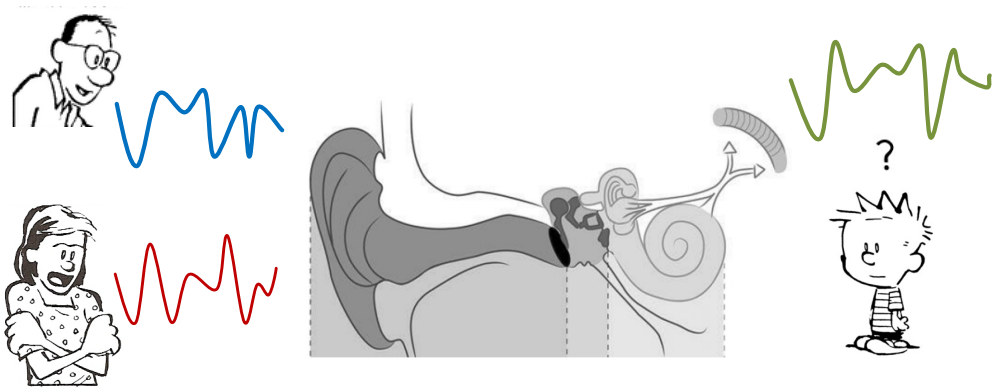
Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 13

Claus Forup Corlin Christiansen

Listening in adverse conditions: Masking release and effects of hearing loss



Centre for Applied Hearing Research
Department of Electrical Engineering

Listening in adverse conditions: Masking release and effects of hearing loss

PhD thesis by
Claus Forup Corlin Christiansen

Technical University of Denmark
2012

Cover illustration: Ear Drawing by *Science Photo Library* and clippings from *Calvin and Hobbes*
Preprint version for the assessment committee.
Pagination will differ in the final published version.

This PhD-dissertation is the result of a research project at the Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark (Kgs. Lyngby, Denmark).

The project was financed partly by the Danish research council and partly by Oticon, Widex and GN Resound through a research consortium. The external stay at the University of Illinois was further supported by travel grants from Mogens Balslevs Fond, Otto Mønsteds Fond, the Oticon Foundation and Brorsons Fond.

Supervisor

Prof. Torsten Dau
Centre for Applied Hearing Research
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Abstract

Speech perception is a complex process involving the ability to detect the speech signal, separate it from interfering sounds and decode the transmitted speech information. In contrast to normal-hearing (NH) listeners, hearing-impaired (HI) listeners often show a large reduction in the masking release (MR), which is the improvement in speech intelligibility when the interferer is different from steady-state noise (e.g., a competing talker). MR is usually measured as the difference in speech reception thresholds (SRTs), the signal-to-noise ratio (SNR) where 50% of the speech is understood, and has mainly been linked to the ability to separate the target from the interferer. However, it is still not clear why HI listeners show a reduced MR and how the ability to decode the speech information is affected by impaired hearing. Thus, the purpose of this thesis was to investigate MR in both NH and HI listeners, to study the effects of hearing loss on the ability to decode speech, and to establish a framework for modeling speech intelligibility based on an auditory processing model.

The first part of the thesis established the modeling framework and showed that, by using a model that captures the processing of the different stages of the auditory system, it is possible to predict speech intelligibility using a very simple back end. Furthermore, the results indicated that the high-energy segments are the most important for speech intelligibility.

The second part focused on recent indications that the large reduction in MR often observed in HI listeners is a result of measuring the MR of HI listeners at a higher signal-to-noise ratio (SNR) in stationary noise relative to NH listeners. The present work presented noise-band vocoded as well as low-pass and high-pass filtered stimuli to NH listeners, thereby decreasing their speech intelligibility and making it possible to compare the MR of NH and HI listeners not only at the same SNR, but also at the same same percent correct, which was not done in previous studies. The MR was found to be only partially related to the SRT obtained in stationary noise. Furthermore, for a competing talker, noise-vocoding strongly reduced the MR of the NH listeners to that obtained with HI listeners. This indicated that deficits in coding of temporal fine structure and fundamental frequency (F0) information may play a critical role for the reduced MR of the HI listeners.

The third part investigated the contribution of high-rate envelope fluctuations, at the output of the auditory filters, to MR. High-rate envelope fluctuations are produced by the interaction between unresolved harmonics and are related to the F0 of voiced speech. A new vocoder technique was developed to effectively attenuate the high-rate envelope fluctuations. Furthermore, high-pass filtering was used to reduce the amount of F0 information from resolved harmonics. The results showed high-rate envelope fluctuations, related to the F0, were sufficient to obtain a large MR. Furthermore, F0-related information from resolved harmonics were also sufficient for MR. However, when both high-rate envelope fluctuations and F0-related information from resolved harmonics were reduced, the MR was strongly reduced. Thus, the results indicated that F0 information is crucial for MR, but that it does not matter if it is obtained from low-order resolved harmonics or from high-rate envelope fluctuations produced by interaction between unresolved harmonics.

The final avenue of investigation focused on the effects of hearing loss on the ability to decode

speech by measuring consonant confusions for both individual HI listeners and also individual utterances of the same consonants. In general, the results showed that individual HI listeners consistently confused the presented utterances with only one other consonant, and that most of the HI listeners actually made the same confusions. The results also indicated that the reason for the large variability in the confusion patterns of HI listeners observed in previous studies is that different utterances of the same consonant promote different confusions and that the HI listeners experience problems with different utterances.

Overall, this thesis provides insights about the large MR observed for NH listeners, why this MR is often reduced for HI listeners and in which way impaired hearing affects the ability to decode speech information.

Resumé

Taleopfattelse er en kompleks proces, der omfatter evnen til at detektere talesignalet, adskille det fra forstyrrende lyde samt at afkode den transmitterede taleinformation. I modsætning til normalthørende udviser hørehæmmede personer ofte en stor reduktion i *masking release* (MR), hvilket er defineret som en forbedring i taleforståeligheden målt med en fluktuerende støjkilde (f.eks. en forstyrrende taler) i forhold til en stationær støjkilde. MR er hovedsagelig blevet forbundet med evnen til at adskille tale fra en forstyrrende lydkilde. Det er imidlertid ikke helt klart, hvorfor hørehæmmede personer har nedsat MR samt hvordan evnen til at afkode taleinformation påvirkes af høretab. Formålet med denne afhandling var derfor at undersøge MR i både normalthørende og hørehæmmede personer, at studere indflydelsen af høretab på evnen til at afkode tale og at udvikle en metode til at forudsige taleforståelighed baseret på en model af hørelsen.

I den første del af afhandlingen blev en taleforståelighedsmodel udviklet og det viste sig, at ved at anvende en model som efterligner signalbehandlingen i de forskellige faser af det auditive system, er det muligt at forudsige taleforståeligheden ved hjælp af en meget enkel kognitiv fase. Endvidere viste resultaterne, at den bedste forudsigelse af taleforståeligheden blev opnået ved kun at betragte de forholdsvis få talesegmenter indeholdende den højeste energi.

Den anden del fokuserede på de seneste indikationer af, at den store reduktion af MR i hørehæmmede personer kan være et resultat af at deres MR ofte bliver målt ved et højere signal-støj-forhold (SNR) i stationær støj sammenlignet med normalthørende personer. I denne del af studiet blev de normalthørende personer testet med *noise-vocoded* såvel som lavpas- og højpas-filtreret stimuli for dermed at mindske deres taleforståelighed. Dette gjorde det muligt at sammenligne MR målt med normalthørende og hørehæmmede personer, ikke kun ved samme SNR, men også ved samme taleforståelighed (procent korrekte sætninger), hvilket ikke er blevet gjort i tidligere undersøgelser. Ved hjælp af denne metode viste det sig at MR kun var delvist relateret til SNR i stationær støj. Resultaterne viste yderligere, at da støjen bestod af en konkurrende taler og stimuli var *noise-vocoded*, blev den målte MR for normalthørende reduceret til de samme niveau som målt med hørehæmmede. Dette indikerede, at en nedsat evne til at gøre brug af den temporale fin-struktur og fundamental frekvensen (F0) af talernes stemmer, formentlig spiller en afgørende rolle for den nedsatte MR hos hørehæmmede personer.

I den tredje del blev det undersøgt hvordan højfrekvente udsving i indhyldningskurven ved udgangen af de auditive filtre bidrager til MR. Højfrekvente udsving i indhyldningskurven opstår på grund af vekselvirkninger mellem uopløste harmoniske komponenter og indeholder information om talens F0. En ny *vocoder* teknik blev udviklet til effektivt at dæmpe de højfrekvente udsving i indhyldningskurven. Endvidere blev et højpasfilter brugt til at reducere den F0 information som kan opnås baseret på opløste harmoniske komponenter. Resultaterne viste at højfrekvente udsving i indhyldningskurven, indeholdende F0 information, var tilstrækkelig til at opnå en stor MR. Derudover var stemme-information baseret på opløste harmoniske komponenter også tilstrækkelig til at opnå en stor MR. Da begge disse typer af F0 information blev kraftigt dæmpet, blev der til gengæld målt en stor reduktion af MR. Dermed viste resultaterne, at F0 information er afgørende for MR, men at det stort set er ligegyldigt, om informationen stammer fra opløste

harmoniske komponenter eller fra højfrekvente udsving i indhyldningskurven opstået på grund af vekselvirkninger mellem uopløste harmoniske komponenter.

Den sidste del af afhandlingen fokuserede på et høretabs indflydelse på evnen til at afkode taleinformation. Det blev udført ved at analysere konsonantforvekslinger for individuelle hørehæmmede personer samt hver enkelt udtalelse af en bestemt konsonant. Overordnet set viste resultaterne at de enkelte hørehæmmede personer forvekslede de præsenterede udtalelser med kun én anden konsonant, og at de fleste af dem faktisk lavede de samme forvekslinger. Resultaterne viste også, at årsagen til den store variation i forvekslings-mønstrene som der tidligere er blevet observeret i forbindelse med hørehæmmede personer skyldes, at forskellige udtalelser af den samme konsonant giver anledning til forskellige forvekslinger, og at de hørehæmmede personer har problemer med forskellige af udtalelserne.

Samlet set giver denne afhandling indsigt i den store MR som normalthørende personer typisk opnår, samt hvorfor denne MR ofte er reduceret hos hørehæmmede personer. Derudover giver afhandlingen også indblik i, hvordan høretab påvirker evnen til at afkode taleinformation.

Preface

This PhD thesis is a result of a little more than 3 years work at the Centre for Applied Hearing Research (CAHR) at the Technical University of Denmark (DTU).

My PhD-study at CAHR have without comparison been the most challenging thing I have ever done. However, I am truly amazed about how much I have grown, both personally and professionally, and I am extremely glad that I took one of the most anxious decisions in my life. I have certainly build op a foundation that I can use for the rest of my life.

The most important reason why I made it all the way is the support and supervision from Torsten. Torsten has an extremely broad knowledge within his field as well as an amazing ability to see the big picture and choose the right directions in a project. At the same time Torsten has a natural understanding of other people, when they need to be challenged and when they need support. Most importantly, Torsten has a huge enthusiasm that influences everybody working with him, and then he just cares very much about all the people in the group, not only professionally, but also personally.

I would like to thank Jont Allen for letting me into his home and family during my external stay at the Human Speech Recognition group at the University of Illinois.

Also big thanks to Christian Lorenzi for inviting me to spend some time with him and his really nice group in Paris.

Furthermore, I'm very grateful to Ewen MacDonald who fortunately decided to travel all the way from Canada to join our group in Copenhagen. All though you have only been a part of my PhD for the last six months, you have already been a great help and support.

Finally, but most importantly, I would like to express my limitless gratitude to my family who have been a big support for me, especially my wife Gry and daughter Alma who came into this world during my time at CAHR. I love you with all my heart.

Claus Christiansen, July 2012.

Contents

Abstract v

Resumé på dansk ix

Preface xiii

Table of contents xiv

List of abbreviations xix

1 General introduction 1

2 Prediction of speech intelligibility based on an auditory preprocessing model 7

2.1 Introduction 8

2.2 Modeling speech intelligibility 11

2.2.1 Speech intelligibility model based on auditory signal processing 11

2.2.2 Speech intelligibility models based on SII and STI 14

2.2.3 Parameters in the speech intelligibility models 15

2.3 Experiment I: Speech intelligibility in noise 16

2.3.1 Experimental data 16

2.3.2 Model predictions 17

2.4 Experiment II: Intelligibility of speech processed with binary masks 18

2.4.1 Experimental data 18

2.4.2 Model predictions 20

2.5 Discussion 22

2.5.1 Capabilities and limitations of the intelligibility models 22

2.5.2 Perspectives 26

2.6 Summary and conclusions 26

3 Relationship between masking release in fluctuating maskers and speech reception thresholds in stationary noise 31

3.1 Introduction 32

3.2 Masking release for hearing-impaired listeners 35

3.2.1 Methods 35

3.2.2 Results 37

3.3 Masking release for NH listeners obtained with processed stimuli 38

3.3.1 Methods 38

3.3.2 Results 40

3.4 Comparison of results for normal and impaired hearing 44

3.5 Discussion 45

3.5.1 Relation between the MR and the SNR in stationary noise 45

3.5.2 Importance of low- and high-frequency information 47

3.5.3 Distortion of carrier information 47

3.5.4 Effects of filtering on SRT and MR 48

3.5.5 Relation between audiometric thresholds and speech perception 48

3.5.6 Effects of the linguistic content 49

3.6 Summary and conclusions 49

4 Contribution of high-rate envelope fluctuations to release from speech-on-speech masking 53

4.1 Introduction 54

4.2 Signal processing 56

4.3 Methods 59

4.3.1 Listeners 59

4.3.2 Speech material 59

4.3.3 Procedure 59

4.4 Results 60

4.5 Discussion 61

4.5.1 Summary of the main results 61

4.5.2 The role of resolved and unresolved harmonics for MR 61
4.5.3 Possible connections to reduced MR in HI listeners 62
4.5.4 Implications for auditory modeling 63
4.6 Summary and conclusions 63

5 Analyzing the variation of consonant confusions in hearing-impaired listeners 67
5.1 Introduction 68
5.2 Method 70
5.2.1 Listeners 70
5.2.2 Stimuli 70
5.2.3 Procedure 70
5.3 Results 71
5.3.1 Consonant recognition scores 71
5.3.2 Consonant confusions 74
5.4 Discussion 79
5.4.1 Summary of main results 79
5.4.2 A possible explanation for different consonant confusions 80
5.4.3 Consonant confusions and auditory functions 80
5.5 Summary and conclusions 81

6 Summary and final thoughts 83

References 85

Bibliography 87

Collection volumes 99

List of abbreviations

AI	Articulation index
AN	Auditory nerve
ANOVA	Analysis of variance
BM	Basilar membrane
CLUE	Conversational language understanding evaluation
CV	Consonant-vowel
CVC	Consonant-vowel-consonant
DT	Danish talker
ERB	Equivalent rectangular bandwidth
ERB_N	Equivalent rectangular bandwidth number scale
ESII	Extended speech intelligibility index
F0	Fundamental frequency
FFT	Fast Fourier transform
HI	Hearing-impaired
HINT	Hearing in noise test
HL	Hearing level
HP	High-pass
IBM	Ideal binary mask
IF	Instantaneous frequency
ISI	Inter-spike interval
ISTS	International speech test signal
ITFS	Ideal time-frequency segregation
LC	Local criterion
LP	Low-pass
MR	Masking release
MSE	Mean-square-error
NH	Normal-hearing
PTA	Pure-tone average threshold
RMS	Root-mean-square
SAM	Sinusoidally amplitude modulated
SDR	Signal-to-distortion ratio

SII	Speech intelligibility index
SNR	Signal-to-noise ratio
SPL	Sound pressure level
SRT	Speech reception threshold
SSN	Speech-shaped noise
STI	Speech transmission index
STMI	Spectro-temporal modulation index
TF	Time-frequency
TFS	Temporal fine structure

1

General introduction

The hearing system is very important for the acquisition of language and the development of speech. Our hearing enables us to communicate with other people in a time where interaction and communication are more important than ever. Communication is the foundation of fellowship and solidarity and, thus, is important for the quality of life and the development of personality. Our hearing provides us with information about our surroundings and warns us against potential dangers from all directions. In addition, hearing enables us to enjoy music, acquire knowledge, listen to radio and follow TV broadcasts.

Hearing impairment has been shown to have a negative effect on a number of psychosocial factors leading to loneliness, depression, low self-esteem, and reduced quality of life (Shield, 2006). According to the World Health Organization (WHO), hearing impairment in children often leads to delayed development of language and cognitive skills and hearing impairment in adults often makes it difficult to obtain, perform, and keep jobs (WHO, 2012). In 2004, the WHO estimated that over 275 million people in the world had moderate-to-profound hearing impairment. Furthermore, it is assumed that approximately 20% of the adult population of Europe (Shield, 2006) and the United States (Lin et al., 2011) have a hearing impairment of 25 dB HL or greater, while nearly 40% of people older than 65 years are estimated to have a disabling hearing impairment (WHO, 2012). Conductive hearing losses, due to problems in the outer or middle ear, can often be treated with surgery. However, most sensorineural hearing losses, caused by damage to the inner ear, auditory nerve or more central auditory stages, can only be compensated for by hearing aids or cochlear implants.

The sound we receive through our ears often consists of a complex mixture of sounds coming from all directions. As described by Helmholtz in 1863, the healthy auditory system possesses a remarkable ability to separate the sounds originating from different sources (von Helmholtz, 1912). Furthermore, normal-hearing (NH) listeners have an amazing ability to follow the conversation of a single speaker in the presence of others, a phenomenon known as the "cocktail-party problem" (Cherry, 1953). Later, speech intelligibility in the presence of fluctuating noise or a competing talker was shown to be much higher than in stationary noise even when the signal-to-noise ratio (SNR) was the same (Duquesnoy and Plomp, 1983; Festen and Plomp, 1990). This effect was called masking release (MR) and explained by the ability to "listen-in-the-dips" of the masker.

The single most common complaint among people with hearing loss is the difficulty in understanding speech in complex acoustic environments, such as background noise, reverberation or competing talkers. Although compensating for the reduced sensitivity (e.g., by hearing aids) largely improves the ability to understand speech in quiet, most hearing-impaired listeners still show great

difficulties in noise (e.g., Duquesnoy and Plomp, 1983; Gustafsson and Arlinger, 1994; Shanks et al., 2002; Hällgren et al., 2005; Metselaar et al., 2008). In contrast to NH listeners, hearing-impaired (HI) listeners do not benefit to the same degree when the masker is fluctuating noise or a competing talker and exhibit very little or no MR (e.g., Festen and Plomp, 1990; Gustafsson and Arlinger, 1994; Peters et al., 1998; George et al., 2006; Lorenzi et al., 2006; Bernstein and Grant, 2009; Strelcyk and Dau, 2009).

When speech is masked by noise or interfering sound sources, such as other speakers, the recognition of the target message relies on a three-step process. First of all, the listener must be able to detect the acoustic energy of the target speech. This is not possible in the time and frequency regions where the masker is much stronger, a situation defined as energetic masking (e.g., Kidd et al., 1998; Freyman et al., 1999; Brungart, 2001). Secondly, the listener must be able to identify the time and frequency regions that belong to the target and the interferer in order to extract the spectro-temporal energy pattern that corresponds to the target speech. The difficulty in extracting the spectro-temporal energy pattern of the target speech due to an interfering sound source has been termed non-energetic or informational masking (Brungart et al., 2006). Thirdly, the listener must be able to decode the meaning of the spectro-temporal energy pattern.

While speech perception in stationary noise has generally been explained in terms of energetic masking (e.g., French and Steinberg, 1947; Steeneken and Houtgast, 1980), speech perception in fluctuating noise and competing speech is often linked to non-energetic masking and problems with target and masker segregation (e.g., Qin and Oxenham, 2003; Hopkins et al., 2008; Brungart et al., 2006, 2009). Voiced speech, generated by vibration of the vocal folds, consists of frequency components (harmonics) that all are integer multiples of the fundamental frequency (F0) which corresponds to the period of the vocal fold vibration. The F0 has been shown to play an important role for the perceptual segregation of concurrent and sequential sources (Brokx and Nooteboom, 1982; Darwin, 1997) and several studies have suggested that the reduced MR exhibited by HI listeners is due to deficits in the processing of F0 information (e.g., Qin and Oxenham, 2003; Lorenzi et al., 2006; Hopkins et al., 2008; Strelcyk and Dau, 2009). Due to the increasing bandwidth of the auditory filters with increasing frequency, the low-order harmonics are considered to be spectrally resolved by the cochlea while the high-order harmonics are considered to be unresolved. While several studies have indicated that F0 information is conveyed primarily by the low-order harmonics, which are resolved by the auditory system (Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994; Bernstein and Oxenham, 2003; Plomp, 1967; Micheyl and Oxenham, 2007; Bird and Darwin, 1998), it is still unclear what role resolved and unresolved harmonics play for speech perception and, in particular, for MR. Recently, it has been proposed that the amount of MR depends on the SNR in stationary noise of which the MR is measured. Bernstein and Grant (2009) suggested that the reduced MR observed in HI listeners is mainly caused by the difference in the SNR in stationary noise used when determining MR for NH versus HI listeners.

While there has been a substantial number of studies investigating the effect of hearing impairment on the ability to detect and segregate speech by measuring speech reception thresholds (SRTs) in individual HI listeners in the presence of various types of interferers, very few studies have

investigated the effect of individual hearing loss on the ability to decode speech. A recent study showed large differences in the patterns of consonant confusions between HI listeners and even between the ears of the same HI listener (Phatak et al., 2009). Furthermore, a study by Phatak et al. (2008) found a large variability of confusions across different utterances of the same consonants in NH listeners. The variability of the confusions for the different utterances was linked to different utterances that have different conflicting cues promoting specific confusions (Li et al., 2010). By extension, the large variation across HI listeners could be caused by the HI listeners having problems with different utterances. Thus, in order to understand how individual hearing loss affects the ability to decode speech, consonant confusions of individual listeners should be investigated on an utterance-by-utterance basis and combined with a spectro-temporal analysis of the presented utterances.

The purpose of the projects presented in this thesis was three fold: to setup a framework for modeling speech perception; to investigate the effect of the stationary-noise SNR for MR in NH and HI listeners; and to investigate the importance of F0 information for MR as represented in resolved and unresolved harmonics. Furthermore, the effect of individual hearing impairment on the ability to decode speech was also studied.

Chapter 2 investigates the ability to predict speech intelligibility by combining a psychoacoustically validated model of auditory preprocessing (Dau et al., 1997a) with a simple central stage that describes the similarity of the internal representation of the test and a reference signal. Specifically, the ability to predict the intelligibility of speech processed with ideal time-frequency segregation (ITFS) is investigated, a technique used to study the effects of energetic and non-energetic masking by removing all the spectro-temporal segments of the noisy speech where the SNR is below given threshold (Brungart et al., 2006). The performance of the developed speech intelligibility model is compared to speech-based versions of the classical speech transmission index (STI) and speech intelligibility index (SII).

Chapter 3 studies whether the reduced MR observed in HI listeners is a result of measuring the MR at a higher SNR in stationary noise than used for NH listeners. In contrast to Bernstein and Grant (2009), the stimuli presented to the NH listeners in the present study are distorted by noise-vocoding as well as low- and high-pass filtering, in order to shift the psychometric functions of the NH listeners to the same range of SNRs that HI listeners are tested at. This makes it possible to measure the MR of the NH and HI listeners both at the same SNR and at the same point on the psychometric function. Furthermore, the distortions make it possible to study whether different frequency regions of the speech or the temporal fine structure are particular important for MR.

Spectrally unresolved harmonics produce high-rate envelope fluctuations at the output of the cochlear filters which are related to the F0 of voiced speech. In **Chapter 4**, it is investigated if these high-rate envelope fluctuations contribute to MR. Oxenham and Simonson (2009) investigated if F0 information provided by the low-order resolved harmonics is important for MR. They used low-pass (LP) and high-pass (HP) filtered stimuli in order to either retain or eliminate low-order harmonics. Both conditions strongly reduced the MR indicating that F0 information from resolved

harmonics alone is not sufficient for MR. However, it is unclear whether the large reduction in the MR found in Oxenham and Simonson (2009) was caused by removing F0 information or due to the filtering process strongly reducing the bandwidth and thereby also the redundancy in the speech signal. A new vocoder technique is developed in this chapter to attenuate the F0-related envelope fluctuations produced by the unresolved harmonics, while preserving the speech message conveyed by the slow-varying envelope. In addition, reduced F0 information from the resolved harmonics is also investigated by HP filtering of the stimuli. In order to preserve as much energy in the speech as possible, the HP filtering uses a much lower cut-off frequency than in Oxenham and Simonson (2009) where the resolved harmonics were removed completely.

