationary noise were found by variation of the absolute threshold (HL  $<50$ dB). Figure 2.12 nevertheless shows that with these fluctuating noise maskers, the effect of absolute threshold can be substantial, especially at lower presentation levels. This could account for the large standard deviation between subjects found by SRT in fluctuating noises (de Laat and Plomp, 1983; Festen, 1987, 1993; Festen and Plomp, 1990; Bronkhorst, 2000; Versfeld and Dreschler, 2002) compared to the small standard deviation between subjects found by SRT in stationary noises (Plomp and Mimpen, 1979).

#### **D. Effect of window length**

With the presentation of the extended SII model, the signals were windowed in time and the length of the time window was frequency dependent. The choice of the time windows was adapted from Moore (1997) and was based on psychophysical data. As discussed above, given these settings, the extended SII model is able to predict the data well. Within a time window, level variations of the signal are averaged. Thus, the longer the time window, the more the signal is smoothed, thus the more the obtained SII will resemble the existing SII (i.e., the SII of stationary noise). On the other hand, if the time windows are taken smaller, all signal variations are caught, which in case of highly fluctuating maskers as interrupted noise results in better SRTs than actually measured. Calculations have been performed to check whether a single fixed window length for all frequency bands could account for the present data set as well. The results of these calculations show that an optimum fit was obtained with a fixed window length of 12 ms, but that this approach could not account for the data as well as the approach with frequency dependent windows. Yet, it remains possible to manipulate the lengths of the individual time windows, in order to reach an even better fit to the data. However, the present choice of parameters does well, and has the advantage that the window lengths are

derived from psychoacoustical measurements. In this paper, rectangular windows have been taken, but future experiments may point at the use of differently-shaped windows, such as an exponential window. The latter shape may be more similar to the shape of the forward-masking function.

## **E. Extensions to the model**

In this paper the authors purposely have tried to stay as close as possible to the existing SII model. Extensions to the existing SII model have been proposed, which seem to work well for the SRT with sentences in a given number of noise maskers. To see to what extent the model can be generalized to other types of speech material and noise maskers, measurements should be performed. Although the basic assumptions regarding the extensions may remain valid, it seems plausible that, as with the existing SII model, different speech materials require different weighting functions or window lengths. With the present data set, an SII of 0.35 corresponded to the amount of information required to reach the SRT. These data were obtained with normal-hearing listeners. As discussed extensively by Noordhoek (2000), hearing-impaired subjects often require more speech information to reach threshold, which she attributed to supra-threshold deficits. These deficits probably deal with a decrease in spectral or temporal resolution. With the extended SII model, both decreases in resolution can in principle be modeled by increasing the width of the different frequency bands, or by increasing the window length or window shape. Perhaps more sophisticated adaptations to the SII model [such as the temporal window model of Oxenham (Oxenham and Moore, 1997, Oxenham and Plack, 1997)] are required. It is left to future research to find out to what extent the model is able to describe the data.

## **F. Other extensions to the SII model**

Another shortcoming of the SII model is its inability to account for synergetic and redundant interactions among the various spectral regions of the speech spectrum (Steeneken and Houtgast, 1999; Müsch and Buus, 2001). Due to fact that the SII uses the long term spectrum of speech and noise (minimum length of 30 sec; ANSI S3.5-1997, 1997), these interactions among the various frequency bands are lost. Nevertheless, speech communication is remarkably robust for

## Chapter 2

normal-hearing listeners and does not have to be broadband to be highly intelligible (Allen, 1994; Warren, et al., 1995; Lippman, 1996; Stickney and Assman, 2000). Steeneken and Houtgast (1999, 2002) implemented a frequency dependent redundancy correction factor to the STI model, which accounts for synergetic and redundant interactions. Since the STI is related to the SII (van Wijngaarden, 2002), it is in principle possible to implement this redundancy correction factor in the SII calculation method.

### Summary

The present paper describes an SII-based approach to model SRTs (Speech Reception Thresholds) for sentences masked by fluctuating noise. The basic principle of this approach is that both speech and noise signal are partitioned into small time frames. Within each time frame the instantaneous SII is determined, yielding the speech information available to the listener at that time frame. Next, the SII values of these time frames are averaged, resulting in the SII for that particular noise type. From the literature many SRT values are available for a variety of noise types. In this paper, it is shown that this approach can give a good account for most existing data. Hence it forms a valuable extension to the existing SII (ANSI S3.5-1997, 1997) model.

*Extended Speech Intelligibility Index*

## *Chapter 3*

# Release from informational masking by time reversal of native and non-native interfering speech

*Koenraad S. Rhebergen, Niek J. Versfeld and Wouter A. Dreschler  
Journal of the Acoustical Society of America (2005), 118 (3), 1274 – 1277.*

## **Abstract**

In many studies, the influence of intelligibility of the interfering speech is avoided by reversing it in time. Usually, intelligibility with time-reversed interfering speech indeed is higher compared to that with normal interfering speech. However, due to the nature of speech, reversed speech also gives rise to increased forward masking. The latter will result in a decrease in intelligibility. Thus, differences in intelligibility as a consequence of reversing speech in time are due to two opposite effects. This paper describes a Speech Reception Threshold (SRT) test with intelligible and unintelligible interfering speech played normally and time-reversed. With Dutch listeners, Swedish reversed interfering speech gave a rise in SRT of 2.3 dB compared with the Swedish interfering speech (played normally). The difference can be attributed to differences in forward masking. Dutch time-reversed interfering speech gave a decrease in SRT of 4.3 dB compared to (intelligible) Dutch interfering speech. The latter is the result of both a release from informational masking and an increase in forward masking. Therefore, the amount of informational masking is larger than 4.3 dB and, if one assumes similar differences in forward masking for Dutch and Swedish speech, may amount to 6.6 dB.

## I. Introduction

There are various papers describing speech intelligibility in the presence of one or more interfering talkers (e.g., Festen and Plomp, 1990; Bronkhorst and Plomp, 1992; Bronkhorst, 2000; Drullman and Bronkhorst, 2000; Brungart, 2001; Brungart *et al.*, 2001, 2002; Summers and Molis, 2004). Remarkably, intelligibility appears to vary greatly between conditions, which in part might be due to the degree of similarity between the target speaker (here: “target” or “signal”) and the interfering speakers (here: “masker” or “noise”). The more similar target and masker are, the more the listener is confused or distracted, which on its turn results in poorer performance. Differences between target and masker with respect to gender (male or female speech) or intelligibility of the interfering speech (native or foreign language) have only a small effect with respect to actual energetic masking of the target, but certainly can have a large effect on intelligibility. The phenomenon of excess masking is often labeled as informational masking (Drullman and Bronkhorst, 2000; Bronkhorst, 2000; Brungart, 2001; Brungart *et al.*, 2001). This paper concentrates on the part of informational masking due to intelligibility of the interfering speech.

In order to study the speech intelligibility in speech-like maskers, and at the same time avoid the informational masking component due to the intelligibility of the interfering speech, many researchers use fluctuating speech-shaped noise (Festen and Plomp, 1990; Bronkhorst and Plomp, 1992; Peters *et al.*, 1998; Versfeld and Dreschler, 2002) or play the interfering speech signal backwards (Festen and Plomp, 1990; Summers and Molis, 2004). Fluctuating speech-shaped noise is made by modulating the long term speech spectrum with the speech envelope of the interfering talker. The envelope can be extracted from broad band speech (one band), two bands (Festen and Plomp, 1990), three bands (ICRA noise; Dreschler *et al.*, 2001), or even more. The fluctuating speech-shaped noise has more or less the same intensity fluctuations in time as real speech, has the same long-term spectrum of speech, but is generally unintelligible because the fine structure is lost.

By using time-reversed speech as a masker, the spectral contents of speech is in essence untouched. The reversed speech is not intelligible and therefore this component of the informational masking by definition is removed. However,

## *Informational Masking*

the envelope of time-reversed speech is reversed as well. The shape of temporal envelope of speech is typically dominated by plosive sounds (Rosen, 1992), and the envelope of these sounds are characterized by quick onsets (steep slope) and slow decays (shallow slope). Reversal of the speech signal thus results in temporal envelopes that have abrupt offsets. Since the auditory system does not follow the offset instantaneously, but rather displays a decay of the envelope across time, abrupt changes cannot be followed accurately, hence soft signals can be easily masked by a preceding strong signal (called forward masking).

Irino and Patterson (1996), Carlyon (1996), Stecker and Hafter (2000), and Schlauch *et al.* (2001) examined the perception of stimuli with ramped envelopes (gradual attack and abrupt decay) and damped envelopes (abrupt attack and gradual decay). The sum of these studies reveal that ramped signals are subjectively judged longer in duration and are perceived louder compared to damped signals. This is explained by the abrupt offset at a high level of a ramped sound which results in a persistence of perception, and more forward masking compared to a damped sound. The same effects could be expected by time reversed speech since its envelope is more ramped compared to the damped normally played speech (Rosen, 1992).

Forward masking has a clear effect on speech intelligibility (Festen, 1987). The recovery time from forward masking is in the range of 100 to 200 ms (Moore, 2003). For completeness, a second type of masking, present in human auditory perception, should be mentioned, *viz.*, backward masking. In this case, a soft signal is masked by a louder signal that follows it. The phenomenon of backward masking is still poorly understood (Moore, 2003). The amount of backward masking obtained with practiced subjects often is little or none (Moore, 2003). Its effect on speech intelligibility is still unclear and probably not very large.

By reversal of the speech masker in time, one expects on the one hand an improvement in intelligibility of the target speech due to the fact that the masking speech becomes unintelligible (partly release from informational masking), and, on the other hand a decrease in performance due to the fact that the temporal envelope of the reversed speech causes an increase of forward masking. Festen and Plomp (1990) and Summers and Molis (2004) found a slight improvement in performance when using reversed speech instead of normal speech as a masker. This improvement is the result of both effects

combined. From these experiments it is not possible to assess the individual contribution of both components.

The goal of the current experiment is to separate informational masking due to intelligibility of the interfering speech from forward masking. Additional masking due to time reversal of the speech is assessed by measuring the speech intelligibility in masking speech of a foreign language. Since the foreign speech is not intelligible (neither forward nor backward), differences in intelligibility due to time reversal must be due to differences in forward masking alone. The same experiment is repeated in native interfering speech. Differences in intelligibility due to time reversal then are the combined effect of an increase in forward masking and a decrease in informational masking. Since the speech materials are similar, and since the single effect of additional masking due to time reversal is known, the effect of informational masking due to intelligibility of the interfering speech can be estimated.

## II. Method

### a. Subjects

Eight normal-hearing subjects (3 male, 5 female) participated. Their mean age was 25 years and ranged from 21 to 39 years. Subjects were native speakers of the Dutch language. Subjects had at least high school education. Each subject had pure-tone thresholds of 15 dB HL or better at octave frequencies from 125 to 8000 Hz (ANSI S3.6, 1996).

### b. Stimuli

The target speech material consisted of short every-day sentences, uttered by a male speaker (Versfeld *et al.*, 2000). The speech material comprised 39 lists of 13 sentences and was developed for a reliable measurement of speech intelligibility in noise. The speech was stored with a sampling rate of 44.1 kHz and 16 bits resolution.

Swedish and Dutch speech was used as interfering speech. The Swedish speech was developed by Hagerman (1982) and consisted of short sentences read by a female speaker. For the Dutch language, a corpus similar to that of Hagerman (1982) was developed. The Dutch sentences were uttered by a different female

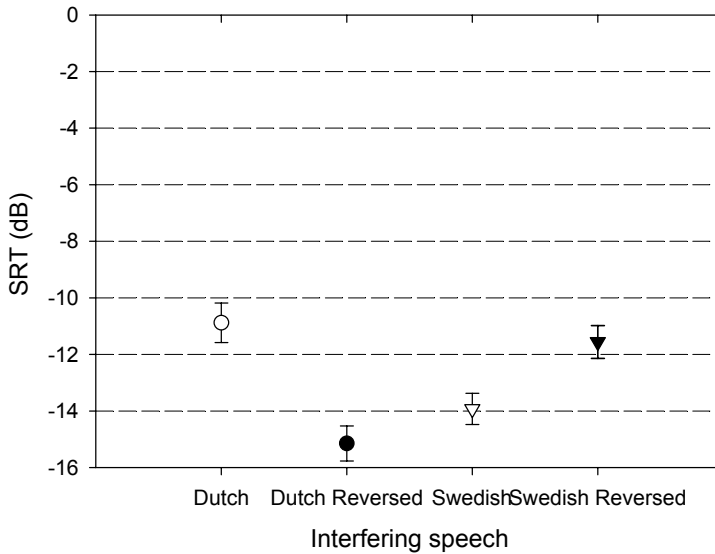
speaker. The Dutch speech was rescaled to obtain the same speaking rate as that of Hagerman's (1982) set, *viz.*, 3.5 seconds per sentence. Scaling was done by use of the PSOLA method (Pitch-Synchronous Overlap Add, a method for manipulating the pitch and duration of an acoustic speech signal, Moulines and Charpentier, 1990). Pseudo-running speech was obtained by concatenation of sentences in succession without pauses.

### **c. Procedure**

Subjects were tested individually in a sound-insulated booth. The monaural speech-reception threshold (SRT) was measured at the better ear for a fixed noise level of 65 dBA. Signals were played out via an Echo soundcard (Gina 24/96) on a PC at a sample frequency of 44.1 kHz and were fed through a TDT Microphone Amplifier (MA2) and a TDT Headphone Buffer (PA4) via THD 39P headphones. After the presentation of a sentence, the subject's task was to repeat the sentence he or she had just been presented. A sentence was scored as correct if all words in that sentence were repeated without any error. A list of 13 sentences, unknown to the subject, was used to estimate the level at which 50 % of the sentences was reproduced without any error, the so-called Speech Reception Threshold, or SRT. For a given condition, the first sentence of the list started far below the expected SRT. The sentence was repeated each time at a 4 dB higher level until the subject was able to reproduce it correctly. The twelve remaining sentences in that list were presented only once, following a simple up-down procedure with a step size of 2 dB. The SRT was estimated according to the procedure described by Plomp and Mimpen (1979), *i.e.*, by taking the mean Signal to Noise Ratio (SNR) of sentence five to thirteen plus the estimated SNR that would have been used for the fourteenth sentence. With each sentence presentation, a random sample of the interfering speech was taken. It started 1200 ms before the start of the sentence and stopped at least 800 ms after the sentence.

In total four masking conditions were tested: Dutch masker forward, Dutch masker reversed, Swedish masker forward, and Swedish masker reversed. The experiment was partitioned into two blocks, a test and a retest. To avoid confounding of measurement condition order and sentence lists, the order of conditions and sentence lists was counterbalanced across subjects according to

an 8 by 8 Latin Square method. In total, each subject received 8 lists of 13 sentences preceded by three practice lists.



*Figure 3.1. Speech Reception Threshold (dB) as function of Interfering Speech type. Dutch (circles), and Swedish (triangles) speech played normal (open symbols), and reversed (filled symbols). Error bars denote the standard deviations between subjects*

### III. Results

Figure 3.1 shows the SRT-values averaged across subjects and test-retest for each of the four conditions of the interfering speech. Error bars denote the standard error of the mean. A 4[condition] x 2[test/retest] x 8[subject] Analysis Of Variance (ANOVA) was performed on the data-set. Of the main effects, only differences in “condition” were significant ( $F[3,21]=24.6, p<0.001$ ). The SRT of the retest was on average 1.2 dB better than the test, but this difference was not significant ( $F[1,7]=4.6, p>0.05$ ). Also, differences between subjects were just not significant ( $F[7,4]=5.6, p>0.05$ ). None of the interactions were significant.

## *Informational Masking*

The difference in SRT between the two conditions with Swedish speech was significant (Tukey's HSD test,  $z=2.4$ ,  $p=0.02$ ) and was on average 2.3 dB. Thus, by time reversing speech, the SRT increases (i.e., performance worsens) from -13.9 dB to -11.6 dB. The difference between the two conditions with Dutch interfering speech was also significant (Tukey's HSD test,  $z=4.28$ ,  $p<0.001$ ). Here, time reversal of the masking speech causes the SRT to improve by 4.3 dB. The latter effect was apparent in all subjects, and the result is in agreement with that of Festen and Plomp (1990). Lastly, the difference between the two conditions with reversed masking speech was also significant (Tukey's HSD test,  $z=3.6$ ,  $p<0.001$ ).

## **IV. Discussion**

The most important finding of the present experiment is that time reversal of foreign, non-intelligible speech that is used as a masker, affects intelligibility and causes the SRT to increase by more than 2 dB. Most likely, reversal of the envelope of the speech signal results in an increased contribution of forward masking. The main conclusion to be drawn from this experiment is that reversal of a speech masker indeed introduces additional masking on top of energetic masking in the form of forward masking.

Time reversal of intelligible (Dutch) masking speech results in a decrease in the SRT of 4.3 dB. This difference is the result of the elimination of intelligibility of the masker on the one hand (enhancing speech intelligibility), and the introduction of additional forward masking on the other hand (decreasing speech intelligibility). If one assumes that the additional amount of forward masking is similar for the Swedish and Dutch masking speech, then in the present experiment the effect of informational masking due to intelligibility of the speech masker can be estimated to be  $4.3 + 2.3 = 6.6$  dB.

If intelligibility of the speech masker were the sole factor determining the amount of informational masking, one would expect the SRT with Swedish reversed speech and Dutch reversed speech to be similar. In the present experiment this is not true; the difference is about 3.6 dB. Apparently, the Dutch reversed speech is a poorer masker (SRT = -15.2 dB) than the Swedish reversed speech (SRT = -11.6 dB). If Dutch interfering speech played in reverse sounds

more similar to Dutch speech than Swedish speech played either forward or in reverse, one would expect the opposite. It is interesting to note that subjects told they could not distinguish between the Swedish reversed speech and the Dutch reversed speech. Also, they misjudged the reversed Dutch speech for Swedish speech and vice versa. Therefore, the effect probably is not very large. As mentioned in the introduction, other factors than intelligibility contribute to informational masking. From the present results, it is difficult to say which of the many potential factors contributed most; perhaps differences in intonation, rhythm, mean pitch, modulation spectrum, or differences in the long-term speech spectrum. If the interfering native- and non-speech were uttered by the same person, with the same pronounced intonation, rhythm and pitch, one would expect the SRTs in the reversed condition to be more alike. But even then, differences might still exist due to specific characteristics of the language. The present experiment is a first attempt to separate energetic masking from informational masking. Future research with different native- and non-native interfering speech, uttered by the same person may give more insight into the effects of forward masking and informational masking on speech intelligibility.

## Conclusions

By using speech from different languages, played both forward or time-reversed, it is possible to disentangle the effects of informational masking caused by intelligibility of the interfering speech and energetic masking. However, time reversal of speech results in an increase in temporal (forward) masking. In the present experiment, this effect is about 2.3 dB. The release from informational masking by making intelligible speech unintelligible by reversing it in time is obscured by an increase in forward masking. In the present experiment, it is shown that informational masking might be as large as 6.6 dB if one assumes the same amount of additional forward masking with Dutch and Swedish reversed speech.

## *Informational Masking*

## *Chapter 4*

# Learning Effect Observed For The Speech Reception Threshold In Interrupted Noise With Normal-hearing Listeners

*Koenraad S. Rhebergen, Niek J. Versfeld and Wouter A. Dreschler  
Submitted to Journal of the Acoustical Society of America.*

## **Abstract**

Traditionally, the Speech Reception Threshold (SRT) is measured in stationary noise. However, non-stationary masking noises are gradually more used, since they seem to be more sensitive to discriminate between conditions and listeners. The results of recent experiments suggest that a learning effect might be present for the SRT in interrupted noise. This paper describes an SRT test with a female or male target speaker with stationary noise and with an 8 Hz interrupted noise. Contrary to repeated SRT measurements in stationary noise, a significant decrease was observed for SRTs in interrupted noise. For both speech materials, after five replications, the SRT improved about 3 dB to 3.5 dB in comparison to that of the first SRT. The outcome of this experiment thus shows that there is a large learning effect present in SRT measurements with interrupted noise but not in stationary noise.

## I. Introduction

Today, the Speech Reception Threshold (SRT) introduced by Plomp and Mimpen (1979) is commonly used in both research and the clinic to measure the 50 % sentence intelligibility level in noise for normal-hearing (NH) or hearing-impaired (HI) listeners. In almost all studies, stationary speech noise is used as a masker. Thus far, no significant learning effects have been observed for speech in stationary noise (Versfeld *et al.*, 2000). In the last years, non-stationary masking noises are progressively more used, because on the one hand these conditions reflect more real-life performance (Festen and Plomp, 1990; Middelweerd, Festen and Plomp, 1990; Versfeld and Dreschler, 2002; Festen and Plomp, 2002), and on the other hand the inter-individual differences are larger, compared to the traditionally used stationary noise. As a result, there is a better discrimination in SRT between NH and HI listeners in fluctuating noise as masking condition (e.g., de Laat and Plomp, 1983; Festen 1987; Festen and Plomp, 1990; Versfeld and Dreschler, 2002; Festen and Plomp, 2002). A recent study by the authors (Rhebergen, Versfeld and Dreschler, 2006b) showed that SRTs for fluctuating noise has an effect of learning: In contrast to SRTs in stationary noise, SRTs in interrupted noise were systematically better in the re-test. In the past, due to the time consuming method and the limited number of sentence lists available, the SRT test mostly is measured only once for a particular condition. However, for a reliable, robust and reproductive result, either for validation purposes of the Extended SII method of Rhebergen and Versfeld (2005), or for clinical use, it is important to explore the possible learning effect for the speech intelligibility in interrupted noise in more detail.

## II. Experiment

Results of a recent study by Rhebergen *et al.*, (2006b) suggested a learning effect for SRT measurements in interrupted noise. It was observed that the mean SRT in interrupted noise of the retest was on average 0.9 dB significant lower than that of the first test whereas there was no learning effect observed in stationary noise. The present experiment was designed to determine if a learning effect is present for speech intelligibility in interrupted noise.

## Method

### a. Subjects

Eight normal-hearing subjects (3 male, 5 female) participated. Their age ranged from 24 to 42 years and was on average 30.9 years. Subjects had at least high school education and were native speakers of the Dutch language. Each subject had pure-tone thresholds of 15 dB HL or better at octave frequencies from 125 to 8000 Hz (ANSI S3.6, 1996).

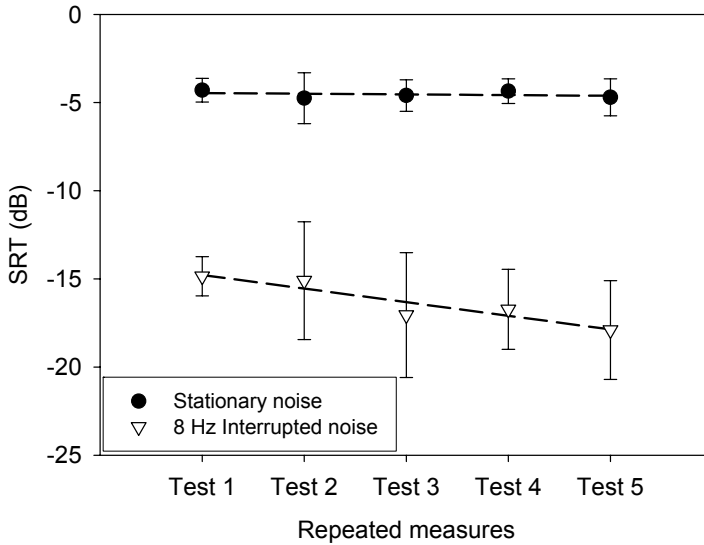
### b. Stimuli

The target speech material consisted of short every-day sentences, uttered by a female or a male speaker (Versfeld *et al.*, 2000). The interfering noise conditions comprised one condition with stationary noise, and one condition with 8 Hz interrupted noise with a duty cycle of 50 %. in all conditions, the noise spectrum was equal to the long-term average spectrum of the target speaker (female or male).

### c. Procedure

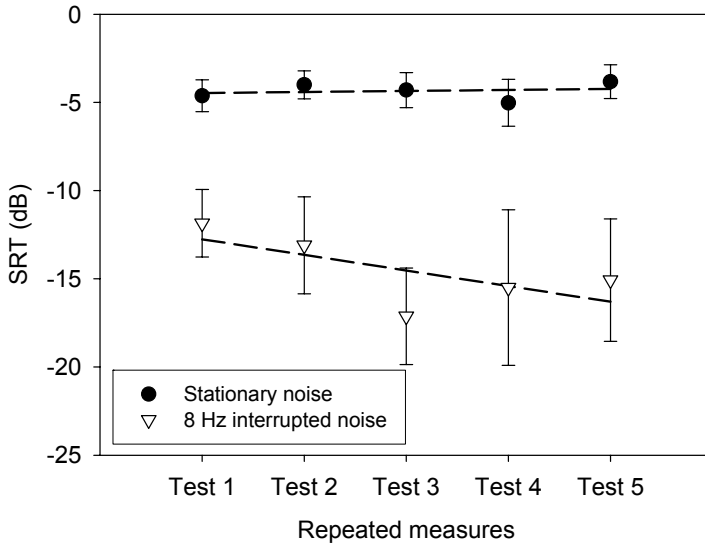
Subjects were tested individually in a sound-insulated booth. Signals were played out via an Echo soundcard (Gina 24/96) on a PC at a sample frequency of 44.1 kHz and were fed through a TDT Headphone Buffer (PA4). Subjects received the signals monaurally at their better ear via TDH 39P headphones at a fixed noise level of 65 dBA. After the presentation of a sentence, the subject's task was to repeat the sentence he or she had just been presented. A sentence was scored correct if all words in that sentence were repeated without any error. A list of 13 sentences, unknown to the subject, was used to estimate the signal-to-noise ratio (SNR) at which 50 % of the sentences was reproduced correctly, the so-called Speech Reception Threshold (SRT). For a given condition, the first sentence of the list started far below the expected SRT. The sentence was repeated each time at a 4 dB higher level until the subject was able to reproduce it correctly. The twelve other sentences in that list were presented only once, following a simple up-down procedure with a step size of 2 dB. The SRT was estimated according to the procedure described by Plomp and Mimpen (1979), i.e., by taking the mean SNR of sentence five to thirteen plus the SNR that would have been used for the fourteenth sentence. With each sentence presentation, a random sample of the interfering noise was taken. It

started 1200 ms before the start of the sentence and stopped at least 800 ms after the end of the sentence. The experiment was partitioned into two blocks, one with a male and one with a female speaker. To avoid confounding effects of measurement condition order and sentence list, the order of conditions and sentence lists was counterbalanced across subjects according to a Latin Square method. In total, each subject received for both speakers 10 lists of 13 sentences.



*Figure 4.1.* SRT (dB) as function of repeated measures for female speech in 8 Hz interrupted noise (triangles) or stationary noise (circles). Error bars denote the standard deviation between subjects. Dashed lines denote the regression fit to the data.

## Learning Effect



*Figure 4.2* SRT (dB) as function of repeated measures for male speech in 8 Hz interrupted noise or stationary noise. Error bars denote the standard deviation between subjects. Dashed lines denote the regression fit to the data.

### III. Results

Figures 4.1 and 4.2 show the SRT-values averaged across subjects as function of repeated measures for the 8 Hz interrupted noise and stationary noise condition. Figure 4.1 and 4.2 display the results obtained with the female and male speaker, respectively. Error bars denote the standard deviation between subjects. A 2[speech corpus]  $\times$  2[noise]  $\times$  5[test/retests]  $\times$  8[subject] Analysis Of Variance (ANOVA) was performed on the data-set. Of the main effects, “speech corpus” was significant ( $F[1,79]=11.14$ ,  $p<0.005$ ), “noise” was significant ( $F[1, 7]=217.1$ ,  $p<0.001$ ), and “test/retest” was significant ( $F[1, 28]=9.05$ ,  $p<0.001$ ). Differences between subjects were not significant ( $F[7, 5.90]=0.93$ ,  $p>0.5$ ). Of the interactions, noise\*test ( $F[4, 28]=5.18$ ,  $p<0.005$ ) and noise \*subjects ( $F[7, 28]=5.59$ ,  $p<0.001$ ) were significant. Regression Analysis was performed on the data-sets.

A straight line was fit through the data of the four subgroups in Figures 4.1 and 4.2.

Post hoc testing showed significant differences in learning effect between stationary noise and interrupted noise. For the SRT in stationary noise with female speech, test number was not significant ( $F[1, 38]=0.11, p>0.7$ ; regression equation:  $SRT = -4.43 - 0.035 * \text{test number}$ ). For the SRT in 8 Hz interrupted noise with female speech, test number was significant ( $F[1, 38]=6.66, p<0.05$ ;  $SRT = -14.01 - 0.773 * \text{test number}$ ): each repetition yields on average a 0.773 dB better SRT.

Similarly, for the SRT in stationary noise with male speech, test number was not significant ( $F[1, 38]=0.26, p>0.6$ ;  $SRT = -4.53 + 0.058 * \text{test number}$ ), whereas for the SRT in 8 Hz interrupted noise with male speech, test number was significant ( $F[1, 38]=05.59, p<0.05$ ;  $SRT = -11.88 - 0.885 * \text{test number}$ ).

#### IV. Discussion

Statistical analysis showed that the SRT measured in interrupted noise improves significantly after retesting, whereas that in stationary noise did not. For both speech materials, the SRT of the fifth SRT test in interrupted noise condition is about 3 (female) to 3.5 (male) dB lower than the SRT of the first test. The repeated measures in stationary noise showed no significant differences after retesting. The latter outcome is in line with Versfeld *et al.* (2000). They repeated eight SRT tests in stationary noise with four different speech materials with a group of twelve normal-hearing listeners. They found no learning effect. The present study shows that there is a substantial learning effect present when testing in interrupted noise. To our knowledge, such a learning effect has not been observed previously. Since there still was an improvement in SRT from the fourth to the fifth re-test, it is not clear when thresholds will stabilize.

The cause of the presence of the learning effect in interrupted noise is still unclear. It is not due to adjustment of the subjects to the experimental procedure, since the effect was not present in stationary noise. Possibly, subjects need time to tune in into the grid of the interrupted noise, such that they become capable to fill in the gaps in the interrupted masked speech signal, much like the continuity effect (Warren, 1970, 1999; Warren *et al.*, 1972;

## *Learning Effect*

Bashford *et al.*, 1988). Another possibility is that subjects require practice to listen into the gaps of the noise. If this were the case, one would expect the learning effect to be dependent on the value of the SRT: Lower SRTs then would be accompanied by larger learning effects. To what degree these largely phenomenological explanations are true is left to future research. In conclusion, future experiments with SRTs in non-stationary noise should contain at least a repeated measure approach and possibly some training to control for learning effects.

## *Chapter 5*

# Validation of the Extended Speech Intelligibility Index for the prediction of the Speech Reception Threshold in fluctuating noise for Normal-hearing Listeners, and suggestions for further improvement

*Koenraad S. Rhebergen, Niek J. Versfeld and Wouter A. Dreschler  
Submitted to Journal of the Acoustical Society of America*

## **Abstract**

The Extended Speech Intelligibility Index (ESII) model proposed by Rhebergen and Versfeld (2005; JASA 117 (4), 2181-2192) forms an extension to the conventional Speech Intelligibility Index model (SII; ANSI S3.5-1997, 1997), and is able to predict for normal-hearing listeners the speech intelligibility in both stationary and non-stationary noise maskers with reasonable accuracy. The ESII model was validated with Speech Reception Threshold (SRT) data from the literature. However, further validation is required and the present paper describes SRT experiments with non-stationary noise conditions that are critical to the extended model. From these data, it can be concluded that the ESII model is able to predict the SRTs for the majority of conditions, but that predictions are better when the ESII model includes a function to account for forward masking.

## I. Introduction

Speech intelligibility decreases due to the presence of a background noise. Parts of the speech signal then are masked by the noise such that not all speech information is available to the listener. French and Steinberg (1947), Fletcher and Galt (1950), and later Kryter (1962) developed a calculation method, known as the Articulation Index (AI), to predict the speech intelligibility under such masking conditions. The AI calculation scheme was re-examined in the Eighties and early Nineties, which led to a new method accepted as the ANSI S3.5-1997 (1997). Since its revision in 1997, the AI is named the Speech Intelligibility Index (SII). A detailed description of the SII can be found in Pavlovic (1987), and the ANSI S3.5-1997 (1997) standard.

To date, the SII model has been designed and validated only for stationary masking noises. In fluctuating masking noises, speech intelligibility is usually much better, since the listener is able to take advantage of the relatively silent periods in the noise masker (Festen and Plomp, 1990; Houtgast *et al.*, 1992; Versfeld and Dreschler, 2002). However, the SII model does not take into account any fluctuation in the masking noise since it uses only the long term speech and noise spectrum. Therefore, it predicts speech intelligibility inaccurately for these conditions. Since many daily-life background noises do fluctuate strongly over time (Koopman *et al.*, 2001), the SII model is unable to predict speech intelligibility in the majority of real-life situations adequately.

Recently, Rhebergen & Versfeld (2005) proposed an extension to the SII model, in order to improve the predictions for speech intelligibility in fluctuating noise. The basic principle of this approach is that both speech and noise signal are partitioned into small time frames. Within each time frame the instantaneous SII is determined, yielding the speech information available to the listener at that time frame. Next, the SII values of these time frames are averaged, resulting in the SII for that particular speech-in-noise condition. With the aid of many data available for a variety of noise types described in the literature, Rhebergen & Versfeld (2005) have shown that their extension allows a good account for most existing data, dealing with the Speech Reception Threshold (SRT) for sentences. However, there still are conditions where the extended SII model (ESII) is unable to give accurate predictions. First, the ESII is unable to predict SRTs for sentences in 100 % Sinusoidally Intensity-Modulated (SIM) speech noise, as measured by Festen (1987). Although the SRT values predicted

## *Validation ESII Model*

by the ESII model yield an improvement over the original SII model, there are still some systematic deviations. Festen found lowest SRTs (i.e., best performance) for modulation frequencies of 16 and 32 Hz, whereas the ESII predicts the best performance for a modulation frequency of 8 Hz.

Second, Rhebergen *et al.*, (2005) measured SRTs with unintelligible interfering speech (foreign language) as a masker played normal and time-reversed. By reversing the unintelligible speech masker in time, the SRT worsened about 2.3 dB. Rhebergen *et al.* (2005) argued that this difference could be attributed to differences in the amount of forward masking: The time-reversed speech masker (having a “ramped”-like envelope, i.e. a gradual increase with a sudden offset) provokes more forward masking than a normal speech masker (being more ‘damped’-like, i.e., a sharp onset followed by a gradual declination). A time-asymmetrical non-speech like noise masker may provide more insight into the effects of temporal forward masking on speech intelligibility. The ESII model is, in essence, a time-symmetrical model. It does not account for the differences in forward and backward masking. The model predicts the same speech intelligibility with a noise masker played normal and time reversed.

Third, the ESII is a model verified with SRT data described in the literature. To enable a fair comparison between the data obtained in different studies, Rhebergen and Versfeld (2005) restricted themselves to the use of SRT data obtained with one set of speech materials, viz., the Dutch speech corpus of Plomp and Mimpen (1979). Even though the corpus is similar, differences between studies sometimes are substantial: some conditions have been measured abundantly (SRT for stationary speech shaped noise), whereas other conditions have been measured sparsely (SIM, Festen, 1987).

Moreover, SRT data have been collected by different researchers in different experimental settings. This introduces additional variance in the data. For example, SRTs in quiet or interrupted noise differ largely between different papers (de Laat and Plomp, 1983; Festen, 1987; Noordhoek, 2000; Duquesnoy, 1983; Plomp, 1978; Plomp and Mimpen, 1979). For a good validation of the ESII model, only SRTs obtained from the same group of subjects and measured in the same experimental conditions should be used for ESII calculations. For instance, the ESII predicts that SRTs in fluctuating noise, unlike stationary

noise, may depend on the subjects absolute threshold, even in normal-hearing listeners.

This paper addresses the problems described above, with the aim to test and, where necessary, refine the ESII. All experiments have been conducted with normal-hearing subjects.

In the first section, SRT tests are performed for nineteen different noise conditions (test and re-test) using the speech material of Versfeld *et al.* (2000). The noise conditions comprise steady state noise, interrupted noise with different modulation frequencies and different duty cycles, SIM noise with different modulation frequencies, and two asymmetrically modulated sawtooth noises. The noise conditions have been selected such to test the ESII critically.

In the next section, the observed SRTs are used to evaluate and refine the ESII model. Notably, the ESII calculations are extended further by using a temporal integration window. Lastly, predictions and limitations of the finally obtained extended SII model will be discussed.

## **II. Experiment**

In this experiment, the SRTs for a number of noise conditions are measured. The results will be used to validate the SII model.