Chapter 5 investigates the effect of hearing impairment on the ability to decode speech in terms of consonant confusions. Based on the finding that different utterances of the same consonant induce different confusions, consonant confusions of individual HI listeners are investigated on an utterance-by-utterance basis in order to get a better understanding of how individual hearing loss affects the ability to decode speech.

Finally, **chapter 6** summarizes the main results and discusses the implications of the findings for auditory modeling, for diagnosing individual hearing loss and for hearing-aid processing.

2

Prediction of speech intelligibility based on an auditory preprocessing model *

Classical speech intelligibility models, such as the speech transmission index (STI) and the speech intelligibility index (SII) are based on calculations on the physical acoustic signals. The present study predicts speech intelligibility by combining a psychoacoustically validated model of auditory preprocessing [Dau et al., J. Acoust. Soc. Am. 102, 2892-2905 (1997)] with a simple central stage that describes the similarity of the test signal with the corresponding reference signal at a level of the internal representation of the signals. The model was compared with previous approaches, whereby a speech in noise experiment was used for training and an ideal binary mask experiment was used for evaluation. All three models were able to capture the trends in the speech in noise training data well, but the proposed model provides a better prediction of the binary mask test data, particularly when the binary masks degenerate to a noise vocoder.

* This chapter is based on Christiansen et al. (2010).

2.1 Introduction

Speech is by far the most important method of communication between humans. However, the transmission of speech can be affected by numerous factors, such as background noise, room reverberation, hearing loss and distortions in hearing aids or other communication devices. Modeling speech intelligibility can help to understand how speech is processed and which parts of the speech signal are important for the successful recognition of the message. Furthermore, a model provides immediate results and can be used continuously in order to examine large sets of data. For many purposes, a model can also replace comprehensive testing with test subjects.

Speech intelligibility was first predicted by French and Steinberg (1947) who introduced the concept of the articulation index (AI). Fundamentally, the AI predicts the speech intelligibility by calculating the signal-to-noise ratio (SNR) between the long-term speech spectrum and the long-term background noise spectrum in a number of frequency bands. In the 1980s and 1990s, the AI was extended in a number of different studies, which were integrated in a new method called the speech intelligibility index (SII; ANSI S3.5, 1997). This version included corrections for hearing sensitivity loss, speech level as well as upward and downward spread of masking.

In the SII, the long-term spectrum of the clean speech and the background noise has to be known in advance. However, in nonlinear systems where distortions are introduced by the processing rather than background noise, it is not possible to apply the SII in the traditional configuration. In order to apply the SII to nonlinear transmission systems, Kates and Arehart (2005) extended it to work with a reference speech signal and a corresponding distorted speech signal and to include the nonlinear distortions peak-clipping and center-clipping.

In order to account for temporal effects on speech intelligibility, such as reverberation in a room, Steeneken and Houtgast (1980) proposed a physical method for evaluating the quality of speech-transmission channels, called the speech transmission index (STI). Similar to the SII method, the STI method is based on the SNR in a number of frequency bands. However, for the STI calculation, the SNR in each band is related to the reduction of amplitude modulations caused by the transmission system. The reduction of modulations is determined by the decrease of the modulation index of sinusoidally modulated noise signals in different modulation frequency bands, divided into different audio frequency bands.

In order to apply the STI to nonlinear transmission systems, such as hearing aids and modern communication systems, several researchers have developed variants of the STI that use speech rather than modulated noise as the probe signal. Payton and Braida (1999) calculated the reduction of modulation power from the modulation spectra of a reference speech signal and a corresponding distorted speech signal. Other speech-based variants of the STI calculated the decrease in modulation power by the cross-spectral density between the reference and the distorted envelope spectrum (Drullman et al., 1994; Payton et al., 2002). Ludvigsen et al. (1990) and Holube and Kollmeier (1996) used the cross-correlation between the reference and the distorted envelope to calculate the SNR in each band directly. Goldsworthy and Greenberg (2004) examined the ability

of the methods of Ludvigsen et al. (1990), Drullman et al. (1994), Holube and Kollmeier (1996) and Payton et al. (2002) to account for nonlinear distortions caused by spectral subtraction and envelope thresholding. Rhebergen and Versfeld (2005) did not predict effects of nonlinear distortions, but developed the extended SII (ESII), where the SII calculation was divided into short time frames in order to account for fluctuating noise types. However, the ESII requires access to the target speech and the interfering noise separately and cannot be used in cases where the speech is degraded or enhanced by some type of signal processing algorithm.

The SII, STI and variants of these all include properties of auditory frequency selectivity in the calculations. Some of the models include ad-hoc corrections to account, to some extent, for upward spread of masking. Still, the models operate on the physical signals and do not consider various aspects and principles of auditory signal processing. Also, they do not take the portions of the speech signal into account that are masked or emphasized by the processing in the auditory system.

The goal of the present study was to use a physiologically motivated model of the auditory processing and to base the speech intelligibility predictions on the internal representations of the stimuli. Models of the auditory periphery have been used relatively extensively for the prediction of audio and speech *quality* (e.g., Beerends and Stemerdink, 1992; Beerends et al., 2002; Huber and Kollmeier, 2006; Karjalainen, 1985; Kim, 2005; Nielsen, 1993; Thiede et al., 2000). In contrast, only a few studies (Elhilali et al., 2003; Holube and Kollmeier, 1996) have attempted to predict speech *intelligibility* based on models of auditory signal processing.

In Holube and Kollmeier (1996), consonant-vowel-consonant (CVC) words were presented after an announcement sentence. The target word was chosen among five alternatives differing only in one of the phonemes. Recognition scores were simulated by processing the test word and the five alternatives with the auditory model of Dau et al. (1996a, 1997a). The alternative with the smallest distance to the test word at the level of the internal representation of the stimuli was considered as the recognized word. This means that the model of Holube and Kollmeier (1996) can only be used in an experimental setup where there are reference words available for the identification of each test word.

Elhilali et al. (2003) developed the spectro-temporal modulation index (STMI) motivated by evidence of spectro-temporal receptive fields of neurons in the primary auditory cortex. The STMI analyzes the temporal and spectral modulations contained in an auditory spectrogram produced by a model of the auditory periphery. By comparing the spectro-temporal modulation content of the distorted speech signal with that of the reference speech signal, the STMI was able to predict speech intelligibility for additive noise, reverberation, phase-jitter and phase-shift. In order to predict effects of presentation level and hearing-impairment on speech intelligibility, Zilany and Bruce (2007) introduced a more physiologically detailed model of the normal and impaired auditory periphery (Zilany and Bruce, 2006) in the STMI.

The auditory processing model used in the STMI was developed to simulate auditory-nerve (AN) fiber responses in cats, but the ability to account for psychoacoustic detection and masking data in humans has not yet been considered in detail. In contrast, the auditory processing model used in

Holube and Kollmeier (1996) was originally developed to account for numerous psychoacoustical detection and masking experiments in humans (e.g., Dau et al., 1997a,b; Derleth and Dau, 2000; Jepsen et al., 2008; Verhey et al., 1999).

The present study presents a new speech intelligibility model where the preprocessing is based on the psychoacoustically validated model of Dau et al. (1996a, 1997a). In contrast to the approach of Holube and Kollmeier (1996), the presented model is based on a comparison of the reference signal and the distorted signal and is not restricted to an experimental setup where the target word is chosen from a number of given alternatives. The speech intelligibility model was trained on data from one experiment and tested on data from another experiment. In the first experiment, which was exclusively used for training, the percentage of correctly identified words was predicted as a function of the signal-to-noise ratio (SNR) for speech-shaped noise, cafeteria noise, car noise and bottle noise, where the cafeteria noise was characterized as fluctuating. In the second experiment, which was exclusively used for testing, the mixtures from the first experiment were processed with a new type of signal processing, termed ideal time-frequency segregation (ITFS), which was introduced by Brungart et al. (2006). This technique can be regarded as an extension of the concept of the ideal binary mask (IBM) (Hu and Wang, 2004; Wang, 2004). Ideal binary masking is essentially a filtering technique that preserves time-frequency segments (TF-units) where the target signal is stronger than the masker (SNR above 0 dB) and eliminates segments where the masker is stronger (SNR below 0 dB). In ITFS, the IBM principle is extended by replacing the constant threshold of 0 dB by a local criterion (LC). The LC can be changed in order to produce an output signal where more or less of the TF units are retained. The present study focuses on the prediction of existing speech intelligibility data by Kjems et al. (2009). In that study, it was shown that, for an intermediate range of LC values (20 - 40% ones in the mask), the intelligibility was close to 100% and independent of the overall mixture SNR, suggesting that the cues from the mask were sufficient for the target speech to be recognized. For high LC values (below 15% ones in the mask), the effect of the mixture SNR was absent or very weak and it was argued that fundamental frequency (F0) tracking, periodicity, and temporal fine structure, did not affect the speech intelligibility in this region. Thus, the ITFS processing tests a given speech perception model's sensitivity to the overall spectro-temporal structure of the signal, rather than the temporal fine structure. Furthermore, the masks with only a few ones investigate if small time-frequency regions of very high energy result in high predicted scores even if the measured speech intelligibility might be low.

The predictions of the developed model were compared to the predictions of a speech-based version of the SII and the STI. In the following, the properties of the proposed model as well as the classical speech intelligibility predictors (SII and STI) are described.

2.2 Modeling speech intelligibility

2.2.1 Speech intelligibility model based on auditory signal processing

The auditory periphery

A schematic overview of the auditory perception model of Dau et al. (1997a) is shown in Fig. 2.1. The model transforms the acoustic signal into a time-varying spectro-temporal internal representation. The incoming acoustic signal is filtered by a linear fourth-order gammatone filterbank (Patterson et al., 1987) in order to roughly simulate the frequency selectivity of the cochlea (Ruggero et al., 1997; Békésy, 1960). The filterbank consists of 32 band-pass filters with center frequencies ranging from 100 to 8000 Hz, equally spaced on the equivalent rectangular bandwidth (ERB) scale, each with a bandwidth of 1 ERB. In the following stages, the output from each frequency band is processed separately. First, the output is half-wave rectified and low-pass filtered at 1 kHz, where the half-wave rectification introduces a DC value in the signal. This stage roughly simulates the transformation of the mechanical basilar membrane oscillations into receptor potentials in the inner hair cells (Palmer and Russell, 1986; Pickles, 1988; Plack, 2005). The low-pass filtering essentially preserves the temporal fine structure of the signal for low frequencies and extracts the envelope of the signal for high frequencies. Effects of nonlinear adaptation as observed in the auditory nerve (e.g., Smith, 1977; Westerman and Smith, 1984) are roughly simulated by a chain of five nonlinear adaptation loops. Each loop consists of a dividing element and a low-pass filter, where the input of the loop is divided by a low-pass filtered version of the output. An initially low value of the low-pass filtered output causes a strong overshoot at the onset of a stimulus. This high onset value is reduced as the low-pass filtered output is raised and eventually reaches a steady-state level where $I/O = O$. Thus, the steady-state response of a single loop to a stationary stimulus can be expressed as $O = \sqrt{I} = I^{\frac{1}{2}}$, which becomes $I^{\frac{1}{2}^5} = I^{\frac{1}{32}}$ for the entire chain of nonlinear adaptation loops. Thus, variations in the input signal that are rapid compared to time constants of the adaptation loops are transformed linearly, whereas slow variations and stationary signals are compressed according to an approximately logarithmic compression, resulting in a higher sensitivity for fast temporal fluctuations. When the stimulus is switched off, the output of the adaptation loops does not immediately return to the initial conditions due to the charge on the capacitors. This property of the nonlinear adaptation stage makes it possible to simulate effects of forward masking. The time constants, which are 5, 50, 129, 253 and 500 ms, were determined in order to account for forward-masking experiments. After the nonlinear adaptation stage, an 8-Hz low-pass filter extracts the envelope of the pre-processed signal. The cut-off frequency of the modulation low-pass filter was determined in simulations of psychoacoustical masking experiments described in Dau et al. (1996b). A modulation filterbank that was developed in later studies (Dau et al., 1997a), was not considered in the present investigation in order to limit the complexity of the simulations. In the final stage, a constant-variance internal noise is added in order to simulate the limited resolution of the auditory system in the framework of this model.

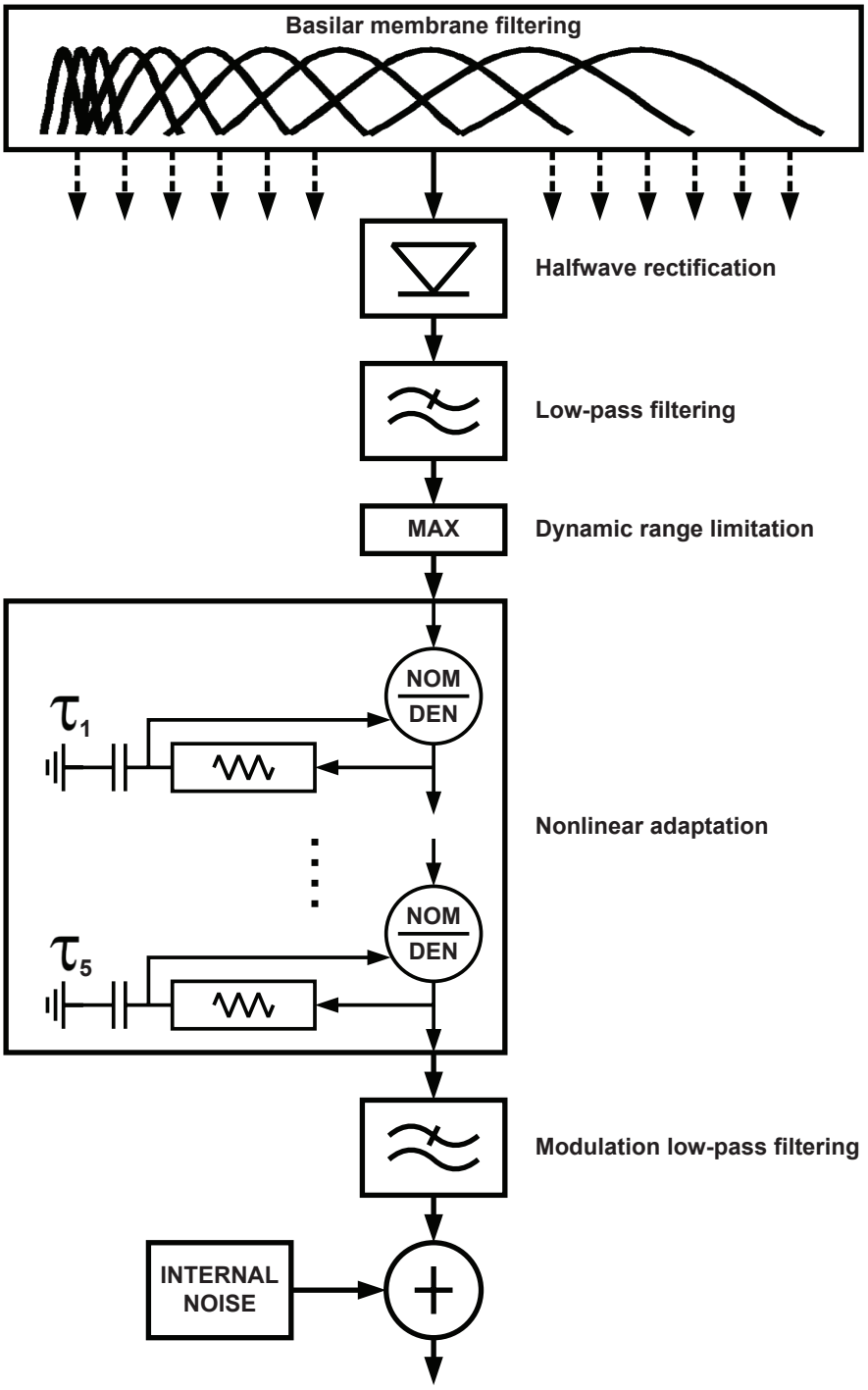


Figure 2.1: The stages of the model auditory perception model Dau et al. (1997a). A gammatone filterbank divides the acoustic signal into several frequency bands each followed by half-wave rectification, 1-kHz low-pass filtering and nonlinear adaptation. Subsequently, modulation low-pass filtering is applied and a constant-variance internal noise is added to limit the resolution.

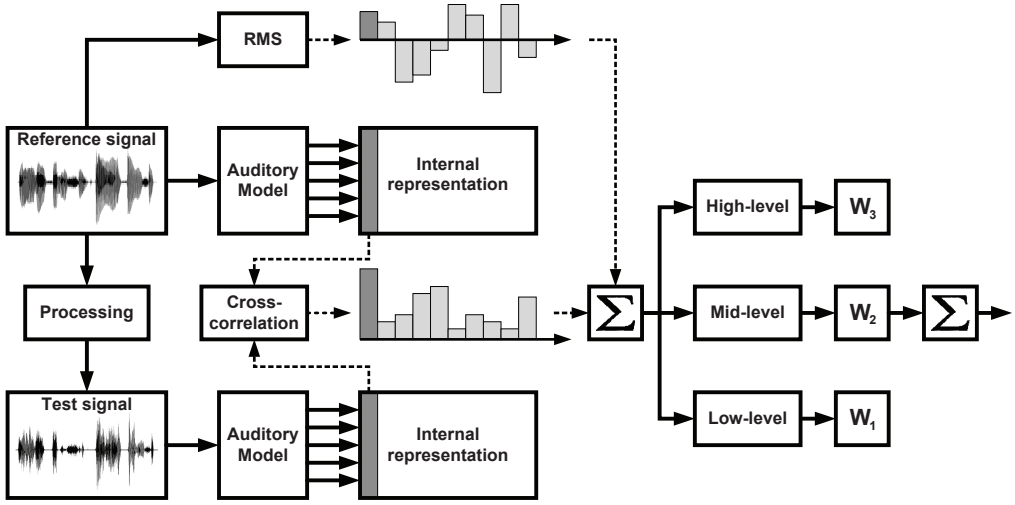


Figure 2.2: Schematic of the proposed speech intelligibility model. The reference and the distorted speech signal are transformed into internal representations by the auditory perception model. Subsequently, the linear cross-correlation coefficient and the root-mean-square (rms) level are calculated in frames of 20 ms, and each frame is classified as high-, mid- or low-level. For each level the average cross-correlation coefficient for the corresponding frames is obtained and the final score is a linear weighting of the three level scores, which is converted to predicted intelligibility by a logistic function.

Central processing

Figure 2.2 shows an overview of the proposed speech intelligibility model. First, the reference speech signal and the corresponding distorted speech signal are processed through the auditory model. The output represents the internal representations of the stimuli. In the following stage, the linear cross-correlation coefficient between the two internal representations is calculated in frames of 20 ms every 10 ms, resulting in an overlap of 50%. The optimum frame length of 20 ms was found empirically based on simulations and corresponds to findings in Huber and Kollmeier (2006). In the same stage, the root-mean-square (rms) level of each frame of the reference signal is determined and compared to the overall rms level of the entire reference signal. Each frame is thereby categorized as high-, mid- or low-level. In the next stage, an overall score is calculated for each level by averaging the cross-correlation coefficients of all frames corresponding to that level. In the final stage, the model output is calculated by a linear weighting of the three level scores and a logistic function is used to transform the model output to predicted intelligibility. The level-based calculation as well as the logistic transformation was motivated in the study by Kates and Arehart (2005) where it was argued that not all segments in a speech signal are equally important. Kates and Arehart found that the mid-level segments were most important whereas low-level segments had very little importance and high-level segments were not relevant. High-level segments are defined here as having an rms level of 0 dB or higher, relative to the overall rms level. The mid-level segments are limited to the range from -5 to 0 dB and the low-level segments are those with a relative rms of -15 dB to -5 dB. These values were chosen in order to obtain an approximately even distribution of the frames among the three levels such that each level covers about the same duration of the sentence. The definition of the three levels is slightly different from the values chosen in

Kates and Arehart (2005). The expression and the parameter values for the linear weighting of the three levels and the logistic conversion are provided in section 2.2.3.

2.2.2 Speech intelligibility models based on SII and STI

Speech-based SII

The speech-based SII model developed by Kates and Arehart (2005) essentially estimates the speech and noise power spectra from a reference speech signal and a corresponding distorted speech signal. These spectra are provided as input to the traditional SII method defined in the ANSI S3.5 (1997) standard. Furthermore, the speech-based SII replaces the SNR in the traditional SII by a signal-to-distortion ratio (SDR) which also includes nonlinear distortions. The speech and noise power spectra are estimated through the calculation of the magnitude squared coherence (MSC; Carter et al., 1973), also known as the normalized cross-power spectrum. In Kates and Arehart (2005), the calculations of the speech and noise power spectra as well as the SDR were performed in short time frames. The procedure is summarized in appendix 2.6. In order to improve the predictions, Kates and Arehart (2005) divided the procedure into three levels in the way described above (section 2.2.1). The final score is calculated by a linear weighting of the three levels and converted to predicted intelligibility by a logistic function. The expression for the logistic function and the parameter values for the linear weighting are given in section 2.2.3.

Speech-based STI

The speech-based STI model, which uses speech as the probe stimulus instead of modulated noise, was developed by Koch (1992) and Holube and Kollmeier (1996). In the speech-based STI, the clean and the degraded speech signals are band-pass filtered using a gammatone filterbank with 32 channels between 100 to 8000 Hz. In each band, the envelope is extracted by half-wave rectification followed by a low-pass filtering with a cutoff frequency of 50 Hz. The average SNR in each band, SNR_k,is calculated via the cross-correlation coefficient r between the reference envelope and the test envelope. Koch (1992) and Holube and Kollmeier (1996) showed that SNR_k is related to r by the expression:

$$SNR_k = 10\log\left(\frac{r^2}{1-r^2}\right) \tag{2.1}$$

From the SNR_k the speech-based STI score is obtained by the same procedure as used in the classical STI method (Steeneken and Houtgast, 1980). The final score is converted to predicted speech intelligibility by a logistic function. The expression and the parameter values for the logistic function are provided in section 2.2.3.

2.2.3 Parameters in the speech intelligibility models

In the case of the proposed model and the speech-based SII method, the final model output is obtained by a linear weighting of the three level scores:

$$c = w_{low}r_{low} + w_{mid}r_{mid} + w_{high}r_{high} \tag{2.2}$$

where w_{low}, w_{mid} and w_{high} are the weights and r_{low},r_{mid} and r_{high} represent the level scores. This step was not used in the speech-based STI method, where the calculation is not divided into three levels. The output of all three models was converted to predicted intelligibility by a logistic function:

$$I = \frac{1}{1 + e^{(O-c)/S}} \tag{2.3}$$

where O and S define the offset and slope of the logistic function, respectively, and c represents the model output.

The parameters were fitted to 2/3 of the data of the first experiment and the fitting was validated on the last 1/3 of the data. The fitting was applied simultaneously to all four noise types and thus not tailored to a single noise type, which should limit the risk of over-fitting. The parameters were then kept constant for the predictions of the data in the second experiment. This was done in order to evaluate the model in an unknown condition. An alternative approach could have been to divide the data sets from both experiments into training subsets and test subsets. In this way, the models could have been trained and tested on both experiments, either separately or collectively. However, due to the high consistency in the data, this probably would not have been a real challenge for the models.

The fitting was performed using a constrained nonlinear optimization, where the sum of the squared errors was minimized. In the cases of the proposed model and the speech-based SII method, where a linear weighting of the three levels was applied, the weights were constrained to be positive and the sum of the weights was defined to be 1. The offset of the logistic function was limited to lie between -1 and 1, whereas the slope was left unconstrained. The parameters of the three models are listed in table 2.1. It can be seen in the table that the speech-based SII results in the best fit to the training data when only the mid-level frames are used in the calculation. This is in agreement with the results of Kates and Arehart (2005) where the best fit was found when the mid-level frames were weighted much higher than the low- and high-level frames. In contrast, the best fit for the proposed model was obtained when only the high-level frames were considered in the calculation. This result is further discussed in section 2.5."

	S	O	w_{low}	w_{mid}	w_{high}
Proposed model	0.056	0.39	0	0	1
Speech-based SII	0.067	0.19	0	1	0
Speech-based STI	0.075	0.20	-	-	-

Table 2.1: Parameters of the three models fitted to the data in the first experiment. The first two columns contain the slope and the offset parameters of the logistic function and the linear weighting of the low-, mid- and high-level frames are shown in columns three to five.

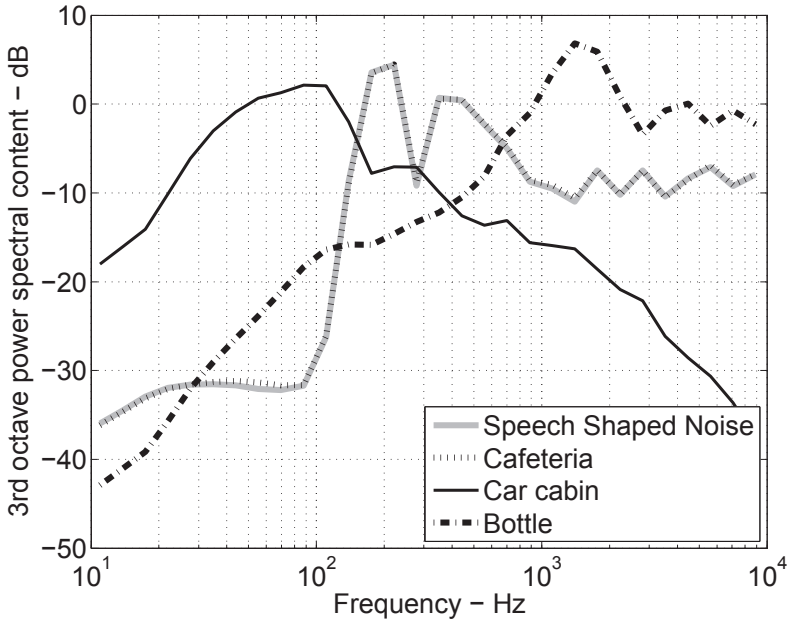


Figure 2.3: Long-term spectrum of the four maskers used in the listening experiments. While the cafeteria masker has been equalized to have the same long-term spectrum as the SSN masker, the car cabin masker and the bottle masker contain high energy at low and high frequencies, respectively. Reproduced from Kjems et al. (2009).

2.3 Experiment I: Speech intelligibility in noise

2.3.1 Experimental data

The psychoacoustical data were collected by Kjems et al. (2009). They measured speech intelligibility using the Dantale II sentence test (Wagener et al., 2003), which is a Danish version of the Hagerman sentence test (Hagerman, 1982a,b, 1984a,b; Hagerman and Kinnefors, 1995). The sentences consist of five words placed in a fixed grammatical structure (name, verb, numeral, adjective, object), such as, "Linda owns eight white boxes". The sentences were generated by choosing each word randomly from a list of 10 alternatives. Kjems et al. (2009) used four different noise types as maskers in the experiments: stationary speech-shaped noise, cafeteria noise, car noise and bottle noise.

Figure 2.3 shows the long-term spectra of the four maskers. The speech-shaped noise (SSN) was created by superimposing 30 Dantale II sentence sequences, with random intervals of separation (Wagener et al., 2003). The cafeteria noise was a recording of a continuous conversation in Danish

between a male and a female talker in a cafeteria environment. As seen in Fig. 2.3, the cafeteria masker was equalized to have the same long-term spectrum as the SSN after recording. However, in contrast to the SSN, it is fluctuating in time. The car cabin noise was a recording from a car driving on a highway and represents a low-frequency masker. Finally, the bottle noise, which is a recording of bottles on a conveyor belt in a bottling hall, represents a high frequency masker. In Kjems et al. (2009), the psychometric functions were measured in order to determine the SNRs of the mixtures used in their second experiment, where they were processed with the binary masks. In the present study, these psychometric functions obtained by Kjems et al. (2009) were used as training data for the speech intelligibility models.

Figure 2.4 (upper left panel) shows the measured psychometric functions obtained by Kjems et al. (2009). The percentage of correctly identified words for each noise type is plotted as a function of the SNR. The SSN masker (open boxes) results in a very steep psychometric function that changes rapidly from 0% to 100% intelligibility when the SNR is increased. The psychometric function for the cafeteria masker (light gray diamonds) has a slightly shallower slope and a slightly lower SRT than the SSN masker. Since the cafeteria masker has the same long-term spectrum as the SSN masker, the better performance for the cafeteria masker is due to its fluctuating nature and the possibility for the listener to listen in the gaps (e.g., Festen and Plomp, 1990; Miller and Licklider, 1950; Peters et al., 1998). For the bottle masker (filled circles), the psychometric function is similar to that obtained for the cafeteria masker but slightly shallower. The large amount of high-frequency energy in the bottle masker does not mask the speech signal as effectively as the SSN masker. Finally, the low-frequency content of the car masker (dark gray triangles) results in a psychometric function which is shifted horizontally towards lower SNRs compared to the SSN masker. This is probably because much of the energy in the car masker is located at frequencies below 200 Hz, which does not overlap strongly with the spectrum of the speech. The psychometric function for the car masker changes rapidly from low to high scores with increasing SNR, and has a slope similar to that of the psychometric function of the SSN masker.

2.3.2 Model predictions

Figure 2.4 also shows the predicted psychometric functions obtained with the proposed model (upper right panel), the speech-based STI (lower left panel) and the speech-based SII (lower right panel). The three models show very similar results. All psychometric functions have the characteristic shape with a relatively sharp transition from 0% at low SNRs to 100% at high SNRs. For the SSN masker, the cafeteria masker and the car masker, the SRTs obtained with the three models lie within a few dB of the measured data. However, for the bottle masker, the three models predict a psychometric function that is shifted horizontally towards higher SNRs, resulting in SRTs of about 10 dB higher than the data. A reason for this discrepancy could be that the high frequency energy of the bottle noise effectively masks most of the speech signal of the high frequencies and the low-frequency bands therefore should be provided with more gain in the models. It is also possible that the rhythm and the impulsive characteristic of the bottle noise make it easier for the

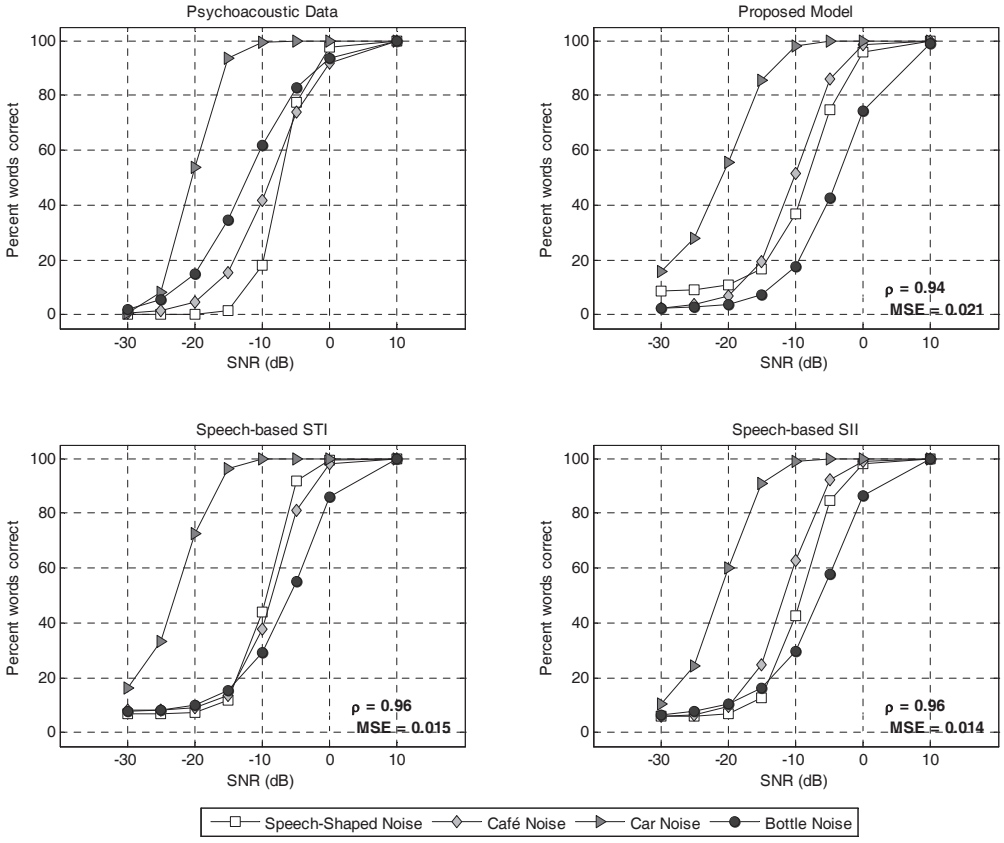


Figure 2.4: Psychometric functions (percentage of correctly identified words as function of SNR) measured experimentally (upper left panel, data reproduced from Kjems et al. (2009)), and predicted by the proposed model (upper right panel), the speech-based STI method (lower left panel) and the speech-based SII method (lower right panel). Each panel shows the psychometric functions obtained using the SSN masker (white boxes), the cafeteria masker (light-grey diamonds), the bottle masker (dark-grey triangles) and the car masker (black circles). The linear cross-correlation coefficient and the mean-square-error (MSE) quantifies the strength of the linear relationship and the error between the predicted and measured scores, respectively, based on the remaining 1/3 of the data used for validation.

listener to perceptually segregate the speech from this noise masker, which is not captured in any of the models.

2.4 Experiment II: Intelligibility of speech processed with binary masks

2.4.1 Experimental data

As in the first experiment, the psychoacoustical data obtained in this experiment were collected by Kjems et al. (2009). They processed the noisy speech from the first experiment with the ideal time-frequency segregation (ITFS) technique (Brungart et al., 2006), where the local SNR criterion (LC) determines whether a given time-frequency region (TF-unit) is preserved in the output signal. Increasing the LC reduces the number of the TF-units where the local SNR is above the threshold

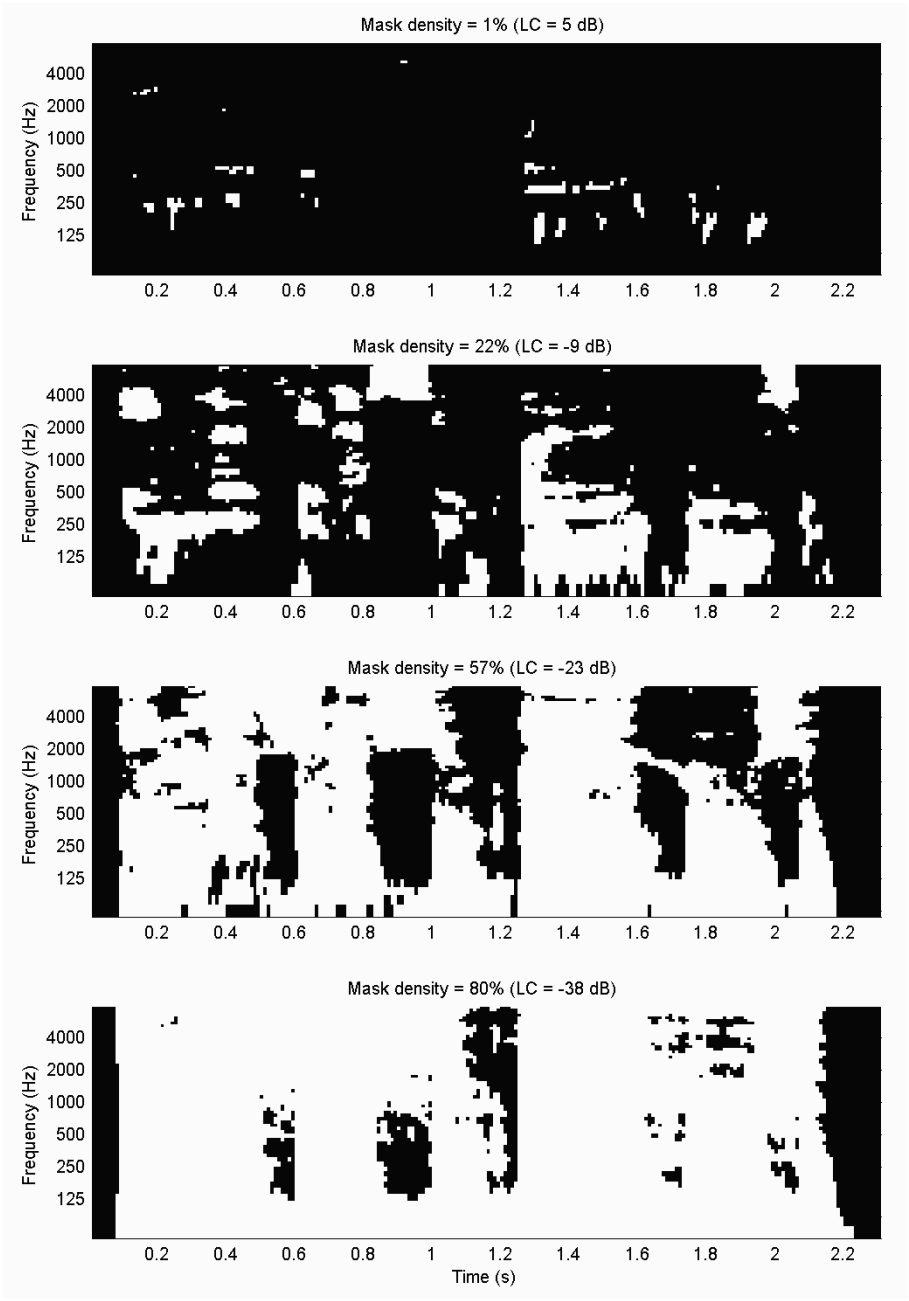


Figure 2.5: Example of four binary masks with densities (percentages of one's) of (a) 1% (b) 22% (c) 57% (d) 80%. The white regions (ones) represent the parts of the signal that are preserved in the processed signal and the black regions represent the parts that are removed (zeros). The corresponding LC values describe the minimum SNR of the mixture in a given region in order to preserve it.

and decreases the number of ones in the binary mask. Similarly, a reduction in the LC increases the number of ones in the mask. The number of ones in the mask is referred to as the mask density. Figure 2.5 illustrates the masks obtained using four different LC values, resulting in four different mask densities. In Kjems et al. (2009), the ITFS processing was performed with 8 different mask densities, including a mask only containing ones that results in an unprocessed mixture. In order to examine the dependency of the overall SNR level, they measured the intelligibility for overall SNRs corresponding to 20% and 50% correctly identified words obtained from the psychometric

functions in the first experiment. Furthermore, to examine the noise vocoding ability of the IBM, Kjems et al. (2009) also measured the speech intelligibility at an SNR of -60 dB, which essentially is pure noise.

The experimentally obtained results for the ITFS processed SSN mixtures are shown in the upper left panel of Fig. 2.6. Only the SSN masker is considered here, since the other maskers show similar results. The percentage of correctly identified words is shown as a function of the mask density (i.e. the percentage of ones in the mask) for the overall SNR values of -7.3 dB (open boxes, 50% SRT), -9.8 dB (light gray diamonds, 20% SRT) and -60 dB (dark gray triangles). The data show a very characteristic pattern as a function of the mask density. Beginning in the right most part of the panel, the mask density is 100% and the mixtures are actually unprocessed. Here, the intelligibility depends on the overall SNR of the mixtures. For the higher SNRs, there are more speech cues (F0, formants and temporal fine structure) present which lead to a higher performance. The results obtained using an overall mixture SNR of -7.3 dB show that, when the density is decreased, the performance is increased. This is because more zeros are included in the mask and the TF-units with the lowest local SNRs are removed. These TF-units contain only very little or no speech information and removing them helps the listener to separate the speech from the masker. As the mask density is decreased further, the performance reaches 100% intelligibility in the region of densities between approximately 60% and 10%. In this region, the binary mask has removed sufficient interfering noise in order for the speech to be identified successfully. For densities below about 10%, an increasing amount of the speech information is removed and intelligibility drops rapidly to a low value of about 35%. The results obtained using mixture SNRs of -9.8 dB and -60 dB also show an increase in performance as the density of the mask is decreased. However, the performance is still better for the higher SNRs due to more speech cues in these mixtures. For the mixtures at -9.8 dB and -60 dB the performance also reaches 100% intelligibility (in a slightly narrower region of densities between 40% and 10%). In this region, the performance is independent of the overall mixture SNR and therefore not affected by the additional speech cues present in the mixtures with higher SNRs. Thus, in this region, the binary mask successfully separates the TF-units containing speech information from the TF-units containing noise. At densities below 10%, the results are almost identical to the results of the -7.3 dB SNR mixtures.

2.4.2 Model predictions

The predictions obtained with the proposed model, the speech-based STI and the speech-based SII are also shown in Fig. 2.6. The simulated functions obtained with the proposed model (upper right panel) are very similar to the measured data for the three mixture SNRs. There are only slight deviations at mask densities above 50% where the model slightly overestimates the performance. The speech-based STI (lower left panel) generally accounts for the intelligibility patterns. However, for the sparsest masks, the simulated intelligibility values are more than 40% higher than the measured data. The patterns predicted by the speech-based SII (lower right panel) deviate strongly from the measured data, particularly in the case of the -60 dB mixtures. For this model, for the

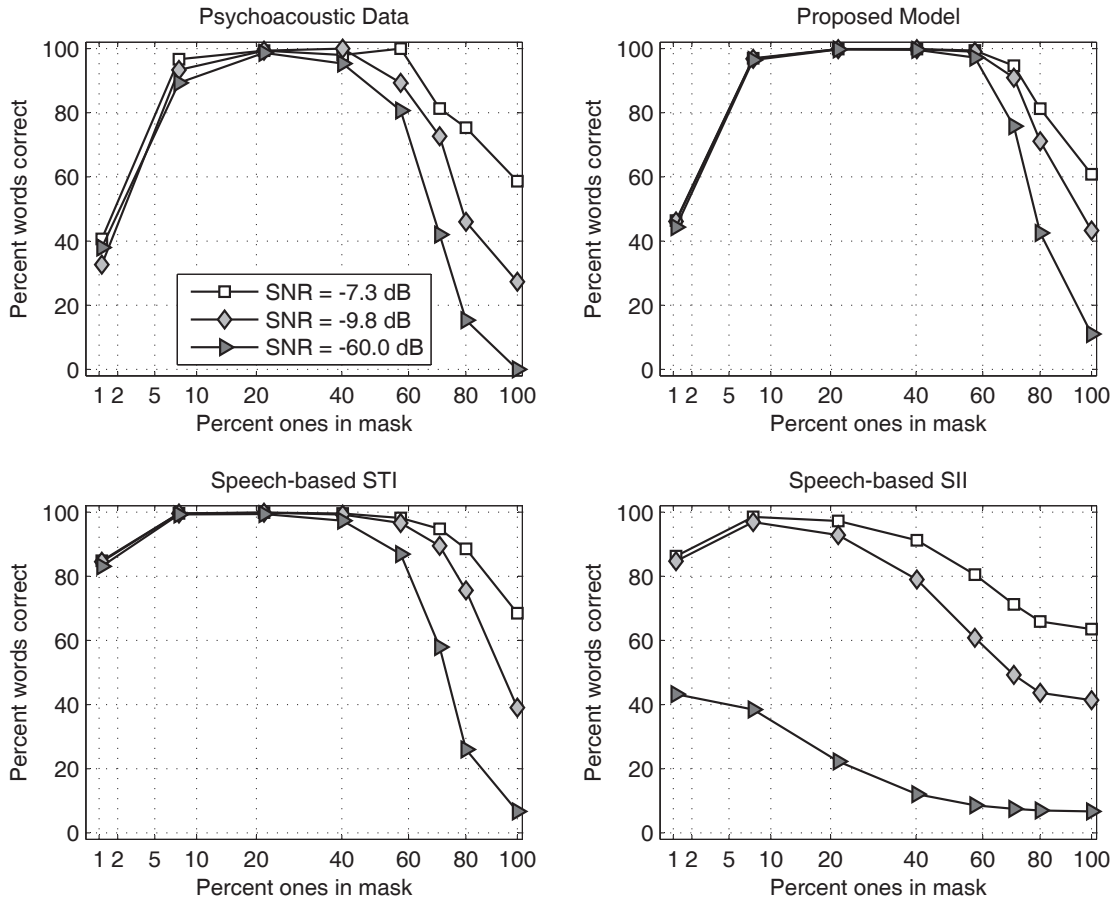


Figure 2.6: Percentage of correctly identified words as a function of mask density (percentage of ones in the mask) measured experimentally (upper left panel), and predicted by the proposed model (upper right panel), the speech-based STI method (lower left panel) and the speech-based SII method (lower right panel). The results in all the panels are based on the SSN as masker signal. The SNRs correspond to the SRTs that result in 50% (white boxes) and 20% (light-grey diamonds) correct responses for the SSN masker and the lowest SNR is -60 dB (dark-grey triangles). Since all four masker types show similar results only the SSN masker is included.

two highest mixture SNRs, the predicted intelligibility is very high for the sparsest masks and reaches a maximum at mask densities around 5%. At densities above 10%, the performance drops slowly in contrast to the measured data where a maximum at densities of about 20% was found. In the experiment with the -60 dB mixtures, the predictions of the speech-based SII only reach 40% compared to nearly 100% in the measured data. Furthermore, this maximum is located at the sparsest mask with a density of 1%.

Figure 2.7 shows the correlation between the predicted and the measured data. The left, middle and right columns present the results for the proposed model, the speech-based STI and the speech-based SII, respectively. This figure summarizes the results of the SSN masker (first row) presented above as well as the results obtained with the cafeteria masker (second row), the car masker (third row), and the bottle masker (fourth row). The squares represent -60 dB SNR and the circles and triangles represent the SNR values corresponding to 20% and 50% correctly identified words, respectively, obtained for the unprocessed mixtures. The dashed lines indicate the diagonal; ideally, the data points would lie on this line. Points above this line reflect an underestimation of

intelligibility in the predictions, whereas points below the line represent an overestimation in the predictions.

The cross-correlation coefficient, ρ, and the mean-square-error (MSE) are indicated in each panel. The results of the proposed model are fairly close to the dashed line for all four noise types (ρ between 0.88 and 0.96, MSE between 0.017 and 0.032). However, the model has a slight tendency to overestimate the scores consistent with the observations from Fig. 2.6 described above. The speech-based STI also shows to be consistent with the data (ρ between 0.84 and 0.89, MSE between 0.038 and 0.06) but overestimates the speech intelligibility consistently in several conditions. This is mainly caused by the very sparse masks as discussed further below. Consistent with the observations from Fig. 2.6, the results obtained with the speech-based SII model are relatively scattered for all four maskers (ρ between 0.47 and 0.66, MSE between 0.79 and 0.13). In particular, for the -60 dB mixture SNR, the predictions strongly deviate from the measured scores. Here, the simulated scores are close to zero in all conditions, including masks where the measured intelligibility is nearly at 100%.

2.5 Discussion

2.5.1 Capabilities and limitations of the intelligibility models

The proposed model accounts reasonably well for the data in both experiments. Additional simulations showed that all auditory preprocessing stages contributed to the accuracy of the model. These simulations were carried out in the exact same way as described earlier in the paper.

The gammatone filterbank with filters equally spaced on the ERB scale, each with a bandwidth of 1 ERB, was crucial in order to obtain a correct weighting of the different frequencies in the signal. If constant-bandwidth filters on a linear frequency scale were used, there would be much more emphasis of the model output at the high frequencies. This would result in a large underestimation of the bottle noise mixtures, where most of the masker energy is at high frequencies.

The nonlinear adaptation loops enhance the envelope fluctuations in the signal. In contrast, stationary portions are compressed which leads to a lower weighting in the model. The importance of these effects in the model can be seen in the bottle noise condition. Without the nonlinear adaptation loops, the large amount of high-frequency energy in the bottle noise would not be compressed and the speech fluctuations and onsets in the low frequency region would not be emphasized. This would also lead to a large underestimation of the bottle noise mixtures in the framework of the model.