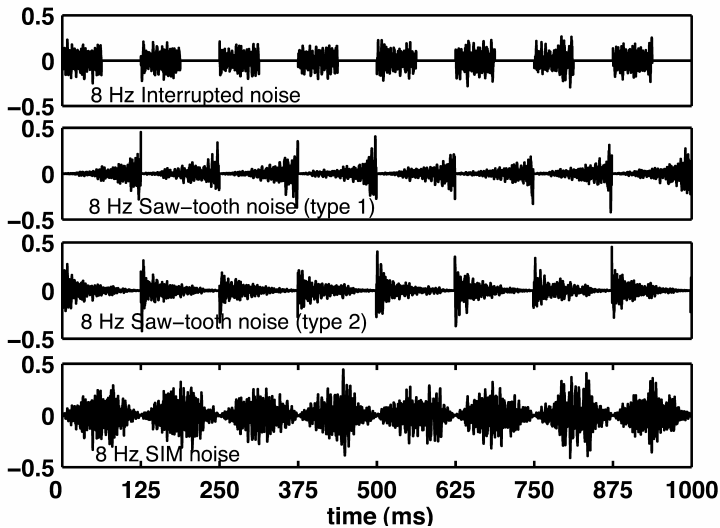
### **Method**

#### **a. Subjects**

Twelve normal-hearing subjects (1 male, 11 female) participated. Their age ranged from 18 to 29 years and was on average 21.5 years. Subjects were native speakers of the Dutch language and had at least high school education. Each subject had pure-tone thresholds of 15 dB HL or better at octave frequencies from 125 to 8000 Hz (ANSI S3.6, 1996).

**b. Stimuli**

The target speech material consisted of short every-day sentences, uttered by a female speaker (Versfeld *et al.*, 2000). The speech material comprises 39 lists of 13 sentences and has been developed for a reliable measurement of the speech intelligibility in noise. The speech was stored at a sample rate of 44.1 kHz and a 16 bits resolution.



*Figure 5.1. Illustration of some masking noises used in the present experiment. In this selection all signals have a modulation frequency of 8 Hz and a spectrum equal to the long-term average spectrum of the female target speech material. The upper panel shows interrupted noise with a duty cycle of 50 %; the second panel saw-tooth noise (Type 1), the third panel saw-tooth noise (Type 2, time reversed version of Type 1)), and at the lower panel a SIM noise.*

**Table 5.1.** Schematic representation of the nineteen noise conditions

Noise condition	Noise type	Modulation frequency	Modulation depth	Duty cycle	Envelope shape
int 4Hz	Interrupted	4	100 %	50 %	square
int 8Hz	Interrupted	8	100 %	50 %	square
int 16Hz	Interrupted	16	100 %	50 %	square
int 32Hz	Interrupted	32	100 %	50 %	square
int 64Hz	Interrupted	64	100 %	50 %	square
int 128Hz	Interrupted	128	100 %	50 %	square
int 8Hz dc40 %	Interrupted	8	100 %	40 %	square
int 8Hz dc45 %	Interrupted	8	100 %	45 %	square
int 8Hz dc55 %	Interrupted	8	100 %	55 %	square
int 8Hz dc60 %	Interrupted	8	100 %	60 %	square
saw-tooth T1	saw-tooth	8	-	-	exponential increasing
saw-tooth T2	saw-tooth	8	-	-	exponential decreasing
SIM 4Hz	SIM	4	100 %	-	sinusoidal
SIM 8Hz	SIM	8	100 %	-	sinusoidal
SIM 16Hz	SIM	16	100 %	-	sinusoidal
SIM 32Hz	SIM	32	100 %	-	sinusoidal
SIM 64Hz	SIM	64	100 %	-	sinusoidal
SIM 128Hz	SIM	128	100 %	-	sinusoidal
steady state	steady state	-	-	-	flat

All 19 interfering noise conditions are given in Table 5.1. The noise conditions comprise one condition with steady state noise, ten conditions with interrupted noise, two conditions with saw-tooth noise, and six conditions with SIM (sinusoidal intensity modulated) noise. Figure 5.1 illustrates the waveforms of four of the noise types. All noise conditions had a long-term average spectrum equal to the long-term average spectrum of the target female speech material. The interrupted noise conditions were modulated with a duty cycle of 50 % and a depth of 100 %, and the modulation frequencies were 4, 8, 16, 32, 64, and 128 Hz. Four conditions had a modulation frequency of 8 Hz, but with a duty cycle of 40, 45, 55, and 60 %. The SIM noises were generated according to Festen (1987). The modulation frequencies were 4, 8, 16, 32, 64, and 128 Hz, and the modulation depth was 100 %. The two saw-tooth noise conditions had a

modulation frequency of 8 Hz and the envelope was exponentially increasing or decreasing in time (Type 1 and Type 2, respectively).

**c. Procedure**

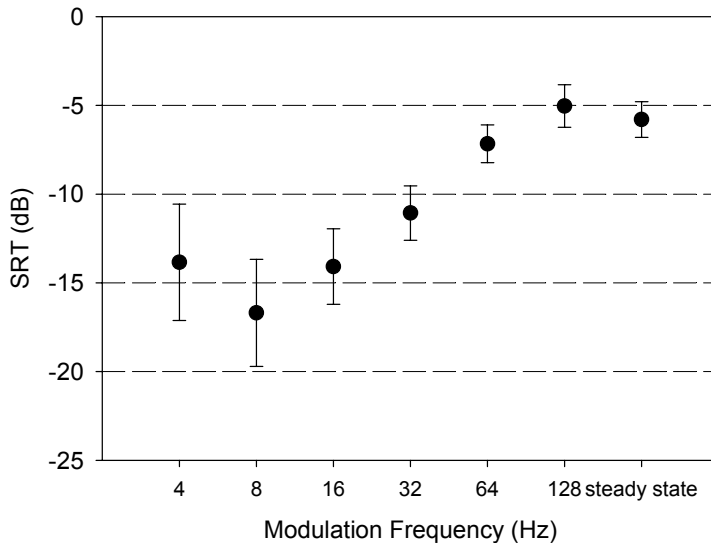
Subjects were tested individually in a sound-insulated booth. Signals were played out via an Echo soundcard (Gina 24/96) on a PC at a sample frequency of 44.1 kHz, and were fed through a TDT Amplifier (MA2) and a TDT Headphone Buffer (PA4). Subjects received the signals monaurally at their best ear via TDH 39P headphones at a fixed noise level of 65 dBA. After the presentation of a sentence, the subject's task was to repeat the sentence he or she had just been presented. A sentence was scored correct if all words in that sentence were repeated without any error. A list of 13 sentences, unknown to the subject, was used to estimate the signal-to-noise ratio (SNR) at which 50 % of the sentences was reproduced without any error, the so-called Speech Reception Threshold, or SRT. For a given condition, the first sentence of the list started far below the expected SRT. The sentence was repeated each time at a 4 dB higher level until the subject was able to reproduce it correctly. The twelve other sentences of that list were presented only once, following a simple up-down procedure with a step size of 2 dB. The SRT was estimated according to the procedure described by Plomp and Mimpen (1979), i.e., by taking the mean SNR of sentence five to thirteen plus the SNR that would have been used for the fourteenth sentence. With each sentence presentation, a random sample of the interfering noise was taken. The noise started 1200 ms before the beginning of the sentence and stopped at least 800 ms after the end of the sentence.

In total, nineteen conditions were tested. The experiment was partitioned into two blocks, a test and a retest block. To avoid confounding of measurement condition order and sentence lists, the order of conditions and sentence lists was counterbalanced across subjects according to a Latin Square method. In total, each subject received 38 lists of 13 sentences preceded by three practice lists.

### III. Results

The results are plotted in Figures 5.2, 5.3, and 5.4. A 19[condition] × 2[test/retest] × 12[subject] Analysis Of Variance (ANOVA) was performed on the data-set. Of the main effects, “condition” was significant ( $F[18,198]=159.08$ ,  $p<0.001$ ), and “test/retest” was significant ( $F[1,11]=14.87$ ,  $p<0.005$ ). The SRT of the retest was on average 0.8 dB better. Differences between subjects were not significant ( $F[11,15.71]=2.14$ ,  $p>0.05$ ). Of the interactions, conditions\*test ( $F[18,198]=1.88$ ,  $p<0.05$ ) and conditions\*subjects ( $F[198,198]=1.36$ ,  $p<0.05$ ) were weakly significant. Below, additional analyses were performed on subsets of the data.

#### A. Interrupted noise

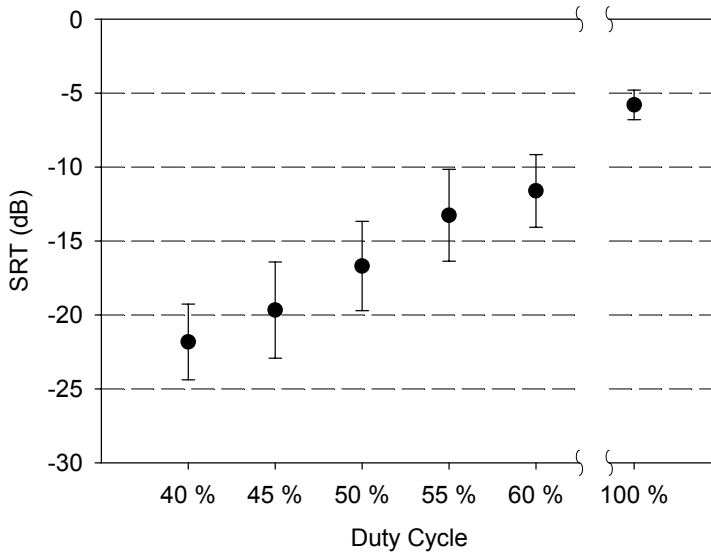


*Figure 5.2. Speech Reception Threshold (dB) as function of modulation frequency (Hz) for the steady state noise and the 4 to 128 Hz modulated interrupted noise conditions. Error bars denote the standard deviations between subjects.*

## Validation ESII Model

Figure 5.2 shows the SRT-values (dB) averaged across subjects and test-retest as function of modulation frequency (Hz) for interrupted noise with a duty cycle of 50 %. Error bars denote the standard deviations between subjects. A 7[condition] x 2[test/retest] x 12[subject] ANOVA showed that the main effect of "condition" was significant ( $F[6, 66]=119.55, p<0.001$ ). Also, the main effect "test/retest" was significant ( $F[1,11]=36.21, p<0.005$ ). The SRT of the retest was on average 0.9 dB better, in agreement with the findings of Rhebergen *et al.* (2006). Differences between subjects were not significant ( $F[11,10.87]=2.51, p>0.05$ ). There were no significant interactions. The SRT is lowest with a modulation frequency of 8 Hz (-16,7 dB). This SRT is comparable, but somewhat higher than that obtained by de Laat and Plomp (1983), who found an SRT of -23 dB at a modulation frequency of 10 Hz. The SRT with a 4 Hz interrupted noise is somewhat higher compared to that with an 8 Hz interrupted noise. This may be accounted for by the fact that at these slow modulation rates noise bursts can mask complete words of a sentence. The trend in these data is consistent with the results of Miller and Licklider (1950), Licklider and Guttman (1957), Gustafsson and Arlinger (1994), Trine (1995), Dubno *et al.* (2002, 2003), and Nelson *et al.* (2003).

Figure 5.3 shows the SRT values averaged across subjects and test-retest as function of duty cycle (%) for interrupted noise with a modulation frequency of 8 Hz. Error bars denote the standard deviations between subjects. The duty cycles were 40, 45, 50, 55, 60, and 100 % (steady state noise). A 6[condition] x 2[test/retest] x 12[subject] ANOVA was performed on these data. Of the main effects, differences in "condition" were significant ( $F[5,55]=124.45, p<0.001$ ). Also, the SRT of the retest was on average 1.3 dB better than the test, which was a significant effect ( $F[1,11]=13.39, p<0.05$ ). Differences between subjects were not significant ( $F[11,10.94]=2.61, p>0.05$ ). There were no significant interactions. The trend of these data show a gradual and almost linear increase in SRT from -21.8 dB up to -11.6 dB as the duty cycle increases from 40 % up to 60 %.



*Figure 5.3. Speech Reception Threshold (dB) as function of duty cycle (%) for interrupted noise with a modulation frequency of 8 Hz. Error bars denote the standard deviations between subjects.*

## B. Saw-tooth noise

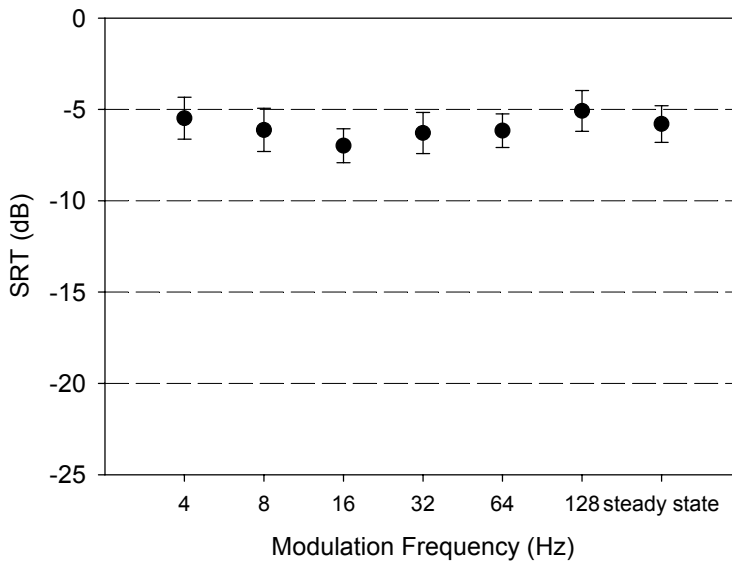
The mean SRT scores of saw-tooth Type 1 and saw-tooth Type 2 were  $-9.0$  dB and  $-12.2$  dB, respectively. A 2[condition]  $\times$  2[test/retest]  $\times$  12[subject] ANOVA was performed on the data with the saw-tooth noise. Of the main effects, differences in “condition” were significant ( $F[1,11]=101.27$ ,  $p<0.001$ ). The SRT of the retest was on average  $0.1$  dB better than the test, which was not significant ( $F[1,11]=0.0001$ ,  $p>0.05$ ). Differences between subjects were not significant ( $F[11,4.8]=1.4$ ,  $p>0.05$ ). There were no significant interactions.

## C. SIM noise

Figure 5.4 shows the SRT-values averaged across subjects and test-retest for each of the six conditions for the interfering SIM noises and for the steady state noise. Error bars denote the standard deviations between subjects. A

### Validation ESII Model

7[condition] x 2[test/retest] x 12[subject] ANOVA showed that of the main effects, only differences in “condition” were significant ( $F[6,66]=10.24$ ,  $p<0.001$ ). The SRT of the retest was on average 0.1 dB better than the test, this difference was not significant ( $F[1,11]=0.36$ ,  $p>0.55$ ). Differences between subjects were not significant ( $F[11,7.56]=2.46$ ,  $p>0.11$ ). There were no significant interactions. Tukey’s HSD tests showed that the 8, 16, 32, and 64 Hz conditions were not significantly different from each other. The trends in the present data are not entirely consistent with the results of Festen (1987). He observed best SRTs for the 32 Hz condition, whereas in this study best SRTs were observed for the 16 Hz condition. Furthermore, the observed SRTs of Festen (1987) were 2 to 3 dB lower than the SRTs observed in this experiment.



**Figure 5.4.** Speech Reception Threshold (dB) as function of modulation frequency (Hz) of the SIM Noise Condition (4 – 128 Hz & steady state noise). Error bars denote the standard deviations between subjects.

## IV. Discussion

The results of the present experiment show some interesting aspects.

First of all, the SRTs of the retest were on average 0.8 dB lower than the SRTs of the first test. After separating the SRT data into subgroups, it was clear that this effect was only present in the interrupted noise conditions. This outcome is in line with Rhebergen, Versfeld, and Dreschler (2006a), who found a clear learning effect for the SRT in 8 Hz interrupted noise but not in steady state noise. Rhebergen *et al.* (2006a) hypothesized that listening into the gaps of the interrupted noise requires practice. Hence, they argued, the learning effect should be more prominent when the gaps are deeper. In other words: Lower (i.e., better) SRTs are accompanied by larger test-retest differences. Figure 5.5 displays the average test-retest difference as a function of the mean SRT for all conditions in the present study. The data in Figure 5.5 form two subgroups: One subgroup is formed by those conditions where the test-retest difference does not exceed 1 dB SRT, and the mean SRT is worse than about -12 dB. The other subgroup is formed by the conditions with relatively good SRTs of -13 dB or better. Here the test-retest differences are larger than 1.5 dB. The latter group consists of conditions with interrupted noise. It seems that the test-retest difference is related to the gap length and not particularly related to a specific modulation frequency. The lower boundary of the gap length where significant learning effects may occur then is in the order of 50 ms.



Dreschler (2005), who described an SRT test with intelligible and unintelligible interfering speech played normal and time-reversed. With Dutch listeners, (unintelligible) Swedish interfering speech gave a rise in SRT of 2.3 dB when played in reverse.

Third, between-subject differences appear to be larger with lower SRT values. Rhebergen and Versfeld (2005) showed with their extended SII method that the psychometric function (i.e., SII as a function of SNR) near SII=0.3 is relatively shallow for speech in interrupted noise compared to that for speech in stationary noise. Consequently, a given variation in SII corresponds to a large variation in the SNR with interrupted noise, but to a smaller variation with stationary noise. In the next section (section V), SII calculations will show whether this explanation can fully account for the differences in variance between these conditions.

Fourth, the large difference in SRT (approximately 6 dB) obtained with interrupted noise between the present results and those of de Laat and Plomp (1983) might be explained by differences in absolute threshold, since de Laat and Plomp (1983) found a high correlation between the SRT and the pure-tone average (PTA, averaged across 500, 1000 and 2000 Hz) in their group of subjects. However, the PTA of the present group of subjects is 7.6 dB better (*viz.*, 2.0 dB HL with a standard deviation of 3.6 dB) compared to the average PTA of 9.6 dB HL (with a standard deviation of 3.6 dB) from the subjects of de Laat and Plomp (1983). Moreover, with the present data, a Pearson correlation coefficient was calculated between the individual PTAs and the noise conditions. Correlations were non-significant ( $r=0.035$ ,  $p>0.05$ ). The differences in observed SRTs must be due to other factors, such as perhaps the use of different speech corpuses (Plomp & Mimpen, 1979, versus Versfeld *et al.*, 2000). Although Versfeld *et al.* (2000) found no significant differences between these sets when comparing them in stationary masking noise, Van Wijngaarden (2003) did, be it in different listening conditions (such as reverberation). It is known that differences in intelligibility between different speech materials become more apparent under increasingly adverse listening situations (Mullennix *et al.*, 1989). But what properties of the speech signal cause these differences is yet unclear. A similar explanation holds for the differences in SRT with SIM noises obtained by Festen (1987) and the present data.

## Validation ESII Model

Next, the increase in SRT with increasing modulation frequency of the interrupted noise (from 8 to 128 Hz) is due to an increase in forward masking. In all conditions, the masker is absent in 50 % of the time, but due to a decrease in “gap” duration (62.5ms down to 3.9 ms), the forward masking induced by each masker pulse becomes more effective. The SRT with a 128 Hz interrupted noise is even worse than with a steady state noise. In this condition, the gaps are very short in duration, such that they are probably entirely masked. At the same time, the masker pulses must be 3 dB higher in level, in order to have the same long term RMS level as stationary noise. Thus, 128 Hz interrupted noise is a more effective masker than is stationary noise.

The increase in SRT with interrupted noise when changing the modulation frequency from 8 Hz to 4 Hz may be accounted for by the fact that with these slow modulation rates, masking of complete words in a sentence can occur. This phenomenon has also been observed by Miller and Licklider (1950), and Nelson *et al.* (2003), who found optimal performance around modulation rates of 10 and 8 Hz, respectively. Because in the SRT procedure every word of the sentence needs to be repeated correctly, it is unsuitable for these low modulation frequencies (less than 8 Hz).

Finally, a slight change in the duty cycle results in a large change in SRT, cf. Figure 5.3. Considering the entire range, a decrease in duty cycle (more pulsed signals with longer gaps) will probably result in increasingly lower thresholds, with the SRT in quiet (absence of a noise masker) as a lower limit. The SRT in quiet is on average about 20 dBA (Duquesnoy, 1983; Duquesnoy and Plomp, 1983; Plomp and Mimpen, 1979b, Noordhoek, 2000), which is, compared to a 65 dBA noise level, equal to an SRT of -45 dB. Additional experiments show that the decrease holds at least to a duty cycle of 5 %, resulting in an SRT of -36 dB. Over a range in the duty cycle of 5 % to 60 %, the increase in SRT is about 0.5 dB/%. An increase in duty cycle (shorter gaps) eventually will result in the SRT in stationary noise, being -5.8 dB. In the region from 60 % to 100 % the increase in SRT as a function of duty cycle thus has to level off. If a linear fit is made through the data in Figure 5.3, the break point is estimated to be near a duty cycle of 70 %, i.e., gaps of about 40ms in duration. One could interpret this as

the point where in successive pulses, the forward masking function is still active when the next pulse occurs.

## V. SII Model predictions

A detailed description of the conventional SII model is given in ANSI S3.5-1997 (1997), and a detailed description of the Extended SII (ESII) model is given in Rhebergen and Versfeld (2005). The basic principle of the conventional SII is that departing from the long term speech spectrum, the long term noise spectrum, and the absolute threshold of hearing, the amount of speech information that exceeds both noise and threshold is calculated. The extension proposed by Rhebergen and Versfeld (2005) is that the calculations are not performed with the long term spectra, but rather that both speech and noise signal are partitioned into small time frames. Within each time frame, the (instantaneous) conventional SII is calculated, representing the speech information available to the listener at that time frame. The ESII for the condition under investigation is obtained by averaging the instantaneous SII across time. With the ESII, the length of the time frames is frequency dependent, and time constants are adapted from gap detection data (Moore, 1997). The length of a time frame ranges from approximately 35 ms in the lowest frequency band up to 9.4 ms in the highest frequency band. The present paper uses the SPIN 21 critical band weighting function (ANSI S3.5-1997, 1997, Table B.1). Furthermore, all SII calculations are conducted with the long term speech spectrum of the female target speaker. In order to approach the sound level at the ear drum (as required by the SII model), all signals are filtered with a fifth order FIR filter with the transfer characteristics of the TDH39P headphone that was used in the experiment. Also, if required, the background noise present in the sound proof booth was added to the noise signal. This seems unnecessary, but the SII model defines silence in each band as  $-50$  dB SPL, whereas in a sound proof booth more realistic numbers are between 0 and 10 dB SPL in the mid and high frequencies and 35 to 50 dB SPL in the frequencies below 100 Hz. These levels have almost no effect with most noise types, except for conditions where noises contain relatively long silent periods, such as is the case with interrupted noise.

Since the purpose of the present paper is to evaluate the model, learning effects were eliminated as much as possible by omitting the first test and considering only the average values of both re-tests. Furthermore, the noise conditions with modulation frequencies below 8 Hz were excluded from the SII calculations. As mentioned earlier, noise conditions with these low modulation frequencies give a rise in SRT due to the masking of complete words.

### **A. Conventional SII calculations**

The fourth column of Table 5.2 shows the results of the calculations with the conventional SII model (ANSI S3.5-1997). By definition, at threshold one would expect the SII to be similar across conditions, since threshold is reached by the availability of a given fixed amount of speech information. Table 5.2 shows that for the conventional SII this certainly is not true, which makes the conventional SII model a poor predictor for speech intelligibility in fluctuating noise. The SII has a mean of 0.17 and a large standard deviation between conditions of 0.13.

### **B. Extended SII calculations**

The fifth column of Table 5.2 shows the Extended SII calculations (Rhebergen & Versfeld, 2005). Here, the mean SII value is 0.35 and its standard deviation is equal to 0.05. This result is far better than the ANSI S3.5-1997 method in the previous section. However, since the ESII is a time-symmetric model, it is unable to account for the difference in threshold between the two conditions with the saw-tooth masker. The ESII always will predict identical thresholds for noises that are each others time reversal.

The present calculations clearly show a second shortcoming of the ESII: Especially for those conditions where the masking noise contains relatively long silent intervals, such as the interrupted noise with a duty cycle of 40 % or 45 %, or the noises with a relatively low modulation frequency, the ESII predictions are relatively high. This indicates that the masking function used in the model underestimates the real masking in these conditions. Simple adaptation of the integration time cannot solve the problem, since this results in deviations for the ESII for the other conditions.

## Chapter 5

In order to overcome these two shortcomings of the model, in the next section a forward masking function is introduced.

*Table 5.2. For each condition of the present experiment, the average SRT values of both re-tests and the results of the various SII calculation schemes are given. Lower rows yield the mean and standard deviation of the SII. FMF denotes the use of the Forward Masking Function, asym.w. denotes the use of a asymmetrical integration window, and Lin.w. denotes the use of a fixed integration window of 4 ms.*

Noise condition	SRT	std	SII	Extended SII	Extended SII & FMF (asym. w.)	Extended SII & FMF (Lin.w.)
int 8Hz	-17.6	3.3	0.000	0.388	0.357	0.3589
Int 16Hz	-15.0	1.8	0.031	0.333	0.2908	0.2993
Int 32Hz	-11.1	1.4	0.012	0.271	0.2604	0.2758
Int 64Hz	-7.5	1.3	0.239	0.309	0.2993	0.2924
int 128Hz	-5.4	1.1	0.297	0.339	0.2993	0.3049
int 8Hz dc40	-22.8	2.3	0.000	0.465	0.3848	0.3938
int 8Hz dc45	-20.8	2.6	0.000	0.432	0.3669	0.3762
int 8Hz dc55	-14.4	2.0	0.044	0.369	0.3544	0.3571
int 8Hz dc60	-11.7	2.0	0.117	0.365	0.3611	0.3642
sawtooth T1	-9.3	1.5	0.178	0.389	0.3687	0.3759
sawtooth T2	-11.9	1.6	0.102	0.310	0.3298	0.3351
SIM 8Hz	-5.8	1.3	0.303	0.381	0.3837	0.3848
SIM 16Hz	-7.1	0.9	0.260	0.322	0.331	0.3347
SIM 32Hz	-6.5	1.3	0.308	0.318	0.3287	0.3347
SIM 64Hz	-6.2	0.9	0.289	0.314	0.3145	0.3226
SIM 128Hz	-5.3	1.0	0.319	0.337	0.3212	0.3259
steady state	-5.5	0.9	0.314	0.316	0.316	0.317

<b>mean SII</b>	0.17	0.35	0.33	0.34
<b>Std SII</b>	0.13	0.05	0.04	0.04

### C. Implementation of forward masking in the Extended SII

A large number of studies have shown the masking of a target signal by a preceding masker (so-called Forward Masking) and its relationship between masker level and the time interval between masker and target signal (Pollack, 1955; Plomp, 1964; Elliot, 1969; Duifhuis, 1973; Widin and Viemeister, 1979; Jesteadt *et al.*, 1982; Moore and Glasberg, 1983; Kidd and Feth, 1982). These studies show a decrease in masking threshold with increased masker-signal delay. The masking threshold returns in about 200 ms to the level of the unmasked target signal threshold (i.e., when no masker is present), regardless of masker level. When plotting the masking thresholds (dB) as a function of masker-target gap on a logarithmic time scale, a linear relationship exists (e.g., Plomp, 1964). Ludvigsen (1985) has modeled this function, which is called in this paper as the forward-masking function (FMF). The model parameters were determined from the results of other studies reported in the literature, and the model predicts forward masking very well, not only for the normal-hearing, but also for the hearing-impaired. Figure 5.6 displays the FMF, i.e., the masked threshold as a function of time (reprinted from Ludvigsen, 1985) for two masker levels (high-level masker and low-level masker) and for a normal-hearing and a hearing-impaired subject. As can be seen in Figure 5.6, the FMF is linear between  $T_0$  and  $T_f$  when time is plotted on a logarithmic axis. Also, the duration of the FMF is always equal, regardless of hearing loss or masker level.

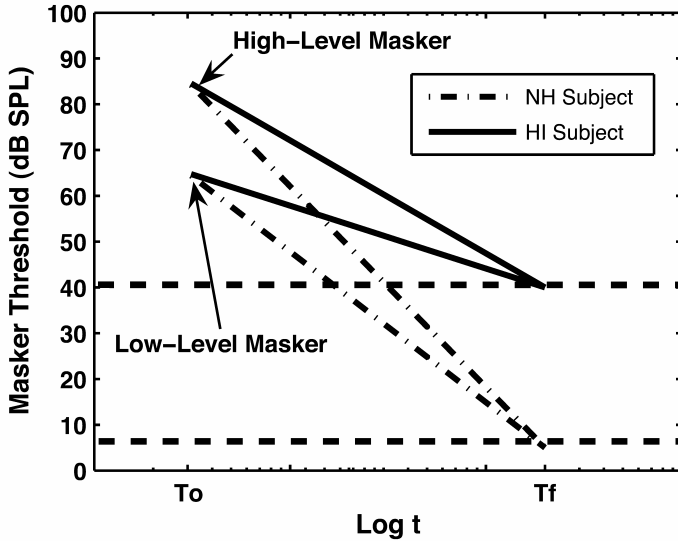
The middle portion of the FMF is a simple linear relationship, and is given by

$$E_{FMF}(t) = E(T_0) - \frac{\log(t / T_0)}{\log(T_f / T_0)} * [E(T_0) - E(T_f)],$$

where at time  $T_0$ , the level of the linear portion  $E(T_0)$  is equal to the level of the envelope of the masker, and where at time  $T_f$ , the linear function intersects with the absolute threshold of hearing  $E(T_f)$ . The values of the parameters  $T_0$  and  $T_f$  are 2 ms and 200 ms, respectively (From Ludvigsen, 1985). To take forward masking into account, the original envelope  $E(t)$  is modified according to:

$E(t) = \max(E(t), E_{FMF}(t))$ . In this manner, sharp onsets in the envelope are followed instantaneously, but sharp offsets make that  $E_{FMF}(t)$  takes over,

resulting in a gradual decline of the envelope (due to forward masking). The FMF does not account for the phenomenon of backward masking, where a soft signal is masked by a louder signal that follows it. Backward masking is still poorly understood (Moore, 2003), and its effect on speech intelligibility is still unclear and probably not very large.



*Figure 5.6. Masked threshold (dB SPL) plotted as a function of time (on a logarithmic axis) for two masker levels (high-level masker and low-level masker) and for a normal-hearing and a hearing-impaired subject. Horizontal dashed lines indicate the absolute threshold of that subject. Figure 5.6 is redrawn from Ludvigsen (1985), figure 4, page 1277.*

Note that the FMF is similar in the low and high frequency bands, whereas temporal integration is not. Both phenomena act separately on the signal, and probably are situated in different places in the auditory system. So far, forward masking was not modeled separately in the ESII, but rather it was taken together with temporal integration. Effectively this gave longer integration times, which can explain that a best fit to the data of Rhebergen and Versfeld

### *Validation ESII Model*

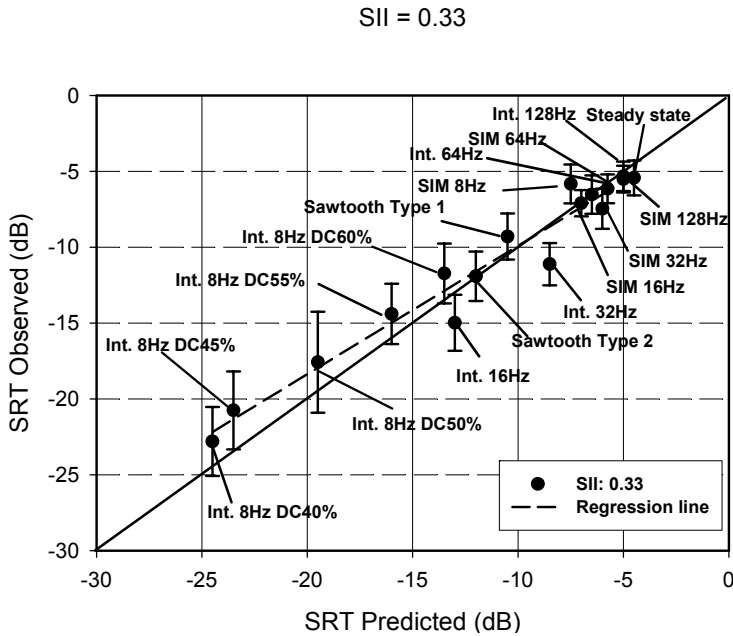
(2005) was obtained by multiplication of the time constants from Moore (1997) by a factor of 2.5. However, when forward masking is modeled separately, the original integration times (the original frequency dependent gap detection lengths) should be used.

The sixth column of Table 5.2 shows the ESII calculations with the FMF included. This addition of forward masking results in better predictions, i.e., SII values are closer together as can be deduced from the lower standard deviation (0.04). Indeed, the SII for those conditions with relatively long silent periods has decreased. In addition, the SII values for the two Saw-tooth conditions are closer together.

When the calculation scheme is simplified by taking all integration times equal to 4ms, the standard deviation in SII for all noise conditions remains the same (0.04), see the seventh column of Table 5.2. This is because the time constants of the FMF are an order of magnitude larger than those of gap detection.

Figure 5.7 displays for all conditions described in the previous section the relationship between the observed SRT and the SRT as predicted by the ESII model with the FMF. Predictions were made under similar assumptions as described above, and the SII was taken equal to 0.33, the average value of the SII in Table 5.2, column 6. If the data of figure 5.7 are considered in detail, some predicted SRTs lie above the diagonal. Since these conditions appear to be the conditions in which a learning effect was observed (see Figure 5.5), it is expected that after all learning effects are overcome, SRTs will become better and hence lie closer to the diagonal. In fact, these SRTs can be predicted better with a threshold SII of 0.35. On the other hand, some conditions are below the diagonal. This implies that subjects perform better than predicted with the SII model.

Several conditions, especially those with low SRTs, show large between-subject differences. As argued above, these large differences are due to the shallowness of the psychometric function (SII as a function of SNR). When converting the observed individual SRTs to SII values, differences between subjects are comparable for all conditions. This indicates that no other factors other than differences in the steepness of the psychometric function play a role.



*Figure 5.7. For all conditions, the observed SRT (dB) is plotted as a function of the predicted SRT (dB), where the prediction has been made with the ESII model with the FMF included. Error bars denote the standard deviation between subjects.*

## VI. Further extensions to the SII model

### A. Hearing-impaired subjects

The addition of a forward masking function has increased the predictive power of the ESII model, at least for normal-hearing subjects. So far, no attempt has been made to use the model with data obtained with hearing-impaired subjects. The results of Ludvigsen (1985) do suspect that the ESII can easily be used for these data. However, an alternative approach is proposed by Oxenham (1995). He states that the FMF is comparable for normal-hearing and hearing-impaired subjects, but that differences arise due to differences in

compressive power of the cochlea. A decrease in compressive power due to increased hearing loss combined with a fixed FMF also results in a growth of forward masking (Oxenham and Bacon, 2003, 2004; Bacon and Oxenham, 2004). Glasberg & Moore (2000) and Wojtczak *et al* (2001) have successfully used this compression function to model their forward masking data measured with normal-hearing and hearing-impaired subjects. Whether the FMF of Ludvigsen (1985) is sufficient for hearing-impaired data, whether the model of Oxenham needs to be implemented, or whether both models are indistinguishable is a question that can only be answered by further research.

## **B. Presentation level**

All experiments have been conducted at a constant level, *viz.*, 65 dBA. It is known that with the increase in level, excitation patterns broaden, hence spectral resolution decreases. The SII model cannot account for the broadening of the auditory filters, nor can it directly account for broadened auditory filters of the impaired auditory system, since filter bandwidths are fixed. If required, extension of the model can be realized by the implementation of a more realistic level dependent auditory filterbank.

## **C. Speaker style**

The SII can account for different speech corpuses and speech sets (speech materials) by means of different weighting functions (i.e., band importance functions). In the present paper, the outcomes of SRTs in interrupted noise or SIM noise obtained with the materials of Versfeld *et al.* (2000) show marked differences with those obtained with the materials of Plomp and Mimpen (1979). Note that these differences are not evident for SRTs in stationary noise. These differences cannot be attributed to differences in thresholds between the subject groups. Differences in speaker rate, articulation, pitch and speaking style (clear versus conversational speech) might have an effect on speech intelligibility. Boothroyd (1990, 2000) and Studebaker & Sherbecoe (2002) proposed an Intensity Importance Function (IIF) for the SII calculation scheme. The IIF indicates to what degree soft, middle loud, and loud portions of the speech signal contribute to intelligibility. The conventional SII model uses a simple linear function, going from -15 dB to +15 dB. It might be that the IIF can

account for differences in speaker style and intelligibility. The degree to which this is true, and the degree to which the IIF is related to the cumulative level distribution of the speech material is left to future research.