In the first experiment, there was no relevant difference in the results obtained with or without the modulation low-pass filter. However, the modulation low-pass filter was very important for the predictions in the second experiment. Without this filter, the simulated scores were close to zero for all conditions of the ITFS processed mixtures at -60 dB SNR. For these stimuli, the speech information is reflected exclusively in the overall spectro-temporal structure (or envelope). The

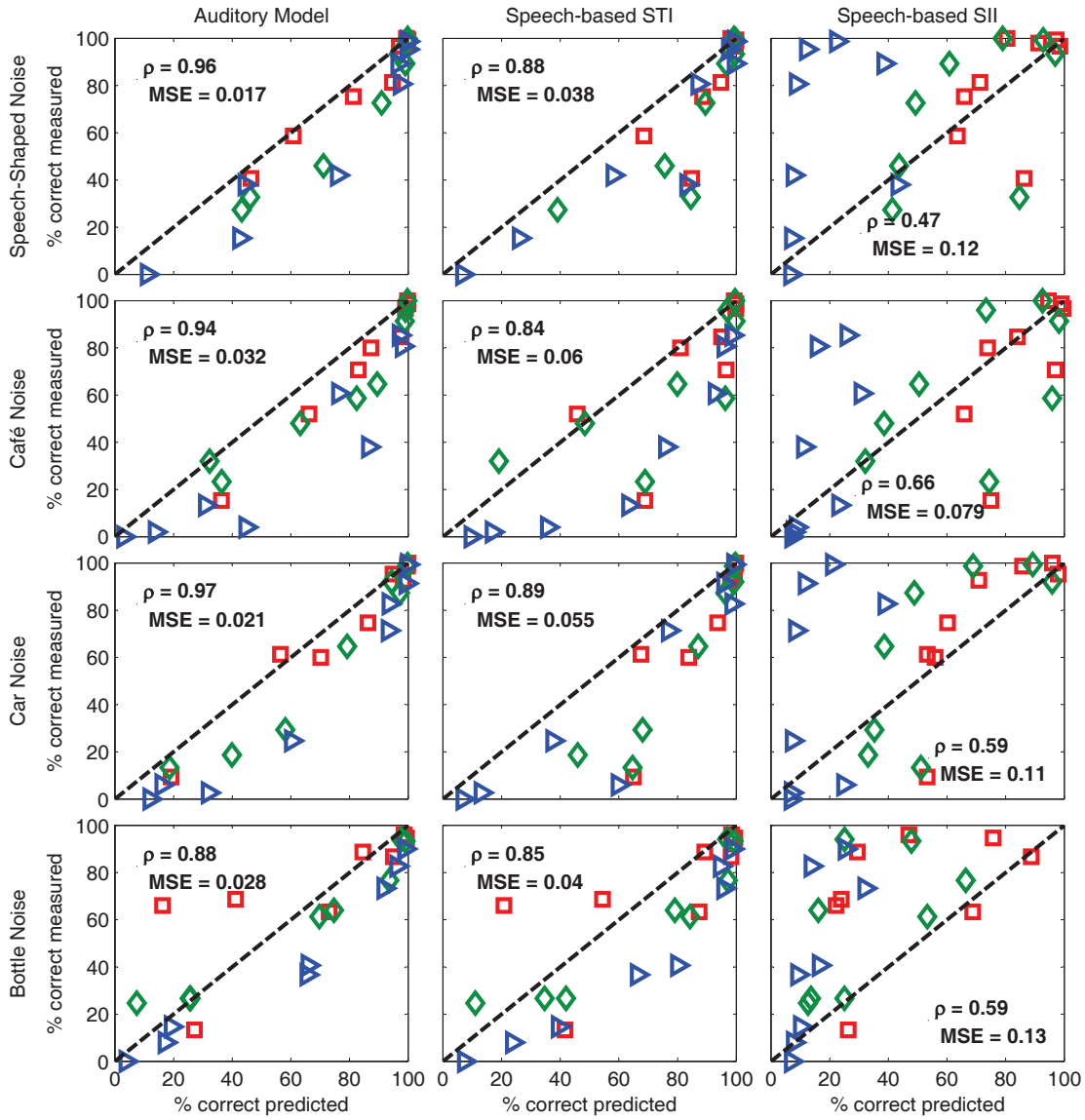


Figure 2.7: Prediction results in the second experiment using ITFS processed mixtures. The left, middle and right column contains the results of the proposed model, the speech-based STI model and the speech-based SII model. For each model the four rows from top to bottom contain the results of the SSN, cafe, car, and bottle masker. The squares represent -60 dB SNR, whereas the circles and the triangles represents the SNR values corresponding to 20% and 50% correct responses in the first experiment. The dashed line illustrates where the measured and predicted scores are the same and ideally the data points should be on this line. Points above the line represent an underestimation in the predictions, whereas points below the line represent an overestimation in the predictions. The linear cross-correlation coefficient and the mean-square-error (MSE) quantifies the strength of the linear relationship and the error between the predicted and measured scores, respectively

modulation low-pass filter essentially extracts the envelope of the signal from the previous stages in the preprocessing and thereby emphasizes the role of the envelope for speech intelligibility within this model.

Also the central processing part of the proposed model was important. The division into short time frames and the use of high-level frames only was found to be critical for the accuracy of the simulations, particularly in the second experiment. The cross-correlation coefficient alone seems to have a limited ability to model speech intelligibility in the presented framework and the central stage is important because it compensates for this limitation. The cross-correlation coefficient, used as an intermediate measure of the speech intelligibility in the model, is an energy-weighted measure of the similarity of the two internal representations. If both internal representations have a large amount of energy in a given time-frequency unit, this single unit has a large influence on the overall cross-correlation coefficient, even if the unit constitutes an extremely small fraction of the signal. This effect is very strong with the ITFS processed mixtures which often contain few regions of large energy. However, this effect is reduced by the short time frames of 20 ms used in the proposed model and this correspondingly reduced the large overestimation observed in the second experiment without the frame calculation. Frames in the order of 20 ms have also successfully been used for the calculation of the SII in fluctuating noise (Rhebergen and Versfeld, 2005), the prediction of audio quality (Huber and Kollmeier, 2006) and speech perception (McClelland and Elman, 1986).

The importance of the level classification scheme is a consequence of the fact that some speech frames only contain speech energy in a narrow frequency region. Non-speech energy in other frequency regions within those frames reduces the simulated speech intelligibility significantly, even though the speech is still detectable. For the ITFS processed mixtures, the energy in the non-speech regions of these frames is removed and the simulated speech intelligibility is therefore not reduced in contrast to the mixtures of the first experiment except for the cafeteria noise, where same effect is also observed. Thus, the simulated intelligibility of these frames is different in the two experiments even though it should be the same. Thus, using these frames either leads to an overestimation in experiment 2 or an underestimation in experiment 1. The frames described here are typically low- and mid-level frames. Thus, in order for the proposed model to successfully simulate the speech intelligibility both in speech-in-noise conditions and for ITFS processed mixtures it is crucial that only the high level frames are used. In this way the classification is also crucial.

However, the level classification of the frames also represents a limitation of the proposed model, because the exact effects of this processing, particularly in connection with the nonlinear transformation between the model output and the predicted intelligibility, are difficult to understand in all aspects. For example, if the central processing part is calibrated to a certain configuration of the auditory preprocessing, it is difficult to evaluate the effects of the changes in this preprocessing, since it requires a recalibration of the central processing part as well. Therefore, it is difficult to evaluate whether the chosen weighting of the different levels is an indication of the general importance of these segments in speech or whether it is a consequence of the specific auditory preprocessing configuration that was chosen. For example, in contrast to Kates and Arehart (2005),

where the *mid-level* frames of the speech almost exclusively determine the simulated intelligibility, the proposed model only considers the *high-level* frames.

Kates and Arehart (2005) argued that the mid-level frames contain most of the information about envelope transients and spectral transitions and that these indicate place and manner of articulation. In contrast, experiment 2 of the present study showed that the relatively sparse binary masks with a density of about 10% typically resulted in intelligibility scores of about 90% and that these almost exclusively preserved the speech information in the high-level frames. However, it is not argued here that the high-level frames are therefore principally more important than the mid-level frames. As mentioned above, it is difficult to make a generalized interpretation based on the fitted parameters of a model which is rather complex.

The speech-based STI accounts reasonably well for the data from the first experiment even though it predicts slightly lower scores in the case of the bottle masker. In the second experiment, the speech-based STI also accounts for most of the conditions, except for the sparsest binary masks where the simulated percent correct were about 40% higher than in the data. This is because the sparsest binary masks produce mixtures where only a few TF-units contain a large amount of energy. This means that the few regions preserved in the ITFS processing have a large influence on the energy-weighted cross-correlation coefficient used to calculate the intelligibility, even though these regions only constitute a very small fraction of the signal. The performance might be improved by dividing the calculation into short time frames as in the proposed model. Thus, the deviations from the data could, to some extent, reflect a limitation in the intelligibility calculation rather than major limitations in the preprocessing.

Similar to the two other models, the simulations of the speech-based SII account well for the data in the first experiment where it also predicted slightly lower scores in case of the bottle masker. In the second experiment, the speech-based SII failed to account for the data in most conditions and failed completely for the lowest SNR where the model predicted an intelligibility of almost 0% compared to 100% in the data. One main reason why the speech-based SII cannot successfully predict the results of the ITFS processed mixtures is the narrow bandwidth of the FFT bins in the spectra used in the speech-based SII. Even if the activity patterns (the envelope) of the reference spectrum and the distorted spectrum are the same, there can be large differences between the values in the individual FFT bins. Since the comparison of the two spectra is based on the values in the individual FFT bins and not the envelope of the spectra, this can lead to very low scores. Additional simulations showed that the performance of this model increases by averaging the FFT bin values into much broader frequency regions. This suggests that a more realistic type of auditory preprocessing is required to account for the data in this experiment.

It was also investigated if the speech-based SII model would perform differently if only the high-level frames were used as in the proposed model. This did not change the results in experiment 2 very much, since the speech-based SII model is not sensitive enough to envelope modulations that are critical in experiment 2. Instead, the ability to account for the training data in experiment 1 was reduced.

2.5.2 Perspectives

The present study showed that the speech-based versions of the STI and SII have limited abilities to predict the intelligibility of speech processed with ideal binary masks including noise vocoded speech, whereas the proposed model performed reasonably well. In order to further examine the generality of this model, it should be tested with other stimuli and additional conditions such as, e.g. filtered speech, peak-clipping, spectral subtraction and reverberation.

The model presented in the present study might be useful as a tool for continuously evaluating signal processing schemes for communication systems and hearing-aids during the development phase. It can also be used to continuously monitor the quality of hearing-aid and speech communication systems when these are in operation. One interesting extension of the current study would be the simulation of intelligibility for hearing-impaired listeners by integrating the nonlinear cochlear processing stage of Jepsen et al. (2008) into the proposed model. Such a model could allow explicitly investigating effects of sensorineural hearing loss on the prediction of speech intelligibility. Finally, another extension would be to include the modulation filterbank of Dau et al. (1997a) in the auditory model.

In the spectro-temporal modulation index (STMI) developed by Elhilali et al. (2003) it is suggested that a joint spectro-temporal modulation analysis is needed in order to predict the impact of phase distortion on speech intelligibility. Phase distortions do not change the envelope in each frequency channel, but change the synchronization of the envelopes across different frequency bands. The traditional STI is therefore not sensitive to these types of distortions. The three models considered in the present study all make a comparison between the reference speech signal and the distorted speech signal in a number of frequency channels. Thus, a phase distortion in some frequency channels is assumed to reduce the calculated similarity between the reference signal and the distorted signal in the models. These models should therefore also, at least to some extent, be able to account for the phase distortions considered in Elhilali et al. (2003), but this would be interesting to test explicitly.

2.6 Summary and conclusions

A speech intelligibility model based on a psychoacoustically validated auditory preprocessing model was presented. The proposed model was used to simulate speech intelligibility data from two different experiments performed by Kjems et al. (2009). The performance of the model was compared to predictions obtained with speech-based versions of the STI and SII. In experiment 1, psychometric functions were simulated for speech masked by speech-shaped noise, cafeteria noise, car noise and bottle noise, where the cafeteria noise was characterized as fluctuating. In experiment 2, speech-in-noise mixtures similar to those of experiment 1 were processed by binary masks with different mask densities and speech intelligibility was simulated as a function of mask density.

The main results of this study are as follows:

(1) The simulated psychometric functions of all three models were comparable to the data, except for the bottle noise, where all three models slightly underestimated intelligibility.

(2) The proposed model accounted well for the data in the second experiment based on ITFS processing. Additional simulations showed that the different stages of auditory preprocessing assumed in the model were important for the successful prediction of speech intelligibility in this experiment. Also the central processing in the model, based on short time frames and the use of high-level frames only, was important for the success of the model. This implies that also the level classification is an essential part of the model.

(3) The speech-based STI produced reasonably good results for the ITFS processed mixtures, except for the sparse masks where the score was dominated by short segments of high energy. Further analysis indicated that the results might be improved by calculating the STI in short time frames and disregarding the low-level segments.

(4) The speech-based SII was a poor predictor of the speech enhancement performed by the ITFS processing and failed completely when predicting the noise vocoding effect of the binary mask. The speech-based SII is limited by the narrow frequency bins of the FFT based calculation, which results in a large dependency on the temporal fine structure of the signals.

(5) In order to examine the generality of the proposed model, it should be tested in additional challenging conditions, such as, e.g. filtered speech, peak-clipping, spectral subtraction and reverberation.

Acknowledgements

We are grateful to Ulrik Kjems for providing all the speech material as well as all the measured data from the psychoacoustic measurements. We also thank two anonymous reviewers for their very helpful comments on an earlier version of this paper. This work has been partly supported by the Danish research council and partly by Oticon, Widex and GN Resound through a research consortium.

Appendix

A. Speech-based SII

The MSC is estimated using the Fast Fourier Transform (FFT). The FFT of the clean speech signal $x(n)$ and the distorted signal $y(n)$ are denoted by $X_m(k)$ and $Y_m(k)$, and calculated in short segments using a hamming window of 32 ms, where k is the FFT bin index and m denotes the segment. The

FFT is calculated with an overlap of 50%. The MSC is defined as:

$$MSC(k) = \frac{\left| \sum_{m=0}^{M-1} X_m(k) Y_m^*(k) \right|^2}{\sum_{m=0}^{M-1} |X_m(k)|^2 \sum_{m=0}^{M-1} |Y_m(k)|^2} \tag{2.4}$$

where M is the total number of segments m and the asterisk denotes the complex conjugate. In Kates and Arehart (2005), the speech part $P(k)$ of the distorted power spectrum $|Y_m(k)|^2$ is the fraction that is linearly related to the reference power spectrum $|X_m(k)|^2$. Based on the results in Carter et al. (1973), this fraction is equal to the MSC and speech power spectrum is therefore estimated by:

$$P(k) = MSC(k) \sum_{m=0}^{M-1} |Y_m(k)|^2 \tag{2.5}$$

The noise part $N(k)$ of the distorted power spectrum is defined as the fraction that is nonlinearly related to the reference signal and is therefore estimated by:

$$N(k) = (1 - MSC(k)) \sum_{m=0}^{M-1} |Y_m(k)|^2 \tag{2.6}$$

In the speech-based SII method, the SDR is calculated in a number of frequency bands by applying the frequency response of a bank of rounded-exponential (Ro-ex) filters to the estimated speech and noise power spectra. The center frequencies and bandwidths of the filters correspond to the critical bands defined in table I of the ANSI S3.5 (1997) standard. By determining the Ro-ex filters for the frequencies used in the FFT calculation, the j^{th} filter can be expressed as $W_j(k)$ and the SNR in the j^{th} band is obtained by:

$$\mathrm{SDR}(j) = \frac{\sum_{k=0}^{M-1} W_j(k) P(k)}{\sum_{k=0}^{M-1} W_j(k) N(k)} \tag{2.7}$$

When the speech and noise power spectra as well as the SDR have been estimated, the remaining part of the speech-based SII calculation follows the steps of the traditional SII.

3

Relationship between masking release in fluctuating maskers and speech reception thresholds in stationary noise †

In contrast to normal-hearing (NH) listeners, hearing-impaired (HI) listeners often show strongly reduced masking release (MR) in fluctuating interferers, which has commonly been associated with spectral and temporal processing deficits. However, it has recently been proposed that the reduced MR could result from an increased speech recognition threshold (SRT) in stationary noise [Bernstein and Grant, J. Acoust. Soc. Am. 125, 3358-3372 (2009)]. This was tested by presenting noise-band vocoded as well as low-pass and high-pass filtered stimuli to NH listeners, thereby increasing their stationary-noise SRTs to those of the HI listeners. If the primary determinant of MR is the SRT in stationary noise then the amount of the MR should be independent of the type of processing used to obtain the stationary-noise SRT. However, the relation between the amount of MR and the stationary-noise SRT depended on the type of processing. For a fluctuating interferer, none of the processing conditions reduced the MR of the NH listeners to that of the HI listeners. In contrast, for an interfering talker, the results for vocoded stimuli were similar to those of the HI listeners. Overall, these results suggest that the observed MR is only partially related to the stationary-noise SRT.

† This chapter is based on Christiansen and Dau (2012).

3.1 Introduction

The primary mode of communication between humans is via speech. Speech communication often takes place in the presence of concurrent talkers, background noise or in a reverberant environment. In such adverse listening conditions, speech intelligibility generally remains high for normal-hearing (NH) listeners, whereas hearing-impaired (HI) listeners often experience major difficulties. In conditions where the interferer fluctuates over time, compared to a situation where it is steady, NH listeners benefit from speech information in the low-intensity parts of the interferer (e.g., Baer and Moore, 1994; Füllgrabe et al., 2006; Rhebergen et al., 2006). This ability has been denoted “listening in the dips" and the improvement in speech intelligibility has been referred to as masking release (MR). In contrast, HI listeners usually show very little benefit from these dips when tested in the same conditions as the NH listeners (e.g., Festen and Plomp, 1990; Gustafsson and Arlinger, 1994; Peters et al., 1998; George et al., 2006; Lorenzi et al., 2006; Bernstein and Grant, 2009; Strelcyk and Dau, 2009).

Compensation for the reduced audibility experienced by HI listeners, e.g. through amplification in hearing aids, largely improves the ability to understand speech in quiet (e.g., Duquesnoy and Plomp, 1983). However, the benefit from amplification is often much smaller in the presence of noise or competing talkers (e.g., Duquesnoy and Plomp, 1983; Gustafsson and Arlinger, 1994; Shanks et al., 2002; Hällgren et al., 2005; Metselaar et al., 2008), although some studies have shown an improvement (e.g., Alcántara et al., 2003). A hearing loss has traditionally been characterized by an attenuation component and a distortion component (Plomp, 1978). The attenuation component is directly related to reduced sensitivity and is the main determinant for speech perception in quiet. The distortion component is often associated with supra-threshold deficits, such as reductions in temporal resolution, spectral resolution and temporal fine structure (TFS) processing and is assumed to be mainly responsible for the speech perception problems in noise (George et al., 2006; Houtgast and Festen, 2008).

It has been argued that reduced frequency selectivity, which is often present among HI listeners, limits the ability to extract speech information from spectral dips in the interferer. However, the literature is not conclusive. Studies simulating reduced frequency selectivity for NH listeners by spectral smearing (ter Keurs et al., 1993; Baer and Moore, 1993) or vocoding (Qin and Oxenham, 2003; Nelson and Jin, 2004) showed a reduced MR, while studies comparing measurements of frequency selectivity and speech recognition for HI listeners did not find any correlation (George et al., 2006; Strelcyk and Dau, 2009). However, Strelcyk and Dau (2009) measured frequency selectivity in low-frequency regions where the pure-tone sensitivity of the HI listeners was only slightly reduced and it is possible that the frequency selectivity in the high-frequency region with strongly reduced sensitivity would have shown a correlation. It seems still unclear to what extent frequency selectivity affects speech recognition in the presence of fluctuating noise or a competing talker.

Reduced MR for HI listeners has also often been associated with decreased temporal resolution in terms of an increased amount of forward masking, which has been argued to reduce the effective

duration of the dips in a fluctuating noise. This has been supported by studies where a correlation between temporal resolution and speech recognition in fluctuating noise was found (Hou and Pavlovic, 1994; Dubno et al., 2003; George et al., 2006). However, it is still not clear how much of the reduced MR for the HI listeners can be accounted for by reduced temporal resolution.

It has also been hypothesized that TFS cues are important to identify speech in the dips of a fluctuating noise or a competing talker and that TFS processing deficits observed in HI listeners might result in a reduced MR (e.g., Lorenzi et al., 2006; Hopkins et al., 2008). There seems to be evidence that TFS cues are crucial for speech recognition in the presence of a competing talker (Qin and Oxenham, 2003; Hopkins et al., 2008; Strelcyk and Dau, 2009), but it is questionable if TFS processing deficits affect speech recognition in fluctuating noise (e.g. Strelcyk and Dau, 2009). To investigate whether pitch information conveyed by the TFS of low-order resolved harmonics is important for MR, Oxenham and Simonson (2009) measured MR for NH listeners obtained with low-pass (LP) and high-pass (HP) filtered speech in the presence of modulated noise and a competing talker. They found similar results for LP and HP filtering, suggesting that low-order resolved harmonics are not more important than high-order unresolved harmonics for MR. They also found that the amount of MR decreased when the SNR in stationary noise was increased and that the MR seemed to vanish when the SNR was 0 dB or greater.

It is well known that HI listeners need a higher signal-to-noise ratio (SNR) than NH listeners in order to understand the same percentage of the speech in the presence of a stationary noise masker. Inspired by Oxenham and Simonson (2009), who showed that the MR for NH listeners is reduced when the SNR in stationary noise is increased, Bernstein and Grant (2009) suggested that the reduced MR experienced by HI listeners might be due to this higher SNR needed for the HI listeners (where the benefit from listening in the dips of the noise might be limited). They measured psychometric functions for NH and HI listeners in stationary noise, modulated noise and an interfering talker. In order to compare the results of the NH and HI listeners at the same SNR in stationary noise, they measured the MR at different points on their psychometric functions. The MR calculated at a high percentage of correct words for NH listeners was compared to the MR calculated at a low percentage of correct words for HI listeners. Bernstein and Grant (2009) found that most of the difference in the MR for speech-modulated noise between the NH and the HI listeners could be accounted for when measuring at the same SNR in the stationary-noise condition (the difference in MR measured at 50% correct was 7 dB and was reduced to only 1 dB when measured at the same SNR ($\approx$ 3 dB) in stationary noise). These results suggested that the smaller MR for the HI listeners resulted from the higher SNR in stationary noise and not necessarily from deficits in supra-threshold auditory processing specifically important for MR. However, for the interfering talker, only about 50% of the difference in the MR between the NH and HI listeners could be accounted for in this way in the same study (the difference in MR measured at 50% correct was 11 dB and was reduced to 5 dB when measured at the same SNR ($\approx$ 3 dB) in stationary noise). Thus, even after compensating for a higher SNR in stationary noise, supra-threshold processing deficits or reduced audibility might have been partly responsible for the reduced speech recognition for the HI listeners in the presence of an interfering talker.

In interpreting their results, Bernstein and Grant (2009) argued that the MR exhibited by individuals at their SRT_{50} is similar to the MR exhibited by the population tested at the same SNR. For example, the MR exhibited by the population tested at an SNR of 5 dB and greater than 50% correct is the same as the MR exhibited by an individual who has an SRT_{50} of 5 dB. This, however, might not generally be valid. Furthermore, measuring the MR at low and high percent correct points on the psychometric function might be problematic since it can be relatively flat. This means that a small deviation in the estimated psychometric function can lead to a large change in the SNR at the given percent correct, which in turn might have a large effect on the calculated MR. Bernstein and Grant (2009) also investigated whether an increase in MR with the availability of visual cues could be attributed to a change in the stationary-noise SNR. However, again MR was compared at different points on the psychometric functions.

In a later study, Bernstein and Brungart (2011) manipulated the word-set size in order to compare the MR of spectrally smeared and noise-vocoded stimuli with unprocessed stimuli at the same stationary-noise SRT. They found no difference between the processed and the unprocessed conditions suggesting that a reduced MR caused by distortions in the TFS and the spectral content of the stimuli was due to increased stationary-noise SRT.

Based on these findings, the present study investigated how increased SRTs of NH listeners obtained using low-pass (LP) filtered, high-pass (HP) filtered and noise-vocoded stimuli influence MR and to what extent the increased SRTs of NH listeners can account for the reduced MR of HI listeners tested with unprocessed stimuli. If the SRT in stationary noise mainly determines the amount of MR, the results should be essentially independent of the processing used to increase the SRTs for the NH listeners. The MR was obtained by comparing the stationary-noise results with those obtained with an 8-Hz sinusoidally modulated noise, a single-talker interferer and the international speech test signal (ISTS; Holube et al., 2010). The ISTS signal is based on speech recordings in six different languages, which were cut into short segments and recombined in a different order making the signal largely unintelligible. The ISTS was considered here in order to investigate if the MR depends on the ability to understand the interfering speech.

In Experiment 1, the MR with unprocessed stimuli was measured for HI listeners. The listeners were chosen to cover a range of different stationary-noise SRTs. In Experiment 2, the MR was measured with processed stimuli for NH listeners. The stimuli were provided with different degrees of processing in order to cover approximately the same range of stationary-noise SRTs as obtained for the HI listeners.

Table 3.1: Audiometric thresholds for the thirteen HI listeners.

				Audiometric thresholds (dB HL)											
ID	Gender	Age	Ear	125	250	500	750	1000	1500	2000	3000	4000	6000	8000	PTA (dB HL)
HI_1	M	68	L	10	15	10	15	15	15	25	60	65	65	65	31.8
			R	5	10	10	15	10	15	30	55	55	60	75	
HI_2	F	72	L	20	30	35	35	35	45	50	60	65	55	55	43.0
			R	20	35	35	35	40	45	45	50	55	50	50	
HI_3	F	66	L	15	15	15	20	20	15	20	50	55	60	70	29.3
			R	15	15	10	10	15	10	15	35	55	50	60	
HI_4	F	51	L	15	25	40	50	45	50	45	50	55	50	65	45.9
			R	20	25	40	60	60	55	60	55	50	40	55	
HI_5	M	57	L	20	25	35	30	30	30	25	30	45	60	70	37.5
			R	20	30	35	35	30	30	25	40	50	60	70	
HI_6	F	64	L	20	35	55	65	65	60	55	55	60	65	70	54.3
			R	25	45	55	65	55	55	50	55	60	55	70	
HI_7	M	63	L	15	20	35	45	50	50	55	60	60	65	75	50.2
			R	20	25	35	50	55	55	60	65	70	70	70	
HI_8	M	66	L	35	35	35	40	35	30	35	55	55	60	65	41.8
			R	25	30	35	40	35	30	30	50	50	55	60	
HI_9	F	66	L	25	25	35	45	50	55^1	65^1	80^1	105	100	100	66.6
			R	25	20	30	45	45	85^1	95^1	115^1	110	105	105	
HI_{10}	F	64	L	25	25	25	25	25	25	30	40	45	55	55	36.8
			R	25	25	30	35	30	30	35	55	55	55	60	
HI_{11}	F	66	L	20	25	30	40	40	55	50	50	60	85	90	55.5
			R	35	40	40	50	50	65	65	60	75	95	100	
HI_{12}	M	61	L	25	40^1	55	50	55	55	55	80	80	70	75	59.5
			R	35	60^1	60	60	55	50	65	70	70	70	75	
HI_{13}	F	64	L	25	35	35	45	45	45	50	70	80	80	80	52.3
			R	25	30	35	35	35	40	55	80	75	75	75	

3.2 Masking release for hearing-impaired listeners

3.2.1 Methods

Listeners

Thirteen HI listeners (five male and eight female) between 55 and 70 years of age (mean age of 64) participated in the experiment. The individual audiometric thresholds are listed in Table 3.1. A difference of less than 15 dB between the pure-tone air and bone-conduction thresholds insured that the hearing losses were of sensorineural origin. Listeners were selected to have different degrees of hearing loss (mild, moderate, moderate-severe and severe), based on their audiograms, assuming that this would be associated with a range of SRTs among the listeners, even though the sensitivity loss in terms of the audiogram is not always closely related to the SRT (Bacon et al., 1998; Houtgast and Festen, 2008; Strelcyk and Dau, 2009). All listeners had gradually sloping and symmetric hearing losses, with differences in audiometric thresholds between left and right ear of typically less than 15 dB at all frequencies. Exceptions are indicated in Table 3.1.

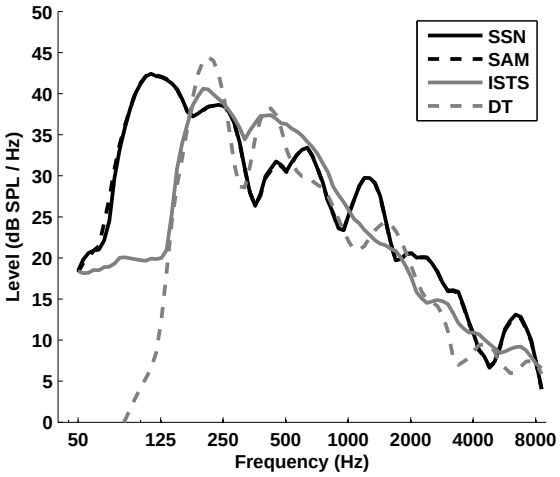


Figure 3.1: Long-term spectra of the four maskers used in the listening experiments, speech-shaped noise (SSN), sinusoidally amplitude modulated noise (SAM), International Speech Test Signal (ISTS) and danish talker (DT).

Stimuli

The SRTs were measured using the Danish speech intelligibility test called conversational language understanding evaluation (CLUE, Nielsen and Dau, 2009), which is very similar to the hearing-in-noise test (HINT) originally developed for English (Nilsson et al., 1994). The CLUE material consists of natural and meaningful sentences representing conversational speech and has a fixed structure consisting of five words per sentence. The sentences were spoken by a male talker with an average fundamental frequency (F0) of 119 Hz. The sentences were presented in four different interferers: (1) A stationary noise with the same long-term spectrum as the sentence material, i.e. a speech-shaped noise (SSN), (2) an 8-Hz sinusoidally amplitude-modulated (SAM) speech-shaped noise, (3) continuous speech produced by a Danish female talker (DT) with an average F0 of 214 Hz and (4) the international speech test signal (ISTS; Holube et al., 2010), which consists of natural speech from six female talkers speaking different languages, whereby the speech material was segmented and remixed using a randomization procedure in order to make it largely unintelligible. The average F0 of the ISTS is 207 Hz. The long-term average magnitude spectra of the four maskers are shown in Fig. 3.1.

Procedure

The experiment was conducted in a double-walled sound insulated booth, where the experimenter controlled the procedure by means of a Matlab application developed specifically for the CLUE test. The digital signals were sampled at 22050 Hz and converted to analog signals by a high-end 24 bit soundcard (RME DIGI96/8). The stimuli were presented diotically over Sennheiser HD580 headphones. The target sentences were presented at a fixed sound pressure level (SPL) of 80 dB, whereas the level of the interferer was determined via an adaptive procedure used to measure the SRTs. The onset and offset of the interferer were 1 s before and 600 ms after the sentence,

respectively, where a ramped squared-cosine function with a duration of 400 ms was applied to the onset and the offset. For each presentation, the interferer was randomly selected from a long sample (SSN: 22 seconds, ISTS: 52 seconds).

The listeners received approximately 30 minutes of training before the SRTs were measured. In the training session, the first sentence was presented at a very low SNR. The SNR was increased in steps of 2 dB until all five words were repeated correctly. The listeners were allowed to guess and the recognized words were repeated verbally to the experimenter and registered without feedback. For the following sentence, the SNR was decreased by 6 dB and again increased in 2 dB steps until all the words were repeated correctly. This was done until 20 sentences were presented, for each of the four interferers.

In the test session, a list of 10 sentences was used to measure the SRT for a given run. The procedure for the presentation of the first sentence was the same as in the training session. However, for the presentation of the remaining nine sentences, the SNR followed a simple adaptive procedure: if all words were repeated correctly, the SNR was decreased by 2 dB, otherwise the SNR was increased by 2 dB. The measured SRT was the average of the last eight SNRs from presentation number 4 to 11 (presentation number 11 results from the response to sentence 10, although the eleventh sentence does not exist). Five runs were conducted for each condition and the average of these SRTs produced the final SRT.

3.2.2 Results

Figure 3.2 shows the MR for the HI listeners as a function of the SRT obtained in stationary noise. Results are shown for the three different interferers: SAM noise (dark-gray circles), ISTS (medium-gray squares) and DT (light-gray diamonds). The MR for each interferer is shown at the SRT in stationary noise obtained for the individual listeners. Thus, all symbols that indicate the same SRT in stationary noise represent the data from one listener. Linear regression lines were fitted to the measured data obtained with the different interferers (SAM: dashed dark-gray, ISTS: dotted medium-gray, DT: dashed-dotted light-gray). For all three interferers, it can be seen that, on average, listeners with small SRTs show a large amount of MR and listeners with larger SRTs show a smaller or no MR. From the lowest to the highest SRT value across the group of HI listeners, the average MR decreased from about 3 dB to -2 dB in the case of the modulated noise (SAM) and from about 9 dB to -1 dB in the case of the two interfering talkers (ISTS and DT). Some listeners showed negative MR values, which indicate that the modulated noise or interfering talker actually masked the speech more effectively than the stationary noise. The results are consistent with the results of Bernstein and Grant (2009) and also agree with the results of George et al. (2006) and Desloge et al. (2010) where the MR for HI listeners decreased with increasing SRT in stationary noise. However, it can also be seen that there are large differences in MR for some of the listeners with similar SRTs in stationary noise. Correlations were calculated in order to quantify the relation between the SRT in stationary noise and the MR. Since listener HI9 showed a much higher SRT in stationary noise than all other HI listeners, the correlations were calculated both with and without listener

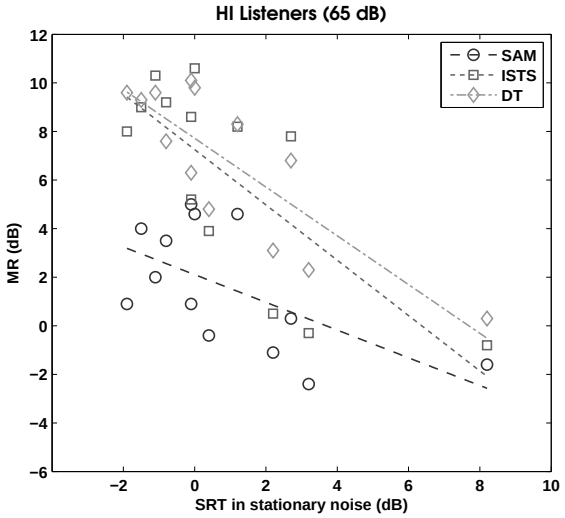


Figure 3.2: MR obtained for the HI listeners measured in the presence of the SAM (dark-gray circles), the ISTS (medium-gray squares) and the DT (light-gray diamonds) interferer, respectively. For each individual listener, the MR is shown at the corresponding SRT in stationary noise. The trends in the data are indicated by linear regression lines for the three interferers.

HI9. When listener HI9 was included, the SRT in stationary noise was significantly correlated with the MR for all three interferes (SAM: $[r = 0.60, p = 0.03]$, ISTS: $[r = 0.75, p = 0.003]$, DT: $[r = 0.84, p = 0.0004]$), even though the correlation for the SAM interferer was relatively low. When the listener HI9 was excluded from the analysis, the SRT in stationary noise was only significantly correlated with the MR obtained with the ISTS $[r = 0.67, p = 0.02]$ and the DT $[r = 0.74, p = 0.005]$ interferers, but not with the MR obtained with the SAM $[r = 0.55, p = 0.06]$ interferer. Based on the relatively low correlation between stationary-noise SRT and MR obtained with the SAM interferer compared to MR obtained with the interfering talkers, it seems that the SRT in stationary noise is only somewhat indicative of the amount of MR.

3.3 Masking release for NH listeners obtained with processed stimuli

3.3.1 Methods

Listeners

Nine NH listeners (two male and seven female) between 40 and 65 years of age (mean age of 50) participated in the experiment. All had hearing thresholds below 20 dB hearing level (HL) at audiometric frequencies between 125 and 4000 Hz. Above 4 kHz, two listeners had thresholds at 25 and 30 dB HL, respectively.

Stimuli

The speech material and the interferers were the same as used in Experiment 1 (Sect. 3.2). However, in this experiment, the sentences were either LP-filtered, HP-filtered or vocoded in order to increase the SRT for the NH listeners in stationary noise. The SRT was gradually modified by changing the cut-off frequency of the LP and HP filtering, and by vocoding the channels below a certain cut-off frequency, which also was modified.

The LP and HP filtering followed closely the procedure used in Oxenham and Simonson (2009). The speech was always presented at a fixed level of 65 dB SPL, whereas the level of the interferer was determined by the adaptive procedure. After setting the levels, the speech was combined with the interferer and the mixture was filtered with an eighth-order Butterworth filter (slope=48 dB/octave) at a given cut-off frequency. An off-frequency noise (SSN) which was filtered at the same cut-off frequency but covering the opposite frequency range was then added to the mixture. Before filtering, the level of the off-frequency noise was 12 dB below the level of the unfiltered speech signal. More details on the filtering procedure can be found in Oxenham and Simonson (2009). A pilot experiment with three young NH listeners showed that LP filtering at the cut-off frequencies 750, 850, 1000, 1750 and 3000 Hz and HP filtering at the cut-off frequencies 250, 1000, 1250, 1500 and 1750 Hz produced SRTs that were roughly evenly distributed between -2 and 8 dB.

Similar to the filtering, speech used in the vocoder was also presented at a fixed level of 65 dB SPL and the level of the interferer was also determined by the adaptive procedure. The vocoded signals were scaled to have the same overall level as the input to the vocoder. The combined speech and interferer signal was decomposed into a number of frequency channels via processing through a gammatone filterbank (Patterson et al., 1987). The filterbank consisted of 32 fourth-order gammatone band-pass filters with center frequencies ranging from 100 to 8000 Hz, equally spaced on an equivalent-rectangular-bandwidth number scale (ERB_N; Glasberg and Moore, 1990), each with a bandwidth of 1 ERB_N. The envelope in each channel was extracted by half-wave rectification and LP filtering with a cut-off frequency of 50 Hz. The LP filter was a sixth-order butterworth filter with a slope of 34 dB/octave. After filtering, the envelope was imposed on a white noise carrier and filtered with the same band-pass filter before all the channels were time aligned and recombined. In many studies, a vocoder has been used to remove TFS information in the speech signal; however, a vocoder also introduces distortions in the spectral domain and in the temporal envelope of the signal. A range of different SRTs for the NH listeners was achieved by only vocoding some of the 32 frequency channels and leaving the remaining channels unprocessed (the number of channels was thus the same in all conditions). The SRT was increased by gradually vocoding more channels, starting with channels at low center frequencies. A pilot study with three young NH listeners indicated that vocoding 4, 14, 22, 28 and 32 channels produced SRTs that were evenly distributed between -3 and 4 dB.

Procedure

SRTs were measured using the same adaptive procedure and experimental setup as in Experiment 1. However, since there were 60 conditions and only 20 sentence lists, only one list was used to measure the SRT in each condition. The listeners were tested with the three types of processing using the same 20 sentence lists, but this was done with 2-3 week intervals between the tests. In order to avoid differences due to training and learning effects, the experiments with the three types of processing were performed in a random order for each listener. The final SRTs were averaged across all NH listeners.

3.3.2 Results

Low-pass filtered stimuli

The left panel of Fig. 3.3 shows the average SRTs for the NH listeners as a function of the cut-off frequency of the LP filter. The four curves represent the results obtained with the SSN (crosses and solid curve), the SAM (circles and dashed curve), the ISTS (squares and dotted curve) and the DT (diamonds and dashed-dotted curve) interferers, respectively. As expected, the highest cut-off frequency led to the lowest SRTs, i.e. the best speech intelligibility. When the cut-off frequency was reduced, a large increase in SRT was observed for all four interferers.

For each cut-off frequency, the MR was obtained by calculating the difference in SRT between the stationary noise and each of the three fluctuating interferers. The obtained MRs are indicated by the open symbols in the right panel of Fig. 3.3 as a function of the SRT obtained in stationary noise. The individual cut-off frequencies corresponding to the respective SRTs are indicated in the figure. Linear regression lines were fitted to the data for each of the interferers to illustrate the overall trend in the data, but were not statistically analyzed. The interferers are indicated by the same symbols and line styles as used in the left panel.

Overall, there is a reduction of the MR with increasing stationary-noise SRT. For the SAM interferer (circles), the MR tends to approach zero at large stationary-noise SRTs. The reduction in the MR obtained with the two interfering talkers (ISTS and DT) seems to be slightly larger than for the SAM interferer, but there is still a considerable amount of MR ($\approx$ 6 dB) at the highest SRT in stationary noise. Most of the MR reduction occurs between the 3000-Hz and the 1000-Hz condition for the SAM interferer and between the 1750-Hz and 1000-Hz condition for the ISTS and DT interferers. When the cut-off frequencies are decreased further (higher stationary-noise SRTs) the MR tends to be more stable. The results are compared to the results of the HI listeners in Section 3.4.

These observations were supported by an ANOVA showing that the MR differed significantly across SRT [$F(4,28) = 12.4, p < 0.0001$], interferer [$F(2,14) = 154.7, p < 0.0001$] and that there was a significant interaction between SRT and interferer [$F(8,56) = 2.9, p < 0.01$]. The trends for each interferer were investigated by performing multiple pairwise-comparisons on the difference

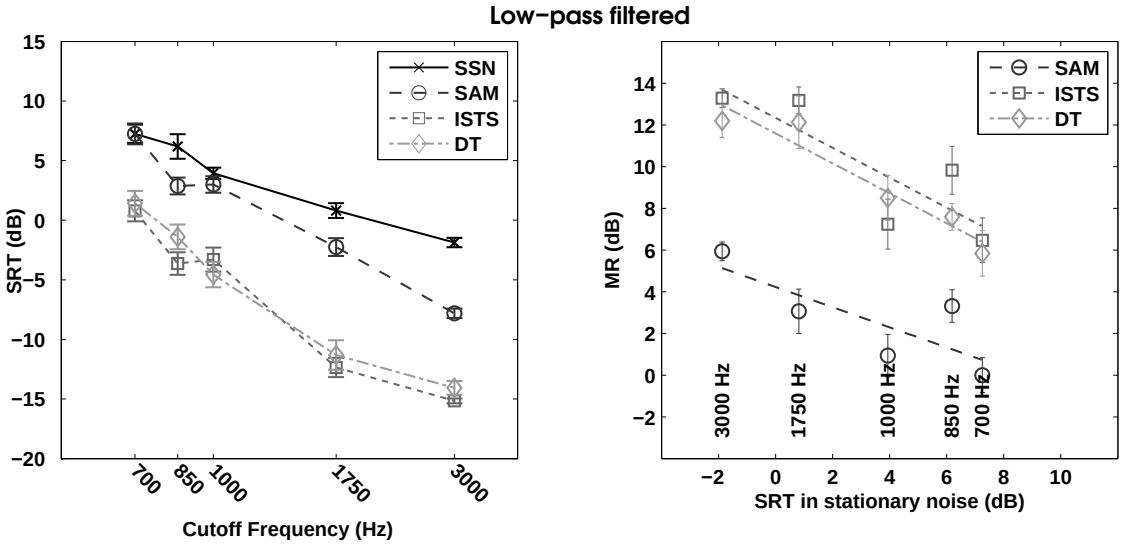


Figure 3.3: The left panel shows the mean SRTs for the NH listeners in the presence of the SSN (crosses and solid black curve), SAM (circles and dashed dark-gray curve), ISTS (squares and dotted medium-gray curve) and DT (diamonds and dashed-dotted light-gray curve) interferer, respectively, for five different cut-off frequencies of the LP filter. Error bars represent $\pm$ 1 standard deviation. The right panel shows the MR for the SAM, ISTS and DT interferers, reflected as the corresponding differences from the SRT obtained with the SSN interferer for each cut-off frequency.

between all the conditions (Holm-Sidak correction was applied). For the SAM and ISTS interferers, this analysis revealed that the MR in the 3000-Hz condition was higher than in the 1000-Hz and 700-Hz conditions [p<0.005]. For the DT interferer, the MR was higher in the 3000-Hz and 1750-Hz conditions than in all the remaining conditions [p<0.008].

High-pass filtered stimuli

The left panel of Fig. 3.4 shows the average SRTs for the NH listeners as a function of the cut-off frequency of the HP filter. For the HP-filtered stimuli, the lowest cut-off frequency led to the lowest SRTs and increasing the cut-off frequency led to a considerable increase in the SRTs.

The right panel of Fig. 3.4 shows the MR as a function of the stationary-noise SRT. Linear regression lines were fitted to the data to illustrate the overall trend in the data, while details in the results are not captured. The cut-off frequencies corresponding to the SRTs are indicated in the figure. Overall, the effect of the HP filtering on the MR is similar to the effect of the LP filtering, showing an overall reduction of the MR with increasing stationary-noise SRT. However, compared to the results from the LP filtering, the reduction in the MR obtained with the SAM interferer seems slightly smaller, whereas the reduction in the MR obtained with the ISTS and DT interferers appears larger. Thus, there appears to be a larger interaction between SRT and MR for the HP filtering. For all three interferers, most of the MR reduction occurs between the 250-Hz and 1000-Hz condition, while a more modest MR reduction is observed when the cut-off frequency is increased above 1000 Hz.

An ANOVA confirmed that the MR differed significantly across SRT $[F(4,28) = 19.5, p <$

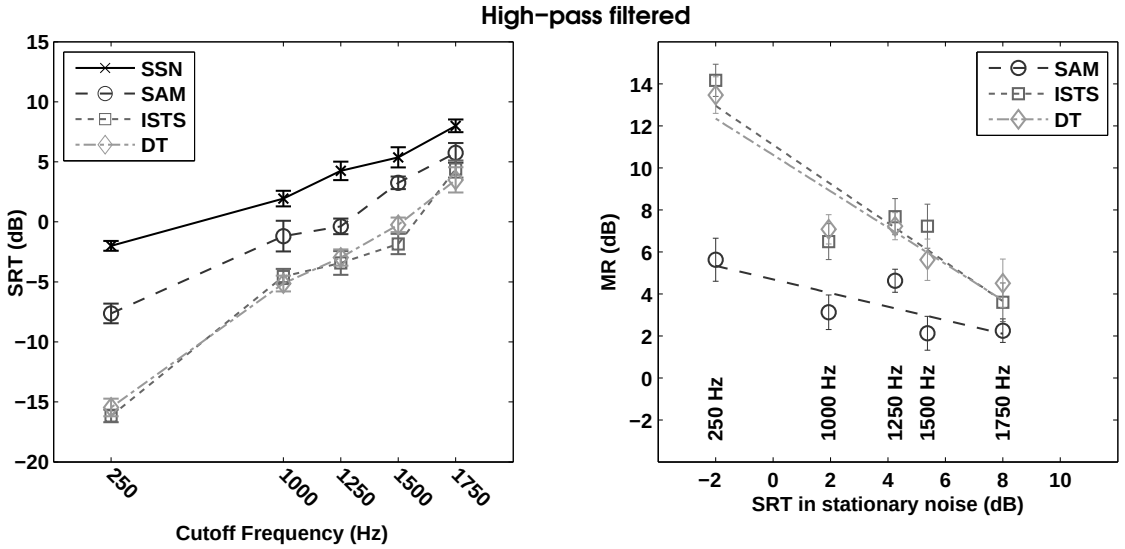


Figure 3.4: The left panel shows the mean SRTs for the NH listeners in the presence of the SSN (crosses and solid black curve), SAM (circles and dashed dark-gray curve), ISTS (squares and dotted medium-gray curve) and DT (diamonds and dashed-dotted light-gray curve) interferer, respectively, as a function of the cut-off frequency of the HP filter. Error bars represent $\pm$ 1 standard deviation. The right panel shows the MR for the SAM, ISTS and DT interferers, reflected as the corresponding differences from the SRT obtained with the SSN interferer for each cut-off frequency.

0.0001], interferer $[F(2,14) = 109.4, p < 0.0001]$ and showed that there was a larger interaction between interferer and SRT $[F(8,56) = 4.2, p < 0.001]$ than for the results from the LP filtering. As for the LP-filtered stimuli, the trends in the data were investigated, for each interferer, by performing multiple pairwise-comparisons on the difference between all the conditions. For the SAM interferer, the MR in the 250-Hz condition was higher than in the 1500-Hz and 1750-Hz [p<0.006], whereas the MRs in the remaining conditions were statistically the same. For the ISTS and the DT interferers, the MR in the 250-Hz condition was higher than in all the other conditions [p<0.0001]. Apart from a significantly higher MR in the 1250-Hz condition compared to the 1750-Hz condition for the ISTS interferer [p<0.005], the MRs in the remaining conditions were statistically the same.

Vocoded stimuli

Figure 3.5 (left panel) shows the average SRTs for the NH listeners as a function of the number of vocoded channels. In general, the SRTs were lowest for the condition with the least number of vocoded channels (4 out of 32) and increased as more channels were included. The effect was much stronger for the two single-talker interferers (ISTS and DT) than for the noise interferers (SSN and SAM).

The right panel of Fig. 3.5 shows MR as a function of the stationary-noise SRT. Linear regression lines were fitted to the data. The number of vocoded channels corresponding to the different SRTs are also indicated in the figure. In the vocoded condition, the change in SRTs for the SSN ($\approx$ 4 dB) and the SAM ($\approx$ 7 dB) interferers were smaller than in the LP and HP-filtered conditions.