## **VII. Limitations of the SII model**

### **A. SRT paradigm and the modulation frequency of the masking noise**

As mentioned above, the SRT paradigm is not particularly well suited for sentences in modulated noise where the modulation frequency is low and masking of whole words may occur. It is possible that SRTs are measured that do not reflect the actual masking situation, since the SRT procedure yields a correct score only when the whole sentence is scored correctly. With the present speech materials, the speech rate is about 4 to 5 syllables per second, hence the duration of a syllable or monosyllabic word is 200 to 250 ms. An 8-Hz interrupted noise contains gaps of 62.5 ms; a 4-Hz interrupted noise contains gaps of 125ms. Apparently, such gap durations already cause a deterioration in intelligibility. Thus, there is a lower limit in the modulation frequency to which the SRT procedure is valid. Rather than taking the modulation spectrum of a noise masker as a starting point to determine whether or not the SRT paradigm can be used, a better approach is to depart from the spectrum of the instantaneous SII: If the modulations in the SII are too low, then there are periods in the speech-in-noise signal where no information is present. For example, with 4-Hz interrupted noise the SII is zero for durations of 125ms. On the other hand, the noise periods in checkerboard-noise (i.e., interrupted noise where in alternating frequency bands the noise is present, while at the same time the noise in other the bands is absent) also are 125 ms in duration. However, the instantaneous SII never is equal to zero. Thus, although there is a strong low frequency component in the modulation spectrum of the noise, the SRT paradigm is still valid because the SII is never equal to zero. Measurements with these specific conditions such as checkerboard noise must be conducted to investigate to what extent this assertion is true.

## B. Effect of informational masking

In cases where one or more interfering talkers are present, the obtained SRT thresholds are worse than predicted by the ESII model (Rhebergen and Versfeld, 2005). It is argued that the masking of speech by speech consists of two parts, *viz.* energetic masking and informational masking (Bronkhorst, 2000; Brungart, 2001; Brungart et al., 2001). Real interfering speech as a masker accounts for an additional rise in SRT. The intelligibility of the interfering speech or differences between target and masker with respect to gender have only a small effect with respect to actual energetic masking of the target speech, but can have a large effect on intelligibility (e.g., Festen and Plomp, 1990; Bronkhorst and Plomp, 1992; Bronkhorst, 2000; Drullman and Bronkhorst, 2000; Brungart, 2001b; Brungart *et al.*, 2001, 2002; Summers and Molis, 2004). The ESII (Rhebergen & Versfeld, 2005) cannot account for informational masking. SRT predictions for speech in the presence of one or more interfering talkers are likely to be unreliable. The informational masking component is largely dependent on the interfering speaker (e.g., gender, articulation, intensity, presents, subject of talk, language, etc.) and likely to be variable between listener to listener as well. The additional informational masking can amount up to 7 dB (Rhebergen *et al.*, 2005).

## VIII. Speech intelligibility and hearing loss.

The SII model in its present form (ANSI S3.5-1997) is able to predict the speech intelligibility in stationary noise for listeners with normal-hearing or with a mild hearing loss (Pavlovic, 1987; Noordhoek, 2000). However, it is not meant to predict the speech intelligibility for listeners with moderate to severe hearing loss (Rankovic, 1998; Noordhoek, 2000). The SII model overestimates the performance of subjects with a pure tone average (PTA) of about 25 dB (HL) and higher (Noordhoek, 2000). Often, hearing-impaired subjects require a better-than-normal speech-to-noise ratio, resulting in a higher calculated SII value. SII values higher than 0.33 indicate that on top of audibility loss that can be compensated by amplification, a supra-threshold deficit is present (Noordhoek, 2000). Indeed, Studebaker & Sherbecoe (2002) note that it is not sufficient to base various aspects of human performance and the fundamental

## *Chapter 5*

nature of speech itself on audibility calculations alone. If some assumptions of the SII model are incorrect or imperfect (e.g., dynamic range of speech; linear intensity importance function), one cannot expect an exact calculation of the audibility for normal-hearing or hearing-impaired listeners in a given condition. However, the ESII model has the potential to adapt the forward masking function to incorporate a compression function to account for the non-linearity in speech perception in fluctuating noise conditions. Future research on speech intelligibility in fluctuating noise conditions for both groups of listeners will provide more insight how to model speech intelligibility in noise.

*Validation ESII Model*

## *Chapter 6*

# Predicting the intelligibility for speech in real-life background noises

*Koenraad S. Rhebergen, Niek J. Versfeld and Wouter A. Dreschler  
Submitted to Ear and Hearing*

## **Abstract**

The Speech Reception Threshold (SRT) is traditionally measured in stationary noise with the long-term average speech spectrum (LTASS) or with the long-term speech spectrum of the target speech. However, in real-life the instantaneous level or the spectrum of the background noise is more likely to be different from stationary LTASS noise. For better understanding the speech intelligibility in real-life background noises, SRTs were measured with normal-hearing listeners with a set of noises which vary in the temporal as well as spectral domain. With the aid of the Extended Speech Intelligibility Index (ESII) model, it will be shown that the SRTs in real-life noises can be predicted reasonably well for most observed SRTs. In this paper, it is shown that in addition to artificial modulated noises, the ESII can give a good account for the present set of real-life noises. Therefore it forms a valuable extension to the existing SII (ANSI S3.5-1997, 1997) model.

## I. Introduction

In many speech intelligibility studies, artificial noises are used to mask the speech signal. The most frequently used masking noise is stationary noise with the long term spectrum equal to that of the target speech (e.g., Smoorenburg, 1992; Noordhoek *et al.*, 1999). As a consequence, in each frequency band, the average signal-to-noise ratio (SNR) is the same, which makes the results less sensitive to the specific frequency characteristic of an experimental set-up. Less often, fluctuating noise is used to mask the speech. Here, too, the noise has the same frequency spectrum as the long term average speech spectrum. Past studies often dealt with interrupted noise with a duty cycle of 50 % and a modulation frequency ranging from 4 to 128 Hz (Miller and Licklider, 1950; de Laat and Plomp, 1983; Dubno *et al.*, 2002, 2003; Nelson *et al.*, 2003 and Rhebergen, Versfeld and Dreschler, 2006), amplitude or intensity modulated noise (Festen, 1987; Bacon *et al.*, 1998; Trine, 1995 and Rhebergen, Versfeld and Dreschler, 2006), or speech modulated noise (Festen and Plomp, 1990; Middelweerd *et al.*; 1990; Peters *et al.*, 1998; Festen and Plomp, 2002; Versfeld and Dreschler, 2002). Noise conditions with temporal gaps are very interesting because they appear to be sensitive to differences between hearing-impaired and normal-hearing listeners. The interrupted noises are very useful in a clinical or experimental setting to examine the speech intelligibility in noise, but it is not clear to what extent they are representative for every-day noises. Real-life noises are more complex in the temporal and in the spectral domain. It is nearly impossible to examine the effect of all existing real-life noises on speech intelligibility. Thus, it is desirable to have a model that is able to predict the intelligibility in these conditions. Since the introduction of the AI (Articulation Index; ANSI S3.5-1969) and its successor, the SII (Speech Intelligibility Index; ANSI S3.5-1997), it is possible to predict speech intelligibility in stationary noises. For a given speech in noise condition, the SII is calculated from the average noise spectrum, the average speech spectrum, and the listener's hearing threshold. The calculated SII score is a number between zero and unity and can be interpreted as the proportion of the total speech information which is perceptually accessible to the listener. Since the SII is calculated from the average spectrum of noise and speech, it does not take into account any fluctuations in the masking noise. As most real-life noises are more fluctuating than stationary, and since the amount of fluctuations can have a large effect on

intelligibility, SII nor AI are capable of predicting speech intelligibility in fluctuating noise. Rhebergen and Versfeld (2005), and Rhebergen *et al.*, (2006b) introduced an extension to the SII calculation scheme with the aim to predict speech intelligibility in stationary and in fluctuating noises. In this method, the concept of an instantaneous SII is introduced and the average across time is taken to come to a overall (dynamic) SII score. Rhebergen and Versfeld (2005) verified their method on the basis of a variety of noise types available in the literature. Rhebergen *et al.* (2006b) refined and validated their method on the basis of experimental data from normal-hearing listeners for a wide range of fluctuating noises. Their Extended SII-model (ESII, Rhebergen *et al.*, 2006b) is able to predict speech intelligibility for a wide range of stationary and fluctuating masking noises. However, all masking noises were artificial in nature. The question is to what extent the model is able to predict speech intelligibility in other than artificial masking noises.

Koopman, Franck and Dreschler (2001) analyzed a large database of noises on the basis of their spectral and temporal contents. The aim of their study was to come to a representative selection of real-life noises for the evaluation of hearing aid signal processing schemes in specific background noise conditions. With the aid of multidimensional scaling techniques, they were able to define four dimensions, such that most sounds in their database could be adequately positioned in this four-dimensional space. Three dimensions correlated well with spectral characteristics of the sounds, one dimension with the temporal characteristics. This classification is particularly suitable for selecting a set of representative real-life sounds. If observed SRTs in a representative selection of real-life background noises can be predicted by the Extended SII method, it is expected that, next to an experimental or clinical setting, the method can be used to predict the speech intelligibility in real-life surroundings such as train stations, public buildings, auditorium, communication channels, etc.

The goal of the current experiment is to determine speech intelligibility in a representative set of real-life noises and compare the observed SRTs with the predicted SRTs calculated with the Extended SII method.

In the first section, SRTs are measured in twelve normal-hearing listeners for twelve different noise conditions (one test and two re-tests per condition). The noise conditions range from a stationary (reference) noise, fluctuating speech

shaped (reference) noise, interfering speech played normal and time-reversed, and a selection of eight real-life noises. The noise conditions are selected in such a way, that they differ in temporal or spectral properties and are representative for most everyday sounds.

In the next section, the observed SRTs are used to evaluate the Extended SII model.

Finally, the predictions and limitations of the extended SII model will be discussed.

## II. Method

### a. Subjects

Twelve normal-hearing subjects (3 male, 9 female) participated. Their age ranged from 19 to 39 years and was on average 26 years. Subjects were native speakers of the Dutch language. They had at least high school education. Each subject had pure-tone thresholds of 15 dB HL or better at octave frequencies from 125 to 8000 Hz (ANSI S3.6, 1996).

### b. Stimuli

The target speech material consisted of the VU98 speech corpus, short everyday sentences, uttered by a female speaker (Versfeld *et al.*, 2000). The speech material comprises 39 lists of 13 sentences and has been developed for a reliable measurement of speech intelligibility in noise. The speech was stored at a sample rate of 44.1 kHz and 16 bits resolution.

Interfering noise stimuli were twelve “real life” noises, selected out of a noise database based on their spectral and temporal differences described by Koopman *et al.*, (2001), and are described in Table 6.1. The personal recordings were made with a Sony PCM-M1 digital audio recorder on DAT (44.1 kHz, 1-channel, 16 bits) using the microphone of a sound level meter (Brüel & Kjær, type 2230). The sound levels were measured in dB SPL. To prevent interference of the wind, a windscreen was placed on the sound level meter. The CD noises were collected from the following sound CD’s: “Sound Ideas-Sound effects library, a 12-CD pack, containing noises of real-life environments”, “Queen of

African Music”, “Noise-rom-0, a CD composed by TNO containing military noises and industrial noises”, “NVA speech material (Smooenburg, 1992).” and “VU98 speech material” (Versfeld *et al.*, 2000). The stationary speech noise spectrum [noise #12] was equal to the long-term average spectrum of the female speech material. The fluctuating speech noise [noise #11] had the long-term average spectrum of the female speech material, and the temporal envelope was that of speech, except for that it was split up in an independent low- and a high frequency part with 1000-Hz crossover frequency, similar to the procedure described by Festen and Plomp (1990). The latter was done to prevent the noise from being intelligible.

**Table 6.1.** 12 Noise signals

Noise	Source
1. Hens & Birds	Personal recording
2. Construction	CD Sound Ideas 2, track 9
3. Machine gun	CD TNO-noise-rom-0, track16
4. Car noise	Personal recording
5. Ramming piles	Personal recording
6. Music	CD Queen of African Music, track 1
7. Crowd	CD Sound Ideas 7, track 16
8. Frogs & Insects	CD Sound Ideas 3, track 5
9. Speech BW (Male Backward)	CD NVA Speech material (Smooenburg, 1992)
10. Speech (Male Forward)	CD NVA Speech material (Smooenburg, 1992)
11. VU98 fluctuating speech spectrum (female)	CD VU98 (Versfeld <i>et al.</i> , 2000)
12. VU98 stationary speech spectrum (female)	CD VU98 (Versfeld <i>et al.</i> , 2000)

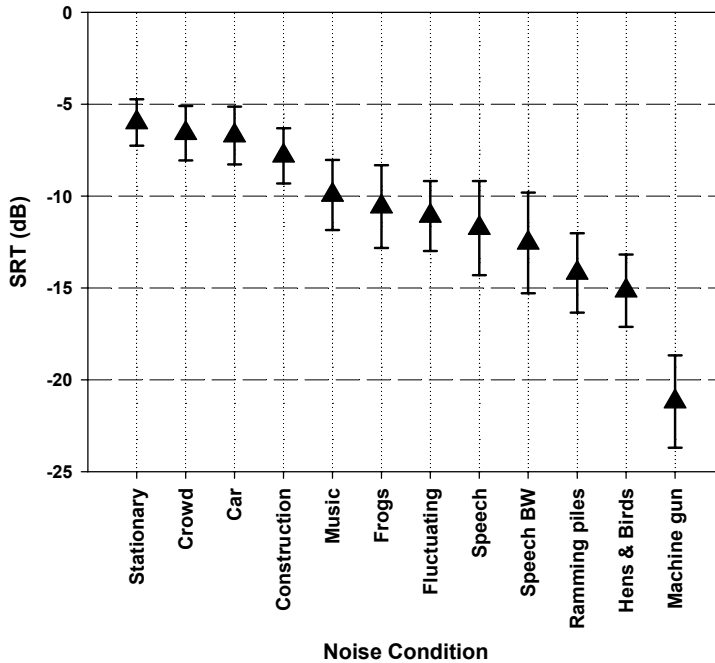
### **c. Procedure**

Subjects were tested individually in a sound-insulated booth. The monaural speech-reception threshold (SRT) was measured at the better ear for a fixed noise level of 65 dBA. Signals were played out via an Echo soundcard (Gina 24/96) on a PC at a sample frequency of 44.1 kHz. The signals were fed through a TDT Microphone Amplifier (MA2) and a TDT Headphone Buffer (PA4) via

## Chapter 6

TDH 39P headphones. After the presentation of a sentence in noise, the subject's task was to repeat the sentence he or she had just been presented. A sentence was scored as correct if all words in that sentence were repeated without any error. A list of 13 sentences, unknown to the subject, was used to estimate the signal-to-noise ratio at which 50 % of the sentences was reproduced without any error, the so-called Speech Reception Threshold (SRT). For a given condition, the first sentence of the list started far below the expected SRT. The sentence was repeated each time at a 4 dB higher level until the subject was able to reproduce it correctly. The twelve remaining sentences in that list were presented only once, following a simple up-down procedure with a step size of 2 dB, that is the level was decrease by 2 dB after a correct response and increased by 2 dB after an incorrect response. The SRT was estimated according to the procedure described by Plomp and Mimpen (1979), i.e., by taking the mean Signal to Noise Ratio (SNR) of sentence five to thirteen plus the estimated SNR that would have been used for the fourteenth sentence, if presented. Note that, since the spectral content of speech and noise differed, the signal-to-noise ratio here is defined as the differences in the A-weighted sound level of both signals. With each sentence presentation, a fixed sample of the interfering masking noise was taken (frozen noise). It started 1200 ms before the start of the sentence and stopped at least 800 ms after the end of the sentence. In total twelve masking conditions were tested. To avoid confounding of measurement condition order and sentence lists, the order of was counterbalanced across subjects according to a 12 by 12 Latin Square method. In total, each subject received 36 lists of 13 sentences. The experiment was preceded by 3 practice lists. A pilot experiment with the same 12 noise conditions provided insight into the SRT for these conditions, such that the appropriate initial speech level could be taken.

## Speech Intelligibility in Real-life Noises



*Figure 6.1. Speech Reception Threshold (dB) as a function of interfering noise condition. Error bars denote the standard deviation between subjects after averaging across the two retests.*

### III. Results

Figure 6.1 shows the SRT-values averaged across the twelve subjects and retests for each of the twelve conditions of the interfering noise. Error bars denote the standard deviation between subjects. A 12[noise condition]  $\times$  3[test/retest/retest]  $\times$  12[subject] Analysis Of Variance (ANOVA) was performed on the data-set. Of the main effects, differences in “noise condition” were significant ( $F[11,121]=287.55$ ,  $p<0.001$ ), differences in “test/retest/retest” were significant ( $F[2,22]=23.178$ ,  $p<0.001$ ), and differences in “subject” were significant ( $F[11,8.151]=26.654$ ,  $p<0.001$ ). The SRT of the first retest was on average 0.94 dB better than the first test; this difference was significant ( $F[1,11]=22.28$ ,  $p<0.001$ ).

The SRT of the second retest was on average 0.2 dB better than the first retest, but this difference was not significant ( $F[1,11]=2.92$ ,  $p>0.1$ ). None of the interactions were significant.

#### IV. Discussion

The overall results show a clear effect of noise type on the SRT: It seems that stationary-like noises are capable of masking the speech more efficiently than fluctuating noises of the same A-weighted RMS level. Also, when speech and noise spectrum become increasingly different, masking efficiency decreases. The values for the SRT in stationary noise and fluctuating speech shaped noise is consistent with other SRT data, approximately -5 dB, and -11 dB, respectively (Festen and Plomp, 1990; Middelweerd *et al.*, 1990; Peters *et al.*, 1998; Festen and Plomp, 2002; Versfeld and Dreschler, 2002). The lowest SRT is observed with the machine gun noise as a masker (-21.2 dB). Due to the interrupted characteristic of this noise sample, it has about the same masking efficiency as an artificial made interrupted noise masker with a modulation frequency of 8 Hz and a duty cycle of 50 % (de Laat and Plomp, 1983; Rhebergen *et al.*, 2006b).

Furthermore, the mean SRTs of the retest were on average 0.94 dB lower than the SRTs of the first test. This effect was only present for the fluctuating noise conditions. These outcomes are in line with Rhebergen *et al.* (2006a), who repeatedly measured the SRT in interrupted noise and stationary noise. Contrary to the SRT in stationary noise, the SRT in interrupted noise was about 0.9 dB improved after one repeated SRT measure.

The SRTs observed in interfering male speech and time-reversed interfering male speech are similar to those obtained by Festen and Plomp (1990), and are in line with Summers and Molis (2004), and Rhebergen *et al.* (2005). The SRT is lower in time-reversed interfering speech compared with normal interfering speech. By reversal of the interfering speech masker in time, an improvement in intelligibility of the target speech is observed due to the fact that the masking speech becomes unintelligible (so-called release from informational masking, the effect that, next to energetic masking, information in the masker increases the masking effectiveness). Rhebergen *et al.*, (2005) showed that by reversal of unintelligible (non-native) speech, the SRT increased (i.e., worsened) with 2.3

dB. This rise in SRT is most likely attributable to an increase in forward masking, due to the nature of the temporal envelope of speech. The observed SRT in time-reversed masking speech thus is a combined effect of a release from informational masking and an increase of forward masking. The difference in the present experiment is about 0.8 dB, which is considerably lower than the difference in SRT of 4.3 dB between normal and time reversed interfering speech, found by Rhebergen *et al.* (2005). Two possible reasons for this discrepancy are the more distracting interfering intelligible speech used by Rhebergen *et al.* (2005) and the possibility that subjects in the former study are easier distracted by interfering speech compared to the subjects in this experiment.

## **V. SII Model predictions**

A detailed description of the Extended SII model is given in Rhebergen *et al.* (2006b). The basic principle of the Extended SII model is that both speech and noise signal are filtered into 21 frequency bands (critical bands). In each band, the envelope of the speech signal and noise signal is shaped with a forward masking function (FMF) algorithm to account for forward masking. Next, speech and noise are partitioned into small time frames of 4 ms in length. Within each time frame, the conventional SII is determined, by determining the audibility of the speech, weighted by the importance of that band for intelligibility. These values are summed across frequency bands, yielding the speech information available to the listener at that time frame. This yields an SII that varies over time. Averaging the SII over time results in the SII for that particular condition.

Both conventional SII (ANSI S3.5-1997) and Extended SII (Rhebergen *et al.*, 2006b) use the SPIN 21 critical band weighting function (ANSI S3.5-1997, 1997, Table B.1). The SII calculations were conducted with the long term speech spectrum of the female target speaker. In order to approach the sound level at the ear drum (as required by the SII model), the speech and noises were filtered with a 5<sup>th</sup> order FIR filter to mimic the characteristics of TDH39 headphone. Also, the noise signals were corrected for the additional background noise,

present in the sound booth. Since the model cannot account for learning effects, only the mean SRT scores of first and second re-test were used.

By definition, at threshold (i.e., at SRT) the same amount of speech information is available to the listener. Thus, when feeding the model with the SRT values, it is expected that all SII values lie close together. Previous research has shown that with normal-hearing subjects, the SII score lies roughly between 0.3 and 0.4 (Versfeld and Dreschler, 2002; Rhebergen and Versfeld, 2005).

### **A. Conventional SII**

The fourth column of Table 6.2 shows the conventional SII calculations (ANSI S3.5-1997). The mean SII value is 0.19 and there is considerable variation between conditions, resulting in a large standard deviation of 0.10. Furthermore, the ANSI S3.5-1997 scheme returns with an SII score of zero for speech in machine gun noise. Generally, conventional SII underestimates the speech in fluctuating noise conditions.

### **B. Extended SII**

The fifth column of Table 6.2 shows the Extended SII (ESII) calculations (Rhebergen & Versfeld, 2005). The mean SII value is 0.35 with a standard deviation of 0.077. This result is far better than the ANSI S3.5-1997.

The sixth column of Table 6.2 shows the results of calculations with a modified ESII, as described by Rhebergen *et al.* (2006b). The mean SII value is 0.34 with a standard deviation of 0.064. When the speech conditions (speech and speech backward) are excluded from the analysis, then the mean SII value is 0.32 with a standard deviation of 0.049. Both analysis results are better than the ANSI S3.5-1997 and are also an improvement compared with the ESII calculations according to Rhebergen and Versfeld (2005).

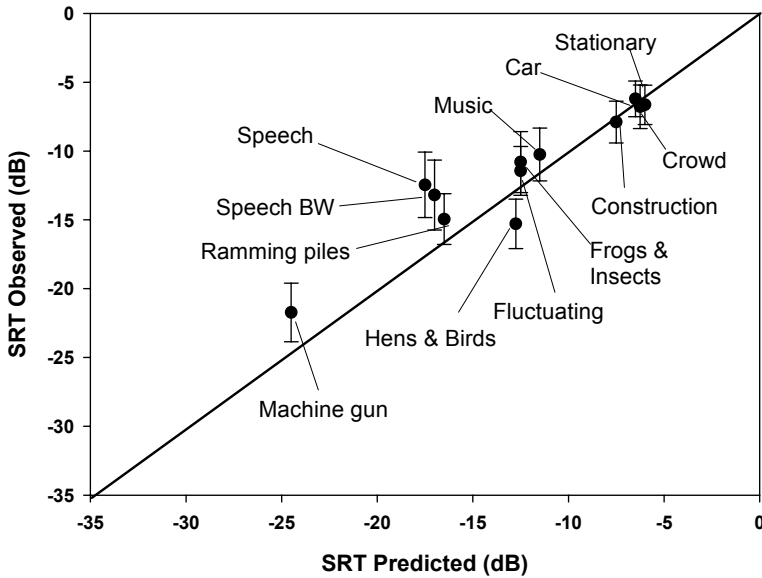
## Speech Intelligibility in Real-life Noises

**Table 6.2.** SRTs and various SII calculations for each of the twelve noise maskers. Lower rows yield average of the SRT and mean and standard deviation of the SII.

Noise condition	SRT	stdv	SII	ESII (2005)	ESII & FMF
Stationary	-6.2	1.3	0.29	0.29	0.29
Crowd	-6.6	1.4	0.30	0.33	0.33
Car	-6.8	1.6	0.32	0.31	0.31
Construction	-7.9	1.5	0.28	0.28	0.28
Music	-10.3	1.9	0.19	0.25	0.25
Frogs	-10.8	2.2	0.28	0.35	0.35
Fluctuating	-11.4	1.8	0.09	0.38	0.36
Speech	-12.5	2.4	0.14	0.47	0.45
Speech BW	-13.2	2.5	0.14	0.45	0.42
Ramming piles	-15.0	1.8	0.10	0.36	0.37
Hens & Birds	-15.3	1.8	0.15	0.26	0.26
Machine gun	-21.7	2.1	0.00	0.46	0.39
<b>All Noises</b>					
<b>mean SII</b>			0.19	0.35	0.34
<b>SII stdv</b>			0.10	0.08	0.06
<hr/>					
<b>Without Speech</b>					
<b>mean SII</b>			0.20	0.33	0.32
<b>SII stdv</b>			0.11	0.06	0.05

Figure 6.2 displays the relationship between the observed SRT and the SRT as predicted by the ESII model (Rhebergen *et al.*, 2006b) with the FMF for all conditions described in the previous section. SRTs were calculated by taking the hearing loss fixed at the mean HL of all twelve subjects, and by setting the threshold value of the SII to the average value of 0.32, as calculated in Table 6.2. Different SRTs were obtained by taking the related sample of the masking noise. The diagonal in Figure 6.2 indicates the points where the observed and predicted SRT are identical. All predicted SRT values are within a few decibels of the diagonal, or even lie on the diagonal, indicating that the model does well with the present set of data. The ESII model yields a substantial improvement

over the ANSI S3.5-1997 SII model. Points above the diagonal indicate that listeners perform worse than predicted by the ESII model. Since the results indicate a learning effect for some conditions, it is expected that the points in Figure 6.2 being above the diagonal will after correction for the learning effect become closer to the diagonal. In some noise conditions there might still be a learning effect present (e.g., machine gun).



**Figure 6.2.** For the masking noises of the present experiment, the observed SRT (dB) is plotted as a function of the SRT (dB) predicted by the Extended SII model (Rhebergen et al., 2006b). Conditions are denoted in short in the figure.

The predicted SRT in interfering speech is about 5 dB higher compared with the observed SRT. The 5 dB difference in SRT is in line with the assumed 6.6 dB informational component measured by Rhebergen et al. (2005). The ESII model cannot account for informational masking, since it only determines the amount of speech information available to the listener. The difference between the observed and predicted SRT is thus influenced by the informational masking on

top of the energetic masking. The problem of the phenomenon informational masking is that it probably is dependent on different characteristics of the interfering speech, such as speaker, articulation style, gender, contents of the message, or dependent on different factors of the listener, such as concentration, interest in conversation of the target or interfering speech. If one considers a range of about 0 to about 7 dB mismatch with the predicted SRT (0 dB: no influence of informational masking on measured SRT, up to about 7 dB due to increasingly more influence of informational masking on measured SRT), one should not use interfering speech in a clinical setting to measure the speech intelligibility in fluctuating noise. Artificial interfering noises (e.g., 8 Hz interrupted noise) will probably give more reliable results due to absence of informational masking (although one should account for learning effects for speech intelligibility in interrupted noise). Apart from the interfering speech maskers, where informational masking may play a role, the ESII model appears to be a good method to predict the SRT in real-life noises.

## **VI. Summary and conclusions**

The present paper describes an SRT (Speech Reception Threshold) experiment with twelve normal-hearing subjects for “real life” masking noises. The noise conditions vary from stationary noise with the same spectrum as the target speech, to fluctuations in the spectral and amplitude domain. The observed thresholds are used to evaluate the Extended SII approach (Rhebergen and Versfeld, 2005; Rhebergen *et al.*, 2006b) to model SRTs for sentences masked by fluctuating noises. For the present set of masking noises, the Extended SII is able to predict the speech intelligibility reasonably accurate.

## *Chapter 7*

# The dynamic range of speech, compression, and its effect on the speech intelligibility in interrupted noise

*Koenraad S. Rhebergen, Niek J. Versfeld and Wouter A. Dreschler*

## **Abstract**

Differences in the dynamic range of the speech signal may play a crucial role for predicting the speech intelligibility in noise. Models set up to predict the speech intelligibility (such as the Speech Transmission Index, STI, the Articulation Index, AI, and the Speech Intelligibility Index, SII) consider the dynamic range of the speech signal fixed regardless of the type of speech material. However, the different models assume different values for the dynamic range (from 30 to 68 dB) and the level of the speech peaks (12 to 15 dB *re*: RMS level). This choice of the dynamic range and the speech peaks has not been based on statistical analyses of the dynamic range of the speech signal itself, but rather on a best fit of the model to the experimental data.

The present paper describes two experiments with normal-hearing subjects to examine the effect of the dynamic range of the speech signal on speech intelligibility in stationary and interrupted noise. The dynamic range has been varied by compression or expansion of the speech signal only, leaving the noise unaltered, or by compression or expansion of the entire speech-in-noise signal. Calculations with the Extended SII model (Rhebergen, Versfeld, and Dreschler, 2006b) yield somewhat better predictions for compressed speech-in-noise when a parameter is included that accounts for variations in the dynamic range of speech.

## I. Introduction

Numerous papers describe speech intelligibility experiments with normal-hearing listeners in stationary noise utilizing the so-called Speech Reception Threshold (SRT) method for sentences, as described by Plomp and Mimpen (1979). The SRT test is an adaptive speech intelligibility test to determine the Signal-to-Noise Ratio (SNR), or so called SRT, required for 50 % sentence intelligibility. Remarkably, the range in observed SRTs varies to some extent between studies. SRT values measured with different Dutch speech corpuses but with the same subjects range from -3.7 dB to -4.5 dB (Versfeld *et al.*, 2000; van Wijngaarden and Houtgast, 2004). Neither the Speech Transmission Index (STI; Steeneken and Houtgast, 1980, 1985), the Articulation Index (AI; ANSI S3.5-1969, 1969), nor its successor the Speech Intelligibility Index (SII; ANSI S3.5-1997, 1997) is able to adequately predict the differences in speech intelligibility for these different speech materials. For example, for a given SNR, the SII is calculated from the listener's hearing threshold, the average noise spectrum, and the average speech spectrum, filtered in frequency bands and weighted by the band importance function which accounts for the type of speech material used (e.g., words, sentences). The calculated SII score is a number between zero and unity and can be interpreted as the proportion of the total speech information which is perceptually accessible to the listener.

Besides the band importance function there is no input variable in the SII model that can account for the differences in intelligibility between speech corpuses. The STI accounts for differences in gender by calculating the speech intelligibility of female speech with six octave bands (250 Hz - 8 kHz) and male speech with seven octave bands (125 Hz - 8 kHz) (Steeneken, 2002; Steeneken and Houtgast, 1999, 2002). However, still no distinction can be made for specific speaker styles.

Differences in the dynamic range of the speech signal may play a crucial role because it is related to the amount of speech information that is elevated above the noise. The dynamic range of the speech signal is fixed for all speech materials in the AI and SII models, but has decreased over the years from 68 dB (Fletcher and Galt, 1950), to 36 dB (Kryter, 1962) down to 30 dB (Houtgast and Steeneken, 1985; ANSI S3.5, 1969,1997). Furthermore, the level of the speech peaks (speech level above the RMS of speech) has been increased from 12 dB

## *Speech Dynamic Range*

(ANSI S3.5-1969) to 15 dB (ANSI S3.5-1997, 1997). This change of the dynamic range and speech peaks has not been based on statistical analyses of the dynamic range of the speech signal itself, but rather on a best fit of the SII model to the Studebaker *et al.* (1993) data (Houtgast, 2005; Pavlovic, 2005, and Studebaker, 2005).

Nowadays, the SII model uses a speech dynamic range of 30 dB with the RMS in the middle (ANSI S3.5-1997, 1997). Already since the introduction of the ANSI S3.5, the 30 dB dynamic range of speech is under dispute (Boothroyd, 1990, 2000; Van Tassell, 1993; Rankovic, 1997, 1998; Studebaker and Sherbecoe, 1999, 2002; Zeng *et al.*, 2002; Molis and Summers, 2003). In addition, the SII and STI do model the cumulative level distribution of speech as a linear function (straight line), starting from 15 dB below the RMS up to 15 dB above the RMS with a slope of 3.3 % speech information per dB. As a result, every decibel increase or decrease in SNR results in an equal amount of change in speech information available to the listener, regardless whether speech is almost fully audible or fully masked. If one assumes that the amount of speech information available to the listener is equivalent or closely related to the amount of audible speech, replacement of the linear weighting function in the SII model by the cumulative level distribution of the speech signal under investigation may result in an improvement of the predictive power of the SII. Drullman (1995a, 1995b) studied the fine structure cues for speech intelligibility and the relative contribution of speech elements above and below the noise level. He concluded that the most important portion of the dynamic range, providing essentially 100 % sentence intelligibility, is between 19 dB below and 1 dB above the RMS. Furthermore, Drullman (1995a) observed that the contribution of the troughs of the speech signal is relatively larger to speech intelligibility than is the contribution of the speech peaks. His results show that the level importance function of speech is non-linearly distributed rather than the linear level importance function of speech that was adopted in the SII model.

Boothroyd (1990, 2000), and Studebaker and Sherbecoe (1999, 2002) suggested to incorporate an Intensity Importance Function (IIF) in the SII calculation, such to account for the non-linear level distribution of speech. Also, they concluded that the effective dynamic range of speech is in the range of 40 dB to 50 dB rather than the commonly used 30 dB. If the effective level distribution (or

dynamic range) of speech has an effect on speech intelligibility in noise, it is important to examine if the shape of the level distribution can account for differences in performance between different speech materials. Rankovic (1998) showed that the AI calculations described by Fletcher and Galt (1950) for predicting the speech intelligibility are better than the method described by ANSI S3.5-1969. Although there are many differences among these two methods, one of the more significant differences is the IIF that each method assumes. The Fletcher and Galt IIF is considerably broader than the ANSI S3.5 IIF, and presumes that the weight of each decibel in the dynamic range of speech decreases nonlinearly in relation to its maximum value. Accordingly, the summarized results in the literature do not support the 30-dB linearly weighted IIF used in AI, SII or STI method and suggest the use of a non-linearly weighted IIF. If the effective dynamic range of speech has an effect on the prediction of the speech intelligibility in stationary noise, it may even be more important to examine the contribution of the effective dynamic range of speech on speech intelligibility in fluctuating noise. By manipulating the dynamic range of speech, the effect of the shape of the IIF might be explored for speech intelligibility in different noise conditions.

The aim of the current study is to examine the effect of the dynamic range of speech on speech intelligibility in fluctuating noise conditions such to improve the SII model by implementation of an IIF function, and examine the speech characteristics of different speech dynamic ranges which could be an important factor in the shape of the IIF for an improved SII model.

Section II describes the methods and results of Experiment I, in which SRTs are measured in interrupted and stationary noise where not the noise, but only the speech signal's dynamic range is varied by means of instantaneous wide dynamic range compression (WDRC). In section III, the spectral and temporal characteristics of the speech conditions are analyzed. With aid of the ESII, the SII values are calculated for all conditions of Experiment I (section IV).

Section V describes the methods and results of Experiment II, in which the same compression scheme is used as in section II, but now applied to the mixed speech-in-noise signal. In Experiment II, SRTs are measured in interrupted and stationary noise. The spectral and temporal characteristics of the speech-in-noise conditions are analyzed in section VI. In section VII, the SII values are

calculated for all conditions of Experiment II with aid of the ESII. Finally, in section VIII, predictions and limitations of the Extended SII model (Rhebergen, *et al.*, 2006b) for compressed speech will be discussed. Furthermore, some suggestions are made for further improvement of the Extended SII to account for the differences between different speech materials.

## II. Experiment I

### A. Method

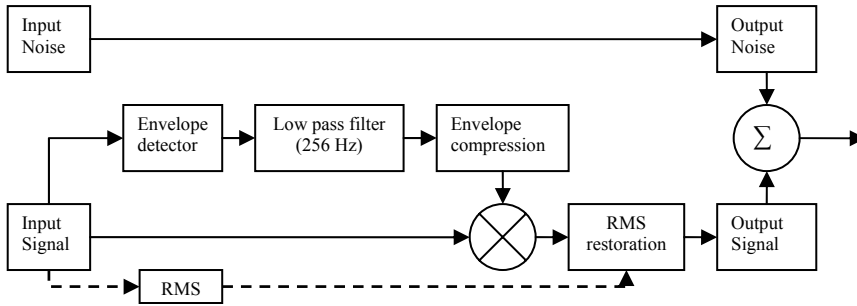
The present experiment was designed to determine if the dynamic range of speech has an effect on the speech intelligibility in interrupted and stationary noise.

#### 1. Subjects

Eight normal-hearing subjects (3 male, 5 female) participated. Their age ranged from 24 to 42 years (average 30.9 years). Subjects were native speakers of the Dutch language. They had at least high school education. Individual subjects had pure-tone thresholds of 15 dB HL or better at octave frequencies from 125 to 8000 Hz (ANSI S3.6, 1996).