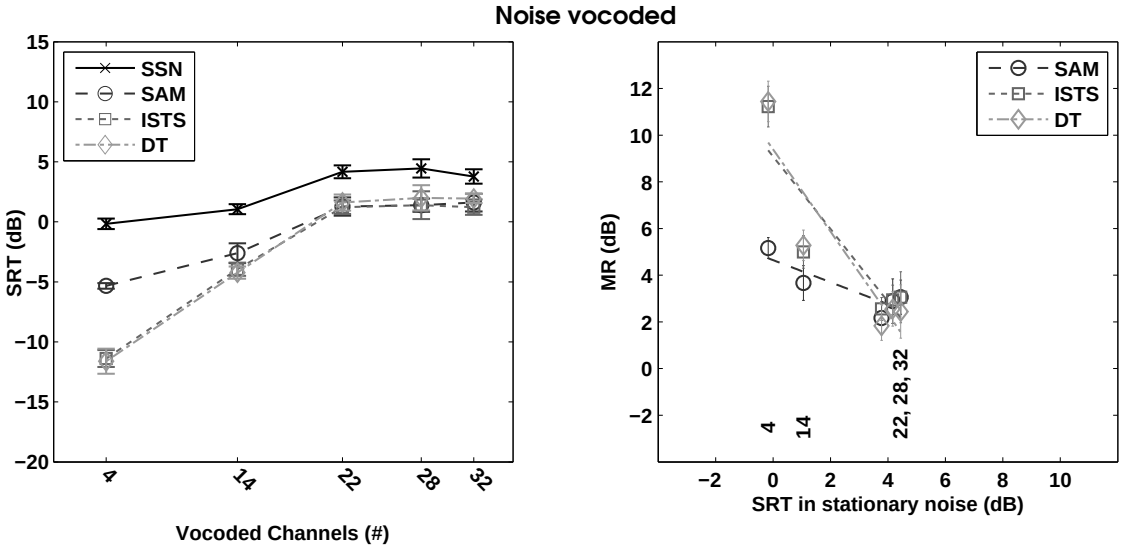


Figure 3.5: The left panel shows the mean SRTs for the NH listeners in the presence of the SSN (crosses and solid black curve), SAM (circles and dashed dark-gray curve), ISTS (squares and dotted medium-gray curve) and DT (diamonds and dashed-dotted light-gray curve) interferer, respectively, as a function of the number of vocoded channels. Error bars represent $\pm$ 1 standard deviation. The right panel shows the MR for the SAM, ISTS and DT interferers, reflecting the corresponding differences from the SRT obtained with the SSN interferer. The number of vocoded channels from the left panel is also indicated at the corresponding stationary-noise SRT values.

However, for the ISTS and DT interferers, the change in SRT was much larger ($\approx$ 13.5 and 12.5 dB, respectively) and close to the change observed in the LP and HP-filtered conditions. Consequently, as can be seen in the right panel of Fig. 3.5, the MRs for the two single-talker interferers were strongly reduced for large stationary-noise SRTs. Specifically, the MR was found to be reduced from about 10 to 2 dB for a small increase of the stationary-noise SRT. In contrast, the MR obtained with the SAM interferer was only reduced from 4.5 dB to 2.5 dB. Interestingly, the figure clearly shows that SRT and MR are only affected when increasing the number of vocoded channels from 4 to 22 and increasing the number of vocoded channels further has no effect at all.

An ANOVA confirmed that the MR differed significantly across SRT $[F(4,28) = 23.9, p < 0.0001]$, interferer $[F(2,14) = 13.4, p < 0.001]$ and showed that the interaction between SRT and MR $[F(8,56) = 5.1, p < 0.0001]$ was larger than the interaction found in the LP and HP filtering conditions. Furthermore, multiple pairwise-comparisons were performed on the difference between all the conditions for each interferer showed that, for the ISTS and DT interferers, the 4-channel condition was significantly different from all other conditions [p<0.0001]. Furthermore, for the DT interferer, the 14-channel condition was also different from the 30-channel and 32-channel conditions [p<0.02]. Thus, the statistical analysis supports the observation that increasing the number of vocoded channels above 22 does not have an effect. The reduction of the MR obtained with the SAM interferer was not significant $[F(4,28) = 2.3, p > 0.08]$.

3.4 Comparison of results for normal and impaired hearing

Figure 3.6 compares the MR results obtained for the NH and HI listeners. The left, middle and right columns present the results for the SAM, ISTS and DT interferers, respectively. The top, middle and bottom rows show the results for the NH listeners obtained with LP filtering, HP filtering and noise vocoding, respectively. It is important to note that the HI listeners were only tested with unprocessed speech, but for comparison with the results obtained for the NH listeners, the HI data for each interferer were replotted for all three types of processing. In each panel, the measured MRs for the NH listeners (open circles) and the HI listeners (crosses) are shown together with the corresponding fitted linear regression lines indicated by the dashed and solid lines. For the NH listeners, each symbol represents the mean across all listeners for a specific condition whereas, for the HI listeners, each symbol shows the mean of five repetitions for a specific listener. The thin vertical and horizontal lines in the figure indicate an SRT and a MR of zero dB, respectively.

If the SRT in stationary noise would determine the amount of MR for a given masker, the MR for the NH listeners should be reduced by the same amount as the HI listeners for a correspondingly increased SRT in stationary noise. In particular, the MR obtained for the NH listeners as a function of SRT in stationary noise should be the same for the three types of processing.

Overall, there is a general reduction in the MR with increasing stationary-noise SRT. However, for the ISTS and DT interferers, the results show that for NH listeners the MR depends both on the SRT in stationary noise and on the type of processing. The results with the LP-filtered stimuli show that NH listeners can achieve an MR of about 4-8 dB higher than HI listeners, with reference to the same stationary-noise SRT. In contrast, for NH listeners presented with vocoded stimuli, the MRs were very similar to the results obtained with the HI listeners, suggesting that the noise-vocoder removes the cues used by the NH listeners to obtain a release from masking. For the SAM interferer, the MR of NH listeners is similar across the three types of processing; however, a small difference in the MR between NH and HI listeners is still observed.

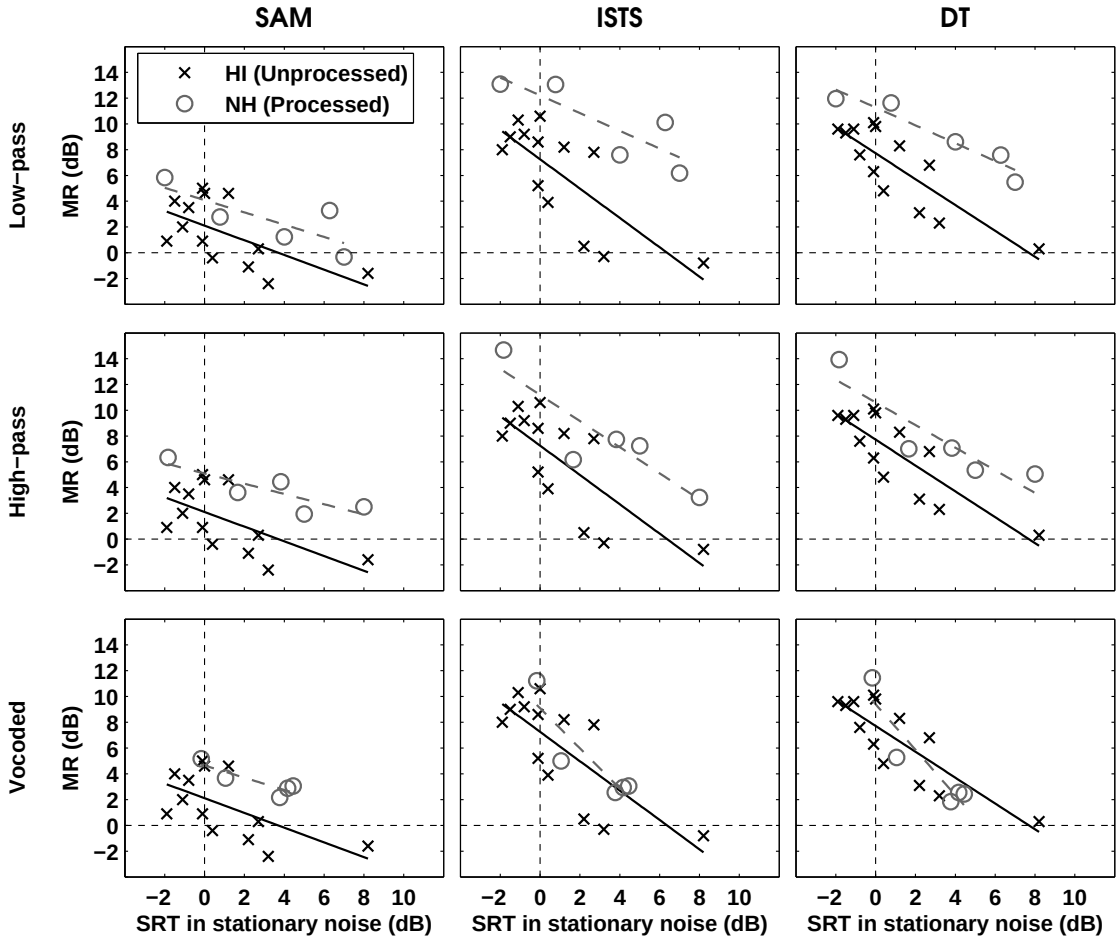


Figure 3.6: Comparison of the MR obtained with the HI listeners using unprocessed stimuli and the NH listeners using processed stimuli. Each panel shows the MR for the HI (crosses and solid line) and NH (circles and dashed line) listeners as a function of the SRT in stationary noise. The columns show the performance in the presence of the SAM, ISTS and DT interferers where each row represents the three types of processing used in the measurements with the NH listeners. The thin dashed vertical and horizontal lines indicate zero SRT and MR, respectively.

3.5 Discussion

3.5.1 Relation between the MR and the SNR in stationary noise

The results from the present study show that the MR obtained by NH listeners is reduced when measured at increased SRTs in stationary noise. The results for the NH listeners demonstrate that the reduction of the MR obtained with processed noisy speech stimuli depends considerably on the type of processing and not only on the SRT in stationary noise. Specifically, the noise-vocoder seem to remove important cues for MR when the masker is an interfering talker whereas other factors play a role for MR in modulated noise. The finding that MR depends on the processing is in contrast Bernstein and Brungart (2011) who found no difference between processed and unprocessed conditions at a given SRT suggesting that reduced MR caused by noise-vocoding and spectral smearing of the stimuli is entirely due to increased stationary-noise SRT.

The results from the present study also show that the NH listeners still perform considerably better than the HI listeners when measured at the same SRT in stationary noise. For the interfering talkers, these results are similar to the findings of Bernstein and Grant (2009). However, in the case of modulated noise they found the MR to be almost entirely predicted by the stationary-noise SRT, which was not found in the present study.

One of the differences between the study of Bernstein and Grant (2009) and the present study is that they reduced the influence of audibility by presenting flat-spectrum speech at 87 dB SPL to the HI listeners, whereas normal-spectrum speech was presented at 80 dB SPL in the present study. In quiet and stationary noise, several studies have shown that most listeners with thresholds greater than 55-60 dB show very little or even a detrimental effect of high-frequency amplification (e.g. Rankovic, 1991; Ching et al., 1998; Hogan and Turner, 1998; Amos and Humes, 2007). For fluctuating noise, Moore et al. (1999) found a small decrease in the SRT of about 2 dB based on the "Cambridge" formula, but they did not measure if the benefit was the same in stationary noise. Peters et al. (1998) found an increase in MR of about 1.5 dB with NAL amplification. Still, the difference in presentation level and spectral shaping between the present study and the study of Bernstein and Grant (2009) probably only has a minor effect on the results. In the present study, a frequency independent gain was chosen in order to deliver an undistorted broadband signal where the inherent component relations were preserved (Halpin and Rauch, 2009b).

Another difference that might have influenced the results is that the present study used sentence scoring whereas the study by Bernstein and Grant (2009) used word scoring. With word scoring, where the listeners need to understand 50% or more of the words in every second sentence, they can probably rely on mostly high-energy information in the speech. With sentence scoring, where the listeners have to understand all the words in every second sentence, they probably also need low-energy information in the speech. At high SNRs, dips in the interferer primarily release low-energy information, which is probably only beneficial when using sentence scoring. This might explain why the NH listeners still showed a large MR at high stationary-noise SRTs in LP and HP filter conditions. However, if the scoring method would have been responsible for the large amounts of MR at high stationary-noise SRTs, it should have affected the results obtained with the vocoded stimuli in the same way, which is clearly not the case.

Finally, the study of Bernstein and Grant (2009) argued that the MR measured at different points on the average psychometric function (i.e., different values of stationary-noise SNR) is the same as the MR exhibited by the individual listeners whose SRT_{50} correspond to the SNRs of the different points, regardless of hearing loss. However, compared to the MRs obtained from the average psychometric functions, the MRs exhibited by the individual listeners at their SRT_{50} showed larger differences between NH and HI listeners. Thus, while their conclusions based on the average psychometric functions differ, their results from the individual listeners are consistent with those of the present study.

3.5.2 Importance of low- and high-frequency information

The coding of pitch and fundamental frequency ($F0$) information by the TFS of low-order resolved harmonics has been suggested to be important for the segregation of the target speech from an interfering noise or talker (e.g., Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994; Micheyl and Oxenham, 2007). In contrast, the results of the present study showed approximately the same reduction in the MR for the LP- and HP-filtered stimuli, except for the SAM interferer where the reduction in the MR was actually larger for the LP-filtered stimuli. Thus, the results of the present study indicate that low-order resolved harmonics are not more important than high-order unresolved harmonics. This is consistent with the results of Oxenham and Simonson (2009) where LP and HP filtering were also used to investigate the effects of low-order resolved harmonics versus high-order unresolved harmonics on MR. Oxenham and Simonson (2009) suggested that a decrease in bandwidth could have negative effect on the MR and a smaller bandwidth in their LP-filtered stimuli could explain that the MR for LP-filtered stimuli was not larger than for HP-filtered. They suggested that it is possible that the MR stems from the inherent redundancy of the a broadband speech signal. In contrast to the results from the present study, Oxenham and Simonson (2009) found that generally the MR approached 0 dB as the SNR approached 0 dB, which was not the case in the present study. A reason for this could be the use of sentence scoring in the present study compared to the use of word scoring in Oxenham and Simonson (2009). SRTs measured with sentence scoring are generally larger than SRTs measured with word scoring. Furthermore, dips in the interferer might provide low-energy speech information which is more crucial when using sentence scoring than when using word scoring.

3.5.3 Distortion of carrier information

For the vocoded stimuli, the SRTs produced by the SSN interferer only increased slightly when the number of the vocoded channels was increased. In contrast, the SRTs for the two single-talker interferers (ISTS and DT) increased strongly with increasing number of vocoded channels, resulting in a large reduction of the MR. Similar to the results from Hopkins et al. (2008), the SRTs primarily changed when channels in the low-frequency range ($<$ 3000 Hz) were vocoded. Hopkins et al. (2008) argued that this strong change is due to the distortion of TFS cues that are assumed to be important for target and masker segregation and which are primarily conveyed in the low-frequency region where phase-locking is most distinct (Palmer and Russell, 1986; Santurette and Dau, 2011). However, it is still not clear to what extent the reduced speech intelligibility of vocoded stimuli actually results from degraded TFS cues, since the vocoder also introduces distortions in the spectral and temporal envelope of the stimuli. Assuming that degraded carrier information is the primary consequence of vocoded stimuli, the good correspondence between the data from the HI listeners (with unprocessed stimuli) and the data from the NH listeners using single-talker interferers suggests that a degraded carrier representation in the HI listeners limits their ability to segregate the target and the interferer. In contrast, the differences between the data from the NH and the HI listeners in the case of the SAM interferer suggest that factors other than TFS coding

may limit the HI listener's ability to utilize the dips in a modulated noise, consistent with Qin and Oxenham (2003) and Strelcyk and Dau (2009).

3.5.4 Effects of filtering on SRT and MR

The plateau in the MR function at higher stationary-noise SRTs observed for NH listeners with the LP filtered and, to some extent, HP filtered stimuli could indicate that the MR reduction is counteracted by an increase in MR due to the filtering process itself. If this is the case it could restrict the comparison between NH and HI listeners. Since higher stationary-noise SRT are related to decreased bandwidth, the plateau in the MR could indicate that a decreasing bandwidth has a positive effect on the MR. However, this is in contrast to Oxenham and Simonson (2009) where a decreasing bandwidth was suggested to have a negative effect on the MR. In the present study, most of the MR reduction occurred in specific frequency regions whereas other regions only caused a small change in the MR. Thus, it is possible that certain frequency regions are more important than others for MR. For the LP filtering, most of the decrease occurred from 3000-Hz to 1000-Hz. An explanation for this could be that the higher frequencies contain more of the low-intensity speech information and that dips in the masker are helpful for low-intensity information. For the HP filtering, most of the decrease occurred from 250-Hz to 1000-Hz. An explanation for this could be that pitch or F0 cues are conveyed primarily by the low-order resolved harmonics and that this information is important for MR.

3.5.5 Relation between audiometric thresholds and speech perception

The results for the HI listeners showed a general reduction in the MR with increasing SRT in stationary noise. However, the results also showed a large variation among the listeners, resulting in a weak correlation between MR in modulated noise and the stationary-noise SRT that was only just significant when listener HI9 was included. For the two interfering talkers (ISTS and DT), the correlation between the MR and the stationary-noise SRT was stronger. In order to investigate if audibility was a better indicator of the MR, the correlation between the MR and the pure-tone average (PTA) for the HI listeners was calculated. The MR obtained with the SAM interferer was more strongly correlated [$r = 0.91, p = 0.00005$] with the PTA than the stationary-noise SRT [$r = 0.55, p = 0.06$]. The reduced sensitivity of the HI listeners limits the amount of audible speech information in the valleys of the modulated noise. The correlations between the PTA and the MR obtained with the ISTS [$r = 0.71, p = 0.009$] and the DT [$r = 0.68, p = 0.01$] interferers were similar to the correlation between the stationary-noise SRT and the MR (ISTS: [$r = 0.67, p = 0.02$], DT: [$r = 0.74, p = 0.005$]). Thus, in the case of the two interfering talkers, the correlation between the MR and the PTA is smaller than for the modulated noise, while the correlation between the MR and the stationary-noise SRT is larger than for the modulated noise. Hence, other factors seem to play a larger role in the case of an interfering talker compared to a modulated noise, which is consistent with the studies of Bernstein and Grant (2009) and Strelcyk and Dau (2009). These

factors could be frequency resolution and TFS processing which might be important for segregating the target from the interferer but less important for the processing of speech in modulated noise.

3.5.6 Effects of the linguistic content

Surprisingly, the DT interferer which was expected to distract the attention of the listeners did not result in higher SRTs than the nonsense ISTS interferer. It is possible that the effects of the linguistic content in the interferers were counteracted by differences in the spectral and temporal characteristics of the interferers (Calandruccio et al., 2010). While the difference in the temporal characteristics (short-term energy distribution) was minimal, there were clear differences in the long-term spectra of the interferers. The calculation of the SNR for the speech material and the interferers (in 1/3 octave bands), weighted by the band-importance function for average speech (ANSI S3.5, 1997), resulted in values of 0.5 and 2.0 dB for the ISTS and the DT interferer, respectively. Thus, based on the long-term spectrum, the ISTS interferer masks the target speech more effectively than the DT interferer. A compensation for this difference would lead to SRTs for the ISTS interferer that are below those obtained with the DT interferer $[F(1,8) = 13.3, p < 0.01]$. Thus, the DT interferer might actually distract the listener more than does the ISTS interferer, but the effect seems to be only marginal.

3.6 Summary and conclusions

The present study investigated whether the reduced MR typically observed for HI listeners compared to NH listeners might result from their larger SNR in the stationary-noise reference condition. By processing the stimuli with LP and HP filtering as well as noise vocoding, and presenting them to NH listeners, the performance in the reference condition was degraded to the same level as for the HI listeners.

The main results of this study were as follows:

(1) For the modulated-noise interferer (SAM), none of the processing techniques reduced the MR for the NH listeners to the same amount as found for the HI listeners. In contrast, for the two speech maskers, the noise vocoder reduced the MR for the NH listeners to an amount similar to that observed with the HI listeners.

(2) The MR for the NH listeners obtained with processed stimuli was found to strongly depend on the type of processing, indicating that the stationary-noise SRT only partly predicts the MR obtained for the NH listeners. Therefore, the reduction in the MR observed for the HI listeners is probably not only determined by the stationary-noise SRT. For the interfering talker, the results are consistent with Bernstein and Grant (2009), but not for the modulated noise where they found the MR to be entirely predicted by the stationary-noise SRT.

(3) Assuming that degraded carrier information represents the primary distortion in the vocoded

stimuli, the good correspondence between the NH results with the vocoded stimuli and the HI results suggests that TFS information might be crucial for segregating the target from the interferer.

(4) Since the vocoder processing did not reduce the MR obtained with the NH listeners to the level obtained with the HI listeners in the condition using the modulated-noise interferer, other factors than TFS information seem to be important for utilizing low-amplitude valleys in the noise. In the condition with the modulated-noise interferer, the MR for the HI listeners was more strongly correlated with their audiometric thresholds than with their stationary-noise SRT, indicating that audibility was an important factor here.

(5) LP and HP filtering reduced the MR by approximately the same amount, indicating that low-order resolved harmonics are not more important for MR than high-order unresolved harmonics. However, this might have resulted from the larger bandwidth of the HP-filtered stimuli.

Overall, the results from the present study show that an increased stationary-noise SRT only partly accounts for reduced MR for the HI listeners and that auditory processing deficits appear to affect the processing of speech particularly in fluctuating interferers. Audibility seems to be an important factor in the case of modulated noise, while intact TFS coding might be more crucial for the processing of competing speech. This study did not examine and characterize different hearing impairment factors. In order to further investigate speech perception in hearing-impaired listeners, a detailed characterization of their individual hearing losses would be required and quantitative models would help to further test hypotheses about the impact of specific impairment factors on speech perception.

Acknowledgments

We wish to thank our colleagues at the Centre for Applied Hearing Research for valuable comments and stimulating discussions. We also thank two anonymous reviewers and the Associate Editor, Emily Buss, for their very helpful comments on an earlier version of this paper. We are grateful to all the listeners for their participation in many hours of testing. This work has been partly supported by the Danish research council and partly by Oticon, Widex and GN Resound through a research consortium.

4

Contribution of high-rate envelope fluctuations to release from speech-on-speech masking [‡]

Masking release (MR) is the improvement in speech intelligibility for a fluctuating interferer compared to stationary noise. Reduction in MR due to vocoder processing is usually linked to distortions in the temporal fine structure (TFS) of the stimuli and a corresponding reduction in the fundamental frequency (F0). However, it is unclear if high-rate envelope fluctuations, produced by the interaction between unresolved harmonics and related to F0, contribute to MR. This was investigated in the present study. Speech reception thresholds (SRT) were measured in the presence of stationary speech-shaped noise and a competing talker, and the corresponding masking release (MR) was determined. Two types of processing were applied to the stimuli. (i) An amplitude and frequency modulated vocoder attenuated the high-rate envelope fluctuations and (ii) high-pass filtering (cutoff = 500 Hz) reduced the influence of F0-related information from low-order resolved harmonics. The results showed that the MR was unaffected by HP filtering, but slightly reduced when high-rate envelope fluctuations were attenuated. When both types of processing were applied, the MR was strongly reduced. Thus, the results indicate that F0-related information is crucial for MR, but that it is not important whether the F0-related information is conveyed by the low-order resolved harmonics or by high-rate envelope fluctuations of unresolved harmonics. This also means that high-rate envelope fluctuations contribute substantially to MR.

[‡] This chapter is based on Christiansen et al. (2012).

4.1 Introduction

The most important mode of communication in our daily life is speech. However, in many situations, speech communication takes place in adverse conditions with high levels of background noise, several interfering talkers or reverberation. Normal-hearing (NH) listeners are typically able to understand speech even at very low signal-to-noise ratios (SNRs). In conditions where the interfering sound is an amplitude modulated noise or a competing talker, NH listeners are commonly able to utilize speech information in the low-amplitude parts of the interferer such that they are able to understand the speech at a much lower SNR than in the case of a stationary-noise interferer. This ability has usually been referred to as "listening-in-the-dips" and the corresponding improvement in speech intelligibility has been termed masking release (MR). Compared to NH listeners, hearing-impaired (HI) listeners often need much higher SNRs to understand speech in noise and often show very little or no MR (e.g., Festen and Plomp, 1990; Gustafsson and Arlinger, 1994; Peters et al., 1998; George et al., 2006; Lorenzi et al., 2006; Bernstein and Grant, 2009; Strelcyk and Dau, 2009). Even after compensating for reduced sensitivity with hearing aids, many HI listeners still show great difficulties in adverse listening conditions (e.g., Duquesnoy and Plomp, 1983; Gustafsson and Arlinger, 1994; Shanks et al., 2002; Hällgren et al., 2005; Metselaar et al., 2008). The reduced MR experienced by HI listeners has traditionally been ascribed to reduced frequency selectivity or an increased amount of forward masking that might limit their ability to benefit from spectral and temporal dips in the masker (e.g., Glasberg et al., 1987; Festen and Plomp, 1990; Baer and Moore, 1993, 1994; Dubno et al., 2003; Nelson and Jin, 2004). Furthermore, it has been proposed that deficits in the processing of the temporal fine structure (TFS) of the stimuli affect the coding of the stimuli's fundamental frequency (F0) in HI listeners (e.g., Qin and Oxenham, 2003; Hopkins et al., 2008; Oxenham and Simonson, 2009).