#### 2. Stimuli

The target speech material consisted of short every-day sentences, uttered by a female speaker (Versfeld *et al.*, 2000). The interfering noise conditions comprised one condition with 8 Hz, 100 % modulated interrupted noise with a duty cycle of 50 %, and one condition with stationary noise. All noise conditions had a spectrum equal to the long-term average spectrum of the female target speech. The dynamic range of the target speech was varied by means of instantaneous Wide Dynamic Range Compression (WDRC) with compression ratios (CRs) of 4:1, 2:1, and 1:1, or with an expansion ratio of 2:3. With CR=1:1, the speech signal is unaltered. A block diagram of the WDRC signal processing on the target speech is shown in Figure 7.1 and is based on the scheme used by van Buuren, Festen, and Houtgast (1999).

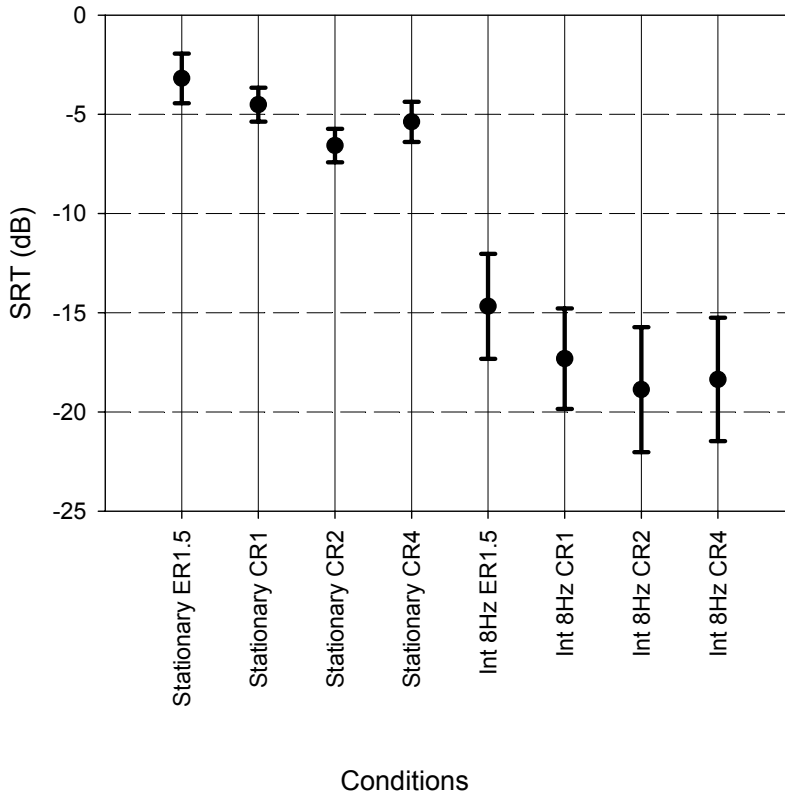


**Figure 7.1.** Block diagram of WDRC signal processing. Dashed lines indicate operations on the RMS level of the speech signal.

Since the aim of this experiment is to examine the effect of the dynamic range of speech on speech intelligibility, the WDRC signal processing operates only on the speech signal. Speech and noise are summed *after* the signal processing. The speech envelope was determined by means of a Hilbert Transform. It was filtered with a 96 dB/oct low pass Bessel filter with a cut-off frequency of 256 Hz to prevent higher frequency envelope modulations from controlling the gain. In contrast to van Buuren *et al.* (1999), who used a 32 Hz low pass filter, here a 256 Hz low pass filter was used to obtain a higher effective compression ratio (Apoux *et al.*, 2004). The low pass filter was applied twice: once to the envelope and once again to the filtered time-reversed envelope to remove the phase shifts introduced in the first filtering. The low passed speech envelope was processed (compression ratio 4:1, 2:1, 1:1, or expansion ratio 2:3) and next divided, sample by sample, by the original envelope, resulting in a multiplication factor (i.e., gain) for every sample of the speech signal. The multiplication factors were applied to the original speech signal and the long term RMS level of the input signal was restored. Finally, the speech and the noise signal were summed. The envelope was only compressed or expanded down to 55 dB below the RMS (knee point). Lower envelope levels were linearly amplified.

### **3. Procedure**

Subjects were tested individually in a sound-insulated booth. The monaural speech-reception threshold (SRT) was measured at the better ear for a fixed noise level of 65 dBA. Signals were played out via an Echo soundcard (Gina 24/96) on a PC at a sample frequency of 44.1 kHz. The signals were fed through a TDT Microphone Amplifier (MA2) and a TDT Headphone Buffer (PA4) via TDH 39P headphones. After the presentation of a sentence in noise, the subject's task was to repeat the sentence word by word. A sentence was scored as correct if all words in that sentence were repeated without any error. A list of 13 sentences, unknown to the subject, was used to estimate the signal-to-noise ratio at which on average 50 % of the sentences was reproduced without any error, the so-called Speech Reception Threshold, or SRT. For a given condition, the first sentence of the list started far below the expected SRT. The sentence was repeated each time at a 4 dB higher level until the subject was able to reproduce it correctly. The twelve remaining sentences in that list were presented only once, following a simple up-down procedure with a step size of 2 dB. The SRT was estimated according to the procedure described by Plomp and Mimpen (1979), i.e., by taking the mean Signal to Noise Ratio (SNR) of sentence five to thirteen plus the estimated SNR that would have been used for the fourteenth sentence, if presented. Note that, due to compression, the spectral content of speech and noise may differ. Therefore, the signal-to-noise ratio here is defined as the difference in sound level of speech and noise in dB rms. With each sentence presentation, a fixed sample of the interfering masking noise was taken (frozen noise). The masking noise started 1200 ms before the start of the sentence and stopped at least 800 ms after the end of the sentence. In total, eight conditions were tested (all conditions with a test and a retest): four conditions with interrupted noise and four with stationary noise, where the four conditions were different with respect to the amount of compression applied to the speech signal. To avoid order effects and to cancel out a possible learning effect, the order of conditions and sentence lists was counterbalanced across subjects according to a Latin Square method. The experiment was preceded by 6 practice lists, to practice the SRT in 8Hz interrupted noise and stationary noise. In total, each subject received 16 lists of 13 sentences within Experiment I (2 noise conditions, 4 speech conditions, and 2 tests).



*Figure 7.2. Speech Reception Threshold (dB) as function of compression ratio and noise condition. Error bars denote the standard deviation between subjects.*

## B. Results

Figure 7.2 displays the SRT-values averaged across subjects as function of compression ratio in 8 Hz stationary noise or in interrupted noise. Error bars denote the standard deviation between subjects. A 2 [noises]  $\times$  4 [speech conditions]  $\times$  2 [test/retest]  $\times$  8 [subject] Analysis Of Variance (ANOVA) was performed on the data-set. Of the main effects, “noises” was significant

## Speech Dynamic Range

( $F[1,7]=294.56$ ,  $p<0.0001$ ), and “speech conditions” was significant ( $F[3,21]=30.79$ ,  $p<0.0001$ ). Differences between “test” and “retest” were not significant ( $F[1,7]=1.33$ ,  $p>0.05$ ), which was also true for “subject” ( $F[7,8.76]=1.03$ ,  $p>0.5$ ). Of the interactions, “noises x subject” ( $F[7,5.85]=5.38$ ,  $p<0.05$ ) was significant.

Also, separate 4[speech conditions] x 2[test/retest] x 8[subject] ANOVAs were performed on the data obtained with stationary noise and with interrupted noise. For the stationary noise conditions, only the main effect of “speech conditions” was significant ( $F[3,21]=51.94$ ,  $p<0.001$ ). Post hoc tests (Tukey HSD) showed that in stationary noise, only speech conditions with CR=1 and CR=4 did not differ significantly. For the interrupted noise conditions, the main effects of “speech conditions” ( $F[3,21]=11.56$ ,  $p<0.001$ ), and “subject” were significant ( $F[7,8.68]=3.69$ ,  $p<0.05$ ). Of the interactions, “speech conditions \*subject” was significant ( $F[7,21]=2.54$ ,  $p<0.05$ ). Post hoc tests (Tukey HSD) showed that the speech condition with ER=1.5 significantly differed from the speech conditions with CR=2 and CR=4. Speech conditions with CR=1, CR=2 and CR=4 did not differ significantly from each other.

## C. Discussion

The present experiment shows that, given the same masking condition, the SRTs with four different compression ratios differ from each other, and that trends are similar in interrupted noise and stationary noise. On average SRTs in interrupted noise are much better than in stationary noise. Starting from the condition with ER=1.5, increasing the CR results in a better SRT, but when the CR is too large the SRT deteriorates again. The optimum value is near CR=2, resulting in an SRT being 1.5 to 2.0 dB better than the uncompressed signal. The increase in speech intelligibility with CR=4 probably is due to the increased deterioration of the fine structure of the speech signal that counteracts the increase in audibility.

The outcomes of the present study are comparable with those from other studies with WDRC speech (Dirks *et al.*, 1986; Boothroyd *et al.*, 1988; Dubno and Dirks, 1989; Kamm *et al.*, 1985; Pavlovic, 1984). WDRC is generally believed to provide better speech recognition in quiet when compared with linear amplification. (Moore and Glasberg, 1986; Van Tasell, 1993; Souza and

Turner, 1998; Souza and Bishop, 1999; Goedegebure, 2005). Due to the compressed dynamic range of speech, weaker parts of the speech signal are easier detectable. Soft parts of the speech signal (notably consonants) that are below the threshold of audibility, become audible due to WDRC. Thus, the consonant to vowel ratio has decreased.

Speech with an expanded dynamic range (expansion ratio 1.5:1) yield poorer SRTs compared with uncompressed speech, both in stationary and interrupted noise. Contrary to compression, expanding the dynamic range of speech has a clearly negative effect on the speech intelligibility. These outcomes are in line with Clarkson and Bahgat (1991), Freyman and Nerbonne (1996), van Buuren *et al.*, (1999), and Apoux *et al.* (2004). Expansion causes the weaker parts of the speech signal to become inaudible. The consonant to vowel ratio increases and thus important (consonant) information disappears. Indeed, subjects reported that the speech signal was easily detectable, but unintelligible.

### **III. Analysis of the speech signal**

In order to understand the implications of the manipulations on the speech signal, the present section examines in detail the spectro-temporal characteristics of the (compressed or expanded) speech signal.

#### **A. Long-term average speech spectrum**

For each condition (where condition denotes the amount of compression or expansion), the long-term average speech spectrum (LTASS) was obtained by concatenation of all sentences in time followed by calculation of the intensity in 21 critical bands according to ANSI S3.5-1997. The LTASS for each of the four compression ratios for the female speech is displayed in Fig 7.3. It can be seen that an increase in the compression ratio results in a small but systematic change in spectral tilt (in the order of 1-2 dB/octave), where the pivoting frequency is near 1200 Hz. To what degree this spectral change has an effect on speech intelligibility is explored in Section IV below (model predictions).

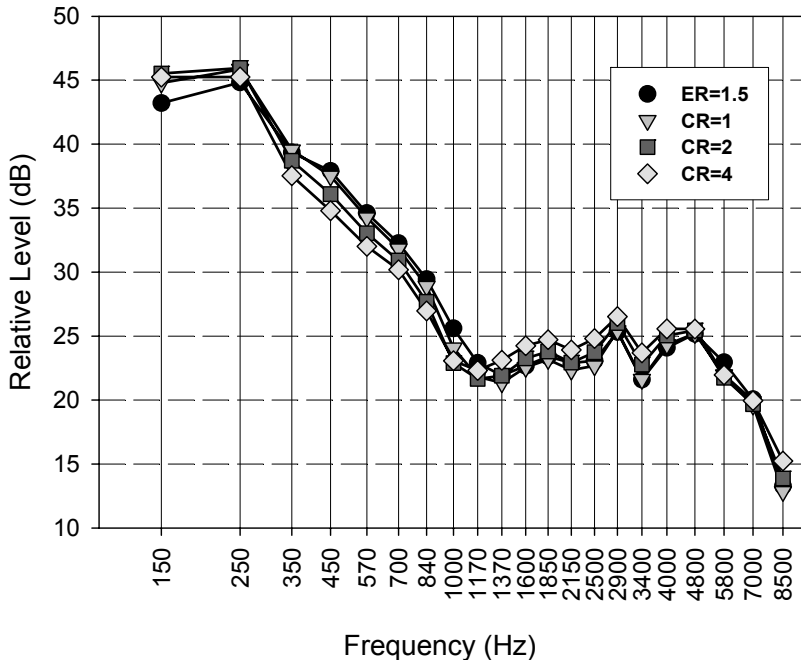
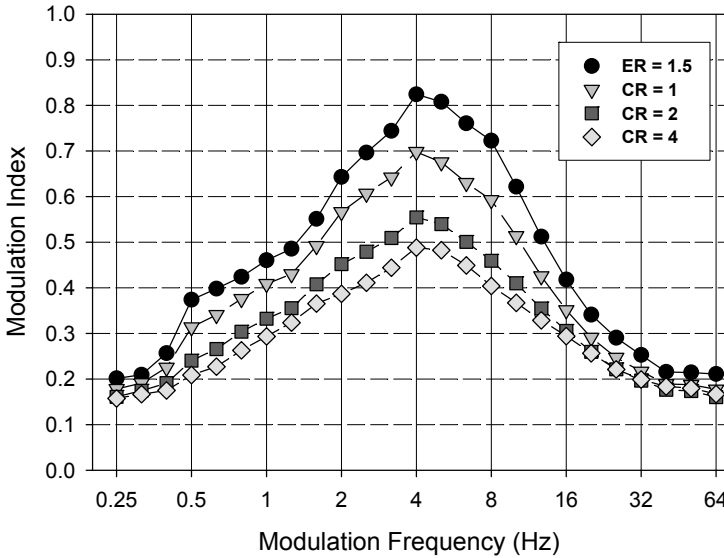


Figure 7.3 Critical band spectrum of speech with compression ratios CR=1, CR=2, CR=4, and expansion ratio ER=1.5.

## B. Modulation spectrum

Figure 7.4 shows for the four conditions the modulation spectrum as calculated in 21 critical bands (150 to 8500Hz; ANSI S3.5-1997), and averaged across the 21 frequency bands for the specific modulation frequency. The modulation index is defined as the RMS of the intensity envelope per 1/3 octave of modulation (0.25 to 64Hz) divided by the average intensity (for example: a 100 % sinusoidal intensity-modulated signal has the modulation index of unity in the 1/3-oct band corresponding to the modulation frequency). The modulation spectrum of the female speech has a peak at about 4 Hz.



*Figure 7.4* Modulation index of speech signals as function of modulation frequency. Compression ratios are  $CR=1$ ,  $CR=2$ ,  $CR=4$ , and expansion ratio  $ER=1.5$ .

Figure 7.4 shows an increased modulation index by expanded speech and a decreased modulation index by compressed speech. Plomp (1988) argued that any reduction of speech modulations result in a reduction of available speech information so that compression cannot lead to improved speech intelligibility. This view is based on the concept of the Modulation Transfer Function (MTF, Steeneken and Houtgast, 1980) and the STI model (Steeneken and Houtgast, 1980; Houtgast and Steeneken 1985) which is based on the assumption that reduction of the modulations in the speech signal leads to reduced speech intelligibility. Plomp's observation was the answer to the question why multi-channel amplitude compression does not lead to improved speech intelligibility in hearing-impaired listeners. In daily life, a hearing aid compresses not only the speech signal but also the modulations of the background noise. Since the STI model uses the MTF, it cannot properly be applied to speech that is contaminated with noise that comprises modulations. Also, Villchur (1989)

disagreed with Plomp's point of view. In the STI model, modulations are reduced by addition of stationary noise to the speech signal, such that the weaker speech parts become inaudible. By compressing or expanding the speech, all speech parts will remain audible, only the level differences are changed. The STI model cannot account for these type of changes, which has experimentally been confirmed by Hohmann and Kollmeier (1995) for compression ratios up to 3:1. Goedegebure (2005) showed that with a SNR of +6 dB, the modulations of the speech signal with a compression ratio of 2:1 remain the same. Souza *et al.*, (1999, 2006) and Souza (2002) noted that compressed speech in quiet improved speech intelligibility, but so far the majority of results do not show improved speech intelligibility with compression on a speech in background noise mix.

### C. Dynamic range of the speech signal

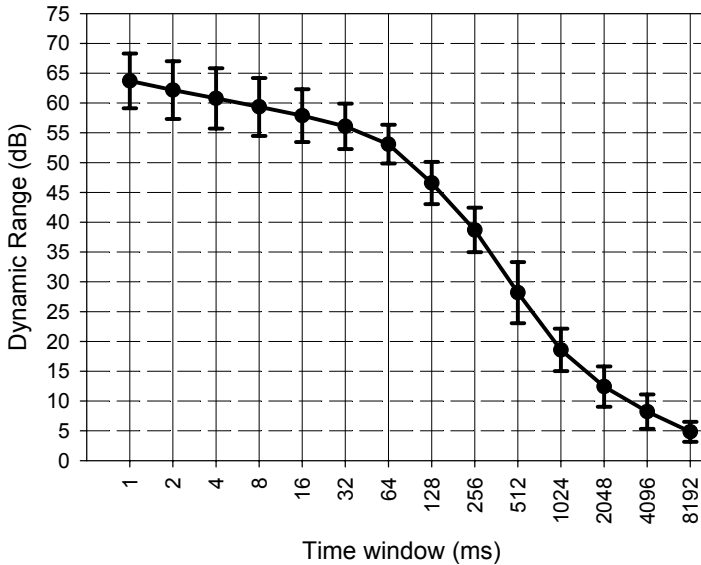
The dynamic range of the speech is defined as the range between the weakest and the loudest parts of the speech and is assumed to be about 30 dB (ANSI S3.5-1997, 1997). As mentioned earlier in this paper, the 30 dB dynamic range of speech is still under dispute. The 30 dB dynamic range is still used in the AI, SII, and STI, and relies heavily on the statistical analysis of running speech by Dunn and White (1940). Cox *et al.* (1988) measured the level distributions with the 1/3-octave short-term rms of normal speech for 30 female and 30 male speakers, and found a 40 to 50 dB dynamic range. Boothroyd *et al.* (1994), Byrne *et al.* (1994), Zeng *et al.* (2002), and Stobich *et al.* (1999) report speech dynamic ranges of about 37, 40, 50, and 70 dB, respectively. The huge differences between these studies probably are not due to differences in speech materials, but rather to differences in the used calculation schemes to obtain the dynamic range of speech. The short-term rms classically is obtained with a sound level meter, which detects the rms envelope of the input signal using a "Fast" time constant. This time constant is nominally 125 ms with 50 % overlap. Some studies use integration time windows ranging from 10 to 200 ms. The short-term rms usually is recorded with broadband, octave, or 1/3 octave bands. The "Peak" time constant refers to the highest instantaneous signal level measured during each passage, and "Max" refers to the highest level of the rms envelope

## Chapter 7

measurement during each passage (Byrne *et al.*,1994). The Leq is the long-term rms level during the total analysis time.

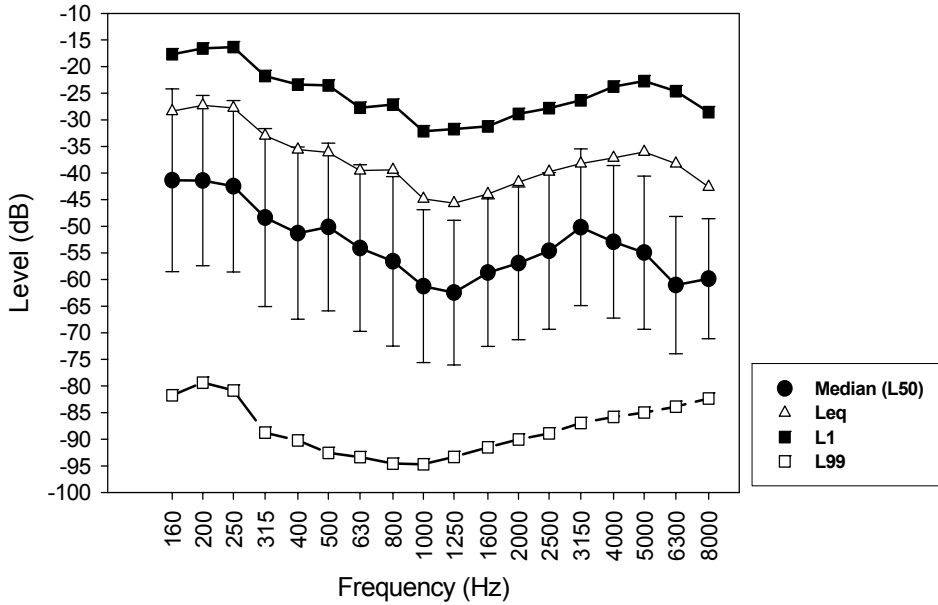
The dynamic range of speech is the short-term speech level range between L1 and L99 (1<sup>st</sup> to 99<sup>th</sup> percentile level) in each band. Speech peaks are situated between L1 and the Leq in each band and refer to the speech information above the Leq. Cox *et al.* (1988) found an 11 dB, and Studebaker and Sherbecoe (1993) found a 10 dB difference between the L1 and the Leq. In their study into the long-term average speech spectrum (LTASS) for 12 different languages, Byrne *et al.*, (1994) also determined some dynamic characteristics of speech and found a 10 dB speech peaks range, which was at that time 1 to 2 dB below the widely accepted value of 12 dB of the AI (ANSI S3.5, 1969), and at present 4 to 5 dB below the 15 dB speech peaks of the SII (ANSI S3.5, 1997). To examine the effect of integration time on the dynamic range of speech, 30 seconds of running speech without any pauses was analyzed. Figure 7.5 shows the mean dynamic range (L1-L99) based on 18 1/3-octave band (160 to 8000Hz; ANSI S3.5-1997), obtained with a sliding rectangular temporal window with an integration time ranging from 1 ms to 8192 ms. The calculated dynamic range here is largest with a sliding temporal window of 1ms and is gradually decreasing to 4.5 dB at 8192 ms. The mean calculated dynamic ranges with a 1ms temporal window is 63.7 dB (standard deviation between 1/3-octave bands is 4.6 dB). The AI, SII, and STI assume a fixed dynamic range across all frequency bands, but the present analyses indicate that this is not true.

## Speech Dynamic Range



*Figure 7.5. The dynamic range averaged across 18 1/3-octave band (160 to 8000Hz; ANSI S3.5-1997), as obtained with a sliding rectangular temporal window with an integration time ranging from 1 ms to 8192 ms.*

Figure 7.6 shows the L1, Leq, median (L50), and L99 of the 18 1/3-octave band (160 to 8000 Hz; ANSI S3.5-1997) of the female speech obtained with a sliding rectangular temporal window with an integration time of 1 ms. The choice for a 1 ms integration time in this paper is not based on psychophysical data, but it is chosen to reveal a clearly larger dynamic range compared with the commonly used integration time of about 125 ms. Furthermore, the mean dynamic range yielded with a 1 ms integration time is closer to the assumed 68 dB dynamic range of the Fletcher AI (Fletcher and Galt, 1950). SII and STI both take the Leq (rms) in the middle of the dynamic range of speech across all frequency bands. Figure 7.6 shows that the Leq is 15.7 dB above the median, which indicates that more speech information is below the Leq than SII and STI assume.



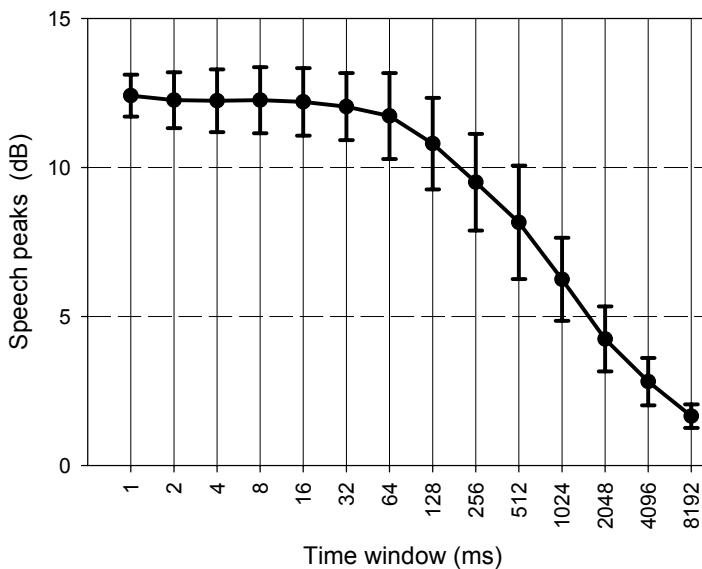
**Figure 7.6** The dynamic range of 18 1/3-octave band (160 to 8000Hz; ANSI S3.5-1997), obtained with a sliding rectangular temporal window with an integration time of 1 ms. Error bars indicate the variation of speech level (i.e. standard deviation) within the frequency band across time.

Figure 7.7 shows the mean speech peaks (L1-Leq) of 18 1/3-octave band of the female speech obtained with a sliding rectangular temporal window with an integration time ranging from 1 ms to 8192 ms. The calculated speech peaks are largest (12.4 dB) with a temporal window of 1ms and is gradually decreasing down to 2 dB at 8192 ms. Speech peaks remain fairly constant at 12 dB for short integration times up to 64 ms and reduce for longer integration times. Both SII and STI assumed fixed speech peaks of 15 dB across all frequency bands, but the present calculations are more in line with the AI (ANSI S3.5-1969) that uses speech peaks of 12 dB. The 3 dB larger speech peaks used in the SII and STI refer to speech information that still can be detected in stationary noise with a SNR of -5 dB. Drullman (1995b) showed that even in noise listeners make use of

## Speech Dynamic Range

information in the troughs of the speech signal below the noise level. To account for this phenomenon in the SII and STI calculation, the introduction of “effective” speech peaks seems to be reasonable (Studebaker & Sherbecoe, 1993, 2002).

The calculations show that the dynamic range of speech is largely determined by the integration time. Experiments on the temporal window of the human auditory system (Moore; 1997; Plack & Oxenham, 1998; Plack & Drga, 2003) revealed an Equivalent Rectangular Duration (ERD) of 11.5 ms. Studies on gap detection (Moore *et al.*, 1997) showed that subjects are able to detection gaps of 2 ms in duration at high frequencies, to about 18 ms at low frequencies. Since listeners are able to detect small changes in signals, it seems to be perceptually relevant to use short integration times to calculate the dynamic range of a speech signal.



**Figure 7.7.** The speech peaks of 18 1/3-octave band (160 to 8000Hz; ANSI S3.5-1997) obtained with a rectangular temporal window with an integration time ranging from 1 ms to 8192 ms.

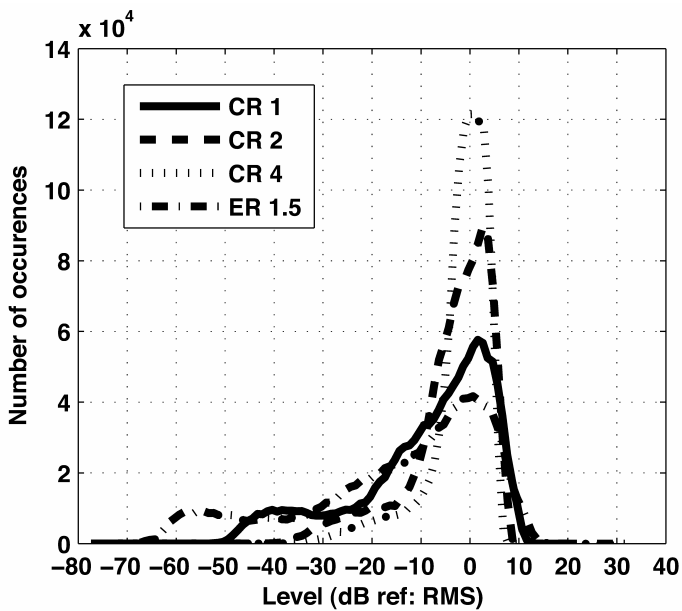
## D. Short-term level distribution of speech

Figure 7.8 shows, for the four conditions used in the listening experiment, the band passed (100-8500Hz) level distribution relative to the rms, obtained with a rectangular temporal window with an integration time of 1 ms. The figure shows that indeed the speech levels are closer together as speech is more compressed. This suggests that the level density function as used in the STI and SII cannot account for differences in compressed speech. Boothroyd (1990, 2000) and Studebaker and Sherbecoe (1999, 2002) proposed an Intensity Importance Function (IIF) in the SII calculation which is closely related to the (mirrored) cumulative level distribution (CLD) of speech. Furthermore, Studebaker and Sherbecoe (1999, 2002) showed with speech intelligibility experiments that the “effective” dynamic range of speech is in the range between 40 dB to 50 dB rather than the commonly used 30 dB. Listeners were unable to reach 100 % speech intelligibility in the presence of noise with an SNR of +16 dB, and only reached 100 % speech intelligibility at an SNR of about +30 dB. The AAI (Aided Articulation Index; Stelmachowicz *et al.*, 1994) accounts for a change in dynamic range (due to compression) by dividing the dynamic range of speech in the SII calculation scheme by the “effective” compression ratio which leads to better prediction of the speech intelligibility of compressed speech in quiet (Souza and Turner, 1999). The distance between the weaker parts in the speech signal and the rms is decreased, which leads to more detectable speech information. Figure 7.9 shows the cumulative level distribution, derived from Figure 7.8. It shows that at a given level (e.g., 0 dB *re*: RMS) the compressed speech with CR=4 yields the highest amount of speech levels above that level (50 %), and gradually decreases with decreasing compression ratio (CR=2: 42 %, CR=1: 35 %, ER=1.5: 32 %). With uncompressed speech, normal-hearing listeners reach an SRT score of about -5 dB in stationary noise, which corresponds to an SII of 0.33 (33 % speech information available). If one assumes that the CLD is closely related to the IIF (Boothroyd, 2000), one should account for 33 % “effective” speech information at threshold in stationary noise. The correction factor for the uncompressed cumulative level distribution is 4.55 dB, i.e., a shift in the curve in Figure 7.9 of 4.55 dB to the left. If all curves in Figure 7.9 are shifted, at an SNR of -5 dB, the cumulative level distribution thus is for CR=1, CR=2, CR=4 and ER=1.5, 33 %, 39 %, 46 %, and 30 %, respectively. The cumulative level

## Speech Dynamic Range

distribution predicts lower SRTs with compressed speech and higher with expanded speech.

Moore, Glasberg, and Stone (2003) examined why commercials are perceived being louder than regular programmes, despite the fact that they have the same rms. Their experiment showed that indeed compressed speech is perceived being louder than the uncompressed speech. Moreover, the present results suggest that also the speech intelligibility in stationary and interrupted noise for radio broadcasts, public amplification systems in trains, stadiums etc can be improved by compressing the speech modestly by a compression ratio of 2:1.



*Figure 7.8. Level distribution of the female speech for the four compression ratios. CR 1, CR 2, CR 4, and ER 1.5 denote compression ratio 1, 2, 4, and expansion ratio 1.5.*

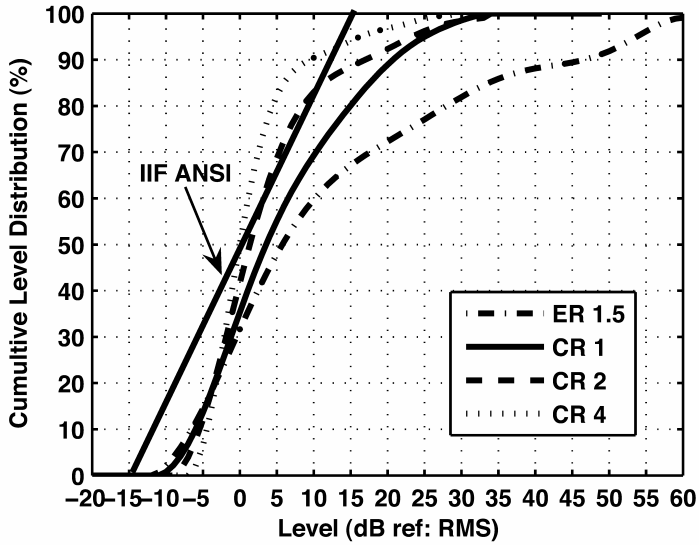


Figure 7.9. Cumulative level distribution of the female speech for the four compression ratios. CR 1, CR 2, CR 4, and ER 1.5 denote compression ratio 1, 2, 4, and expansion ratio 1.5. IIF-ANSI denotes the original IIF of the SII.

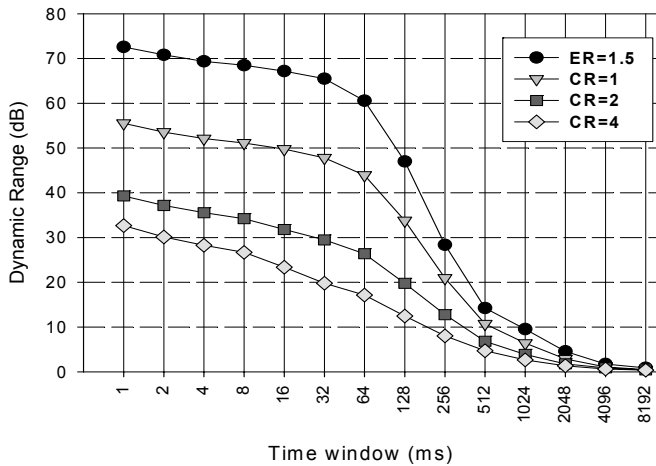


Figure 7.10. The broadband (100-8500Hz) dynamic range for the four conditions with compressed speech obtained with a sliding rectangular temporal window with an integration time ranging from 1 ms to 8192 ms. CR=1, CR=2, CR=4, and ER=1.5 denote compression ratio 1, 2, 4, and expansion ratio 1.5.

#### **IV. SII Model predictions**

A detailed description of the Extended SII (ESII) model is given in Rhebergen, *et al.*, (2006b). The basic principle of the ESII model is that it determines the amount of speech information that is still accessible when speech is partly masked by noise or partly inaudible due to the threshold of hearing. In the model, both speech and noise signal are filtered in 21 frequency bands. The envelope of the speech signal and noise signal are modified such to account for forward masking. Forward masking implies that rapid offsets in the envelope are smoothed, since the auditory system is incapable of following such fast transitions. After this, the speech and noise signal are partitioned into small time frames of 4 ms in length. Within each time frame, the conventional SII is determined, yielding the speech information available to the listener at that time frame. This results in an SII that changes over time, the so-called instantaneous SII. The averaged instantaneous SII finally results in a number between zero and unity, where  $SII=0$  means that no speech information is available to the listener, and  $SII=1$  means that all speech information is audible.

Here, the speech intelligibility has been calculated with the ESII (Rhebergen, *et al.*, 2006b), using the SPIN 21 critical band weighting function (ANSI S3.5-1997, 1997, Table B.1). The ESII calculations have been conducted with the long term average speech spectrum of the (compressed if appropriate) female target speaker. In order to approach the sound level at the ear drum (as required by the SII model), the speech and noise have been filtered with a 5<sup>th</sup> order FIR filter with the characteristics of TDH39P headphones. Also, background noise present in the sound proof booth has been added as an additional noise signal.

If one assumes that, at the threshold of intelligibility (i.e., the SRT) for all conditions the same proportion of speech information is available to the listener, then a correct model should yield calculated SII values that are the same for these different conditions. With normal-hearing listeners, and with everyday sentences as speech materials, the SII score is between 0.3 and 0.4 (Versfeld and Dreschler, 2002; Rhebergen and Versfeld, 2005).

## A. ESII Model calculations

To examine the effect of the dynamic range of speech, SII calculations have been performed with three different Intensity Importance Functions (IIF). First, the default IIF of the ANSI S3.5-1997 model has been used. This IIF does not account for any changes in the dynamic range of speech, and is a simple linear function with a dynamic range fixed at 30 dB, ranging from -15 dB to +15 dB *re* the rms of speech. Second, the IIF approach of the AAI model has been used (Stelmachowicz *et al.*, 1994). A detailed description of the AAI model is found in Stelmachowicz *et al.*, (1994) or Souza and Turner (1999). The IIF of the AAI is obtained by dividing the original IIF of the SII (the simple linear function) by the effective compression ratio of the compressed speech signal. The third IIF is obtained by taking the cumulative level distribution (CLD) of the speech signal (see Figure 7.9), together with a shift in rms of 4.55 dB, such to ensure that the SRT for uncompressed speech in stationary speech noise results in -5 dB.

The dynamic range of the speech signal is calculated with broadband speech (100-8500 Hz). Figure 7.10 shows the dynamic range of the four different compression conditions as a function of time window. Note that the dynamic range for the original speech signal (CR=1) is lower than that in Figure 7.5. This is due to the fact that here the dynamic range has been determined for the broadband signal, whereas in figure 7.5 it has been determined for the 1/3 octave band signal.

For the AAI model, the effective compression ratio was obtained by dividing the dynamic range at 1 ms time window of the original speech signal (CR=1) by the dynamic range (also 1 ms time window) of the compressed signal. The effective compression ratio for every condition is given in the fourth column of Table 7.1. The ESII calculations (Rhebergen, *et al.*, 2006b) with the three different IIFs are displayed in Table 7.1, column 5 to 7. The mean SII score for the default IIF (5<sup>th</sup> column) is 0.36, with a standard deviation of the SII between conditions of 0.04. Similarly, the AAI-IIF (6<sup>th</sup> column) results in an average SII of 0.34, and standard deviation of 0.09. The SII results with the CLD-IIF are given in the seventh column, and are on average 0.36 (standard deviation of 0.05).