Coding of F0 plays an important role for the perceptual segregation of concurrent and sequential sources (Brokx and Nooteboom, 1982; Darwin, 1997) and may underly the observed MR when the masker is a competing talker (e.g. Qin and Oxenham, 2003; Bernstein and Grant, 2009; Bernstein and Brungart, 2011; Christiansen and Dau, 2012). In general, there are two different theoretical concepts describing how the F0 of a stimulus can be extracted by the auditory system; via place or temporal coding. In terms of place coding (pattern matching), the F0 of a stimuli can be extracted by matching harmonic templates to basilar membrane (BM) excitation pattern. (Wightman, 1973; Terhardt, 1974; Cohen et al., 1995). In terms of temporal coding, the firing of auditory-nerve cells synchronous with BM vibration can be used to extract the F0 of the input stimulus via inter-spike intervals (ISIs) (Licklider, 1951; de Cheveigné, 1998; Meddis and Hewitt, 1991). At low frequencies, the spectral harmonics of voiced speech are spatially resolved by the auditory system and the F0 of the input stimuli can be extracted from the BM excitation pattern or from the the period of individual frequency components in the corresponding channels. At high frequencies, the harmonics are considered to be spatially unresolved due to the increasing bandwidth of the auditory filters with increasing center frequency. However, interaction between harmonics within the same auditory filter gives rise to high-rate envelope fluctuations related to the F0 of the stimuli. Since the

ability of auditory-nerve cells to phase lock to the vibration of the BM is progressively reduced for increasing frequency, it is generally assumed that at high frequencies, the ISIs of the auditory nerve cells reflect the periodicity of the envelope fluctuations (e.g., Palmer and Russell, 1986).

Several studies have shown that low-order harmonics provide better F0 discrimination performance than high-order unresolved harmonics (Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994; Bernstein and Oxenham, 2003) and dominate the perceived F0 in the case of conflicting cues (Plomp, 1967; Micheyl and Oxenham, 2007; Bird and Darwin, 1998). However, the contribution of high-rate envelope fluctuations to MR has not been investigated explicitly.

Oxenham and Simonson (2009) investigated if pitch information provided by the low-order resolved harmonics is important for MR. They measured MR for NH listeners using low-pass (LP) and high-pass (HP) filtered stimuli in order to either retain or eliminate low-order harmonics, while achieving the same speech intelligibility in steady-state noise. In both conditions, MR was greatly reduced. Oxenham and Simonson (2009) suggested that MR might be determined mainly by the perceptual redundancy of the target speech instead of the F0 of resolved harmonics. Interestingly, they also noted that although the MR was relatively small, the pitch of unresolved high-order harmonics was sufficient for source segregation in the case of speech stimuli.

Stone et al. (2008) investigated the role of high-rate envelope fluctuations for speech perception using vocoder processing and found that NH listeners showed a speech intelligibility benefit with an interfering talker. In contrast, Xu and Zheng (2007) found that NH listeners showed no speech intelligibility benefit from high-rate envelope fluctuations in stationary noise. Combined, these results indicates that high-rate envelope fluctuations might be important for MR.

The F0 of voiced speech might be important for speech-on-speech masking in two ways: through the identification of the time intervals that contain target speech versus those that contain competing speech, and through the identification of the frequency regions that contain target speech versus those that contain competing speech. In the higher frequencies this information may be provided by the high-rate envelope fluctuations produced by unresolved harmonics.

In order to test if high-rate envelope fluctuations contribute to MR, a novel signal processing technique was developed to attenuate these envelope fluctuations. The technique is based on a traditional tone-vocoder but maintains the instantaneous frequency in each channel. Briefly, the input signal was divided into 16 frequency channels where both the amplitude and the instantaneous frequency (IF) course were estimated. In each channel, a LP filtered version of the IF course was used to drive a sine generator which was then modulated by a LP filtered version of the estimated envelope. Finally, all channels were recombined to generate the output signal. Using frequency modulated tone carriers, a natural sounding speech output can be obtained using a relatively small number of channels. This allows for a relatively large separation of carrier frequencies in adjacent channels so that high-rate envelope fluctuations are not reintroduced in each channel due to interaction between carriers, when the channels are recombined. The attenuation of high-rate envelope fluctuations was combined with a reduction in the F0-related information from low-order resolved harmonics obtained via LP filtering with a relatively low cut-off frequency of 500 Hz. It

was thus possible to investigate separately the effect of reduced resolved harmonics and reduced high-rate envelope fluctuations as well as the effect of reducing both.

In the present study, speech intelligibility was measured in four different processing conditions. The cutoff-frequency of the envelope low-pass filter was chosen to be either 30 Hz or 300 Hz in order to attenuate or retain high-rate envelope fluctuations. After vocoding, the stimuli were either HP filtered at 500 Hz or unprocessed in order to either reduce or retain F0-related information conveyed by the low-order resolved harmonics.

4.2 Signal processing

The stimuli were processed by a vocoder with amplitude and frequency modulated tone carriers, as illustrated in Fig. 4.1. First, the signal was decomposed into 16 frequency channels using a gammatone filterbank (Patterson et al., 1987). The filterbank consisted of fourth-order gammatone band-pass filters with center frequencies ranging from 50 to 7500 Hz, equally spaced on an equivalent-rectangular-bandwidth (ERB_N) number scale (Glasberg and Moore, 1990), each with a bandwidth of 1 ERB_N. In each channel, the envelope and instantaneous frequency (IF) were estimated in two parallel paths. The envelope was calculated by the absolute value of the analytical signal (via the Hilbert transform) and LP filtered with a fourth-order butterworth filter (24 dB/octave slope). The IF was estimated using the algorithm described in Nguyen et al. (2009), which is a Kalman smoother based dynamic autoregressive model developed for tracking the IF of noisy and non-stationary sinusoids. The estimated IF was smoothed with a 50-ms median filter and a 50-ms moving-average filter. The estimated IF was used to drive a sine generator and the output signal was amplitude modulated by the envelope signal. Before recombining all the channels, the root-mean-square (RMS) value in each channel was normalized to the input RMS in the corresponding channel. The final signal was scaled to have the same overall level as the input to the vocoder.

In order to remove or retain F0-related envelope modulations, the cut-off frequency of the envelope LP filter was either 30 Hz or 300 Hz. The speech and interferer were processed independently and mixed after the processing. The mixture was either presented to the listeners without any further processing or HP filtered with a cut-off frequency of 500 Hz. Thus, four different conditions were considered in the experiment: Two broadband conditions with a 30-Hz (BB30) or a 300-Hz (BB300) envelope filter in the vocoder, and two HP filtered conditions with a 30-Hz (HP30) or a 300-Hz (HP300) envelope filter in the vocoder.

The HP filtering procedure was conducted in the same manner as described in Oxenham and Simonson (2009). The signals were mixed at the appropriate SNR and then HP filtered at 500 Hz with a fourth-order butterworth filter. An off-frequency masker was generated by LP filtering the speech-shaped noise at 500 Hz (fourth-order butterworth filter), and the RMS level was adjusted to 12 dB below the level of the target sentence.

By using a vocoder with frequency modulated carriers it is possible, to some extent, to preserve the original temporal and spectral structure in the speech signal and obtain a naturally sounding

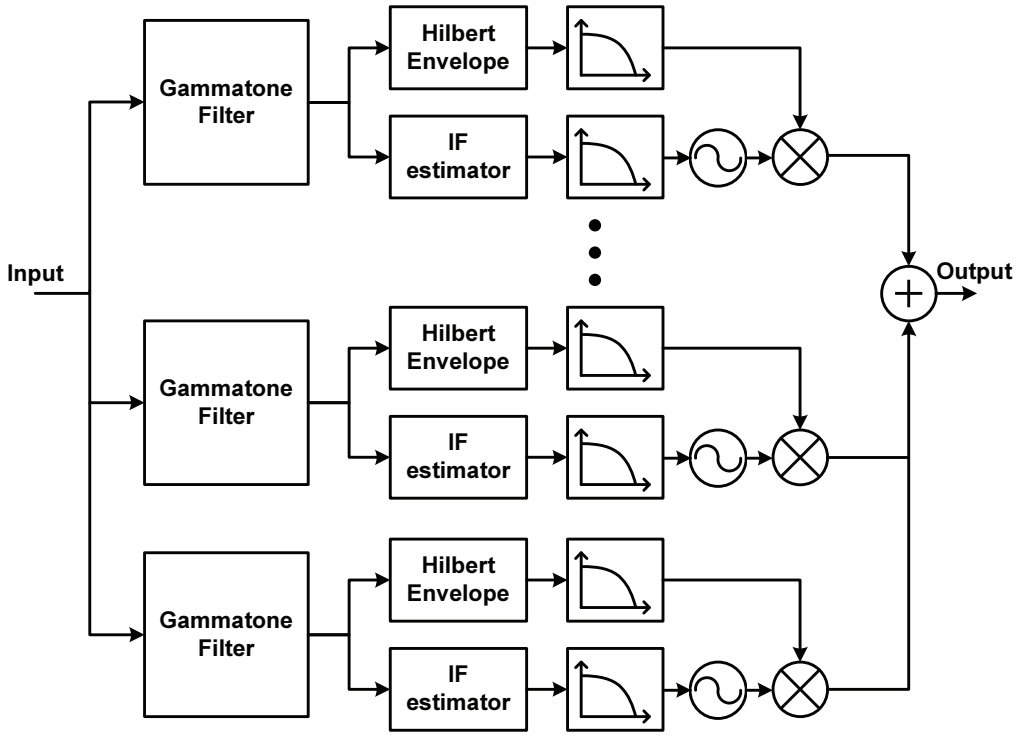


Figure 4.1: Schematic of the amplitude and frequency modulated vocoder. The input signal is divided into 16 frequency channels. In each channels the envelope and IF is calculated and low-pass filtered. The IF drives a sine generator where the amplitude is modulated by the estimated envelope.

representation of the speech using only 16 channels. In the 16-channel filterbank, the separation between the center frequencies is relatively large ($\approx$ 2 ERB_N). This is advantageous because it avoids interactions between the carriers when the channels are recombined. Thus, using this processing technique, it is possible to generate stimuli that have the same overall spectro-temporal energy pattern but with reduced high-rate envelope fluctuations. This is illustrated in Fig. 4.2 that shows the auditory spectrogram of a sentence processed by the vocoder with an envelope cut-off frequency of 30 Hz (left panel) and 300 Hz (right panel), respectively. The auditory spectrogram was produced using 128 fourth-order gammatone filters, ranging from 0 to 8000 Hz and equally spaced on an ERB_N number scale. The output of each channel is the Hilbert envelope low-pass filtered at 500 Hz. The two panels show that the overall spectro-temporal structure of the processed signals is very similar. The main difference is found in the higher frequency channels where high-rate envelope fluctuations can be seen in the right panel (300-Hz envelope LP filter) but not in the left panel (30-Hz envelope LP filter).

Figure 4.3 shows a more detailed comparison of the envelopes obtained with the 30-Hz and the 300-Hz envelope cut-off frequency in the frequency channels at 119-Hz, 1085-Hz and 2958-Hz, respectively. The envelopes shown in the figure are based on an analysis of the *processed* signal (i.e., the vocoder output), using the same gammatone filterbank as in the vocoder. This was done to verify that high-rate envelope fluctuations were not reintroduced in the processed signal due to interaction between the carriers. The envelopes were low-pass filtered Hilbert envelopes similar to the processing in the vocoder.

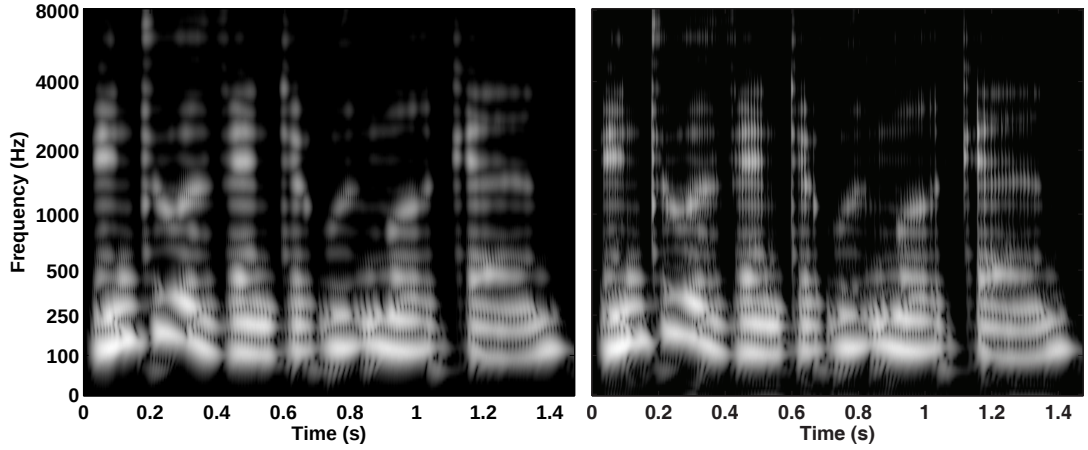


Figure 4.2: Auditory spectrogram of the sentence ("Han hoppede op på cyklen"), processed by the vocoder with a 30-Hz (left panel) and a 300-Hz (right panel) envelope low-pass filter. The spectrograms were obtained using a 128-channel gammatone filterbank, with 1-ERB_N wide filters, equally spaced on an ERB_N number scale.

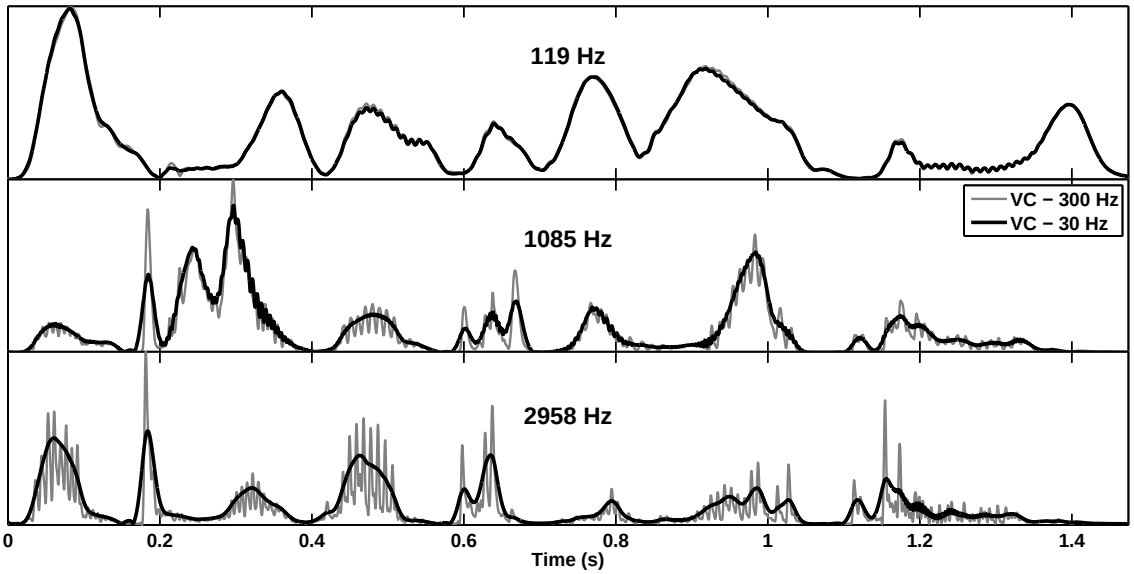


Figure 4.3: Analysis of the envelopes of the vocoded signals. The processed signals were analyzed using the same front end as the vocoder (16-channel gammatone filterbank and low-pass filtered Hilbert envelopes). The three panels show the output of the channels with the center frequencies: 119-Hz (top), 1085-Hz (middle) and 2958-Hz (bottom). In each panel, the envelope of the vocoded signal with a 300-Hz (dark-gray) and a 30-Hz (black) low-pass filter is shown, respectively.

In the top panel, it is clear that the two envelopes are almost identical due to the slow fluctuations in the amplitude of the signal. At 1085 Hz (middle panel), the 300-Hz condition shows a small amount of high-rate envelope fluctuations related to the F0 of the talker, which is not reflected in the 30-Hz condition. At 2958 Hz (bottom panel), similarly, no high-rate envelope fluctuations are represented in the 30-Hz condition. However, in this case, they are even more pronounced in the 300-Hz condition. Thus, the reanalysis clearly shows that high-rate envelope fluctuations in the mid- to high-frequency channels are attenuated or completely removed when the 30-Hz low-pass filter is applied in the vocoder processing.

4.3 Methods

4.3.1 Listeners

Five NH listeners with audiometric thresholds of 20 dB or less at all measured frequencies between 125 and 8000 Hz participated in the experiment. The age of the listeners ranged between 22 and 32 years with a mean age of 27.

4.3.2 Speech material

Speech reception thresholds (SRTs) were measured using the Danish speech intelligibility test called conversational language understanding evaluation (CLUE, Nielsen and Dau, 2009), which is very similar to the hearing-in-noise test (HINT) originally developed for English (Nilsson et al., 1994). The CLUE material consists of natural and meaningful sentences representing conversational speech and has a fixed structure consisting of five words per sentence. The sentences were spoken by a male talker with an average fundamental frequency (F0) of 119 Hz.

The maskers were an unintelligible single talker and a stationary noise. The single talker was the international speech test signal (ISTS; Holube et al., 2010), which consists of natural speech from six female talkers speaking different languages that have been segmented and remixed using a randomization procedure in order to make it largely unintelligible. The average F0 of the ISTS signal was of 207 Hz. The stationary noise was equalized to have the same long-term spectrum as the ISTS signal.

4.3.3 Procedure

The experiment was conducted in a double-walled sound insulated booth, where the experimenter controlled the procedure by means of a Matlab application developed specifically for the CLUE test. The digital signals were sampled at 22050 Hz and converted to analog signals by a high-end 24 bit soundcard (RME DIGI96/8). The stimuli were presented diotically over Sennheiser HD580 headphones. The target sentences were presented at a fixed level of 65 dB SPL, whereas the level of the interferer was determined via an adaptive procedure used to measure the SRTs. The onset and offset of the interferer were 1 s before and 600 ms after the sentence, respectively, where a ramped squared-cosine function with a duration of 400 ms was applied to the onset and the offset. For each presentation, the interferer was randomly selected from a long sample (SSN: 22 seconds, ISTS: 52 seconds).

The listeners received approximately 20 minutes of training before the SRTs were measured. In the training session, the first sentence was presented at a very low SNR. The SNR was increased in steps of 2 dB until all five words were repeated correctly. The test subjects were allowed to guess and the recognized words were repeated verbally to the experimenter and registered without

feedback. For the following sentence, the SNR was decreased by 4 dB and again increased in 2 dB steps until all the words were repeated correctly.

In the test session, a list of 10 sentences was used to measure the SRT for a given run. The procedure for the presentation of the first sentence was the same as in the training session. However, for the presentation of the remaining nine sentences, the SNR followed a simple adaptive procedure: If all words were repeated correctly, the SNR was decreased by 2 dB; otherwise the SNR was increased by 2 dB. The measured SRT was the average of the last eight SNRs from presentation number 4 to 11, where the last presentation is the SNR determined after the last sentence although there is no sentence presented. Five runs were conducted for each condition and the average of these SRTs produced the final SRT.

4.4 Results

The left panel of Fig. 4.4 shows the average SRTs obtained with the two maskers (SSN and ISTS) in the four different experimental conditions. As expected, the SRTs for the ISTS masker (gray lines) are much lower than for the SSN masker (black lines). For the SSN masker, there was a small ($\approx$ 1 dB) but significant difference between the 300-Hz (solid line) and the 30-Hz (dashed line) envelope conditions [p<0.05], indicating that high-rate envelope fluctuations contribute by a small amount to speech perception in stationary noise. For the ISTS masker, the SRT in the BB30 condition was about 3 dB higher than for the BB300 condition [p<0.0001], while the SRT in the HP30 condition was 7-8 dB higher than in the HP300 condition [p<0.0001]. Thus, there is a clear interaction between envelope filtering and the reduction of resolved harmonics via HP filtering. The results for the ISTS masker indicate that, for a competing talker, high-rate envelope fluctuation plays a substantial role for MR. However, the contribution of high-rate envelope fluctuations is considerably smaller when the listeners can also rely on F0-related information from resolved harmonics.

The right panel of Fig. 4.4 shows the MR for the ISTS in the four conditions, representing the difference in SRT between the SSN and ISTS maskers. The conditions are indicated by the same symbols and line styles as used in the left panel. A repeated measures ANOVA confirmed that the MR differed significantly across envelope filter conditions $[F(1,4) = 70.83, p < 0.002]$, HP filter conditions $[F(1,4) = 10.01, p < 0.04]$ and that there was a significant interaction between envelope and HP filter conditions $[F(1,4) = 14.99, p < 0.02]$. There was no effect of reducing the F0-related information from low-order resolved harmonics on the MR obtained with a competing talker [p=0.72]. In contrast, there was a small but significant reduction of the MR ($\approx$ 1.5 dB) when high-rate envelope fluctuations were attenuated [p<0.005]. However, when F0-related information from both resolved and high-order unresolved harmonics was reduced, a large reduction in the MR was observed ($\approx$ 5-7 dB) [p<0.0001].

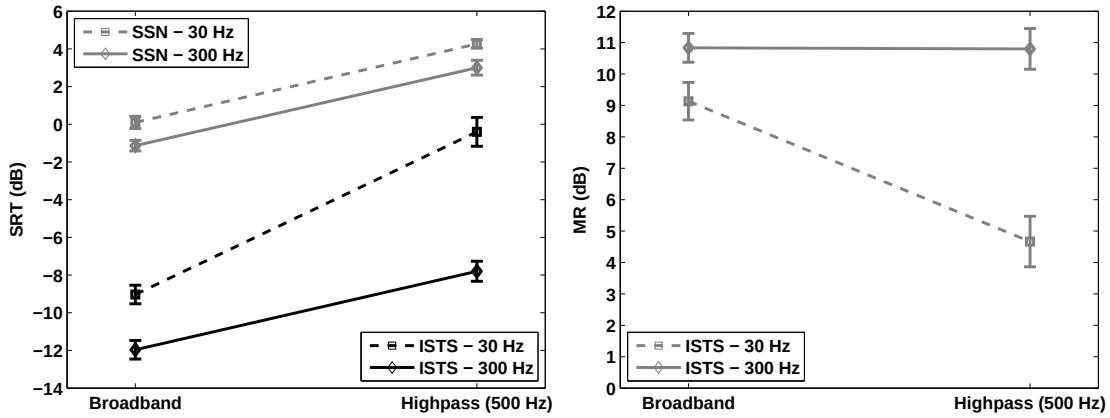


Figure 4.4: The left panel shows the average SRTs obtained with SSN masker (black lines) and the ISTS masker (gray lines) in the four different conditions. The results obtained using the 300-Hz and the 30-Hz envelope filter in the vocoder are represented by the solid and dashed lines, respectively. Whether the stimuli were broadband or high-pass filtered is denoted on the abscissa. The right panel shows the MR calculated as the difference in SRTs between the SSN and ISTS maskers from the left panel. Errorbars represent $\pm$ 1 standard deviation

4.5 Discussion

4.5.1 Summary of the main results

The results from the present study show that HP filtering the overall stimuli at 500 Hz had no effect on MR. However, LP filtering the high-rate envelope fluctuations at 30 Hz reduced MR by approximately 1.5 dB. Importantly, when both processing schemes were applied, MR was strongly reduced (approximately 6.5 dB). Thus, as expected, the results suggest that F0 information from low-frequency harmonics contribute significantly to MR. However, more importantly, the results suggests that high-rate envelope fluctuations are even more important for MR. Thus, although several studies have shown that low-order resolved harmonics dominate over high-order unresolved harmonics in pitch perception (Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994; Bernstein and Oxenham, 2003; Plomp, 1967; Micheyl and Oxenham, 2007; Bird and Darwin, 1998), the results from the present study suggest that the high-rate envelope fluctuations produced by unresolved harmonics may be more important for speech perception.

4.5.2 The role of resolved and unresolved harmonics for MR

The finding that the listeners were able to achieve a large MR mainly based on F0 information from unresolved harmonics supports the interpretation of the results in Oxenham and Simonson (2009) that the pitch of high-order harmonics was sufficient for source segregation in the case of speech stimuli. It is also consistent with the indications of Stone et al. (2008) who used a noise and a tone vocoder to investigate the importance of F0-related high-rate envelope cues for speech perception. They measured the speech intelligibility with a competing talker and found that the

high-rate envelope cues contributed to the speech intelligibility. However, they did not explicitly investigate MR, i.e., the advantage of a competing talker relative to a stationary interferer.