The SII model with the original IIF predicts the data best: variations (i.e., standard deviation) in SII values between conditions are smallest. Next best is

## Speech Dynamic Range

the SII model with the CLD-IIF, and the SII model with the AAI-IIF gives the worst predictions. Thus, with the present data, refinement of the IIF does not result in better predictions.

*Table 7.1.*

Noise condition	SRT	stdv	Eff.CR	Linear IIF	AAI-IIF	CLD-IIF
Stationary ER 1.5	-3.2	1.3	0.76	0.39	0.42	0.36
Stationary CR 1	-4.5	0.9	1.00	0.35	0.35	0.35
Stationary CR 2	-6.6	0.8	1.41	0.28	0.19	0.29
Stationary CR 4	-5.4	1.0	1.70	0.33	0.21	0.44
Interrupted ER 1.5	-14.7	2.6	0.76	0.40	0.41	0.32
Interrupted CR 1	-17.3	2.5	1.00	0.38	0.38	0.36
Interrupted CR 2	-18.9	3.2	1.41	0.36	0.37	0.37
Interrupted CR 4	-18.4	3.1	1.70	0.37	0.38	0.40
<b>Mean ALL</b>				0.36	0.34	0.36
<b>stdv</b>				0.04	0.09	0.05
<b>Mean stationary</b>				0.34	0.29	0.36
<b>stdv</b>				0.05	0.11	0.06
<b>Mean Interrupted</b>				0.38	0.39	0.36
<b>stdv</b>				0.02	0.02	0.03

## V. Experiment II

### A. Method

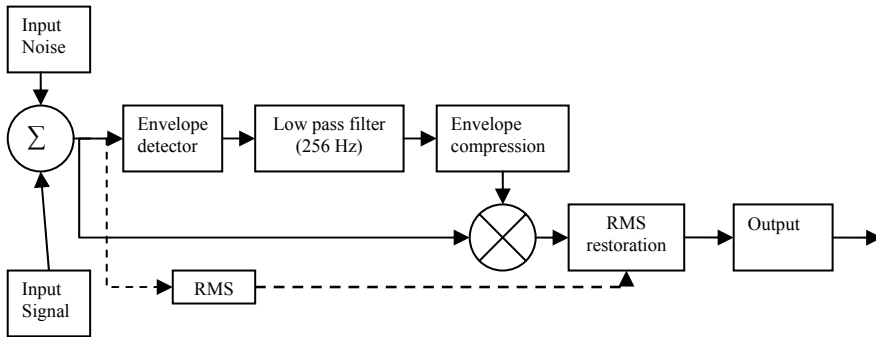
Compressing only the speech while leaving the noise unaltered is a condition that can only be achieved in a laboratory setting. In real life, speech and noise are mixed before they are fed to the compressor (e.g., hearing aid). The present experiment was designed to assess how the SRT changes when both speech and noise are compressed simultaneously.

### 1. Subjects

The eight normal-hearing subjects who participated in Experiment I (3 male, 5 female) also participated in Experiment II.

### 2. Stimuli

The target speech material consisted of short every-day sentences, uttered by the same female as in Experiment I (Versfeld *et al.*, 2000). The interfering noise conditions were the same as in Experiment I, and comprised one condition with 8Hz interrupted noise with a duty cycle of 50 % and one condition in stationary noise. All noises had a spectrum equal to the long-term average spectrum of the target speaker. The speech-in-noise mix was compressed by means of instantaneous Wide Dynamic Range Compression (WDRC) with compression ratios (CRs) of 4:1, 2:1, and 1:1, or with an expansion ratio of 1:1.5. With CR=1:1, the signal remained unaltered. The same WDRC signal processing is used in Experiment I, except the compression now is conducted on the speech-in-noise mix (see Figure 7.11).



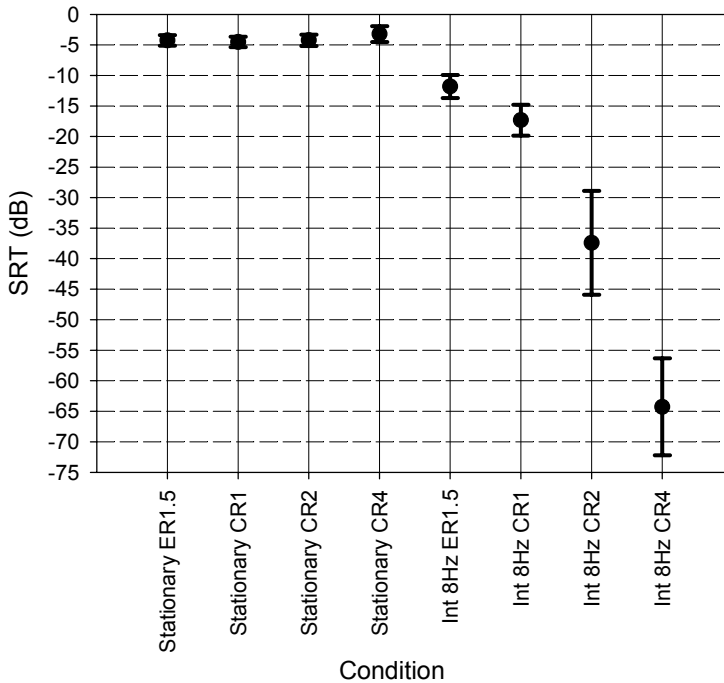
*Figure 7.11. Block diagram of WDRC signal processing. Dashed lines indicate operations on the RMS level of the speech signal.*

### 3. Procedure

The procedure has been described in Experiment I. In total, eight conditions were tested (test & retest): four conditions in stationary noise with WDRC on the speech-in-noise mix and four interrupted noise conditions (8 Hz interrupted noise) with WDRC on the speech-in-noise mix. To avoid confounding of

## Speech Dynamic Range

measurement condition order and sentence lists, the order of conditions and sentence lists was counterbalanced across subjects according to a Latin Square method. In total, each subject received 16 lists of 13 sentences.



*Figure 7.12. Speech Reception Threshold (dB) as a function of compression ratio and noise condition. Error bars denote the standard deviation between subjects.*

## B. RESULTS

Figure 7.12 displays the SRT-values averaged across subjects as a function of compression ratio in 8 Hz interrupted noise or in stationary noise. Error bars denote the standard deviation between subjects. A 2 [noises]  $\times$  4 [speech conditions]  $\times$  2 [test/retest]  $\times$  8 [subject] Analysis Of Variance (ANOVA) was performed on the data set. Of the main effects, “noises” was significant ( $F[1,7]=423.69$ ,  $p<0.001$ ), and “speech conditions” was significant

( $F[3,21]=194.29$ ,  $p<0.001$ ). Differences between “test” and “retest” were not significant ( $F[1,7]=2.227$ ,  $p>0.05$ ), which was also true for “subject” ( $F[7,7.05]=0.95$ ,  $p>0.5$ ). Of the interactions, “noises x subject” ( $F[7,14.3]=2.85$ ,  $p<0.05$ ) was significant.

Also, separate 4[speech conditions] x 2[test/retest] x 8[subject] ANOVAs were performed on the data obtained with stationary noise and with interrupted noise. For the stationary noise conditions, only the main effect of “speech conditions” was significant ( $F[3,21]=6.271$ ,  $p<0.005$ ). Post hoc tests (Tukey HSD) showed that the speech condition with CR=4 significantly differed from all other speech condition. For the interrupted noise conditions, the main effects of “speech conditions” was significant ( $F[3,21]=232.75$ ,  $p<0.001$ ). There were no significant interactions. Post hoc tests (Tukey HSD) showed that the speech conditions with CR=1, CR=2, and CR=4 significant differ, and the speech condition with ER=1.5 significantly differed from the speech conditions with CR=2, and CR=4.

### C. Discussion

The present experiment shows that, given the same masking condition, the SRTs with four different compression ratios differ from each other. On average SRTs in interrupted noise are much better than in stationary noise. The exact size of the effects are complicated by the fact that the SRT's displayed in Figure 7.12 are expressed in SNRs *before* the compression scheme was applied to the speech-in-noise. The “apparent” SNR of the speech-in-noise *after* compression depends on the used compression scheme and masking condition. Souza *et al.*, (2006) published a method introduced by Hagerman & Olofsson (2002) to calculate the “apparent” SNR after the speech-in-noise is compressed. For a detailed description, see Souza *et al.*, (2006). The basic technique is as follows: Three speech-in-noise signals are fed to the compressor; (1) the original speech and original noise mixed at a given SNR, (2) the original speech and phase inverted noise mixed at the same SNR as the original signal, and (3) the original noise and phase inverted speech mixed at the same SNR as the original signal. *After* signal compression, files 1 and 2, are added to cancel the noise, thus leaving only the speech (“apparent” speech), and files 1 and 3 are added to cancel the speech, thus leaving only the noise (“apparent” noise). The

“apparent” SNR then is defined as the level difference between the “apparent” speech and “apparent” noise. With this technique, Souza *et al.*, (2006) showed that the “apparent” SNR for speech in stationary noise is changed after compression. Of course, this approach is only a linear approximation, since (non-linear) compression causes distortion terms. As long as these distortion products are relatively small, this technique is allowed. With the same method the “apparent” SRT was calculated for the present data. The observed SRTs in stationary noise for ER=1.5 (-4.3 dB), CR=1 (-4.5 dB), CR=2 (-4.3 dB) and CR=4 (-3.3 dB) yield “apparent” SNRs of -3.6, -4.5, -4.7, and -3.9 dB, respectively. As with the observed SRTs in stationary noise with compressed speech, the lowest SRT is obtained for CR=2. As expected, CR=1 yield the same SRT, but CR=4 and ER=1.5 result in higher SRTs. Apparently, the compressed speech-in-stationary noise mix with these compression schemes decreases (i.e., worsens) the SNR, except for CR=1.5.

For the conditions with speech in interrupted noise, the “apparent” SNR has been calculated with the same procedure. The observed SRTs for ER=1.5 (-11.8 dB), CR=1 (-17.3 dB), CR=2 (-37.4 dB) and CR=4 (-64.3 dB) transform to “apparent” SNRs of -13.9, -17.3, -21.9, and -31.5 dB, respectively. The lowest SRT is obtained with CR=4, followed by CR=2, CR=1, and ER=1.5. The “apparent” SNRs show less spread/variability than the original SNRs.

## VI. Analysis of the speech-in-noise signal

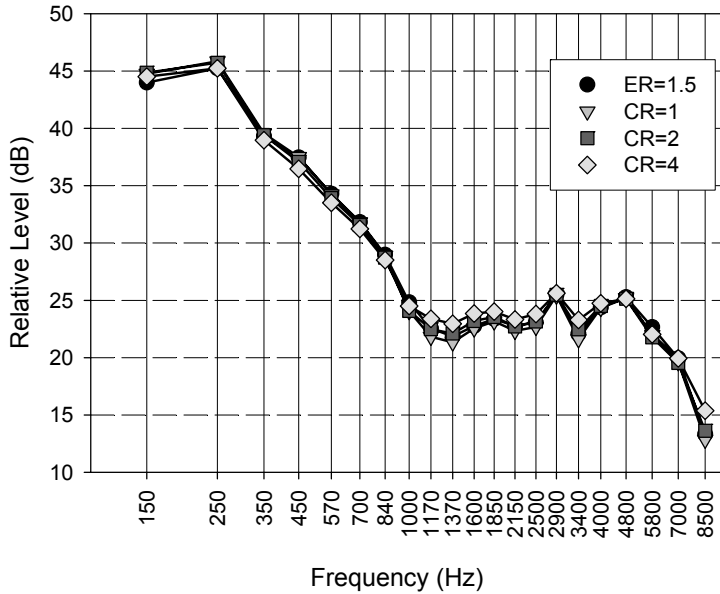
In order to understand the implications of the manipulations on the speech-in-noise signal, the present section examines in detail the spectro-temporal characteristics of the (compressed or expanded) speech signal.

The speech-in-stationary noise is compressed with the four compression schemes at the mean SRT. The compressed speech signal is obtained with the technique of Souza *et al.* (2006). The analysis of the speech signal is conducted with the same technique as explained in section III.

The speech-in-interrupted noise is not examined in this section. 50 % of the time it has resemblance with compressed speech in stationary noise, and 50 % of the time with compressed speech as in Experiment I.

## A. Long-term average speech spectrum after Speech-in-Noise compression

The LTASS for each of the four compression ratios for the female speech is displayed in Fig 7.13. Contrary to the effects of compression and expansion on the speech signal only (see Figure 7.3), the spectrum changes only marginally when the compression and expansion process the speech-in-noise signal.



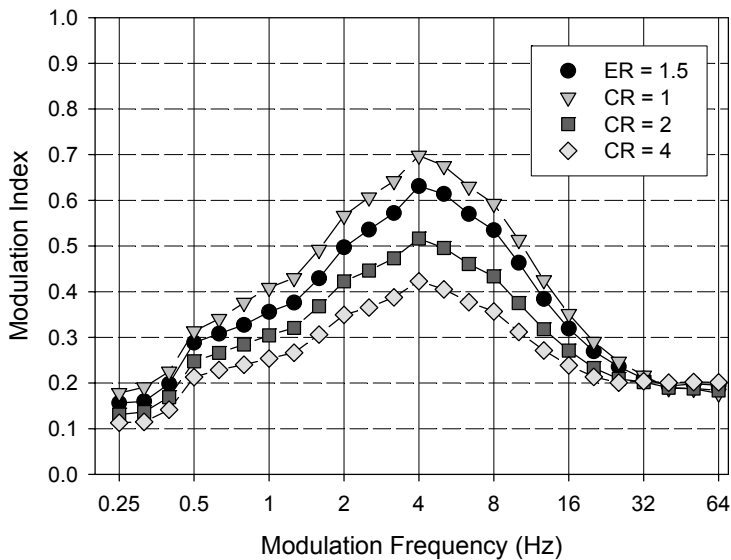
*Figure 7.13.* Critical band spectrum of speech with a CR of 1:1, 2:1, 4:1 or expansion ratio of 1:1.5.

## B. Modulation Index of speech after Speech-in-Noise compression

Figure 7.14 shows for the four conditions the modulation spectrum as calculated in 21 critical bands (150 to 8500 Hz; ANSI S3.5-1997), and averaged across the 21 frequency bands for the specific modulation frequency. The mean

## Speech Dynamic Range

modulation spectrum of the female speech has a peak at about 4 Hz, and has the same shape as the modulated speech signal in Experiment I. The obtained speech signal with CR=1 (no compression) has the same modulation spectrum as in Experiment I. The method introduced by Souza *et al.* (2006) thus leaves the speech signal unaltered. If one assumes that the method also has no effect on compressed speech-in-noise signal, then for all compressed speech-in-noise signals the modulation index of the speech is decreased compared with the compressed and expanded speech in Experiment I. In this respect, Plomp's (1988) hypothesis that intelligibility is coupled to the modulation index is valid: for speech in noise uncompressed speech results in the highest modulation index, and even expansion seems to decrease the modulation index.



**Figure 7.14.** Modulation index of speech-in-noise signals as function of modulation frequency. Compression ratios are CR=1, CR=2, CR=4, and expansion ratio ER=1.5.

### **C. Short-term level distribution of speech after Speech-in-Noise compression**

Figure 7.15 shows, for the four conditions, the band passed (100-8500Hz) level distribution relative to the rms, obtained with a rectangular temporal window with an integration time of 1 ms. In spite of the reduction of the modulation index due to compression and expansion of the four speech-in-noise conditions, the level distribution for all four speech signals is more or less the same. This is especially clear in the cumulative level distribution (CLD) of the four speech signals (Figure 7.16). Compared with the compressed speech signals in Experiment I, the present four CLD functions are very similar between 0 to about 70 % cumulative speech level. The expected differences for speech intelligibility in stationary noise are small on the basis of these cumulative level distribution curves.

The outcomes of the long-term average speech spectrum, the modulation spectra, and the level distributions raise the question whether the modulation index is changed by the effectiveness of the compression scheme, or whether it is changed due to increased distortion of the speech signal, hence an artifact of the calculation scheme of Souza (2006). Informal listening to the compressed speech signals of Experiment I (only speech compressed) and the extracted compressed speech signals of Experiment II (speech-in-noise compressed) revealed clear speech signals in Experiment I, and with the exception of CR=1, more noisier speech signals in Experiment II. Thus, with exception of uncompressed speech, the speech signals of Experiment I and Experiment II are dissimilar. The noisier speech signals in Experiment II are probably a result of the interaction between the (non-linear) speech-in-noise compression, and the linear approximation technique of Souza (2006).

## Speech Dynamic Range

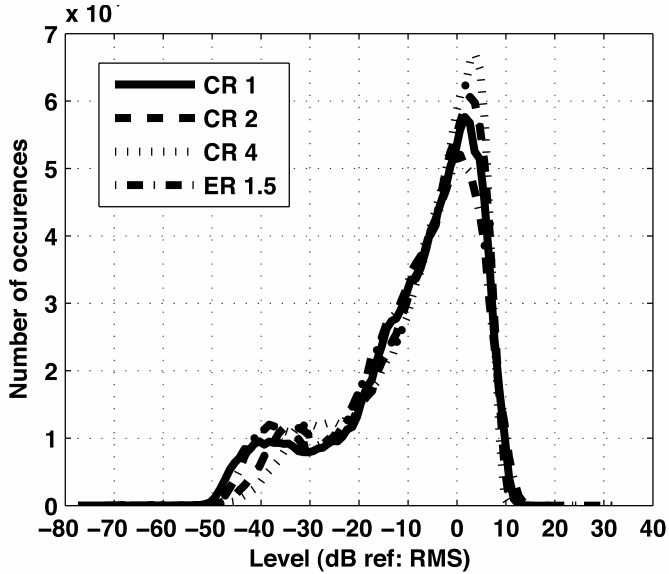


Figure 7.15. Level distribution of the female speech for the four compression ratios. CR 1, CR 2, CR 4, and ER 1.5 denote compression ratio 1, 2, 4, and expansion ratio 1.5.

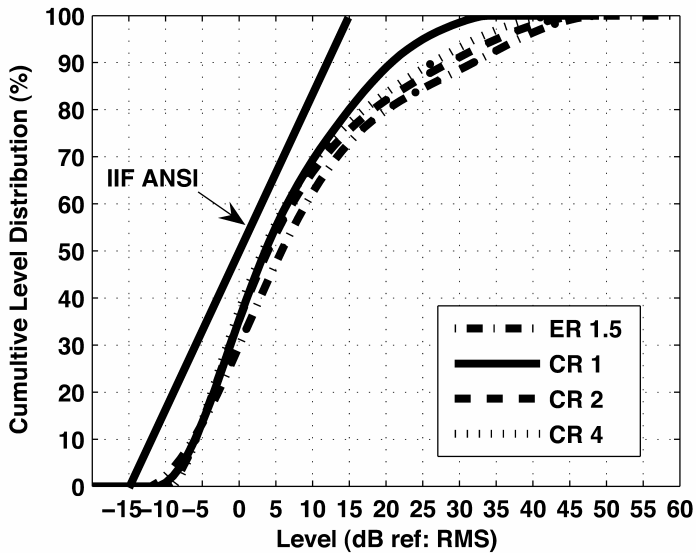
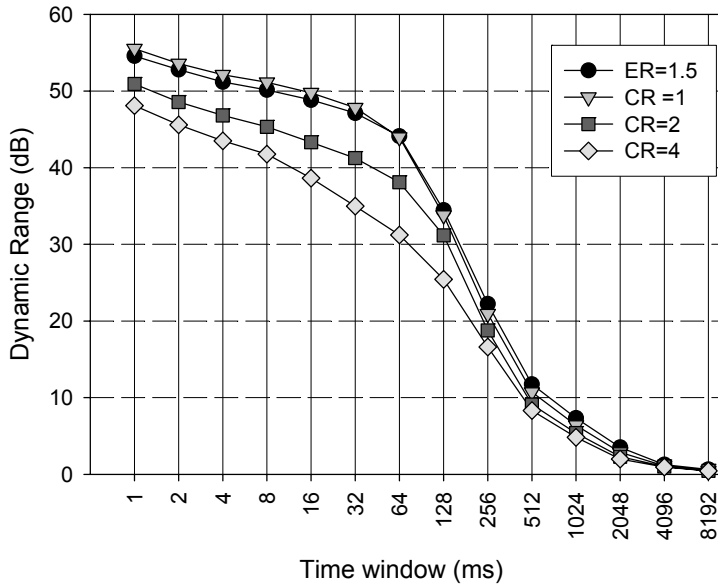


Figure 7.16. Cumulative level distribution of the female speech for the four compression ratios. CR 1, CR 2, CR 4, and ER 1.5 denotes compression ratio 1, 2, 4, and expansion ratio 1.5. IIF-ANSI denotes the original (linear) IIF of the SII.



*Figure 7.17.* The broadband (100-8500Hz) dynamic range for the four conditions with compressed speech-in-noise obtained with a sliding rectangular temporal window with an integration time ranging from 1 ms to 8192 ms. CR=1, CR=2, CR=4, and ER=1.5 denote compression ratio 1, 2, 4, and expansion ratio 1.5.

## VII. ESII Model calculations

The ESII calculations are the same as described in Experiment I.

The ESII for stationary noise has been calculated with three different Intensity Importance Functions (IIF) to account for the dynamic range of speech. First, the original (linear) IIF of the SII (ANSI S3.5-1997), second, the IIF approach of the AAI model (Stelmachowicz *et al.*, 1994), and third the IIF is obtained by taking the cumulative level distribution (CLD) of the speech signal (see Figure 7.16), together with a shift in rms of 4.55 dB, such to ensure that the SRT for uncompressed speech in stationary speech noise results in -5 dB.

## *Speech Dynamic Range*

The ESII calculations in interrupted noise are calculated in a similar way, except for that the AAI-IIF and CLD-IIF have been derived from the speech signal as present in the gaps of the interrupted noise. The shape of the IIFs thus is the same as that of Experiment I. The calculated broadband dynamic range of speech for speech-in-noise is shown in Figure 7.17. Compared with the dynamic range of the compressed speech signals in Experiment I, the dynamic range of the present speech signal is less affected by compression and expansion. For the AAI model, the effective compression ratio was obtained by dividing the dynamic range at 1 ms time window of the original speech signal (CR=1) by the dynamic range (also 1 ms time window) of the compressed signal. The effective compression ratio for every condition is given in the fifth column of Table 7.2.

The ESII calculations (Rhebergen, Versfeld, and Dreschler, 2006b) with the three different IIFs are displayed in Table 7.2, column 6 to 8. The mean SII score for the default IIF (6<sup>th</sup> column) is 0.34, with a standard deviation of the SII between conditions of 0.067. Similarly, the AAI-IIF (7<sup>th</sup> column) results in an average SII of 0.34, and standard deviation of 0.072. The SII results with the CLD-IIF are given in the seventh column, and are on average 0.33 (standard deviation of 0.046).

The SII model with the CLD- IIF predicts the data most consistently: variations (i.e., standard deviation) in SII values between conditions are smallest. Next best is the SII model with the default IIF, and the SII model with the AAI-IIF gives the poorest predictions. The total mean ESII calculations for Experiment I & II are displayed at the bottom of Table 7.2. The ESII calculations with the CLD- IIF predict the data most consistently: variations (i.e., standard deviation) in SII values between conditions are smallest, followed by the default IIF, and the ESII calculations with the AAI-IIF gives the poorest predictions.

Table 7.2

Noise condition	SRT	stdv	App.SNR	Eff.CR	Linear IIF	AAI - IIF	CLD- IIF
Stationary ER 1.5	-4.3	0.9	-3.6	1.02	0.38	0.37	0.33
Stationary CR 1	-4.5	0.9	-4.5	1.00	0.35	0.35	0.35
Stationary CR 2	-4.2	0.9	-4.7	1.09	0.34	0.33	0.34
Stationary CR 4	-3.2	0.9	-3.9	1.15	0.37	0.34	0.40
Interrupted ER 1.5	-11.8	1.9	-13.7	0.76	0.39	0.41	0.31
Interrupted CR 1	-17.3	2.5	-17.3	1.00	0.38	0.38	0.36
Interrupted CR 2	-37.4	8.5	-21.9	1.41	0.34	0.36	0.36
Interrupted CR 4	-64.3	8.0	-31.5	1.70	0.18	0.17	0.24
<b>Mean ALL</b>					0.34	0.34	0.33
<b>stdv</b>					0.08	0.07	0.05
<b>Mean stationary</b>					0.36	0.35	0.35
<b>stdv</b>					0.02	0.02	0.03
<b>Mean Interrupted</b>					0.32	0.33	0.32
<b>stdv</b>					0.10	0.11	0.06
<b>Mean Exp. I &amp; II</b>					0.35	0.34	0.35
<b>stdv</b>					0.05	0.08	0.05

## VIII. Discussion

### Dynamic range of speech

As mentioned earlier in this paper, there is still no consensus on the physical dynamic range of the speech signal. For a large part this is due to the specific parameter choice to estimate the speech dynamic range. Psychological measurements by Studebaker and Sherbecoe (2002) indicate that the effective dynamic range of uncompressed speech is between 40 to 50 dB.

For compressed speech, it is important to account for the dynamic range of the speech signal used. The dynamic range, the level distribution and its cumulative level distribution have been calculated with a 1 ms sliding temporal window. It is not evident that this window length is the best choice to obtain

## *Speech Dynamic Range*

the physical properties of the speech signal. On the other hand, normal-hearing listeners can detect gaps of about 2 ms at high frequencies to about 18 ms at low frequencies. A rectangular sliding temporal window with a length of 125 ms will give a physical dynamic range of about 32 dB with broad band speech and about 43 dB with 1/3 octave filtered speech. With an exponential sliding temporal window with the same lengths the calculated physical dynamic range will decrease even more. If one considers that the effective dynamic range is about 45 dB (Studebaker and Sherbecoe, 2002), then the physical dynamic range of speech may be expected to be much larger, and probably is more in line with the 68 dB range of Fletcher and Galt (1950). For model predictions, a speech dynamic range larger than the assumed 30 dB is needed to improve the speech intelligibility predictions (Rankovic, 1998; Studebaker and Sherbecoe, 2002; Dubno *et al.*, 2005).

## **Intensity Importance Function**

The IIF functions used in the AI (ANSI S3.5-1969), STI, and SII models are linear functions and have been based on a first order best fit, not on psychophysical measurements (Houtgast, 2005; Pavlovic, 2005; and Studebaker, 2005). Studebaker and Sherbecoe (2002) calculated the IIF function for NU#6 (Studebaker and Sherbecoe, 1993) speech material and concluded that the shape of the IIF is non-linear. Boothroyd (1990, 2000) suggested that the IIF function of a speech signal is related to its cumulative level distribution. Indeed, the cumulative level distribution of the NU#6 speech material has a high resemblance with the obtained IIF function by Studebaker and Sherbecoe (2002).

In this paper, the IIF is equal to the cumulative level distribution of the compressed speech signals. Although these IIFs are not based on the psychophysical measurements as conducted by Studebaker and Sherbecoe (2002), the predicted SII scores for the compressed speech conditions are more consistent than with the linear IIF function of the SII and AAI. Further research must be conducted to examine the relation between the cumulative level distribution of (compressed) speech and the IIF. If one considers the fact that the level distribution is not identical for different speakers, this approach has the potential to model the speech intelligibility as function of speaker more

accurately. For this reason it is worthwhile to examine the effect of compression of the dynamic range of speech for different speakers.

## **Speech-in-noise compression**

The method introduced by Hagerman and Olofsson (2002) and published by Souza *et al.* (2006) can give insight into the effectiveness of compression on a speech-in-noise signal. The present paper shows with the aid of this technique that the dynamic range of speech alone is more effectively compressed than speech-in-noise. Furthermore, compression reduces the modulation index for all speech conditions, whereas the CLD is more or less the same for all speech conditions. This observation raises the question whether this effect is due to the different techniques applied for the calculations of the modulation index (filtered in 1/3 octave bands) and the CLD (broadband speech), or due to the possible limitations in the technique of Souza *et al.* (2006) to analyze compressed speech. Compression or expansion introduces distortion products in the speech signal and these distortion products presumably have a larger effect on the technique of Souza *et al.* (2006) to calculate the speech signal than on the speech signal itself. Even expanded speech has a lower modulation index compared with the other speech conditions. This observation suggests that this (linear) technique should be used with caution in case of non-linearly processed speech in noise.

With regard to the calculated “apparent SNR”, the same consideration should be taken into account. Given the basic assumption that at threshold the SII-values for different noise conditions should be the same, then the calculated “apparent SNR” in interrupted noise conditions with CR=4 is too low. For noise conditions with CR=2 the SII value is about the same as in noise conditions with CR=1. Thus, with an instantaneous WDRC scheme (CR=2), the technique of Souza *et al.* (2006) appears to result in a reasonably good prediction of the “apparent SNR”, but for higher compression ratios deviations may be found.

## **Effects on speech intelligibility**

For the SRT in interrupted noise the speech intelligibility increases with compression. In stationary noise, subjects perform equally well for CR=2 and tend to perform worse for CR=4. These results showed that better speech

intelligibility is possible in interrupted noise for normal-hearing subjects and the compression scheme used (in this chapter WDRC with CR=2). Whether this compression scheme can improve the speech intelligibility for hearing-impaired subjects is left for further research.

## **IX. Summary & conclusions**

The present paper describes two SRT experiments with normal-hearing subjects to examine:

- a) The effects of the dynamic range of the speech signal on speech intelligibility in stationary and interrupted noise. With aid of an adequate Intensity Importance Function, the Extended SII model is capable to predict the speech intelligibility for compressed or expanded speech-in-noise.
- b) The effects of the compressed speech-in-noise on speech intelligibility in stationary and interrupted noise. The speech intelligibility for compressed speech is not improved compared to uncompressed, or expanded speech-in-stationary noise, but speech intelligibility does improve significantly for speech-in-interrupted noise. In stationary noise, expansion has a negative effect on the speech intelligibility.

## *Chapter 8*

# Predicting the Speech intelligibility in fluctuating noise in hearing-impaired listeners

*Koenraad S. Rhebergen, Niek J. Versfeld and Wouter A. Dreschler*

## **Abstract**

The validated Extended Speech Intelligibility Index (ESII) model of Rhebergen, Versfeld, and Dreschler (2006b) forms an extension to the conventional Speech Intelligibility Index (SII) model, and is able to predict for normal-hearing listeners the speech intelligibility in both stationary and non-stationary noise maskers with sufficient accuracy. The ESII model has been validated with the aid of Speech Reception Threshold (SRT) experiments in normal-hearing listeners. In the present paper, a first attempt is made to evaluate the ESII with SRT data of normal-hearing and hearing-impaired listeners in stationary noise and 10 Hz interrupted noise measured at three different noise levels. Data have been measured by de Laat and Plomp (1983). The results show that the ESII model is able to predict the speech intelligibility reasonably well, but that predictions are better when the ESII model includes a function that takes cochlear compression into account.

## I. Introduction

It is well known that hearing-impaired listeners have more difficulties in following a conversation in background noise compared to normal-hearing listeners. This difference in speech intelligibility is most prominent in fluctuating background noises (de Laat and Plomp, 1983; Festen and Plomp, 1990; Hygge *et al.*, 1992; Peters *et al.*, 1998; Versfeld and Dreschler, 2002; Festen and Plomp, 2002; Nelson *et al.*, 2003). One way to determine the ability to understand speech in these situations is to measure the speech reception threshold (SRT) for sentences in noise (Plomp and Mimpen, 1979). The SRT is defined as the signal-to-noise ratio (SNR) needed for 50 % sentence intelligibility. Normal-hearing listeners are able to reach lower (i.e., better) SRTs in fluctuating noise compared to the SRTs in stationary noise with the same RMS level. Normal-hearing listeners are able to “listen into the gaps of the fluctuating background noise”. Hearing-impaired listeners perform hardly better in fluctuating noise conditions compared with stationary noise conditions. Contrary to normal-hearing listeners, hearing-impaired listeners are unable to benefit from the gaps in the fluctuating background noise to gain a better SRT score. Until recently, the AI (Articulation Index; ANSI S3.5-1969) and its successor the SII (Speech Intelligibility Index; ANSI S3.5-1997) were unable to predict the speech intelligibility in fluctuating masking noises. Since the SII calculations are based on the long-term average spectrum of the noise and the speech, it does not take into account any fluctuations in the masking noise. As most real-life noises are more fluctuating than steady state in nature, a good method was needed to calculate the speech intelligibility in fluctuating noise. Rhebergen and Versfeld (2005) and Rhebergen, *et al.*, (2006b) introduced an extension to the SII model to predict the speech intelligibility in stationary and in fluctuating noises. This new calculation scheme computes the instantaneous SII in small time frames and takes the average to come to a final SII score. Rhebergen and Versfeld’s (2005) method is based on a variety of noise types available from the literature, whereas Rhebergen, *et al.*, (2006b) validation study is based on their own SRT data obtained with normal-hearing listeners for a range of fluctuating noises. The latter study led to an upgrade of the method proposed by Rhebergen and Versfeld (2005), such that it now can account for effects due to forward masking. This new SII extension has also the potential to

model hearing impairment, and to model the effect of overall noise level on speech intelligibility.

The goal of the current paper is to re-examine and model the SRT data reported by de Laat and Plomp (1983), who measured SRTs in young normal-hearing and young hearing-impaired listeners in stationary and interrupted noise at different noise levels.

## **II. Method**

In this section, the SRT data from normal-hearing (NH) and hearing-impaired (HI) subjects have been taken from the study of de Laat and Plomp (1983), and the raw data (individual SRTs and absolute hearing levels) of this study have been taken from the Ph.D thesis of de Laat (1989). The description of the SRT experiment has been taken from the method section of de Laat (1989), and extended, where needed, to give the reader an impression how the SRT data have been obtained.

### **a. Subjects**

Twenty hearing-impaired subjects participated in this experiment. Their age ranged from 13 to 19 years. Subjects were native speakers of the Dutch language and pupils of a highschool the for hearing-impaired. The pure-tone thresholds were measured for nineteen 1/3 octave frequencies with an adapted Békésy procedure (125 to 8000 Hz). The Pure Tone Average (PTA: 500, 1000, 2000, 4000 Hz) of the HI subject was on average 53 dB HL (with a standard deviation of 7.7 dB). Furthermore, ten NH subjects participated. They were matched to the hearing-impaired group with respect to age and education. Each normal-hearing subject had pure-tone thresholds of 15 dB HL or better at octave frequencies from 125 to 8000 Hz (PTA: 10 dB HL with a standard deviation of 2.7 dB).

### **b. Stimuli**

The target speech material consisted of short every-day sentences, uttered by a female speaker (Plomp & Mimpen, 1979). The speech material comprised 10 lists of 13 sentences that has been developed for a reliable measurement of the

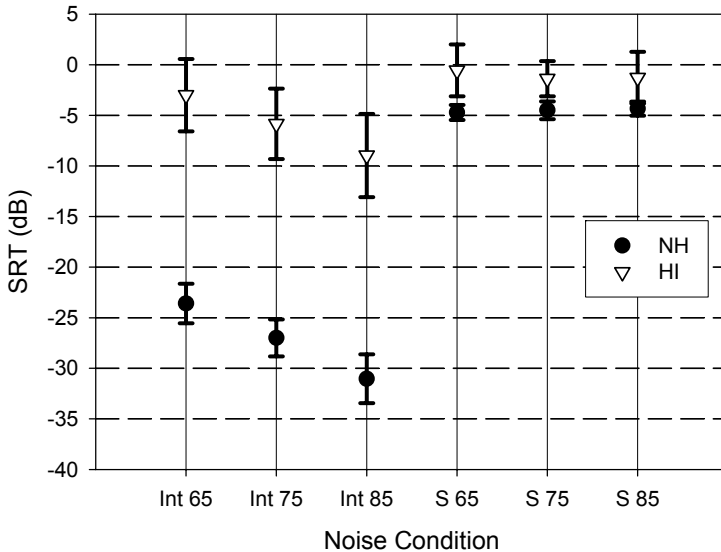
speech intelligibility in noise. The speech was stored at a sample rate of 15625 Hz with an 8 bits resolution. The interfering noise conditions consisted of one condition with stationary noise, and one condition with 100 % modulated 10 Hz interrupted noise with a duty cycle of 50 %. Both noise conditions had a spectrum equal to the long-term average spectrum of the target speaker.

### c. Procedure

Subjects were tested individually in a sound-insulated booth. The monaural speech-reception threshold (SRT) was measured at the better ear for a fixed noise level fixed at 65, 75, and 85 dBA. The speech and noise were presented via Beyer Dynamic DT 48 headphones. After the presentation of a sentence, the subject's task was to repeat the sentence he or she had just been presented. A sentence was scored as correct if all words in that sentence were repeated without any error. A list of 13 sentences, unknown to the subject, was used to estimate the level at which 50 % of the sentences was reproduced without any error, the so-called Speech Reception Threshold, or SRT. For a given condition, the first sentence of the list started far below the expected SRT. The sentence was repeated each time at a 4 dB higher level until the subject was able to reproduce it correctly. The twelve remaining sentences in that list were presented only once, following a simple up-down procedure with a step size of 2 dB. The SRT was estimated according to the procedure described by Plomp and Mimpen (1979), i.e., by taking the mean Signal to Noise Ratio (SNR) of sentences five to thirteen plus the estimated SNR that would have been used for the fourteenth sentence.

## III. Results

Figure 8.1 shows the SRT-values averaged across subjects for each of the six conditions of the interfering noise. A 2[subject group]\* 2[noise condition]\*3[level] Analysis Of Variance (ANOVA) was performed on the SRT dataset. Of the main effects, differences in "subject group" ( $F[1,168]=832.40$ ,  $p<0.0001$ ), differences in "noise condition" ( $F[1,168]=1032.88$ ,  $p<0.0001$ ), and differences in "level" were significant ( $F[2,168]=21.43$ ,  $p<0.0001$ ). The interactions "subject group\*noise condition" ( $F[1,168]=432.42$ ,  $p<0.0001$ ), and "level\*noise condition" ( $F[2, 168]=19.49$ ,  $p<0.0001$ ) were significant.



**Figure 8.1.** Speech Reception Threshold (dB) as function of interfering noise condition for normal-hearing (circles) and hearing-impaired (triangles) listeners. Int 65, 75 and 85 denote the conditions with 10 Hz interrupted noise fixed at 65, 75, and 85 dBA, respectively. S 65, 75 and 85 denote the conditions with stationary noise fixed at 65, 75, and 85 dBA, respectively. Error bars denote the standard deviations between subjects.

A 10[subject]\* 2[noise condition] \*3[level] ANOVA was performed on the dataset obtained with the normal-hearing subjects. Of the main effects, differences in “noise condition” ( $F[1,9]=2043.04, p<0.0001$ ), and differences in “level” were significant ( $F[2, 18]=39.70, p<0.0001$ ). Of the interactions, only “level\* noise condition” ( $F[2, 18]=35.22, p<0.0001$ ) was significant. Post Hoc tests (Tukey HSD) showed significant differences between the three interrupted noise conditions, and no significant differences between the three stationary noise conditions. Furthermore, SRTs observed in interrupted noise were significantly different from the stationary noise conditions.

A 20[subject]\* 2[noise condition] \*3[level] ANOVA was performed on the dataset obtained with the group of hearing-impaired subjects. Of the main effects,

differences in “subject” ( $F[19, 21.13]=5.95, p<0.0001$ ), differences in “level” ( $F[2, 38]=38.68, p<0.0001$ ), and differences in “noise condition” were significant ( $F[1, 19]=114.56, p<0.0001$ ). The interactions “subject \*noise condition” ( $F[19, 38]=3.03, p<0.002$ ), and “level\*noise condition” ( $F[2, 38]=34.25, p<0.0001$ ) were significant. Post Hoc tests showed no significant differences between the three stationary noise conditions. A significant difference was found between the 10 Hz interrupted noise conditions at 65 dBA and 85 dBA. Furthermore, with the 10 Hz interrupted noise, SRTs obtained at 75 dBA and 85 dBA were significantly different from the other noise conditions.

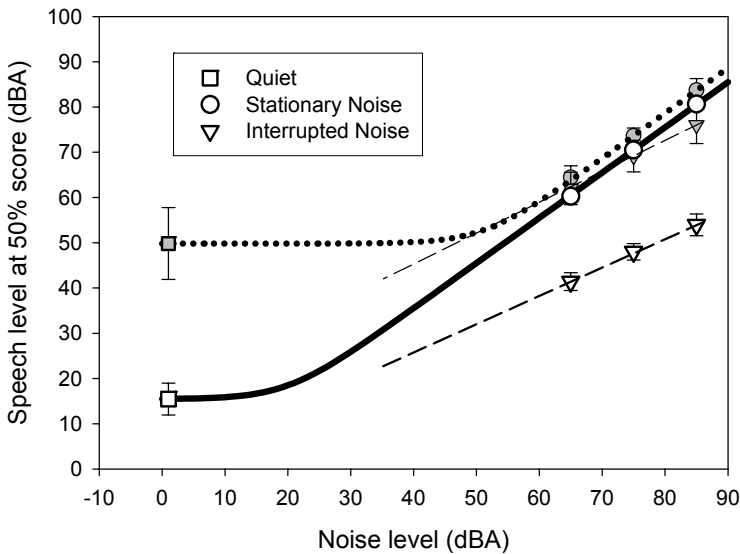
The results show that with interrupted noise, each 10-dB increase of the noise level results in an about 4 dB decrease in SRT for normal-hearing subjects, and 3 dB decrease for hearing-impaired subjects. With stationary noise, the increase in noise level did not produce a significant effect.

#### IV. Discussion

One of the main findings from this experiment is the relative poor performance of hearing-impaired subjects in interrupted noise. In stationary noise the average loss in terms of the critical SNR is about 3.1 dB, whereas the difference between normal-hearing and hearing-impaired subjects can be as large as 22.1 dB in 10-Hz interrupted noise at high presentation levels.

Figure 8.2 shows the effects of noise level and noise type (quiet, stationary noise, and 10Hz interrupted noise with a duty cycle of 50 %) on the speech levels required for 50 % sentence score. The lower (solid) curve and upper (dotted) curve represent the predicted speech levels in stationary noise, for normal-hearing listeners and hearing-impaired listeners, respectively, according to the SRT model of Plomp (1986). Where the SRT in quiet (SRT<sub>q</sub>) is expressed as the speech level in dBA, the SRT in noise can be measured expressed as the “critical” SNR in dB (e.g., SRT measured in stationary noise level of 65 dBA with speech level of 60 dBA gives an SRT score (SNR) of -5 dB). The two curves are calculated with the mean SRT<sub>q</sub> and the SNR at a stationary noise level of 75 dBA for the normal-hearing and hearing-impaired subjects. For both normal-hearing and hearing-impaired subjects, the measured speech levels in stationary noise can be predicted accurately. When the stationary noise level

is beyond a critical level (about 35 dBA for NH, and 65 dBA for HI subjects), for every dB increase of the noise level, one dB increase in speech level is required to yield 50 % sentence score. Thus by increasing the stationary noise level, the SNR to maintain 50 % sentence score is constant. Contrary to the SNR in stationary noise, the SNR in interrupted noise is dependent on the noise level. For both groups, SRT thresholds decrease with 3 to 4 dB per 10 dB increase in noise level for interrupted noise, resulting in a decrease of SRT in terms of the critical SNR, whereas SNR was essentially unchanged in stationary noise. The slope of the regression line for speech level in interrupted noise for NH and HI subjects is much shallower compared to the slope of the speech level in stationary noise. The SRT model for stationary noise (Plomp, 1986) is thus not suitable for SRT predictions in interrupted (non-stationary) noise.



**Figure 8.2.** Speech level required for a 50 % sentence score as function of noise level. Symbols denote the speech level measured in quiet, stationary noise, and 10 Hz interrupted noise for normal-hearing (open symbols), and hearing-impaired (filled symbols) listeners. Error bars denote the standard deviation between subjects. The solid curve represents the predicted speech level in stationary noise for normal-hearing listeners, and the dotted curve represents the predicted speech level in stationary noise for hearing-impaired listeners.

## Chapter 8

The latter effect might be a result of two factors: audibility and steeper temporal slopes at higher noise levels. For hearing-impaired subjects in interrupted noise, audibility may have played a certain role. To that end, de Laat and Plomp measured the SRT<sub>q</sub>, which for normal-hearing subjects turned out to be 15.5 dBA (standard deviation 3.5 dB), and for hearing-impaired subjects 49.8 dBA (standard deviation 7.9 dB). The fact that the speech levels at threshold in interrupted noise are 25 to 38 dB higher than in quiet for normal-hearing subjects, and 12 to 26 dB higher than in quiet for hearing-impaired subjects suggests that audibility in hearing-impaired subjects may have had an effect, but generally did not play a major role. Moreover, both groups benefited about an equal amount from the increase in noise level, whereas one would have expected a larger benefit if audibility would have been the most important factor.

An alternative manner to examine the effects of audibility is to calculate the correlation between the SRT and the PTA, given the individual data. For the entire group of subjects, correlations between PTA and SRT in interrupted noise were 0.95, 0.97, and 0.96 for noise levels of 65 dBA, 75 dBA, and 85 dBA, respectively. These correlations were highly significant ( $p < 0.001$ ), but when the groups were separated into a normal-hearing and a hearing-impaired group, the correlations dropped and were only significant for the hearing-impaired group (0.56, 0.73, and 0.67 for noise levels of 65 dBA, 75 dBA, and 85 dBA, respectively;  $p < 0.01$ ,  $p < 0.001$ , and  $p < 0.001$ , respectively). After regrouping the hearing-impaired group into moderate (40 – 60 dB HL;  $n=14$ ), and severe hearing impairment (60 – 80 dB HL;  $n=5$ ), a significant correlation was found only in the group with severe hearing impairment (0.92, 0.91, and 0.96 for noise levels of 65 dBA, 75 dBA, and 85 dBA, respectively;  $p < 0.028$ ,  $p < 0.034$ , and  $p < 0.011$ , respectively). Apparently, audibility due to the fact that portions of the speech fall below the absolute threshold plays only a significant role in the group with severe hearing impairment. Since this group of hearing-impaired subjects is rather small, individual calculated SII values for all subjects might give more insight in this matter.

## **V. Model predictions**

A detailed description of the Extended SII (ESII) model is given in Rhebergen, *et al.*, (2006b). As in the SII, the Extended SII predicts speech intelligibility on the basis of the auditory threshold, the spectrum of the noise and the spectrum of speech, but the basic principle of the Extended SII model is that fluctuations of the noise are taken into account. The envelope of the speech signal and noise signal is calculated in 21 frequency bands, and modified with a FMF (forward masking function) algorithm to account for forward masking. The slope of the FMF is dependent on the amount of hearing loss. As a result, the FMF is steeper for normal-hearing subjects than for subjects with a hearing loss. After the addition of forward masking, the speech and noise are partitioned into small time frames of 4 ms in length. Within each time frame, the conventional SII is determined, yielding the speech information available to the listener at that time frame. Next, the SII values of these (about 500) time frames are averaged, resulting in the SII for that particular condition.

The speech intelligibility was predicted with the ESII (Rhebergen, *et al.*, 2006b) using the SPIN 21 critical band weighting function (ANSI S3.5-1997, 1997, Table B.1). The SII calculations were conducted with the long term speech spectrum of the female target speaker. In order to approach the sound level at the ear drum (as required by the SII model), the speech and noises were filtered with a 5<sup>th</sup> order FIR filter with the characteristics of the Beyer Dynamic DT48 headphone.

### **Extended SII calculations**

The 3<sup>rd</sup> and 5<sup>th</sup> column of Table 8.1 show the results of the ESII calculations (Rhebergen, *et al.*, 2006b) based on the individual data for the normal-hearing and hearing-impaired subjects, respectively. Absolute thresholds at each 1/3-octave frequency were input to the ESII model. For the group of normal-hearing subjects, the mean SII across all noise conditions is 0.28, with a standard deviation of 0.06. Similarly, for the group of hearing-impaired subjects the SII is equal to 0.30, with a standard deviation of 0.09.

Chapter 8

**Table 8.1.** SRTs (dB) and the ESII calculations for each group and each noise masker. Lower rows yield mean and standard deviation of the ESII.

	NH (N=10)		HI (N=20)	
	SRT (dB)	ESII	SRT (dB)	ESII
<b>stationary @ 65 dBA</b>	-4.7 (0.7)	0.34 (0.03)	-0.7 (2.6)	0.21 (0.08)
<b>stationary @ 75 dBA</b>	-4.5 (0.9)	0.34 (0.03)	-1.4 (1.7)	0.28 (0.09)
<b>stationary @ 85 dBA</b>	-4.3 (0.7)	0.32 (0.02)	-1.3 (2.6)	0.34 (0.09)
<b>10Hz Int @ 65 dBA</b>	-23.6 (2.0)	0.21 (0.03)	-3.0 (3.6)	0.31 (0.09)
<b>10Hz Int @ 75 dBA</b>	-27.0 (1.1)	0.23 (0.03)	-5.8 (3.5)	0.35 (0.07)
<b>10Hz Int @ 85 dBA</b>	-31.0 (2.4)	0.24 (0.03)	-9.0 (4.1)	0.34 (0.06)
<b>mean stationary noise</b>		0.33		0.28
<b>Stdv</b>		0.03		0.10
<b>Mean 10 Hz int noise</b>		0.23		0.33
<b>Stdv</b>		0.03		0.07
<b>mean all noises</b>		0.28		0.30
<b>Stdv</b>		0.06		0.09

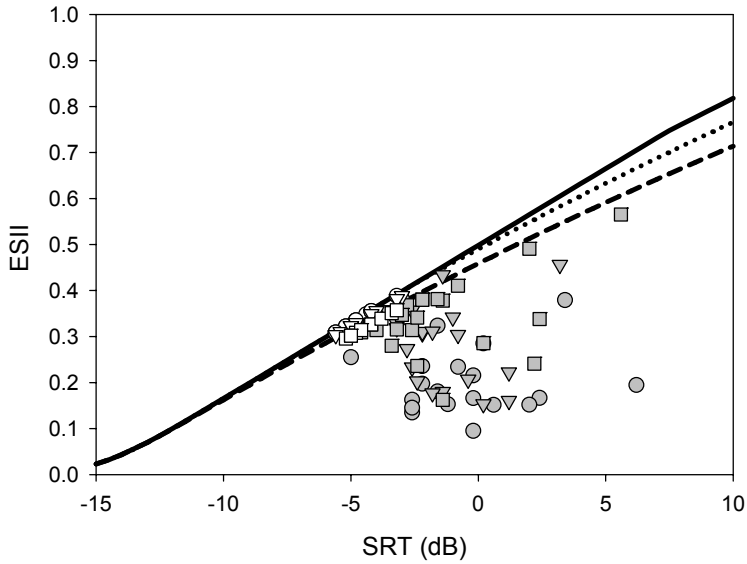
A 2[subject group]\* 2[noise condition] \*3[level] Analysis Of Variance (ANOVA) was performed on the calculated SII values. Of the main effects, differences in “subject group” (NH vs. HI) ( $F[1,168]=4.98, p<0.05$ ), differences in “noise condition” ( $F[1,168]=6.30, p<0.05$ ), and differences in “level” were significant ( $F[2,168]=5.72, p<0.005$ ). For all groups and noise conditions, the SII increases with increasing noise level from 65 dBA to 75 dBA. The interactions “subject group\*noise condition” ( $F[1, 168]=55.22, p<0.0001$ ), and “subject group\*level” ( $F[1, 168]=4.31, p<0.05$ ) were significant. It reflects the difference in average SII between noise conditions and subject group: Normal-hearing subjects obtain lower SIIs in interrupted noise, whereas for hearing-impaired subjects the opposite is true.

A 10[subject]\* 2[noise condition] \*3[level] ANOVA was performed on the SII data-set obtained with the normal-hearing subjects. Of the main effects, differences in “noise condition” ( $F[1,9]=417.67, p<0.0001$ ) was significant. Of the interactions, only “level\*noise condition” ( $F[2, 18]=3.62, p<0.05$ ) was significant.

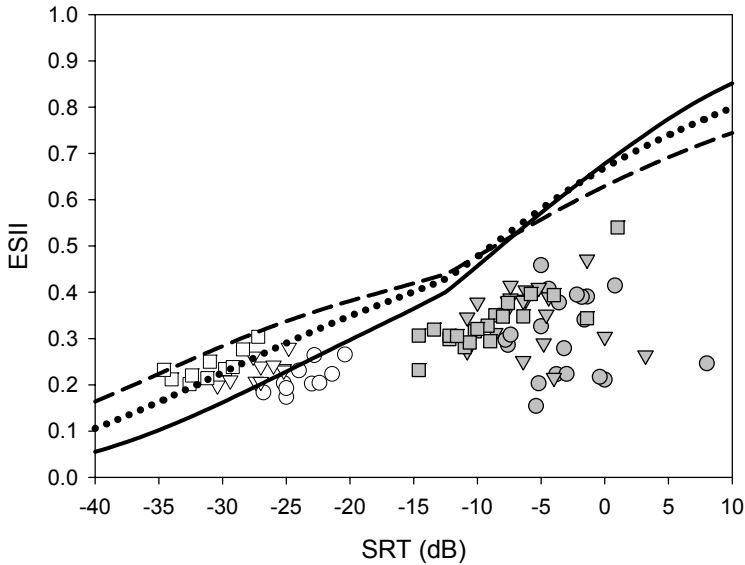
Post Hoc tests (Tukey HSD) showed no significant differences between the three interrupted noise conditions, and no significant differences between the three stationary noise conditions. Furthermore, the 10 Hz interrupted noise conditions were significantly different from the stationary noise conditions.

A 20[subject]\* 2[noise condition] \*3[level] ANOVA was performed on the SII data-set obtained with the group of hearing-impaired subjects. Of the main effects, differences in “subject” ( $F[19, 8.44]= 7.40, p<0.005$ ), differences in “level” ( $F[2, 38]=18.96, p<0.0001$ ), and differences in “noise condition” were significant ( $F[1, 19]=30.86, p<0.0001$ ). Of the interactions, only “level\*noise condition” ( $F[2, 38]=8.76, p<0.005$ ) was significant. Post Hoc tests (Tukey HSD) showed no significant differences between the three interrupted noise conditions. A significant difference was observed between the stationary noise conditions at 65 dBA and the other two stationary noise conditions. Furthermore, the 10 Hz interrupted noise at 65 dBA, 75 dBA, and 85 dBA were significantly different from the stationary noise condition at 65 dBA.

Figures 8.3 and 8.4 show the individual SII values for all subjects in stationary and 10 Hz interrupted noise, respectively. The lines in the figures represent the theoretical maximum SII value for the SRT in a specific noise condition at a specific noise level if audibility of the speech is not influenced by the hearing level. Contrary to the theoretical maximum SII in stationary noise at lower speech levels ( $< 73$  dB SPL; the level distortion function in the SII model is active above the 73 dB SPL), is the maximum SII in interrupted noise dependent on the noise level. The FMF accounts for less forward masking at higher levels, which results in about 4 dB SRT decrease with 10 dB increased noise level. This 4 dB shift is also observed with the SRT measurements in interrupted noise for the NH subjects.



**Figure 8.3.** The Extended Speech Intelligibility Index (ESII) as function of SRT in stationary noise for normal-hearing (open symbols) and hearing-impaired (filled symbols) listeners. Circles, triangles, and squares denote the noise condition fixed at 65, 75 and 85 dBA, respectively. The normal, dotted, and dashed lines represent the maximum ESII as a function of SRT in stationary noise fixed at 65, 75, and 85 dBA, respectively.



**Figure 8.4.** The Extended Speech Intelligibility Index (ESII) as function of SRT in 10 Hz Interrupted noise for normal-hearing (open symbols) and hearing-impaired (filled symbols) listeners. Circles, triangles, and squares denote the noise conditions fixed at 65, 75 and 85 dBA, respectively. The normal, dotted, and dashed line represent the maximum ESII as a function of SRT in 10 Hz interrupted noise fixed at 65, 75, and 85 dBA, respectively.

For normal-hearing listeners, the average SII in interrupted noise is about 0.11 lower compared with that in stationary noise. A clear explanation for this difference is not yet found. One of the reasons might be that the ESII model has been validated with the speech material of Versfeld *et al.* (2000), and that under critical conditions, differences between speech corpuses express themselves more in interrupted noise than in stationary noise (cf. van Wijngaarden and Houtgast, 2004). In that respect, with the present speech material of Plomp and Mimpen (1979) subjects might need less speech information in interrupted noise compared to the material of Versfeld *et al.* (2000). Indeed, Rhebergen, *et al.*, (2006b) measured about 6 dB higher SRTs with the latter speech material in about the same noise condition (8 Hz interrupted noise fixed at 65 dBA). Since

forward masking is not speech material dependent, another explanation might be, as suggested in Chapter 7, that the linear intensity importance function (IIF) of the SII (ANSI S3.6-1997) is not sufficiently accurate to account for speech intelligibility predictions in fluctuating noise for specific speech material (e.g., clear vs. conversational speech). It is expected that a speech material dependent IIF for this speech corpus (clear speech) will account for higher SII values for SRT in interrupted noise for normal-hearing listeners. The FMF is assumed to be related to the hearing loss of the listener (Ludvigsen, 1985). Further research on the relation between speech in interrupted noise for NH and HI subjects and performances in forward masking experiments could give more knowledge about the relationship between the FMF and the hearing loss of the listener.

HI subjects who are unable to understand speech properly, even when it is presented sufficiently audible (e.g., with a hearing aid), need in general more speech information compared with other NH or HI listeners to understand speech. Although speech is audible, the auditory system apparently is unable to process it properly (so-called supra-threshold deficits, Noordhoek, 2000). In general, these listeners require a higher SNR (thus higher SII values) to reach the threshold level of intelligibility. The ESII model has the potential to model some factors that might result in better understanding supra-threshold deficits, such as reduced spectral and temporal resolution, and decreased basilar membrane compression.

#### **D. Basilar membrane compression and the ESII**

Oxenham and co-workers (see Oxenham and Bacon, 2003, 2004, and Bacon and Oxenham, 2004, for a review) showed that amplitude compression by the basilar membrane plays an important role in forward masking. They showed that when the input signal first is compressed and next is convolved by a temporal window, it can adequately describe the data of many studies dealing with forward masking. They also showed that a decrease in temporal resolution is not due to an increase in the size of the temporal window, but rather to a decrease in compression. The SII model assumes normal basilar membrane compression. Thus, when the compressive power of the basilar membrane is decreased, for example due to a significant hearing loss, the SII model will yield less reliable predictions. Then, the dynamic range of speech, as represented at

the output of the basilar membrane, will be larger than when no hearing loss is present. One way to account for loss in compressive power is to expand the signal with a fixed amount and compress it again to a degree that corresponds to the amount of basilar membrane compression in the (hearing-impaired) cochlea. With normal-hearing subjects, the net result will be zero, and the original SII results are obtained. However, with hearing-impaired subjects, it will result in a signal that is expanded to some extent, which will result in different SII values, that may be expected to be worse.

Hornsby and Ricketts (2001) used the SII to calculate the speech intelligibility of compressed speech with different compression ratios from 1:1 to 6:1. The SII values showed no significant effect as function of compression ratio. Compression probably will act differently on speech in interrupted noise. The nonlinear behavior of the basilar membrane enables increased gain of the speech during the absence of the noise, allowing increased audibility and less forward masking. To model compression in the cochlea, the input-output function of the basilar membrane proposed by Oxenham (1995) (Eq 1.) can be used:

$$\text{Gain} = -0.1 * L + A + B * (1 - 1 / (1 + \exp(0.05 * (50 - L)))) \quad , \quad (\text{Eq 1.})$$

where  $L$  is the input level in dB SPL;  $A$  and  $B$  are defined as:

$$A = -0.0894 * G_{\text{max}} + 10.894, \quad B = 1.1789 * G_{\text{max}} - 11.789.$$

$G_{\text{max}}$  represents the maximum gain (dB).

The maximum gain of the input-output function of the basilar membrane proposed by Oxenham (1995) has a range of 0 to 60 dB (e.g., 0 dB: no basilar membrane compression left).

Bacon and Oxenham (2004) note a best fit for NH listeners with a maximum gain of 48 dB; and hearing-impaired listeners on average maximum gain of 8 dB (range between 0 and 15 dB), but this maximum gain for HI subjects is dependent on the type of hearing loss.

## Chapter 8

The instantaneous gain on the signal and noise can be calculated as follows: For a given speech-in-noise condition, the signal-to-noise ratio at threshold (i.e., the SRT) is determined and the signal with this SNR is filtered into 21 different critical bands (with the aid of 200<sup>th</sup> order FIR filters). Next the intensity envelope in each band is determined and fed to the compressive function of Oxenham (1995) with a maximum gain of 48 dB if normal-hearing listeners are involved (Bacon and Oxenham, 2004). The instantaneous gain is obtained by dividing the compressed envelope by the original envelope, this instantaneous gain is used to modify the envelope of speech and noise individually. With these functions, the remaining SII calculations are performed, which are the same as the Extended SII.

*Table 8.2. SRTs and Extended SII calculations with basilar membrane compression. Lower rows yield mean and standard deviation of the ESII.*

	NH (N=10)		HI (N=20)			
	SRT (dB)	stdv	ESII	SRT (dB)	stdv	ESII
<b>10Hz Int @ 65 dBA</b>	-23.6	2.0	0.29	-3.0	3.6	0.30
<b>10Hz Int @ 75 dBA</b>	-27.0	1.1	0.32	-5.8	3.5	0.35
<b>10Hz Int @ 85 dBA</b>	-31.0	2.4	0.33	-9.0	4.1	0.35
<b>stationary @ 65 dBA</b>	-4.7	0.7	0.34	-0.6	2.6	0.30
<b>stationary @ 75 dBA</b>	-4.5	0.9	0.35	-1.4	1.7	0.37
<b>stationary @ 85 dBA</b>	-4.3	0.7	0.35	-1.3	2.6	0.40
<b>mean int noise</b>			0.31			0.33
<b>Stdv</b>			0.02			0.02
<b>mean stationary noise</b>			0.35			0.36
<b>Stdv</b>			0.01			0.05
<b>mean all noises</b>			0.33			0.35
<b>Stdv</b>			0.03			0.04

## **Calculations with the Extended SII with compression**

The 4<sup>th</sup> and 7<sup>th</sup> column of Table 8.2 show the Extended SII calculations (Rhebergen, *et al.*, 2006b) where the algorithm of Oxenham (1995) has been implemented. Absolute threshold values at each 1/3 octave frequency, averaged across subject group (normal-hearing and hearing-impaired) were input to the model. For all listeners (i.e., also the normal-hearing group), the slope of the FMF was kept fixed, corresponding to a hearing loss of 50 dB for all frequencies. The maximum gain was fixed at 48 dB for the normal-hearing subjects. For the NH subjects, the mean SII value for all noise conditions is 0.33 with a standard deviation of 0.026. For the group of HI subjects, the best fit was found with a maximum compression gain of about 20 dB. For the HI subjects, the mean SII is 0.35 with a standard deviation of 0.037. The results for the NH subjects are more or less the same as the Extended SII model without the compression function. The fixed FMF in combination with a 48 dB maximum compression gain for NH subjects yields about the same SII values. The assumed fixed FMF corresponding to a hearing loss of 50 dB HL thus appears to be a proper masking function for all subjects. As mentioned earlier, there are no specific data to check at which slope the FMF for the HI subjects should be fixed. This is thus also the case for the maximum gain for the compression for the HI subjects. For this reason, the gain was found with the best fit for the interrupted noise conditions (assuming an SII value between 0.3 and 0.4). The maximum gain of 20 dB gave good SII values for the interrupted noise conditions. Additionally, the SII values for the stationary noise conditions improved as well. The 20 dB gain is thus sufficient to explain the speech intelligibility at lower levels for this group of HI subjects. There seems to be no direct relationship between the maximum gain and the PTA for HI subjects. The PTA of the HI subjects was on average 53 dB. If there is a linear relationship between the PTA and the maximum gain, one would expect a maximum gain of about 7 dB for the HI subjects (60 dB (maximum gain compression function)-53 dB = 7 dB).

## IV. Discussion

The most important observation of this paper is that the Extended SII is able to predict the speech intelligibility in interrupted noise for HI subjects at different noise levels. The predictive power of the SII model improves when a compression function is included in the Extended SII model. The FMF in the Extended SII works reasonably well, but further improvement remains possible. For this reason, more research is needed to examine the relation between hearing impairment, speech intelligibility and other auditory functions such as forward masking and auditory filtering. The SII model is a simple model that does not account for filter shape at different levels or for different hearing impairments. There are many factors that can influence the performance of a hearing-impaired or normal-hearing subject. It is difficult to account for all possible factors that contribute to the speech intelligibility for a single subject. This is also the case for the Extended SII method with compression. It can give good predictions for normal-hearing subjects, and it is an improvement for hearing-impaired subjects compared with the Extended SII without compression. The predictions for these hearing-impaired subjects are based on a best fit with a maximum gain of 20 dB. As long as there is no method to predict the maximum gain for the individual listener, it is a method to fit the observed SRTs rather than a method to predict the SRTs. Furthermore, this approach is computationally more complicated and time consuming than the Extended SII calculation without the compression function. The FMF adapted from Ludvigsen (1985) has the potential to model the forward masking for hearing-impaired subjects in the Extended SII model individually. Further research is needed to improve this function in order to allow better predictions for individual hearing-impaired subjects.

## **IX. Summary & conclusions**

The present paper describes an evaluation of the Extended SII approach (Rhebergen, *et al.*, 2006b) to model SRTs (Speech Reception Thresholds) for sentences masked by interrupted noises for normal-hearing and hearing-impaired subjects.

The Extended SII approach is able to predict the speech intelligibility well for groups of normal-hearing and hearing-impaired subjects in different noise levels. The inclusion of a compression function in the Extended SII method can give even better predictions for the hearing-impaired subjects. More research is needed to examine the relation between the speech intelligibility in fluctuating noise of hearing-impaired subjects and their performance in forward masking experiments with aim to model the speech intelligibility for individual hearing-impaired listeners in a variation of noise conditions.

## *Chapter 9*

### General Discussion

## **General Discussion**

This thesis dealt with the modeling of the speech intelligibility in fluctuating noise for normal-hearing listeners. To that end, the Speech Reception Threshold (SRT) for sentences, as introduced by Plomp and Mimpen (1979), for normal-hearing listeners was determined for a wide range of noises that differed with respect to spectro-temporal characteristics. Modeling the speech intelligibility in noise may give insight into the mechanisms of speech perception for normal-hearing listeners, and may be useful to understand the decrease in performance of hearing-impaired listeners. Hopefully, it can help in the search of signal processing to optimally use the residual capacities of the impaired auditory system.

### **The Extended SII model**

A new approach has been introduced in Chapter 2 to model the SRT in fluctuating noise for normal-hearing listeners. The new method is an extension of the Speech Intelligibility Index (SII) model (ANSI S3.5-1997, 1997), and is called the ESII (Extended Speech Intelligibility Index) model. The SII has been developed to predict the speech intelligibility in noise, but it has been validated for speech in stationary noise only, and is not suited to predict the SRT in fluctuating noise. The SII does not take into account the fluctuations in the masking noise, since it calculates the speech intelligibility by means of the long-term spectrum of the speech and noise at a given SNR, and the subject's hearing level. The ESII accounts for the modulations in the masking noise by calculating instantaneous SIIs over a two-second period (about sentence length). The "instantaneous" SII is calculated in 21 critical bands with a frequency-dependent time window ranging from 35 to about 9.3 ms from the lower to the higher frequency critical bands, respectively. The window lengths were taken from frequency-dependent gap detection data from the literature, and were multiplied by a factor 2.5 to account for forward masking. This approach can give a good account for most existing data with fluctuating background noises, described in the literature. The mean SII (ESII) thus is a good predictor of the speech intelligibility in fluctuating noise, and therefore a valuable extension to the SII-model.

Chapter 5 reports on a study with normal-hearing listeners to validate the ESII model introduced in Chapter 2. For a range of masking conditions, critical to the ESII model, SRTs have been measured. The results have been used to test and refine the model. A revision has been proposed, such that a better prediction is obtained for speech intelligibility in fluctuating noise. The previous model as described in Chapter 2 does not properly account for forward masking. The implementation of a forward masking function (FMF; Ludvigsen, 1985) in the ESII, being dependent on presentation level and hearing loss, improved the predictions for speech intelligibility in fluctuating noise. The FMF accounts for increased temporal resolution at higher levels, and decreased temporal resolution with increased hearing loss. The FMF enables the ESII not only to model the speech intelligibility as function of audibility, but also to account for the effects of forward masking on speech perception. Moreover, the FMF can better account for differences in SRT due to time-asymmetrical envelope shapes of the masking noises (i.e. saw-tooth noise, played normal or in reverse). By addition of forward masking by the FMF in the ESII model the factor 2.5 needed in Chapter 2 to account for forward masking becomes obsolete. The integration time thus is reduced to 3.7 ms at high frequencies.

In Chapter 6, a study with normal-hearing listeners has been described, to further validate the ESII model of Chapter 5. SRTs have been measured for a range of real-life background masking conditions. The results of Chapter 6 show that it is valid to use the ESII model to predict the speech intelligibility in real-life background noises. The model apparently is capable to account for sounds that comprise more complex spectro-temporal variations than do artificial masker signals used in most speech intelligibility experiments.

### **Limitations of the ESII model**

One of the aspects discussed in Chapter 2 is the rise in SRT due to a second interfering speaker. Since the interfering speaker masks not only in a physical manner, but also by means of speech information, the additional masking is denoted as informational masking.

In Chapter 3, a novel method has been presented to examine the effect of informational masking. The study shows that with non-intelligible interferers, listeners suffer less from informational masking than with an intelligible interferer. In the present group of subjects, informational masking adds about 6

to 7 dB to the SRT. As a result, the SRT in interfering speech, as predicted with the ESII model, is lower than the actually observed SRT: The ESII model is unable to account for informational masking, because it only accounts for physical masking. Therefore, if the ESII is used to predict the speech intelligibility in the presence of interfering speech, the actual SRT can be higher than predicted, the difference depending on the degree to which the listener is affected by informational masking.

A phenomenon observed in the listening experiments was the difference in learning effect between different noise conditions. The results in Chapter 4 have shown that with stationary noise as a masker, no learning effect is observed, whereas with interrupted noise there is quite a strong effect. These findings suggest that future experiments with SRTs in non-stationary noise should contain at least a repeated-measure approach and possibly some training to control for learning effects. The ESII model does not account for learning effects, but predicts the best speech intelligibility possible in a given SNR. Consequently, when modeling SRT data measured with only one observation in non-stationary noise, one should keep in mind that for some non-stationary noise conditions for some listeners a learning effect may be present. It might also be possible that the learning effect for speech intelligibility in non-stationary noise is dependent on speaker style (i.e. clear vs conversational speech). Further research on this matter is needed to come to a final conclusion.

## **The dynamic range of speech**

In Chapter 7, the relationship between the dynamic range of the speech signal and the intelligibility in stationary and interrupted noise for normal-hearing listeners has been examined. The results show increased speech intelligibility when the speech signal is instantaneously compressed to a modest degree (compression ratio CR=2:1) while the noise remains uncompressed. Intelligibility decreases when the speech is expanded with an expansion ratio of 1.5:1 (CR=2:3). These findings support the idea that the SRT in noise also depends on the distribution of the speech information along the intensity range of speech. The distribution of speech information along the intensity axis can be modeled by an Intensity Importance Function (IIF, Studebaker, 2002), and in this thesis it has been examined to what degree the IIF of the SII (ANSI S3.5-

1997) model corresponds to the actual level distribution. The ESII is less successful in predicting the speech intelligibility for speech compressed to a larger degree (CR=4:1). Compression not only causes weak parts of the signal to become audible, but also introduces noticeable distortions due to its nonlinear character. Apparently the latter has an increasingly stronger effect on intelligibility, causing the SRT to increase again at higher compression ratios. The results suggest that speech intelligibility in stationary and interrupted noise for radio broadcasts, public amplification systems in trains, stadiums etc can be improved by compressing the speech modestly by a compression ratio of 2:1. In conjunction with the experiment with compressed speech in otherwise unaltered noise, Chapter 7 also reports on an experiment where both speech and noise have been compressed simultaneously (as is the case in real-life situations). In stationary noise, speech intelligibility is negatively affected by compression, whereas in interrupted noise SRT strongly decreases. The latter is due to the gain of the speech signal in the gaps of the interrupted noise. Hence, there are situations where instantaneous compression indeed can increase intelligibility, despite the many negative reports. It is challenging to determine the characteristics of the masking noise where compression may be beneficial.

### **Exploring the use of the ESII model for hearing-impaired listeners**

In Chapter 8, the ESII model is used to predict the speech intelligibility for hearing-impaired listeners in stationary or interrupted noise. Again, based on the data from de Laat and Plomp (1983), there are opportunities to improve the ESII model. Instead of using Ludvigsen's (1985) FMF, a method is suggested where forward masking is a combined effect of cochlear compression followed by a fixed (i.e., similar for normal-hearing and hearing-impaired listeners) forward masking function. Both methods imply that speech intelligibility for hearing-impaired subjects not only is reduced due to audibility, but also as a result of reduced temporal resolution. Whereas the present articulation models are unable to account for factors such as temporal resolution, the ESII model has the potential to explore the predictability of speech perception in hearing-impaired listeners with a supra-threshold deficit.

## Further suggestions for research on the speech intelligibility in fluctuating background noise

### *Speech intelligibility as function of speaking style*

One of the most striking aspects of SRT results is the difference in SRT between the different speech materials. Observed SRTs measured in stationary noise with the long-term spectrum of the target speaker range from -3.0 dB (Nilsson *et al.*, 1994) to -7.8 dB (van Wieringen and Wouters, 2006). SRT scores measured with different speech corpora but with use of the same subjects range from -3.66 dB to -4.5 dB (Festen and Plomp, 1990; Versfeld *et al.*, 2000; Versfeld and Dreschler, 2002, van Wijngaarden and Houtgast, 2004). Differences in SRTs probably are not primarily due to lexical differences in corpora or languages but rather due to speaking style (Nilsson *et al.*, 1994). Over 20 years, in several studies differences in speech intelligibility have been examined between clearly articulated speech (clear speech) and normally articulated speech (conversational speech). The mean finding is that clear speech yields better SRTs for normal-hearing and hearing-impaired listeners in quiet, in background noise, and in reverberation compared with conversational speech (Picheny *et al.*, 1985, 1986, 1989; Payton *et al.*, 1994, Uchanski *et al.*, 1996, Mullennix *et al.*, 1989; Krause and Braidá, 2002; Bradlow *et al.*, 2003; van Wijngaarden and Houtgast, 2004). It showed that, in general, clear speech is more intelligible than conversational speech.

Intonation in speech contains various cues that can serve as an indication how to interpret the perceived speech (e.g., order-, question-, exclamation-, and round off sentence). The intonation of speech is defined as the course of the fundamental frequency F0 over time (Collier, 1975). The F0 of female speech usually is higher than that of male speech, F0 ranging from approximately 120 to 360 Hz and 100 to 220 Hz for female and male speech, respectively (Picheny *et al.*, 1985, 1986, 1989; Bradlow *et al.*, 1996; 2003, Green *et al.*, 2005). Speaking style (e.g., clear or conversational speech) seems to be related to the variance of F0. If subjects are asked to speak clearly, then the mean F0 frequency and F0 variance is increased compared with their conversational speech. Also, the speaking rate is decreased (Picheny *et al.*, 1985; Faulkner and Rosen, 1999; Bradlow *et al.*, 2003; Krause and Braidá, 2002, 2004; Goberman and Elmer, 2004). Furthermore, acoustical analysis of clear and conversational speech showed for

clear speech a higher long-term spectral energy between 1000 Hz and 3150 Hz (Payton *et al.*, 1994, 1999; Kraus and Braidá, 2002, 2004), and a higher modulation index in the modulation spectrum (Payton *et al.*, 1994, 1999, 2002; Kraus and Braidá, 2002, 2004; Liu *et al.*, 2004; van Wijngaarden and Houtgast, 2004) than for conversational speech. Differences in SRT due to speaking style have implications for the predictions of the speech intelligibility in quiet, stationary noise, and in reverberation (Payton *et al.*, 1994, 2002; van Wijngaarden and Houtgast, 2004). Neither the Speech Transmission Index (STI; Steeneken and Houtgast, 1980, 1985), the Articulation Index (AI; ANSI S3.5-1969, 1969) nor the SII are able to properly discriminate between conversational or clear speech in quiet or other masking conditions. Since most common speech is more “conversational” than “clear” in nature, and since most speech intelligibility tests use clear speech, research effort is required to close the gap between laboratory predictions with clear speech and real world experience with conversational speech.

### *Acoustical analysis of Dutch speech materials*

Table 9.1 shows the acoustical analysis of the four commonly used Dutch speech corpuses for SRT experiments (first 130 sentences). The female speech has a higher mean F0 frequency and larger F0 variance compared with the male speech. The Plomp and Mimpen (1979) speech has the lowest articulation rate, the highest F0 and the largest F0 variance, followed by VU98 female (Versfeld *et al.*, 2000), VU98 male (Versfeld *et al.*, 2000), and Smoorenburg (1992) male speech. A perceptually more relevant measure of F0 variation expressed in Hertz is the variance expressed in semitones (Maasen and Povel, 1984, Bradlow *et al.*, 2003). A semitone is defined as the logarithmic difference between two frequencies:  $\text{semitones} = 12 * \log(f_2/f_1)/\log(2)$ , where  $f_1$  and  $f_2$  represent the two frequencies. The acoustical analysis of the four speakers suggest on basis of the F0 variance in semitones and speaking rate that the Plomp and Mimpen (1979) speech might be characterized as more clearly articulated, followed by the VU98 male speech, VU98 female speech, and the Smoorenburg speech. Versfeld *et al.* (2000) measured for all four corpuses the SRT in stationary noise. These SRTs are given in the last row of Table 9.1. Indeed, although only four sets are involved, and differences are relatively small, the trend in SRT is similar to that in F0 variance. More SRT experiments are needed with a wider range of speech

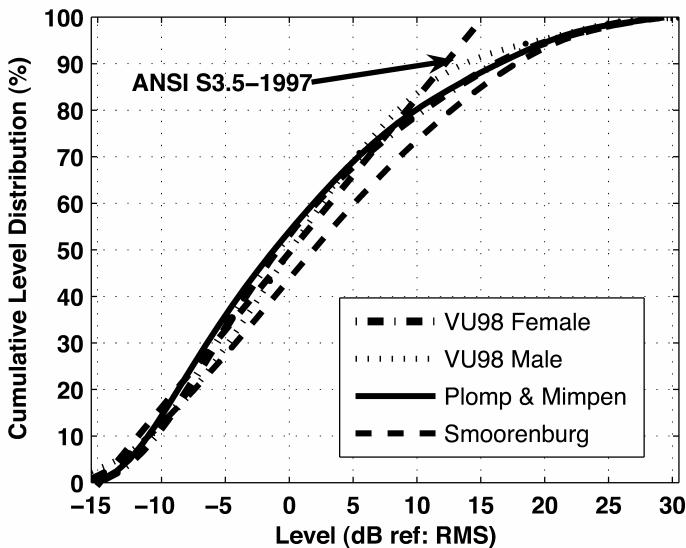
## General Discussion

corpuses, measured with the same group of listeners, preferably in fluctuating background noise to gain more insight into the dominant factors governing speaking style.

*Table 9.1. Acoustical analysis of four Dutch speech corpuses. Between brackets is denoted the standard deviation between (130) sentences.*

Speech corpus	VU98 Female	VU98 Male	Plomp & Mimpen	Smoorenburg
Gender	Female	Male	Female	Male
Average sentence duration (sec)	1.8 (0.3)	2.0 (0.3)	2.4 (0.1)	1.8 (0.2)
Minimal sentence duration (sec)	1.3	1.2	2.0	1.5
Maximal sentence duration (sec)	2.7	3.0	2.9	2.3
Average syllables per second	4.61 (0.6)	4.20 (0.5)	3.61 (0.9)	4.58 (0.9)
F0 mean (Hz)	177 (16.9)	110 (14.3)	231 (11.8)	115 (5.4)
F0 min (Hz)	124 (11.2)	76 (10.1)	147 (29.7)	83 (5.8)
F0 max (Hz)	243 (32.3)	120 (24.6)	316 (25.7)	150 (10.4)
F0 range in semitones (Hz)	11.6	11.8	13.2	10.2
F0 variance (Hz)	1270 (796)	395 (320)	1833 (732)	356 (117)
SRT (dB) (Versfeld, and Dreschler, 2000)	-4.1	-4.0	-4.5	-3.7

Figure 9.1 shows the broadband (100-8500 Hz) smooth cumulative level distribution (CLD) of the same four corpuses, relative to their respective RMS, based on 30 seconds of running speech. The four CLDs are corrected with 4.55 dB to account for effective speech peaks (chapter 7). The figure shows marked differences in level distribution, where none of the level distributions resembles the level distribution given by the SII (ANSI S3.5-1997; gray dashed line). In the range of -10 dB to about 5 dB, the CLD is steepest for the Plomp and Mimpen corpus, followed by the VU98 female, VU98 male, and Smoorenburg corpus. It is worthwhile to examine in what way the differences in cumulative level distribution are related to differences in intelligibility, as has been proposed in Chapter 7. Note that, even after the modifications of the original SII model, the speech still is put into the model as stationary speech noise. Eventually, it may be necessary to replace this speech noise by real speech. Not only will this approach account for variations in speech articulation rate and intonation, but also for the interaction between the speech rate and modulation frequencies of the masking noise. Unfortunately, insertion of real speech into the SII model will create many conceptual and computational problems. Also, the link to the original SII model will be lost



*Figure 9.1. The Cumulative level distribution of speech as function of speech corpus.*

*Impact for predicted speech perception in Hearing-impaired*

In principle the articulation schemes, such as the Fletcher AI, its successors ANSI AI, (ANSI S3.5-1969), and the STI have been designed to predict the speech intelligibility for normal-hearing listeners. Many individuals with sensorineural hearing impairment, in addition to an increased absolute threshold (which all AI schemes take into account), also exhibit deterioration in supra-threshold sound and speech processing capabilities (Pavlovic, 1989). Pavlovic, Studebaker, and Sherbecoe (1986) developed a method to incorporate this deterioration into the AI model, and in its present form ( SII; ANSI S3.5-1997) has been taken into account by means of a level distortion factor. Although these additions to the articulation theory improve the speech intelligibility predictions for the *average* mild to moderate hearing-impaired listener, it is still difficult to predict accurately the performance for individual hearing-impaired listeners (Ching *et al.*, 1998, 2001; Hogan and Turner, 1998). The SII model does not take into account the listener's auditory temporal or spectral resolution. Furthermore, the performance of listeners with dead regions is also overestimated by the articulation theory (Moore, 2002; Hornsby and Ricketts, 2006). Subjects with higher thresholds probably have stronger supra-threshold deficits (Noordhoek, 2000) compared to subjects with mild hearing loss. In fact, Noordhoek claims that about 40 % of the hearing-impaired listeners suffer from supra-threshold deficits. The measured speech processing abilities of these hearing-impaired listeners are poorer than expected on basis of their threshold (i.e. higher SII values). As long as the SII model, or any other calculation scheme, is not extended to account for factors covering the supra-threshold deficits, the ability to predict the speech intelligibility for hearing-impaired listeners, aided hearing-impaired listeners, or the benefit of a hearing aid in terms of an increase in performance will be insufficient and unreliable. The ESII model in its present form has the potential to entangle this problem. Contrary to the other calculation schemes it has the ability to account for temporal processing by means of a variation of the integration time (e.g., as function of age). Furthermore it has a Forward Masking Function (FMF) that accounts for decreased temporal resolution by increased hearing level. Lastly, due to its instantaneous type of calculation, it has the ability to add other dynamic calculations such as basilar membrane compression (Chapter 8). Although these solutions have not been studied yet, the expectation is that they

will yield better speech intelligibility predictions for hearing-impaired listeners compared to the other articulation models. Further research will reveal to what extent the ESII model can account for supra-threshold deficits.

### *Speech intelligibility and hearing aid algorithms*

In Chapter 7, the method of Hagerman and Olofsson (2002) has been used to extract the speech from the speech-in-noise mix after compression, which is a non-linear process. Although this method is only a first-order approach, which means that it will give valid results only for linear or close-to-linear processes, it can be used for other signal processing algorithms besides compression (Souza *et al.*, 2006). The combination of this technique with the ESII model might give more insight into the effects of signal processing in different types of hearing aids on speech perception in noise. Unfortunately, many hearing aids contain signal-processing algorithms that are kept secret for commercial reasons. Thus, these hearing aids can only be considered as black boxes. Often, only subjective listening experiments have been performed to evaluate speech intelligibility of a signal-processing algorithm. To get an impression about the effects of hearing aid algorithms (noise reduction, compression, filtering) on speech intelligibility, the method of Hagerman and Olofsson (2002) may turn out to be useful. Still, as pointed out in chapter 7, compression or expansion introduces distortion products in the speech signal and these distortion products presumably have a larger effect on the technique of Hagerman and Olofsson (2002) to reconstruct the speech signal than on the speech signal itself. The outcomes of chapter 7 suggest that this (linear) technique should be used with caution in case of non-linearly processed speech in noise, and the same consideration should be taken into account with regard to the calculated “apparent SNR”. Assuming that at threshold for 50 % sentence correct, the calculated SII-values for different compression algorithms in different noise conditions should be the same, then the calculated “apparent SNR” for interrupted noise conditions with higher compression ratios is too low.

## A model approach is still a simplified version of reality

This thesis comprises several studies in relation with predicting the speech intelligibility in fluctuating noise. With an extension to the SII model it is possible to predict the *average* speech intelligibility in a variety of fluctuating noise conditions. Although the ESII is a valuable addition to the conventional calculation schemes, it still is just a simplified version of reality. With this model, it is not possible to draw definite conclusions for individuals or for individual conditions. This is because the observed SRT data always will be related to individual cognitive performance (e.g., concentration, fatigue, IQ, language, age, temporal processing, intellect), which is undefined and not a variable in the SII model. Furthermore, some assumptions of the articulation theory that have been changed to simplify the model may give ambiguous results (e.g., dynamic range of speech, intensity importance function, and auditory filter type). Some researchers have used audibility calculations to draw conclusions about various aspects of human auditory performance and the fundamental nature of speech itself (e.g., Rankovic, 2002; Noordhoek, 2000). If some assumptions in those calculations schemes were incorrect, then it may be necessary to revise at least some of the conclusions that were derived using them (Studebaker and Sherbecoe, 2002). Rankovic (2002) compared predictions for the consonant recognition test scores of hearing-impaired subjects irrespective of the presence or absence of dead regions. The results showed that Fletcher's AI is generally accurate for these subjects. She calls into question whether a clinical test for dead regions would offer additional predictive information since the results of the AI analysis are sufficiently accurate. Moore (2002) argues that, in fact, the AI is not accurate in predicting *incremental* benefit of amplifying frequencies well above the estimated edge frequency ( $F_e$ ) of a dead region. The Fletcher AI predicts 10-15 % improved speech scores for amplification of frequencies well above the  $F_e$ , were in fact the largest improvement was only 7 %. Thus, even Fletcher's AI cannot account for dead regions, since the predictions are based on the subject's audiogram.

Conclusions drawn from different calculation schemes (Fletcher AI, ANSI AI, SII, STI) can also lead to confounding results. Rankovic (1998) compared Fletcher's (1953) AI method to Kryter's simplified AI (ANSI S3.5-1969) version, in predicting results for the Consonant Vowel (CV) identification task. For these

## Chapter 9

CV data, Rankovic's calculations showed higher predictive power with the Fletcher method than the standardized AI (ANSI S3.5-1969) method.

The Fletcher method is hardly ever used after its publication in 1953, apparently because of its complex calculation scheme. Today, with availability of powerful computers, this is not a valid argument anymore (e.g., Müsch, 2002). Since the Fletcher AI is based on more than 30 years of research, it might be interesting to re-review the different methods for the prediction of speech intelligibility in noise.

This thesis showed that the ESII model is a good approach to model the speech intelligibility in non-stationary noise for normal-hearing listeners. It has the potential to better account for supra-threshold deficits by hearing-impaired listeners. Furthermore, due to its dynamic calculations is the ESII model particularly suitable for hearing aid algorithms evaluation. The ESII model has a considerable larger predictive power with respect the other articulation models. Therefore the ESII model is a valuable extension to de SII (ANSI S3.5-1997) model.

## *General Discussion*

## Summary

## **Summary**

One of the striking capabilities of the human auditory system is that it can decode a spoken message embedded in a background noise. In some cases, speech can still be intelligible when it is 20 to 30 dB lower in level than that of the background noise. To date, no speech recognition system or algorithm is able to even approach this human achievement. It appears that listeners with normal hearing are, in one way or another, able to take advantage of the relative silent periods in the noise. Thus, fluctuations in the background noise result in better speech intelligibility. Many studies show that hearing impairment decreases the ability to make use of the fluctuations in the background noise. This is why listeners with hearing loss experience a handicap when they have to understand speech under such adverse conditions.

The performance of listeners with respect to speech intelligibility in different background noises can be assessed with the Speech Reception Threshold (SRT) test. The SRT test is an adaptive speech intelligibility test with which the Signal-to-Noise Ratio (SNR), or so-called SRT, can be determined that is required for 50% sentence intelligibility. Using short, everyday sentences and noise with the average spectrum of speech, normal-hearing subjects reach an SRT of about -12 dB in fluctuating noise, whereas their SRT in stationary noise is about -5 dB. Subjects with hearing impairment still may be able to reach SRTs near -5 dB, but perform worse in fluctuating noise, resulting in SRTs of about equal to SRTs in stationary noise. Consequently, a good performance in fluctuating noise implies a good performance in stationary noise, whereas the opposite is not true.

Intelligibility in stationary noise can be predicted by the commonly used Speech Intelligibility Index (SII) model. The SII model calculates the percentage of speech information that is available to the listener for a given speech-in-noise condition.  $SII=0$  means that no speech information is available to the listener, and  $SII=1$  means that all speech information is audible. For speech (i.e., short everyday sentences) in stationary noise, normal-hearing listeners require about 33% of all speech information (i.e.  $SII=0.33$ ) to understand half of the sentences without any errors. If one assumes that at the SRT, all listeners need an about equal percentage of speech information, the value of the SII must be about 0.33

## Summary

for all noise conditions. The SII is calculated from and the threshold, the average spectrum of noise and speech, hence it does not take into account any fluctuations of the masking noise. Accordingly, the model will predict equal SRTs for speech in stationary and fluctuating noise. Vice versa, an SII calculated for an SRT in fluctuating noise will be too low (i.e., an SRT in fluctuating noise of -12 dB gives an SII of about 0.09 instead of 0.33).

In this thesis the speech intelligibility in fluctuating noise conditions for normal-hearing listeners has been investigated. The aim was to predict for normal-hearing listeners the data obtained with the SRT test for sentences in a wide range of fluctuating noises. A model approach for the speech intelligibility in noise can give insight in auditory speech perception in normal-hearing listeners, and might explain the reduced performance in hearing-impaired listeners.

In Chapter 2, a novel method has been introduced to model the SRT in fluctuating noise for normal-hearing listeners. The new approach is an extension of the commonly used SII model, which is a validated calculation scheme to predict the SRT in stationary noise conditions. The basic principle of the SII model is that it determines the amount of speech information that is still audible when speech is partly masked by noise or partly inaudible due to the threshold of hearing. To that end, both speech and noise signal are filtered into 21 frequency bands. In each frequency band, the remaining speech information is determined, eventually resulting in the SII. The extension to the existing model is that after filtering, the speech and noise signal are partitioned into small time frames. Within each time frame, the conventional SII is determined, yielding the speech information available to the listener at that time frame. This results in an SII value that changes over time, the so-called instantaneous SII. The averaged instantaneous SII finally results in a number between zero and unity, the average amount of available speech information. The Extended SII (ESII) model, as described in Chapter 2, is able to give a good account for most existing SRT data described in the literature.

In the specific situation that the speech intelligibility is hampered by interference of a second speaker, SRTs generally are worse than predicted. In this thesis, this phenomenon is called “informational masking”. In Chapter 3, a

## *Summary*

novel method is presented to examine the magnitude of informational masking on the speech intelligibility in interfering speech. The study shows that with non-native (i.e. unfamiliar with the language) speech as an interferer, listeners suffer less from informational masking than with a native (intelligible) interfering speaker. For the present experiment, the amount of informational masking is estimated to be 6.6 dB. As a result, the predicted SRT in interfering speech is 6 to 7 dB lower than the observed SRT, since the ESII model is unable to account for informational masking. Hence, when the ESII is used to predict the speech intelligibility in presence of interfering speech, one should keep in mind that actual SRTs can be several decibels higher than predicted.

Another aspect that can influence the observed SRT is a learning effect for speech intelligibility in fluctuating noise. The results in Chapter 4 show that with stationary noise as a masker, no learning effect is observed, whereas with interrupted noise there is a very strong learning effect present. These findings suggest that future experiments with SRTs in non-stationary noise should contain at least a repeated measure approach and possibly some training to control for learning effects.

Chapter 5 describes a study with normal-hearing listeners to further validate the ESII model as introduced in Chapter 2. A large range of fluctuating masking noises, critical to the ESII model, was used to measure SRTs in normal-hearing subjects. The results have been used to refine the model. By introducing a forward masking (FMF), model predictions could be significantly improved. For instance, the FMF now can account differences in SRT due to asymmetrical envelope shapes of the masking noises.

Another evaluation study with normal-hearing listeners is reported on in Chapter 6. Here, the ESII model as introduced in Chapter 5 of this thesis has been used to predict SRTs obtained for speech in real-life background noises. The results of Chapter 6 show that the model is capable to account for sounds that show more complex spectro-temporal variations than artificial masker signals used in most speech intelligibility experiments.

The dynamic range of speech may play an important role in modeling speech intelligibility in noise, because at a given signal-to-noise ratio, it largely

## *Summary*

determines the amount of speech information that exceeds the noise, hence is audible. It is asserted that a change in dynamic range (e.g., by instantaneous compression) changes the information distribution along the intensity axis, thus affecting the SRT. In Chapter 7, for normal-hearing listeners, the effect of the dynamic range of the speech signal on the intelligibility in (uncompressed) stationary and interrupted noise has been examined. The results show increased speech intelligibility when the speech signal is compressed with a compression ratio (CR) of 2:1. Intelligibility decreases when the speech signal is compressed further to 4:1 (probably due to distortion products), or is expanded to a ratio of 1:1.5. In addition to this experiment, speech intelligibility for simultaneously compressed speech and stationary noise, as well as speech and interrupted noise has been measured. In stationary noise, the speech intelligibility is negatively affected at CR=4:1, whereas in interrupted noise, the speech intelligibility is positively affected. The latter is caused by the extra gain of the speech signal in the gaps of the interrupted noise. The distribution of speech information can be described by an Intensity Importance Function (IIF). Inclusion of the IIF in the SII model results in slightly better predictions.

The model as introduced in Chapter 2 and refined in Chapter 5 now can account for forward masking. The model now also predicts changes in temporal resolution as function of masking level or as a function of hearing loss. The forward masking function accounts for increased temporal resolution at higher levels, and decreased temporal resolution with increased hearing loss. In Chapter 8, the ESII model is used to predict the speech intelligibility for hearing-impaired listeners in stationary and interrupted noise. Next to the forward masking function, a method is proposed to model the effect of cochlear compression for normal-hearing and hearing-impaired listeners for speech intelligibility in noise. The method shows that speech intelligibility in hearing-impaired subjects is not only reduced due to audibility, but also due to reduced temporal resolution. Consequently, hearing-impaired subjects cannot reach SRTs similar to those of normal-hearing subjects in fluctuating noise, even at equal audibility.

The model approach applied in this thesis to predict the speech intelligibility in fluctuating noise was used to obtain more insight in the working of the human auditory system. Not only have we gained better understand why listeners with

## *Summary*

normal-hearing can reasonably well follow a conversation in non-stationary background noise conditions, in addition we have some starting indications why listeners with a hearing-impairment are unable to perform as good as a normal-hearing in non-stationary noise. Further research in this field might lead to better understanding the disability for listeners with a hearing-impairment in real-life noise conditions.

The ESII model has a significant larger predictive power with respect the other articulation models because it can be used in stationary, and non-stationary noise conditions. Therefore is the ESII model a valuable extension to de SII (ANSI S3.5-1997) model. Furthermore, the ESII model fits, due to its instantaneous calculation of the speech intelligibility, perhaps good for hearing aid, and communication algorithms evaluation. Besides the much needed, but time-consuming subjective testing, is it with aid of the ESII model in the near future quite possible for more objective testing of hearing aid algorithms. This will make it easier to get a quick impression how effective the hearing aid algorithms are for a certain type of hearing loss.

## References

## References

- Allen, J. B. (1994). "How do humans process and recognize speech," *IEEE Trans. Speech and Audio Processing*, 2, 567-577.
- Allen, J. B. (1996). "Harvey Fletcher's role in the creation of communication acoustics," *J. Acoust. Soc. Am.* 99, 1825-1839.
- ANSI (1969). ANSI S3.5-1969, "Methods for the calculation of the articulation index," (American National Standards Institute, New York).
- ANSI (1996). "ANSI S3.6-1996, "American national standard methods for Specification for audiometers" (American National Standards Institute, New York).
- ANSI (1997). ANSI S3.5-1997, "American national standard methods for calculation of the speech intelligibility index, "(American National Standards Institute, New York).
- Apoux, F., Tribut, N., Debrulle, X., and Lorenzi, C. (2004). "Identification of envelope-expanded sentences in normal-hearing and hearing-impaired listeners," *Hear. Res.* 189, 13-24.
- Bacon, S. P., Opie, J. M., and Montoya, D. Y. (1998). "The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds," *J. Speech Lang. Hear. Res.* 41, 549-563.
- Bacon, S. P., and Oxenham, A. J. (2004). "Psychophysical manifestations of compression: Hearing-impaired listeners," in *Auditory Compression*, Eds. S. P. Bacon, A. N. Popper, and R. R. Fay (Springer, New York), pp. 107-152.
- Bashford, J.A., Jr., Meyers, M.D., Brubaker, B.S., & Warren, R.M. (1988). "Illusory continuity of interrupted speech: Speech rate determines durational limits", *J. Acoust. Soc. Am.* 84, 1635-1638.
- Boothroyd, A., Springer, N., Smith, L., and Schulman, J. (1988). "Amplitude compression and profound hearing loss," *J. Speech Lang Hear. Res.*, 31, 362-376.
- Boothroyd, A. (1990). "Articulation index: Importance function in the intensity domain," *J. Acoust. Soc. Am. Suppl.* 1 88 S31.
- Boothroyd, A., Erickson, F. N., and Medwetsky, L. (1994). "The hearing aid input: A phonemic approach to assessing the spectral distribution of speech," *Ear Hear.* 6, 432-442.
- Boothroyd (2000). "Thar's gold in them thar hills: mining the P/I function Carhart Memorial," Lecture, delivered to the Annual Convention of the American Auditory Society, Scottsdale Arizona.
- Bosman, A. J., and Smoorenburg, G. F. (1995). "Intelligibility of Dutch CVC syllables and sentences for listeners with normal-hearing and with three types of hearing impairment," *Audiology* 34, 260-284.
- Bradlow, A.R., Torretta, G.M., and Pisoni, D.B. (1996). "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," *Speech com.* 20, 255-272.
- Bradlow A.R., Kraus N., and Hayes E. (2003). "Speaking clearly for children with learning disabilities: Sentence perception in noise," *J. Speech Hear Res.* 46, 80-97.

## References

- Bronkhorst, A. W., and Plomp, R. (1992). "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," *J. Acoust. Soc. Am.* 92, 3132-3139.
- Bronkhorst, A. W. (2000). "The Cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions," *Acustica* 86, 117-128.
- Brungart, D. S. (2001). "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am.* 109, 1101-1109.
- Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (2001). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *J. Acoust. Soc. Am.* 110, 2527-2538.
- Brungart, D. S., and Simpson, B. D. (2002). "The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal," *J. Acoust. Soc. Am.* 112, 664-676.
- Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Heto, R., Kei, J., Lui, C., Kiessling, J., Kotby, M.N., Nasser, N.H.A., El Kholly, W.A.H., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G.I., Westermann, S., and Ludvigsen, C. (1994). "An international comparison of long-term average speech spectra," *J. Acoust. Soc. Am.*, 96, 2108-2120.
- Carhart, R., Tillman, T. W., and Greetis, E. S. (1969). "Perceptual masking in multiple sound backgrounds," *J. Acoust. Soc. Am.* 45, 694-703.
- Carlyon, R. P. (1996). "Spread of excitation produced by maskers with damped and ramped envelopes," *J. Acoust. Soc. Am.* 99, 3647-3655.
- Ching, T.Y.C., Dillon, H., and Byrne, D. (1998). "Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification," *J. Acoust. Soc. Am.* 103, 1128-1140.
- Ching, T.Y.C., Dillon, H., Katsch, R., and Byrne, D. (2001). "Maximising effective audibility in hearing aid fitting," *Ear Hear.* 22, 212-224.
- Clarkson, P.M., and Bahgat, S.F., (1991). "Envelope expansion methods for speech enhancement," *J. Acoust. Soc. Am.* 89, 1378-1382.
- Collier, R. (1975). "Physiological correlated of intonation patterns," *J. Acoust. Soc. Am.* 58, 249-255.
- Cox, R. M., Matesich, J. S., and Moore, J. N. (1988). "Distributions of short-term rms levels in conversational speech," *J. Acoust. Soc. Am.* 84, 1100-1104.
- de Laat, J. A. P. M., and Plomp, R. (1983). "The reception threshold of interrupted speech for hearing-impaired listeners," in *Hearing - Physiological Bases and Psychophysics*, edited by R. Klinke and R. Hartman, Springer Verlag (Berlin), 359-363.
- de Laat, J. A. P. M. (1989). "The perception of fluctuating sounds by hearing-impaired listeners," Doctoral thesis, Free University, Amsterdam.
- Dirks, D. D., Bell, T. S., Rossman, R. N., and Kincaid, G. E. (1986). "Articulation index predictions of contextually dependent words," *J. Acoust. Soc. Am.* 80, 82-92.

## References

- Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (2001). "ICRA Noises: Artificial noise signals with speech-like spectral and temporal properties for hearing aid assessment," *Audiology* 40, 148-157.
- Drullman, R. (1995a). "Temporal envelope and fine structure cues for speech intelligibility," *J. Acoust. Soc. Am.* 97, 585-592.
- Drullman, R. (1995b). "Speech intelligibility in noise: Relative contribution of speech elements above and below the noise level," *J. Acoust. Soc. Am.* 98, 1796-1798.
- Drullman, R., and Bronkhorst, A. W. (2000). "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *J. Acoust. Soc. Am.* 107, 2224-2235.
- Dubno, J. R., and Dirks, D. D. (1989). "Auditory filter characteristics and consonant recognition for hearing-impaired listeners," *J. Acoust. Soc. Am.*, 85, 1666-1675.
- Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (2002). "Benefit of modulated maskers for speech recognition by younger and older adults with normal-hearing," *J. Acoust. Soc. Am.* 111, 2897-2907.
- Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (2003). "Recovery from prior stimulation: masking of speech by interrupted noise for younger and older adults with normal-hearing," *J. Acoust. Soc. Am.* 113, 2084-2094.
- Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (2005). "Word recognition in noise at higher-than-normal levels: Decreases in scores and increases in masking," *J. Acoust. Soc. Am.* 118, 914-922.
- Duifhuis, H. (1973). "Consequences of peripheral frequency selectivity for nonsimultaneous masking," *J. Acoust. Soc. Am.* 54, 1471-1788.
- Dunn, H. K., and White, S. D. (1940). "Statistical measurements on conversational speech," *J. Acoust. Soc. Am.* 11, 278-288.
- Duquesnoy, A. J. (1983). "Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons," *J. Acoust. Soc. Am.* 74, 739-743.
- Eddins, D. A., Hall, J. W., III, and Grose, J. H. (1992). "The detection of temporal gaps as a function of frequency region and absolute noise bandwidth," *J. Acoust. Soc. Am.* 91, 1069-1077.
- Elliott, L. L. (1969). "Masking of tones before, during, and after brief silent periods in noise," *J. Acoust. Soc. Am.* 45, 1277-1279.
- Faulkner, A., and Rosen, S. (1999). "Contributions of temporal encodings of voicing, voicelessness, fundamental frequency, and amplitude variation to audio-visual and auditory speech perception," *J. Acoust. Soc. Am.* 106, 2063-2073.
- Festen, J. M. (1987). "Speech-perception threshold in a fluctuating background sound and its possible relation to temporal resolution," in *The Psychophysics of Speech Perception*, edited by M.E.H. Schouten, Martinus Nijhoff Publishers (Dordrecht), 461-466.
- Festen, J. M., and Plomp, R. (1990). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal-hearing," *J. Acoust. Soc. Am.* 88, 1725-1736.

## References

- Festen, J. M. (1993). "Contributions of comodulation masking release and temporal resolution to the speech-reception threshold masked by an interfering voice," *J. Acoust. Soc. Am.* 94, 1295-1300.
- Festen, J. M., and Plomp, R. (2002). "Application of the speech transmission index to the hearing-impaired," in *Past, present and future of the speech transmission index*, edited by S. J. van Wijngaarden, TNO Human Factors (Soesterberg), 69-78.
- Fletcher, H., and Galt, R. H. (1950). "The perception of speech and its relation to telephony," *J. Acoust. Soc. Am.* 22, 89-151.
- Fletcher, H. (1953). "*The ASA Edition of Speech and Hearing in Communication*" Originally Published in 1953, edited by Allen, J. B. in 1995, *Acoust. Soc. Am.*
- French, N. R., and Steinberg, J. C. (1947). "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.* 19, 90-919.
- Freyman, R.L., and Nerbonne, G.P. (1996). "Consonant confusions in amplitude-expanded speech," *J. Speech Hear. Res.* 39, 1124-1137.
- Glasberg, B. R., and Moore, B. C. (1992). "Effects of envelope fluctuations on gap detection," *Hear. Res.* 64, 81-92.
- Glasberg, B.R., and Moore, B. C. A. (2000). "Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise," *J. Acoust. Soc. Am.* 108, 2318-2328.
- Goberman A.M, and Elmer L.W. (2005). "Acoustic analysis of clear versus conversational speech in individuals with Parkinson disease," *J Commun Disord.* 38, 215-230.
- Goedegebure, A. (2005). "Phoneme Compression – Processing of the speech signal and effects on speech intelligibility in hearing-impaired listeners," Ph-D. thesis Erasmus University Rotterdam.
- Green, T., Faulkner, A., Rosen, S., and Macherey, O. (2005). "Enhancement of temporal periodicity cues in cochlear implants: effects on prosodic perception and vowel identification," *J. Acoust. Soc. Am.* 118, 375-85.
- Gustafsson, H. A., and Arlinger, S. D. (1994). "Masking of speech by amplitude-modulated noise," *J. Acoust. Soc. Am.* 95, 518-529.
- Hagerman, B. (1982). "Sentences for testing speech intelligibility in noise," *Scand. Audiol.* 11, 79-87.
- Hagerman, B., and Olofsson, A. (2002). "A method to measure the effects of noise reduction algorithms using simultaneous speech and noise," presentation at International Hearing Aid Research Conference, Lake Tahoe, CA.
- Hogan, C. A., and Turner, C. W. (1998). "High-frequency audibility: Benefits for hearing-impaired listeners," *J. Acoust. Soc. Am.* 104, 432-441.
- Hohmann V., and Kollmeier, B. (1995). "The effect of multichannel dynamic compression on speech intelligibility," *J. Acoust Soc Am.* 97,1191-1195.
- Hornsby, B.W., and Ricketts, T.A. (2001). "The effects of compression ratio, signal-to-noise ratio, and level on speech recognition in normal-hearing listeners," *J. Acoust. Soc. Am.* 109, 2964-73.
- Hornsby, B. W. Y., and Ricketts, T. A. (2003). "The effects of hearing loss on the contribution of high- and low-frequency speech information to speech understanding," *J. Acoust. Soc. Am.* 113, 1706-1717.

## References

- Hornsby, B.W., and Ricketts, T.A. (2006). "The effects of hearing loss on the contribution of high- and low-frequency speech information to speech understanding. II. Sloping hearing loss," *J Acoust Soc Am.* 119, 1752-63.
- Houtgast, T., and Steeneken, H. J. M. (1985). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.* 77, 1069-1077.
- Houtgast, T., Steeneken, H. J., and Bronkhorst, A. W. (1992). "Speech com. in noise with strong variations in the spectral or the temporal domain," *Proceedings of the 14<sup>th</sup> International Congress on Acoustics, Vol 3, pp. H2-6.*
- Houtgast, T.(2005). *Personal communication.*
- Howard-Jones, P. A., and Rosen, S. (1992). "The perception of speech in fluctuating noise," *Acustica* 78, 258-272.
- Howard-Jones, P. A., and Rosen, S. (1993). "Uncomodulated glimpsing in "checkerboard" noise," *J. Acoust. Soc. Am.* 93, 2915-2922.
- Hygge, S., Ronnberg, J., Larsby, B., and Arlinger, S. (1992). "Normal-hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds," *J. Speech Hear. Res.* 35, 208-215.
- Irino, T., and Patterson, R. D. (1996). "Temporal asymmetry in the auditory system," *J. Acoust. Soc. Am.* 99, 2316-2331.
- Jesteadt, W., Bacon, S. P., and Lehman, J. R. (1982). "Forward masking as a function of frequency, masking level, and signal delay," *J. Acoust. Soc. Am.* 71, 950-962.
- Kamm, C. A., Dirks, D. D., and Bell, T. S. (1985). "Speech recognition and the Articulation Index for normal and hearing-impaired listeners," *J. Acoust. Soc. Am.*, 77, 281-288.
- Kates, J. M. (1987). "The short-time articulation index", *J. Rehabil. Res. Dev.* 24, 271-276.
- Kates, J. M., and Arehart, K. H. (2005). "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.* 117, 2224-2237.
- Kidd, G., and Feth, L. L. (1982). "Effects of masker duration in pure-tone forward masking," *J. Acoust. Soc. Am.* 75, 1384-1386.
- Koopman, J., Franck, B. A., and Dreschler, W. A. (2001). "Toward a representative set of "real-life" noises," *Audiology* 40, 78-91.
- Krause, J. C., and Braida, L. D. (2002). "Investigating alternative forms of clear speech: the effects of speaking rate and speaking mode on intelligibility," *J. Acoust. Soc. Am.* 112, 2165-2173.
- Krause, J.C., and Braida, L.D (2004). "Acoustic properties of naturally produced clear speech at normal speaking rates," *J. Acoust. Soc. Am.* 115, 362-378.
- Kryter, K. D. (1962). "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Am.* 34, 1689-1697.
- Kryter, K. D. (1962). "Validation of the articulation index," *J. Acoust. Soc. Am.* 34, 1698-1702.
- Larsby, B., and Arlinger, S. (1994). "Speech recognition and just-follow-conversation tasks for normal-hearing and hearing-impaired listeners with different maskers," *Audiology* 33, 165-176.

## References

- Licklider, J. C. R., and Guttman, N. (1957). "Masking of speech by line-spectrum interference," *J. Acoust. Soc. Am.* 29, 287-296.
- Lippmann, R. P. (1996). "Accurate consonant perception without mid-frequency speech energy," *IEEE Trans. Speech and Audio Processing* 4, 567-577.
- Liu S., Del Rio E, Bradlow, AR., and Zeng F.G. (2004). "Clear speech perception in acoustic and electric hearing," *J. Acoust. Soc. Am.* 116, 2374-2383.
- Ludvigsen, C. (1985). "Relations among some psychoacoustic parameters in normal and cochlearly impaired listeners," *J. Acoust. Soc. Am.* 78, 1271-1280.
- Ludvigsen, C. (1987). "Prediction of speech intelligibility for normalhearing and cochlearly hearing-impaired listeners," *J. Acoust. Soc. Am.* 82, 1162-1171.
- Marcell, M.M. and Cohen, S. (1992). "Hearing abilities of Down syndrome and other intellectually impaired adolescents," *Research in Developmental Disabilities.* 13, 533-551.
- Middelweerd, M. J., Festen, J. M., and Plomp, R. (1990). "Difficulties with speech intelligibility in noise in spite of a normal pure-tone audiogram," *Audiology* 29, 1-7.
- Miller, G. A. (1947). "The masking of speech," *Psycho.Bull.* 44, 105-129.
- Miller, G. A., and Licklider, J. C. R. (1950). "The intelligibility of interrupted speech," *J. Acoust. Soc. Am.* 22, 167-173.
- Molis, M.R., and Summers, V. (2003). "Effects of high presentation levels on recognition of low-and high-frequency speech," *ACOUSTICAL RESEARCH LETTERS ONLINE* 4, 124-128.
- Moore, B. C. J., and Glasberg, B. R. (1983). "Growth of forward masking for sinusoidal and noise maskers as a function of signal delay; implications for suppression in noise," *J. Acoust. Soc. Am.* 73, 1249-1259.
- Moore, B. C. J., and Glasberg, B. R. (1986). "A comparison of two-channel and single-channel compression hearing aids," *Audiology*, 25, 210-226.
- Moore, B. C., Peters, R. W., and Glasberg, B. R. (1996). "Detection of decrements and increments in sinusoids at high overall levels," *J. Acoust. Soc. Am.* 99, 3669-3677.
- Moore, B. C. (1997). *An Introduction to the Psychology of Hearing*, fourth edition, Academic Press (London).
- Moore, B. C. J. (2002). "Response to "Articulation index predictions for hearing-impaired listeners with and without cochlear dead regions," *J. Acoust. Soc. Am.* 111, 2549-2550.
- Moore, B. C. (2003). *An Introduction to the Psychology of Hearing*, fifth edition, Academic Press (London).
- Moore, B.C.J., Glasberg, B.R., and Stone, M.A. (2003). "Why are commercials so loud? - Perception and modeling of the loudness of amplitude-compressed speech," *J. Audio Eng. Soc.* 51, 1123-1132.
- Moulines, E and Charpentier, F. (1990). "Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones," *Speech Com.* 9, 453-467.

## References

- Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (1989). "Some effects of talker variability on spoken word recognition," *J. Acoust. Soc. Am.* 85, 365–378.
- Müsch, H. (2001). "Review and computer implementation of Fletcher and Galt's method of calculating the Articulation Index," *ACOUSTICAL RESEARCH LETTERS ONLINE* 2, 25-30.
- Müsch, H., and Buus, S. (2001). "Using statistical decision theory to predict speech intelligibility. II. Measurement and prediction of consonant-discrimination performance," *J. Acoust. Soc. Am.* 109, 2910-2920.
- Neijenhuis, K., Sink, A., Priester, G., van Kordenoordt, S., and van der Broek, P. (2002). "Age effects and normative data on a Dutch test battery for auditory processing disorders," *Int. J. Audiology* 41, 334-346.
- Nelson, P. B., Jin, S. H., Carney, A. E., and Nelson, D. A. (2003). "Understanding speech in modulated interference: cochlear implant users and normal-hearing listeners," *J. Acoust. Soc. Am.* 113, 961-968.
- Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.* 95, 1085-1099.
- Noordhoek, I. M., Houtgast, T., and Festen, J. M. (1999). "Measuring the threshold for speech reception by adaptive variation of the signal bandwidth. I. Normal-hearing listeners," *J. Acoust. Soc. Am.* 105, 2895-2902.
- Noordhoek, I. M. (2000). "Intelligibility of narrow-band speech and its relation to auditory functions in hearing-impaired listeners", Doctoral thesis, Free University, Amsterdam.
- Noordhoek, I. M., Houtgast, T., and Festen, J. M. (2000). "Measuring the threshold for speech reception by adaptive variation of the signal bandwidth. II. Hearing-impaired listeners," *J. Acoust. Soc. Am.* 107, 1685-1696.
- Oxenham, A. J., and Moore, B. C. (1994). "Modeling the additivity of nonsimultaneous masking," *Hear. Res.* 80, 105-118.
- Oxenham, A. J. (1995). "Psychophysical consequences of peripheral auditory nonlinearity," Unpublished Ph.D. Thesis, Cambridge.
- Oxenham, A. J., and Moore, B. C. (1997). "Modeling the effects of peripheral nonlinearity in normal and impaired hearing". In W. Jesteadt (Ed.), *Modeling Sensorineural Hearing Loss* (pp. 273– 288). Mahwah, NJ: Erlbaum.
- Oxenham, A. J. and Plack, C. J. (1997). "A behavioral measure of basilar-membrane nonlinearity in listeners with normal and impaired hearing," *J. Acoust. Soc. Am.* 101, 3666 -3675.
- Oxenham, A. J., and Bacon, S. P. (2003). "Cochlear compression: Perceptual measures and implications for normal and impaired hearing," *Ear Hear.* 24, 352-366.
- Oxenham, A. J., and Bacon, S. P. (2004). "Psychophysical manifestations of compression: Normal-hearing listeners," in *Auditory Compression*, Eds. S. P. Bacon, A. N. Popper, and R. R. Fay (Springer, New York), pp. 62-106.
- Oxenham, A. J., Rosengard, P.S. and Braida, L.D. (2004). "Perceptual consequences of normal and abnormal peripheral compression: Potential links between psychoacoustics and speech perception," *J. Acoust. Soc. Am.* 115, 2421.

## References

- Pavlovic, C. V. (1984). "Use of the articulation index for assessing residual auditory function in listeners with sensorineural hearing impairment," *J. Acoust. Soc. Am.*, 75, 1253-1258.
- Pavlovic, C. V., and Studebaker, G. A. (1984). "An evaluation of some assumptions underlying the articulation index," *J. Acoust. Soc. Am.* 75, 1606-1612.
- Pavlovic, C. V., Studebaker, G. A., and Sherbecoe, R. L. (1986). "An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals," *J. Acoust. Soc. Am.* 80, 50-57.
- Pavlovic, C. V. (1987). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," *J. Acoust. Soc. Am.* 82, 413-422.
- Pavlovic, C. V. (1989). "Speech spectrum considerations and speech intelligibility predictions in hearing aid evaluations," *J. Speech Hear. Res.* 54, 3-8.
- Pavlovic, C. V. (2005). *Personal communication*.
- Payton, K. L., Uchanski, R. M., and Braidia, L. D. (1994). "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," *J. Acoust. Soc. Am.* 95, 1581-1592.
- Payton, K L., and Braidia, L. D. (1999). "A method to determine the speech transmission index from speech waveforms," *J. Acoust. Soc. Am.* 106, 3637-3648.
- Payton, K L., Braidia, L. D., Chen, S., Rosengard, P., and Goldsworthy, R. (2002). "Computing the STI using speech as a probe stimulus," in van Wijngaarden, S. J. (eds) *Past, present, and future of the Speech Transmission Index*, TNO Human Factors, Soesterberg, The Netherlands, 125-138.
- Peters, R. W., Moore, B. C., and Baer, T. (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, 577-587.
- Picheny, M., Durlach, N. and Braidia, L. (1985). "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," *J. Speech Hear. Res.* 28, 96-103.
- Picheny, M., Durlach, N. and Braidia, L. (1986). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *J. Speech Hear. Res.* 29, 434-446.
- Picheny, M., Durlach, N. and Braidia, L. (1989). "Speaking clearly for the hard of hearing. III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech," *J. Speech Hear. Res.* 32, 600-603.
- Plack, C. J., and Oxenham, A. J. (1998). "Basilar-membrane nonlinearity and the growth of forward masking," *J. Acoust. Soc. Am.* 103, 1598-1608.
- Plack, C. J., and Drga, V. (2003). "Psychophysical evidence for auditory compression at low characteristic frequencies," *J. Acoust. Soc. Am.* 113, 1574-1586.
- Plomp, R. (1964). "Rate of decay of auditory sensation," *J. Acoust. Soc. Am.* 36, 277-282.
- Plomp, R., and Mimpen, A. M. (1978). "Speech-reception threshold for sentences as a function of age and noise level," *J. Acoust. Soc. Am.* 66, 1333-1342.
- Plomp, R., and Mimpen, A. M. (1979). "Improving the reliability of testing the speech reception threshold for sentences," *Audiology* 18, 43-52.

## References

- Plomp, R., and Mimpen, A. M. (1979). "Speech-reception threshold for sentences as a function of age and noise level," *J. Acoust. Soc. Am.* 66, 1333-1342.
- Plomp, R. (1986). "A signal-to-noise ratio model for the speech-reception-threshold of the hearing-impaired," *J. Speech Hear. Res.*, 29, 146-154.
- Plomp R (1988). "The negative effect of amplitude compression in multi-channel hearing aids in the light of the modulation-transfer function," *J Acoust Soc Am* 83, 2322-2327.
- Pollack, I. (1955). "Masking by a periodically interrupted noise," *J. Acoust. Soc. Am.* 27, 353-355.
- Rankovic, C. M. (1998). "Factors governing speech reception benefits of adaptive linear filtering for listeners with sensorineural hearing loss," *J. Acoust. Soc. Am.* 103, 1043-1057.
- Rankovic, C. M. (2002). "Articulation index predictions for hearing-impaired listeners with and without cochlear dead regions," *J. Acoust. Soc. Am.* 111, 2545-2548.
- Rhebergen, K.S., and Versfeld, N.J. (2005). "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.* 117, 2181-2192.
- Rhebergen, K. S., Versfeld, N. J., and Dreschler, W.A. (2005). "Release from informational masking by time reversal of native and non-native interfering speech," *J. Acoust. Soc. Am.* 118, 1274-1277.
- Rhebergen, K. S., Versfeld, N. J., and Dreschler, W.A. (2006a). "Learning effect observed for the speech reception threshold in interrupted noise with normal-hearing listeners," submitted to *J. Acoust. Soc. Am*
- Rhebergen, K. S., Versfeld, N. J., and Dreschler, W.A. (2006b). "Validation of the extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise for normal-hearing listeners, and suggestions for further improvement ," submitted to *J. Acoust. Soc. Am.*
- Rhebergen, K. S., Versfeld, N. J., and Dreschler, W.A. (2006c). "Predicting the intelligibility for speech in real-life background noises ," submitted to *Ear Hear.*
- Rosen, S. (1992). "Temporal information in speech: acoustic, auditory and linguistic aspects," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 336, 367-373.
- Shailer, M. J., and Moore, B. C. (1983). "Gap detection as a function of frequency, bandwidth, and level," *J. Acoust. Soc. Am.* 74, 467-473.
- Shailer, M. J., and Moore, B. C. (1987). "Gap detection and the auditory filter: phase effects using sinusoidal stimuli," *J. Acoust. Soc. Am.* 81, 1110-1117.
- Schlauch, R. S., Ries, D. T., and DiGiovanni, J. J. (2001). "Duration discrimination and subjective duration for ramped and damped sounds," *J. Acoust. Soc. Am.* 109, 2880-2887.
- Smits, J. C. M. (2006). "Hearing screening by telephone, fundamentals & applications," Doctoral thesis, Free University, Amsterdam.
- 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.

## References

- Souza, P. E., and Turner, C. W. (1998). "Multichannel compression, temporal cues and audibility," *J. Speech Lang. Hear. Res.* 41, 315-326.
- Souza, P. E., and Turner, C. W. (1999). "Quantifying the contribution of audibility to recognition of compression-amplified speech," *Ear Hear.* 20, 12-20.
- Souza, P. E. (2002). "Effects of compression on speech acoustics, intelligibility, and speech quality," *Trends Amplification* 6, 131-165.
- Souza, P. E., Jenstad, L. M., and Boike, K. (2006). "Measuring the acoustic effects of compression amplification on speech in noise," *J. Acoust. Soc. Am.* 119, 421-437.
- Stecker, G. C., and Hafter, E. R. (2000). "An effect of temporal asymmetry on loudness," *J. Acoust. Soc. Am.* 107, 3358-3368.
- Steeneken, H. J., and Houtgast, T. (1980). "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.* 67, 318-326.
- Steeneken, H. J. M., and Houtgast, T. (1980). "A physical method for measuring speech transmission quality," *J. Acoust. Soc. Am.*, 69, 318-326.
- Steeneken, H. J., and Houtgast, T. (1985). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.* 77, 1069-1077.
- Steeneken, H. J. (1992). "On measuring and predicting speech intelligibility," Doctoral thesis, University of Amsterdam.
- Steeneken, H. J., and Houtgast, T. (1999). "Mutual dependence of the octave-band weights in predicting speech intelligibility," *Speech com.* 28, 109-123.
- Steeneken, H. J., and Houtgast, T. (2002). "Validation of the revised STI<sub>r</sub> method," *Speech com.* 38, 413-425.
- Stelmachowicz, P., Lewis, D., Kalberer, A., and Creutz, T. (1994). "Situational hearing aid response profile users manual \_SHARP, v 2.0\_" Boys Town National Research Hospital, Omaha, NE.
- Stickney, G. S., and Assmann, P. F. (2001) "Acoustic and linguistic factors in the perception of bandpass-filtered speech," *J. Acoust. Soc. Am.* 109, 1157-1165.
- Stobich, B., Zierhofer, C. M., and Hochmair, E. S. (1999). "Influence of automatic gain control parameter settings on speech understanding of cochlear implant users employing the continuous interleaved sampling strategy," *Ear Hear.* 20, 104-116.
- Studebaker, G. A. (2005). *Personal communication*.
- Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. (1987). "A frequency importance function for continuous discourse," *J. Acoust. Soc. Am.* 81, 1130-1138.
- Studebaker, G. A., Sherbecoe, R. L., and Gilmore, C. (1993). "Frequency importance and transfer functions for the Auditec of St. Louis recordings of the NU-6 word test," *J. Speech Hear. Res.* 36, 799-807.
- Studebaker, G. A., Taylor, R., and Sherbecoe, R. L. (1994). "The effect of noise spectrum on speech recognition performance-intensity functions," *J. Speech Hear. Res.* 37, 439-448.
- Studebaker, G. A., Sherbecoe, R. L., McDaniel, D. M., and Gwaltney, C. A. (1999). "Monosyllabic word recognition at higher-than-normal speech and noise levels," *J. Acoust. Soc. Am.* 105, 2431-2444.

## References

- Studebaker, G. A., and Sherbecoe, R. L. (2002). "Intensity-importance functions for bandlimited monosyllabic words," *J. Acoust. Soc. Am.* 111, 1422-36.
- Summers, V. and Molis, M.R. (2004). "Speech Recognition in fluctuating and continuous maskers: effects of hearing loss and presentation level," *J. Speech Lang. Hear. Res.* 47, 245-256.
- ter Keurs, M., Festen, J. M., and Plomp, R. (1993). "Limited resolution of spectral contrast and hearing loss for speech in noise," *J. Acoust. Soc. Am.* 94, 1307-1314.
- Trine, T. D. (1995). "Speech recognition in modulated noise and temporal resolution: Effects of listening bandwidth," Unpublished doctoral dissertation, University of Minnesota, Twin Cities.
- Turner, C. W., and Henry, B. A. (2002). "Benefits of amplification for speech recognition in background noise," *J. Acoust. Soc. Am.* 112, 1675-1680.
- Uchanski R.M., Choi S.S., Braid L.D., Reed C.M., Durlach N.I. (1996). "Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate," *J. Speech Hear. Res.* 39, 494-509.
- van Buuren, R.A., Festen, and J.M., Houtgast, T., (1999). "Compression and expansion of the temporal envelope: evaluation of speech intelligibility and sound quality," *J. Acoust. Soc. Am.* 105, 2903-2913.
- van Tasell, D. J. (1993). "Hearing loss, speech, and hearing aids," *J. Speech Lang Hear. Res.*, 36, 228-244.
- van Wieringen, A. and Wouters, J. (2006). "De LIST en LINT: Nederlandstalige spraakaudiometrielijsten met zinnen en getallen," *Proceedings Nederlandse Vereniging voor Audiologie (Dutch Society of Audiology)*, 27-01-2006.
- van Wijngaarden, S.J. (2002). "*Past, Present and future of the speech transmission index*," *Proceedings of the International Symposium on STI, TNO Human Factors, (Soesterberg, The Netherlands)*.
- van Wijngaarden, S. J. (2003). "The intelligibility of non-native speech," *Doctoral thesis, Free University, Amsterdam*.
- van Wijngaarden, S. J. and Houtgast, T. (2004). "Effect of talker and speaking style on the SpeechTransmission Index," *J. Acoust. Soc. Am.* 115 38-41.
- Versfeld, N. J., Daalder, L., Festen, J. M., and Houtgast, T. (2000). "Method for the selection of sentence materials for efficient measurement of the speech reception threshold," *J. Acoust. Soc. Am.* 107, 1671-1684.
- Versfeld, N. J., and Dreschler, W. A. (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.
- Villchur, E. (1989). "Comments on "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function" [*J. Acoust. Soc. Am.* 83, 2322-2327 (1988)]," *J. Acoust. Soc. Am.* 86, 425-427.
- Warren, R.M. (1970). "Perceptual Restoration of Missing Speech Sounds", *Science, New Series*, 167, No. 3917. 392-393.
- Warren, R.M., Obusek, C. J., and Ackroff, J. M. (1972). "Auditory Induction: Perceptual Synthesis of Absent Sounds," *Science, New Series*, 176, No. 4039. 1149-1151.

## *References*

- Warren, R.M., Riener, K. R., Bashford, J. A., and Brubaker, B. S. (1995). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," *Perception & Psychophysics* 57, 175-182.
- Warren, R.M. (1999). "Auditory Perception: A New Analysis and Synthesis". New York: Cambridge University Press.
- Widin, G. P., and Viemeister, N. F. (1979). "Intensive and temporal effects in pure-tone forward masking," *J. Acoust. Soc. Am.* 66, 388-395.
- Wojtczak M, Schroder AC, Kong YY, and Nelson DA. (2001). "The effect of basilar-membrane nonlinearity on the shapes of masking period patterns in normal and impaired hearing," *J. Acoust. Soc. Am.* 109, 1571-86.
- Zeng, F. G., Grant, G., Niparko, J., Galvin, J., Shannon, R., Opie, J., and Segel, P. (2002). "Speech dynamic range and its effect on cochlear implant performance," *J. Acoust. Soc. Am.* 111, 377-386.

## *References*

## Samenvatting

## **Samenvatting**

Eén van de meest opvallende eigenschappen van het menselijke gehoor is dat het in staat is om gesproken taal omgeven door achtergrondgeluid te ontcijferen. In sommige gevallen kan spraak zelfs nog verstaanbaar zijn als het in niveau 20 tot 30 dB zachter is dan het achtergrondgeluid. Tot op heden is nog geen spraakherkenner in staat gebleken het menselijke gehoor op dit punt te evenaren. Het lijkt er op dat luisteraars met een normaal gehoor op een of andere manier in staat zijn gebruik te maken van de relatief stille periodes in het achtergrondgeluid. Fluctuaties in het achtergrond geluid resulteren dus in een betere spraakverstaanbaarheid. Verschillende studies hebben aangetoond dat luisteraars met een slecht gehoor minder goed in staat zijn om gebruik te maken van de relatief stille periodes in een fluctuerend achtergrondgeluid. Luisteraars met een gehoorverlies ervaren dit als een handicap als ze een gesprek willen volgen onder zulke omstandigheden.

De spraakverstaanbaarheid van een luisteraar in verschillende achtergrondgeluiden kan gemeten worden met de Speech Reception Threshold (SRT) test. De SRT test is een adaptieve spraakverstaanbaarheidstest die de signaal-ruisverhouding (Signal-to-Noise Ratio, SNR) bepaalt die nodig is om 50% van het zinsmateriaal te kunnen verstaan. Voor korte alledaagse zinnen halen luisteraars met een normaal gehoor een SRT van ongeveer -12 dB in fluctuerende ruis (met het gemiddelde spectrum van de spreker), terwijl hun SRT in stationaire ruis ongeveer -5 dB is. Luisteraars met een slecht gehoor zijn lang niet altijd in staat om een SRT te halen van ongeveer -5 dB in stationaire ruis, en bereiken in fluctuerende ruis vaak geen betere score dan hun SRT in stationaire ruis. Hierdoor kan de prestatie in fluctuerende ruis een goede indicatie zijn voor de prestatie in stationaire ruis, terwijl het omgekeerde niet geldt.

De spraakverstaanbaarheid in stationaire ruis kan voorspeld worden met het doorgaans veel gebruikte Speech Intelligibility Index (SII) model. Het SII model berekent het percentage spraakinformatie dat beschikbaar is voor de luisteraar in een gegeven spraak-in-ruis conditie.  $SII=0$  betekent dat er geen spraakinformatie beschikbaar is voor de luisteraar, terwijl  $SII=1$  betekent dat alle spraakinformatie hoorbaar is. Voor spraak (korte alledaagse zinnen) in stationaire ruis hebben luisteraars met een normaal gehoor ongeveer 33% van de totale spraakinformatie nodig ( $SII=0.33$ ) om de helft van de zinnen foutloos te verstaan. Als men aanneemt dat luisteraars op drempelniveau (de SRT) bij allerlei condities ongeveer een gelijke hoeveelheid spraakinformatie nodig

## *Samenvatting*

hebben, dan moet de SII voor alle die condities ongeveer 0.33 zijn. De SII wordt berekend uit de gehoordrempel van de luisteraar, het gemiddelde lange termijn spectrum van de ruis en het gemiddelde lange termijn spectrum van de spreker waardoor er geen rekening wordt gehouden met fluctuaties van het achtergrondgeluid. Hierdoor zal het SII model gelijke SRTs voorspellen voor spraak in stationaire en fluctuerende ruis. Vice versa, een SII berekend voor de werkelijke SRT in fluctuerende ruis zal onwaarschijnlijk laag zijn (een SRT in fluctuerende ruis van -12 dB geeft een SII van ongeveer 0.09 in plaats van 0.33).

In dit proefschrift is de spraakverstaanbaarheid in condities met fluctuerende ruis bij luisteraars met een normaal gehoor onderzocht. Het doel was om voor luisteraars met een normaal gehoor de gemeten SRT data voor zinsmateriaal te kunnen voorspellen in een groot bereik van fluctuerende ruis condities. Een modelmatige benadering voor spraakverstaanbaarheid in ruis kan meer inzicht geven in het systeem van spraakperceptie bij luisteraars met een normaal gehoor en kan mogelijk verklaren waardoor luisteraars met een gehoorverlies slechter presteren in fluctuerende ruis condities.

In hoofdstuk 2 wordt een nieuwe methode geïntroduceerd om de SRT in fluctuerende ruis voor luisteraars met een normaal gehoor te modelleren. De nieuwe benadering is een aanvulling op het veelvuldig gebruikte SII model, dat is gevalideerd voor de berekening van de spraakverstaanbaarheid in stationaire ruiscondities. Het principe van het SII model is dat de hoeveelheid spraakinformatie bepaald wordt die hoorbaar is wanneer spraak voor een deel gemaskeerd wordt door ruis of onder de gehoordrempel van de luisteraar valt. Het spraak- en ruissignaal wordt gefilterd in 21 frequentiebanden. In elke frequentieband wordt bepaald hoeveel spraakinformatie beschikbaar is hetgeen uiteindelijk resulteert in een SII waarde. De uitbreiding op het bestaande SII model is dat na het filteren in frequentiebanden de spraak en ruis opgedeeld worden in kleine tijdsintervallen. Binnen elke tijdsinterval wordt de conventionele SII berekend, die de hoeveelheid spraakinformatie representeert voor een luisteraar binnen dat tijdsinterval. Dit resulteert in een SII waarde die varieert in de tijd, ook wel de instantane SII genoemd. Het gemiddelde van de instantane SII resulteert uiteindelijk in een waarde tussen nul en één en dit geeft de gemiddelde hoeveelheid spraakinformatie weer. De uitbreiding op het SII model, het Extended SII (ESII) model, zoals beschreven in hoofdstuk 2, is in staat om de meeste SRT data uit de literatuur te beschrijven.

In specifieke situaties waar het spraakverstaan wordt gehinderd door interferentie van een tweede spreker, is de SRT gewoonlijk slechter dan

## *Samenvatting*

voorspeld. In dit proefschrift wordt dit fenomeen “informational masking” genoemd. In hoofdstuk 3 is een nieuwe methode geïntroduceerd die de hoeveelheid “informational masking” op de spraak meet bij interferentie van een tweede spreker. De studie toont aan dat met buitenlandse (niet verstaanbare) spraak als stoorspreker, luisteraars minder last hebben van “informational masking” dan met een verstaanbare stoorspreker. Bij deze studie is de hoeveelheid “informational masking” geschat op 6.6 dB. Omdat het ESII model geen rekening houdt met “informational masking”, zal de voorspelde SRT bij een verstaanbare stoorspreker 6 tot 7 dB lager liggen dan de gemeten SRT. Als men het ESII model gebruikt voor het voorspellen van de spraakverstaanbaarheid in aanwezigheid van interfererende spraak, moet men dus rekening houden dat de voorspelde SRT lager kan uitvallen dan de gemeten SRT.

Een ander aspect dat invloed kan hebben op de gemeten SRT is een leereffect bij het spraakverstaan in fluctuerende ruis. De resultaten in hoofdstuk 4 tonen aan dat er geen leereffect aanwezig is voor het spraakverstaan in stationaire ruis maar dat er wel een leereffect aanwezig is in blokruis. Deze uitkomsten suggereren dat bij toekomstige SRT experimenten in niet-stationaire ruis men op zijn minst herhaalde metingen moet doen en zo mogelijk wat training voorafgaand aan de SRT-test om te controleren voor leereffecten.

In hoofdstuk 5 wordt een studie beschreven met normaalhorenden voor verdere validatie van het ESII model zoals beschreven in hoofdstuk 2. Een serie van fluctuerende maskeerruizen waarmee het ESII model kritisch getest kan worden, is gebruikt om SRTs te meten bij luisteraars met een normaal gehoor. Met het toevoegen van een voorwaartse maskeerfunctie (FMF) kan het model significant worden verbeterd. De FMF kan bijvoorbeeld rekening houden met SRT verschillen als gevolg van de asymmetrische vorm van “omhullende” van de maskeerruizen.

Het dynamische bereik van het spraaksignaal kan een belangrijke rol spelen bij het modelleren van de spraakverstaanbaarheid in ruis omdat het bij een gegeven SNR de hoeveelheid spraakinformatie bepaalt die boven het ruisniveau uitkomt. Er wordt aangenomen dat een verandering van het dynamische bereik van spraak, bijvoorbeeld door compressie, de informatiedistributie verandert en derhalve invloed heeft op de SRT.

In hoofdstuk 7 is bij luisteraars met een normaal gehoor het effect van het dynamische bereik van het spraaksignaal in (ongecomprimeerde) stationaire- en blokruis bestudeerd. De uitkomsten van de studie laten een toename van de

## *Samenvatting*

spraakverstaanbaarheid zien als het spraaksignaal instantaan gecomprimeerd is met een compressie ratio (CR) van 2:1. De spraakverstaanbaarheid neemt weer af als het spraaksignaal verder wordt gecomprimeerd naar 4:1 (waarschijnlijk door toename van vervorming van het signaal), of wordt geëxpandeerd met een ratio van 1:1.5. Naast dit experiment is de spraakverstaanbaarheid bij simultane compressie op zowel de spraak en stationaire ruis als wel de spraak en blokruis gemeten. In stationaire ruis is de spraakverstaanbaarheid negatief beïnvloed door CR=4:1, terwijl in blokruis de spraakverstaanbaarheid positief beïnvloed wordt. Dit laatste is het gevolg van extra versterking van het spraaksignaal in de gaten van de blokruis. De distributie van de spraakinformatie kan beschreven worden door een Intensity Importance Function (IIF). De toevoeging van de IIF in het ESII model draagt bij tot iets betere voorspellingen van de spraakverstaanbaarheid.

Het ESII model zoals geïntroduceerd in hoofdstuk 2 en verfijnd in hoofdstuk 5 met de FMF kan nu ook rekening houden met voorwaartse maskering. Het model kan ook een verandering in temporele resolutie meenemen als een functie van maskeerniveau of als functie van gehoorverlies. De FMF beschrijft een toename van de temporele resolutie bij hogere niveaus en een afname in temporele resolutie bij een toenemend gehoorverlies. In hoofdstuk 8 is het ESII model gebruikt om de spraakverstaanbaarheid te voorspellen voor luisteraars met een gehoorverlies in stationaire ruis en blokruis. Naast het gebruik van de FMF is een methode geïntroduceerd om het effect van cochleaire compressie bij normaal- en slechthorenden te modelleren op de spraakverstaanbaarheid in ruis. De methode toont aan dat niet alleen de afname van de hoorbaarheid van het spraaksignaal het spraakverstaan reduceert, maar ook een afname van de temporele resolutie. Hierdoor zullen luisteraars met een gehoorverlies niet dezelfde SRT waarde kunnen halen als luisteraars met een normaal gehoor, zelfs bij gelijke hoorbaarheid van het spraaksignaal.

In dit proefschrift is een modelmatige benadering gebruikt om meer kennis te krijgen in de werking van het systeem van spraakperceptie. Door het modelleren van de spraakverstaanbaarheid in fluctuerend achtergrondgeluid kunnen we nu beter begrijpen waarom luisteraars met een normaal gehoor een gesprek nog goed kunnen volgen en hebben daarnaast meer inzicht gekregen waarom luisteraars met een slecht gehoor dit meestal niet goed kunnen. Meer onderzoek naar de werking van het menselijke gehoor in

## *Samenvatting*

fluctuerende (alledaagse) achtergrondgeluiden kan mogelijkwerwijs meer begrip geven over de handicap die een luisteraar met een slecht gehoor heeft.

Het ESII model is een beduidend betere methode om de spraakverstaanbaarheid te voorspellen dan de andere modellen doordat het in zowel stationaire als non-stationaire condities gebruikt kan worden. Daarom is het ESII model een waardevolle aanvulling op het bestaande SII (ANSI S3.5-1997) model. Door de instantane berekening van de spraakverstaanbaarheid, is het ESII model in potentie uitermate geschikt voor het evalueren van communicatie en hoortoestel algoritmes. Hierdoor ontstaat, naast het broodnodige, maar helaas tijdrovende, subjectieve testen, de mogelijkheid het ook mogelijk om in de toekomst met het ESII model de nieuwe hoortoestellen objectief te evalueren. Het zal dan wellicht makkelijker worden om sneller een indruk te kunnen krijgen over de effectiviteit van de diverse instelmogelijkheden van een hoortoestel bij een bepaald type gehoorverlies.

## Dankwoord



## Dankwoord

Het in dit proefschrift beschreven onderzoek is verricht op de afdeling KNO, Klinische & Experimentele Audiologie van het AMC, in de periode van 1 juli 2002 tot en met 30 juni 2006.

Promoveren doe je niet alleen. Ik wil daarom iedereen graag bedanken die op zijn of haar manier een inhoudelijke bijdrage heeft geleverd aan de totstandkoming van dit proefschrift. Mijn promotor Wouter Dreschler wil ik hartelijk bedanken voor zijn support en advies. Hij heeft in de afgelopen jaren een mooie researchgroep opgebouwd waar een goede teamgeest heerst en ik in alle vrijheid mijn onderzoek mocht doen. Mijn co-promotor Niek Versfeld wil ik in het bijzonder bedanken voor zijn begeleiding. De deur stond altijd open voor het stellen van vragen en bediscussiëren van ideeën. Verder heeft hij mij geweldig geholpen bij het gladstrijken van mijn soms kromme schrijfstijl.

De leden van de promotiecommissie Prof. dr. E. de Boer, Prof. dr. A.W. Bronkhorst, Prof. dr. W.J. Fokkens, Prof. dr. ir. T. Houtgast, Prof. dr. ir. L.C.W. Pols en Prof. dr. J. Wouters ben ik zeer erkentelijk voor de genomen moeite om dit proefschrift te beoordelen en te becommentariëren. Hun vriendelijke woorden, suggesties en bedenkingen heb ik erg gewaardeerd.

Lázló Körössy, Maarten van Beurden, Jan Koopman en Bas Franck wil ik hartelijk bedanken voor de technische ondersteuning en de vele inspirerende discussies. Daarnaast wil ik alle collega's van het AC bedanken voor de goede sfeer en gezelligheid. Extra dank gaat uit naar al diegenen die als proefpersoon zijn opgetreden. Zonder hun medewerking had het hele onderzoek nooit kunnen plaatsvinden. Voorts wil ik de "spraak in fluctuerende ruis" werkgroepleden, Tammo Houtgast, Joost Festen, Erwin George, Gaston Hilkhuisen, en "A.T.O." leden Johannes Lijzinga en Rolph Houben bedanken voor de vele discussies en overdenkingen.

Tot slot wil ik mijn familie en vrienden hartelijk bedanken voor hun welgemeende belangstelling. In het bijzonder wil ik Corine bedanken voor alle zorgen, geduld en de vele praktische tips in laatste face van mijn promotietraject. Ze heeft mij geweldig geholpen om dit proefschrift de juiste vorm te geven.



## Curriculum Vitae



## Curriculum vitae

Koenraad Sjoerd Rhebergen werd geboren op 24 april 1970 om 3:34 in het VU ziekenhuis te Amsterdam. Na de peuter, kleuter en lagere school in Ermelo en de MAVO in Harderwijk, behaalde hij in 1990 zijn VHBO diploma aan 't Knooppunt te Zwolle. In hetzelfde jaar begon hij aan de studie Fysiotherapie aan de Leidsche hogeschool. Via een kort uitstapje naar de studie Bewegingstechnologie aan de Haagsche hogeschool en een jaar Bewegingswetenschappen aan de Vrije Universiteit in Amsterdam is hij uiteindelijk in het voorjaar van 1995 Psychologie gaan studeren aan de Universiteit Leiden. Tijdens zijn studie heeft hij zich in het bijzonder gericht op de psychofysische keuzevakken door de geestdrift van Dr. G ten Hoopen. Zijn stage liep hij bij TNO Technische Menskunde (TNO Human Factors; thans TNO Defence, Security and Safety) in Soesterberg op de afdeling perceptie, sectie gehoor onder begeleiding van Dr. R. Drullman en Dr. J Vos. Zijn afstudeeronderzoek werd gedaan op het slaap en waak laboratorium van de Universiteit Leiden bij de vakgroep Functieleer aan de faculteit Sociale Wetenschappen (o.b.v. Drs. M. Varkevisser en Prof. dr. G. Kerkhof). In maart 2001 behaalde hij zijn doctoraalexamen Psychologie met specialisatie Experimentele en Theoretische Functieleer (Cognitieve Psychologie). Na zijn afstuderen is hij werkzaam geweest als onderzoeker bij het Slaap en Waakcentrum in het Westeinde ziekenhuis (MCH) te Den Haag (o.b.v. Prof. dr. G.A.Kerkhof). Van juli 2002 tot en met juni 2006 was hij aangesteld als assistent in opleiding bij de afdeling Keel, Neus en Oorheelkunde van het Academisch Medisch Centrum in Amsterdam bij de vakgroep Klinische en Experimentele Audiologie van Prof. dr. ir. W.A. Dreschler onder begeleiding van Dr. ir. N.J. Versfeld, alwaar het in dit proefschrift beschreven onderzoek werd verricht. Sinds juli 2006 is hij werkzaam als onderzoeker op dezelfde afdeling en doet onderzoek op het gebied van modelleren van de spraakverstaanbaarheid bij slechthorenden.