Oxenham and Simonson (2009) used LP filtering with a cutoff frequency of 1200 Hz and HP filtering with a cutoff frequency of 1500 Hz to isolate the effects of resolved and unresolved harmonics. In both conditions, the MR was found to be greatly reduced indicating that resolved and unresolved harmonics separately did not contribute substantially to MR. This differ from the results of the present study where F0-related information from resolved harmonics and the high-rate envelope fluctuations of unresolved harmonics considered separately, contributed substantially to MR. However, these differences in results are likely due to differences in the procedures. Oxenham and Simonson (2009) suggested that the greatly reduced MR in both of their conditions could have been caused by a large reduction of the perceptual redundancy of the target speech due to the filtering and that MR might be determined mainly by the redundancy instead of the F0 of resolved harmonics. In the present study, the filtering of the envelope fluctuations did not reduce the bandwidth of the overall spectro-temporal energy pattern of the speech signal and the high-pass filtering used a relatively low cut-off frequency. Thus, the redundancy of speech information in the processed stimuli in present study was relatively well preserved. Based on this, it seems that as long as redundancy is preserved, F0-related information from both resolved harmonics and high-rate envelope fluctuations of unresolved harmonics contribute substantially to MR.

The finding that F0-related high-rate envelope fluctuations produced by unresolved harmonics contribute slightly more to MR than F0-related information from resolved harmonics could indicate that unresolved harmonics are distributed across a slightly larger frequency range than resolved harmonics and thereby contributing more to across frequency segregation.

4.5.3 Possible connections to reduced MR in HI listeners

Since the results of the present study showed that F0 information plays a crucial role for speech perception in the presence of a competing talker, it is likely that HI listeners, who often experience great difficulties in such a condition, might have difficulties in the processing of F0 information. However, we can only speculate about why these difficulties arise. Deficits in F0 processing could be a general deficit in extracting F0 information from the input stimuli, even if the stimuli only consist of a single talker in quiet conditions. Since high-rate envelope fluctuations were found to be sufficient for MR in NH listeners, difficulties in the processing F0, is probably related to deficits in the ability to extract F0 information from the temporal response of the auditory nerve fibers. This would indicate problems with phase-locking at frequencies as low as 100-250 Hz. However, this seems unlikely since NH listeners show robust phase-locking up to about 1200 Hz (Lindgreen, 2009; Santurette and Dau, 2012; Heinz, 2012). However, it could also be that the extraction of F0 information from a single source is more or less intact, but that HI listeners have deficits in the processing of F0 information from simultaneous sources and thereby have difficulties separating the target speech from an interfering talker. This could be due to interaction of source carriers within an auditory filter, which might be more pronounced in HI listeners with reduced frequency selectivity.

A larger interaction of source carriers due to reduced frequency selectivity could probably be simulated in NH listeners in future experiments using broader filters in the vocoder presented in the current study. Further experiments with both NH and HI listeners are needed in order to determine if HI indeed have deficits in the processing of F0 information.

4.5.4 Implications for auditory modeling

The finding that high-rate envelope fluctuations contribute substantially to MR indicate that the auditory system analyses modulation frequencies well beyond 30 Hz. If these fluctuations are indeed used to extract F0 information this suggest that the modulation frequencies up to about 300 Hz are analysed, which is approximately the upper limit for female speech. If MR also can be achieved with child speech it could suggest an analysis of even higher frequencies. The auditory system has been shown to perform a frequency selective analysis of envelope fluctuations which have been modeled by a modulation filterbank similar to the modeling of cochlear filters (Dau et al., 1997a,b; Ewert and Dau, 2000). A modulations-frequency specific analysis has also been found to be crucial for the prediction of speech intelligibility (Steeneken and Houtgast, 1980; Elhilali et al., 2003; Jørgensen and Dau, 2011) and speech quality (Kim, 2005). However, these speech perception models only consider frequency modulations up to 32 or 64 Hz. The results from the present study indicate that, in certain conditions, auditory models should include an analysis of relatively high modulation frequencies, possibly up to several hundred of Hz.

4.6 Summary and conclusions

The present study investigated the contribution of F0-related high-rate envelope fluctuations to MR obtained using a competing talker and stationary speech-shaped noise as maskers. This was done by LP filtering the envelope fluctuations using an amplitude and frequency modulated tone-vocoder. In addition, the contribution of F0 information from resolved harmonics was also investigated by removing some of these using a HP filter.

High-rate envelope fluctuations were found to be important for MR obtained with a competing talker. Indeed, the results suggest that they were equally as important as low-order resolved harmonics, if not even more important. The presence of high-rate envelope fluctuations or resolved harmonics were both found to be sufficient for MR. These findings suggest that, for some situations, auditory models may require an analysis of modulation frequencies spanning the range of F0 produced by the talkers of the stimuli. Although F0 information was found to be important for MR, further work is needed to determine if the reduced MR exhibited by HI listeners is indeed due to deficits in the processing of F0 information.

Acknowledgments

We wish to thank our colleagues at the Centre for Applied Hearing Research for valuable comments and stimulating discussions. We are grateful to all the listeners for their participation in many hours of testing. This work has been partly supported by the Danish research council and partly by Oticon, Widex and GN Resound through a research consortium.

5

Analyzing the variation of consonant confusions in hearing-impaired listeners [§]

The patterns of consonant confusions obtained with hearing-impaired (HI) listeners have been shown to vary strongly between individual listeners and even between the left and the right ear (Phatak et al., 2009). However, measurements with normal-hearing (NH) listeners have also shown a large variability in confusion patterns across different utterances of the same consonants. Thus, the variation across HI listeners could be due to each listener making errors on only one specific utterance. To further understand the problems of the individual HI listeners, the present study investigated confusion patterns of HI listeners on an utterance-by-utterance basis. Each listener showed very consistent responses, confusing most utterances with only one specific consonant. Furthermore, while the confusions for each utterance were often found to be the same across listeners and ears, the confusions depended strongly on the utterance. A possible explanation for the results is that all utterances contain a primary acoustic cue leading to correct recognition of the consonant. In addition, some utterances also contain a secondary cue that leads to correct recognition even though the primary cue is masked or inaudible. However, most of these utterances also contain a conflicting cue that can lead to confusion with another consonant when the primary cue is masked or inaudible, since the conflicting cue often is stronger than the secondary cue. As the combination of primary, secondary and conflicting cues is different for each utterance, the confusion patterns are also different. Thus, the variation of the confusion patterns across HI listeners for a specific consonant seems to be a result of the HI listeners making errors on different utterances promoting different confusions.

[§] This chapter is based on Christiansen et al. (2012).

5.1 Introduction

Compared to NH listeners, HI listeners often experience great difficulties in understanding speech in complex acoustic environments and different types of background noise. Although compensating for reduced sensitivity largely improves the ability of HI listeners to understand speech in quiet (e.g., Duquesnoy and Plomp, 1983), they still experience difficulties in the presence of noise (e.g., Duquesnoy and Plomp, 1983; Gustafsson and Arlinger, 1994; Shanks et al., 2002; Hällgren et al., 2005; Metselaar et al., 2008). In particular, HI listeners have problems with fluctuating noise or interfering talkers (e.g., Festen and Plomp, 1990; Gustafsson and Arlinger, 1994; Peters et al., 1998; George et al., 2006; Lorenzi et al., 2006; Bernstein and Grant, 2009; Strelcyk and Dau, 2009).

When speech is masked by noise or other interferers such as competing speakers, the recognition of a target message roughly relies on a three-step process. First, the listener must be able to detect the acoustic energy of the target speech. Second, the listener needs to be able to separate the spectro-temporal energy pattern of the target speech from the masker. Third, the listener must be able to decode the meaning of the spectro-temporal energy pattern. Speech perception in the presence of steady-state noise is mainly considered to be limited by the amount of speech rendered inaudible by the noise, i.e., the ability to detect the target speech (e.g., French and Steinberg, 1947; Steeneken and Houtgast, 1980; ANSI S3.5, 1997). In contrast, speech perception in the presence of fluctuating noise and competing speech is mainly considered to be limited by the ability to segregate the target speech from the masker (e.g., Qin and Oxenham, 2003; Hopkins et al., 2008; Brungart et al., 2006, 2009). While several studies have investigated how the hearing loss in individual HI listeners affects their ability to detect and segregate speech by measuring speech reception thresholds (SRTs) in different types of noise, only a few studies have examined how the hearing loss of individual listeners effects their ability to decode the available speech information.

The purpose of the present study was, therefore, to investigate the decoding errors of HI listeners by measuring consonant confusions in individual listeners. Consonant recognition studies have shown that HI listeners make significantly more errors than NH listeners in quiet (Walden and Montgomery, 1975; Bilger and Wang, 1976) and in the presence of a noise (Dubno et al., 1982; Gordon-Salant, 1985). However, these studies have only analyzed the average confusions made by a group of HI listeners and have not investigated the large variability across HI listeners often found when measuring SRTs.

Perceptual confusions of consonants were first investigated by Miller and Nicely (1955), who found a very systematic pattern in the confusions of the different consonants. They measured the confusions of 16 different consonants masked by white noise or exposed to low-pass filtering. Analyzing the confusions in terms of articulatory features showed that place of articulation was strongly affected by low-pass filtering, while voicing and nasality were only weakly affected. Furthermore they found that: "*When a perceptually relevant acoustic feature of a speech sound is masked by noise, that sound becomes confused with related speech sounds. Such confusions provide vital information about the human speech code, i.e., the perceptual feature representation of speech sounds in the auditory system.*"

Many of the studies investigating the perceptual cues of speech have used synthetic speech stimuli, where the different cues are easily controlled. However, a major limitation of synthetic speech is that it requires *a priori* knowledge about the speech cues and that it is only possible to find the cues that have been hypothesized.

As part of a larger study, Phatak and Allen (2007) and Phatak et al. (2008) repeated the classical confusion matrix experiment of Miller and Nicely (1955) with white noise and speech-weighted noise, respectively. The overall aim was to combine the measured confusions with a spectro-temporal analysis of the presented stimuli in order to identify the acoustic features, extracted by the human auditory system, which form the basis for perception of different speech sounds. In order to capture the natural variability in speech production, the consonant confusions were measured with 18 different talkers. For some consonants, the confusion patterns varied significantly from utterance to utterance, while other consonants showed a large variability in the number of errors across the different utterances.

Régnier and Allen (2008) combined the consonant confusions from Phatak and Allen (2007) and Phatak et al. (2008) with a 4-step spectro-temporal analysis of the stimuli used to produce the consonant confusions. The analysis was performed on all the utterances of the consonant /t/ and reliably identified the primary feature of the /t/ as a synchronous 2-8 kHz temporal burst occurring approximately 50 ms before the onset of the following vowel. Li et al. (2010) combined the spectro-temporal analysis of Régnier and Allen (2008) with consonant confusions obtained using time-truncated as well as high-pass and low-pass filtered stimuli. This was done as part of the development of a general psychoacoustic method to find the perceptual cues of all the stop consonants in natural speech. Besides identifying the perceptual cues of the stop consonants, they also found that many of the natural speech sounds contained perceptually conflicting cues triggering the recognition of another consonant than the presented one. Kapoor and Allen (2012) investigated the relative importance of the burst and formant transition for the recognition of the stop consonants /t,k,d,g/ using consonant-vowel stimuli. By amplifying or attenuating the burst feature, they found a strong relation between the change in the strength of the burst feature and the change in the SNR needed to obtain a score of 90% correct, indicating that the bursts are the primary perceptual cues. The results also showed that when the bursts were attenuated the presented consonants were often confused with a specific other consonant due to conflicting cues, supporting the results of Li et al. (2010). Furthermore, in some cases, when the primary feature was removed, the listeners were still able to correctly identify the presented consonant due to a formant-onset cue that was stronger than the conflicting cues and acted as a secondary cue. Thus, the large differences in the consonant confusions obtained with different utterances of the same consonant (Phatak and Allen, 2007; Phatak et al., 2008) indicate that different utterances have different conflicting cues.

Recent studies investigating consonant confusion of HI listeners have shown that there is not only a large variation in the number of errors across HI listeners, but also between the left and right ear of the same listener (Phatak et al., 2009; Han, 2011; Allen and Han, 2011). Similarly, differences in the pattern of confusions were observed across the HI listeners (Phatak et al., 2009) as well as between the ears of the same listener (Han, 2011; Allen and Han, 2011). Based on these findings,

it appears that combining the knowledge about the acoustic features (that form the basis for the perception of the different consonants) with measures of consonant confusions made by individual HI ears (on an utterance-by-utterance basis) can provide valuable information about the effect of hearing loss on speech decoding. In this framework, the present study exclusively investigated the consonant confusions made by individual HI ears on an utterance-by-utterance basis.

5.2 Method

5.2.1 Listeners

Eight HI listeners between 65 and 84 years of age (mean age of 74) participated in the experiment. All listeners were native American English speakers with sensorineural hearing loss indicated by type A tympanogram. The individual audiometric thresholds are listed in table 5.1.

5.2.2 Stimuli

The speech stimuli was consonant-vowel (CV) sounds consisting of 14 different consonants (/b/,/d/,/g/,/p/,/t/,/k/,/ʃ/,/s/,/ʒ/,/z/,/f/,/v/,/m/,/n/) followed by the vowel sound /ɑ/, from the LDC2205S222 database (Fousek et al., 2004), recorded at the Linguistic Data Consortium (University of Pennsylvania). Phatak and Allen (2007) measured consonant confusions in NH listeners for a subset of the CV sounds in the LDC2205S222 database consisting of 18 different speakers and found that for each of the CV sounds, the utterances of specific speakers were very robust to noise. The utterances of each CV sound used in the present study were chosen from these noise robust utterances, in a way, such that each CV sound was represented by a male and a female utterance.

5.2.3 Procedure

The experiment was conducted in a sound insulated booth, where the digital signals that had been sampled at 16 kHz were converted to analog signals by a 16 bit soundcard (Soundblaster Live). The stimuli were presented monaurally through an Etymotic ER-2 insert earphone. The target speech was presented at a fixed sound pressure level (SPL) adjusted to the most comfortable level for each listener using an external TDT PA5 attenuator, while the level of the noise masker was varied according to the SNR for each presentation. The listeners were seated in front of computer monitor where they responded by choosing the the perceived CV from a graphical user interface showing all the possible responses. The listeners were able to play each CV as many times as they liked before responding. The presentation of the different consonants, talkers and SNRs were performed in a random order for each listener. Consonant confusions were measured independently for the left and right ear of the 8 HI listeners at four different signal-to-noise ratios (SNRs), Quiet, 12 dB, 6 dB and 0 dB. Each utterance was presented several times following a semi-adaptive procedure. In

Table 5.1: Audiometric thresholds for the eight HI listeners.

				Audiometric thresholds (dB HL)									
ID	Age	Ear	125	250	500	1000	1500	2000	3000	4000	6000	8000	PTA (dB HL)
HI_1	82	L	40	40	45	45	45	45	45	45	65	75	52.0
		R	45	45	45	50	45	45	55	65	80	110	60.6
HI_2	66	L	30	30	25	30	25	35	55	65	70	80	44.5
		R	25	25	25	25	25	30	55	60	90	80	44.0
HI_3	74	L	30	30	30	30	30	45	40	45	55	55	39.0
		R	30	30	30	20	25	30	45	50	55	60	37.5
HI_4	84	L	40	35	35	25	30	35	35	60	95	95	48.5
		R	30	25	20	25	30	40	45	65	75	90	44.5
HI_5	72	L	20	10	15	30	30	35	35	35	20	45	27.5
		R	25	25	25	25	30	35	45	50	40	55	35.5
HI_6	79	L	20	15	25	20	30	20	35	50	45	65	32.5
		R	20	10	25	15	30	30	40	35	45	50	30.0
HI_7	65	L	15	10	5	5	20	20	35	55	20	25	21.0
		R	15	10	10	15	15	20	15	45	25	30	20.0
HI_8	67	L	15	15	10	5	10	10	40	55	35	60	25.5
		R	20	25	25	20	15	5	15	40	35	60	26.0

the first round of the experiment, each utterance was presented four times. If all four presentations were correctly identified, the utterance was only presented once in the second round, resulting in 5 presentations in total. If there was an error in the first round the utterance was presented six times in the second round, resulting in 10 presentations in total.

All the consonant confusions were measured by Woojae Han at the Speech and Hearing Science Department of the University of Illinois at Urbana-Champaign (Han, 2011).

5.3 Results

The consonant recognition scores in terms of the percentage of correctly identified consonants are described first, followed by a more detailed investigation of the confusions made by the individual listeners for different utterances.

5.3.1 Consonant recognition scores

The left panel of Fig. 5.1 shows the percentage of correctly identified consonants as a function of the SNR for each of the HI listeners, averaged across all the presented consonants. For comparison, the grand average of all the listeners is also included. As expected, there was a large variation in the performance across listeners. Listener HI1 showed the lowest overall recognition, with scores of less than 60% correct in quiet conditions, while listeners HI6 and HI8 showed the highest overall recognition with scores close to 100% in quiet and close to 90% at the lowest SNR. This corresponds well to the audiograms showing very mild hearing losses for listener HI6 and HI8

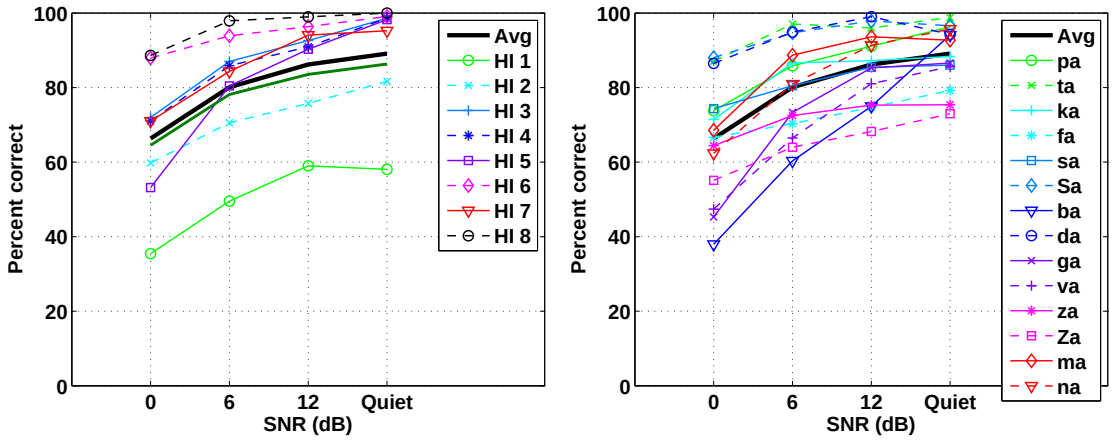


Figure 5.1: Percentage of correctly identified presentations as function of SNR for each HI listener, averaged across all the different consonants (left panel) and and for each consonant, averaged across all the different listeners (right panel). The different listeners and consonants are indicated in the two panels.

and the most severe hearing loss for listener HI1. However, the results also show that listener HI7, who has the mildest hearing loss of the group, performed very similarly to listeners HI3 and HI4 even though these two listeners have much higher audiometric thresholds. Another interesting observation is that, although listeners HI5 through HI8 have relatively similar audiograms and achieve close to a 100% correct perfomance in quiet, listener HI5 is much more adversly affected by an increase in the amount of masking noise. Indeed, only listener HI1 exhibited lower recognition performance at the lowest SNR than listener HI5. Also listener HI7, who has the lowest audiometric thresholds, is considerably more affected by noise than either listener HI6 and HI8. These results clearly show that the recognition performance of individual HI listeners cannot be accounted for solely by their individual audiometric thresholds. Furthermore, the results show that some listeners exhibit very high performance in the presence of noise, while other listeners have problems even in quiet conditions. Finally, some listeners exhibit a high performance in quiet but were very sensitive to background noise.

The right panel of Fig. 5.1 shows the percentage of correctly identified presentations for each of the presented consonants, averaged across all HI listeners. Similar to the results for the individual listeners, there is a large variation of the recognition scores across the consonants. The consonants /ta/, /da/ and /Sa/ form a group of easily recognized consonants with close to 100% correct responses for all but the lowest SNR. At the other end of the range, two groups emerge. The first group consits of /Za/, /za/ and /fa/. The recognition performance for this group is poor across all SNRs, including the quiet condition. The recognition performance for this group decreases rapidly with increasing noise. Thus, based on these results, some consonants are more easily recognized than others and the recognition some consonants is more robust to background noise. However, as illustrated in the left panel, this pattern might differ dramatically across the HI listeners.

In order to further analyze the data, the consonant recognition needs to be considered for each listener individually. Figure 5.2 shows such an analysis performed on the results from listener HI6.

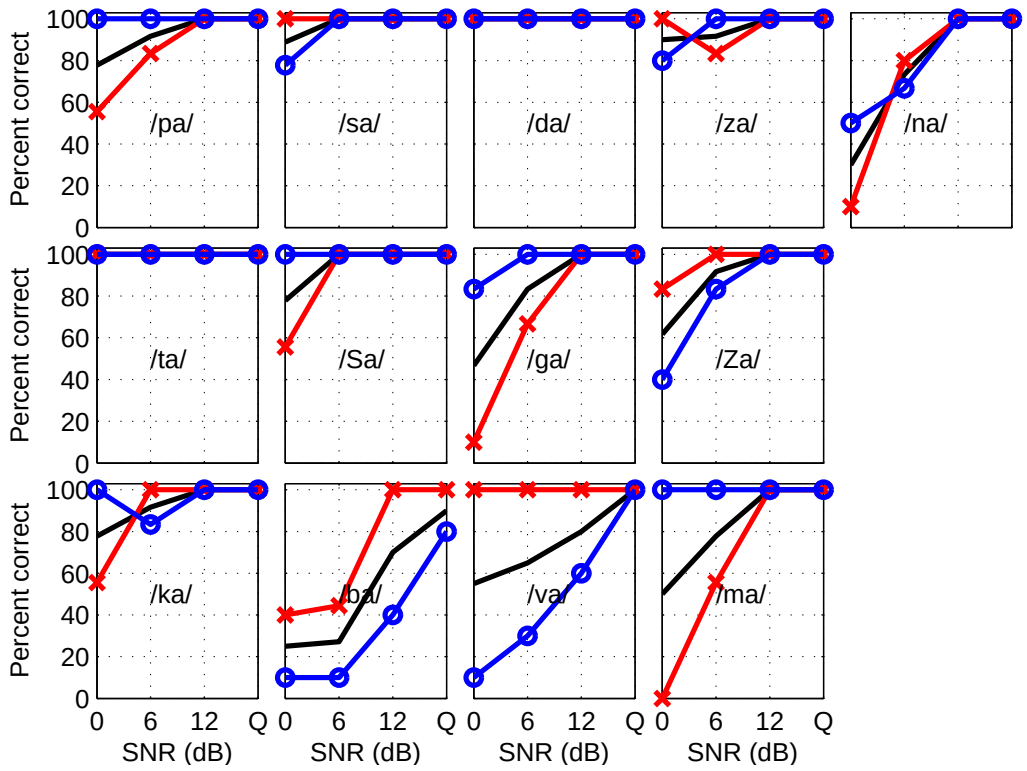


Figure 5.2: The panels in the figure show the consonant recognition for each of the consonants presented in the experiment obtained for the left ear of listener HI6. Each panel shows the percentage of correct responses as a function of SNR for the male utterance (blue line), for the female utterance (red line) as well as the average of both the utterances (black line). The consonant corresponding to each panel is indicated in panels.

The figure shows the recognition scores for all the presented consonants in separate panels. In each panel, the average score (black line) is shown together with the scores for the female (red line) and male (blue line) utterances. Although listener HI6 shows the highest average scores, with only about 10% errors at the lowest SNR, the errors are not evenly distributed across all the different consonants as might be expected. In contrast, listener HI6 has difficulty identifying several specific consonants (/ba/, /ga/, /ma/, /na/, /va/). Based on the overall average, one would not expect listener HI6 to have major communication difficulties. However, depending on the nature and consequences of the confusions, this listener could potentialy have much larger communication difficulties. In particular, Fig. 5.2 also shows that, for the /ga/, /va/ and /ma/, listener HI6 responds correctly to almost all the presentations of one of the utterances across all SNRs, while the scores of the other utterances approach 0% at low SNRs. Thus, listener HI6 does not generally have a problem recognizing /ga/, /va/ and /ma/, but difficulties with specific utterances. Similar observations can be made for most of the other HI listeners. Combined with a spectro-temporal analysis of the utterances of such consonants, these results may provide valuable information about the hearing loss of the listeners.

5.3.2 Consonant confusions

Figure 5.3 shows the major confusions for the utterances produced by the female (top panel) and male (bottom panel) talkers. In each panel, the bars show the proportion of each of the confusions relative to the total error. Only confusions contributing more than 15% are included. Thus, the difference between a value of 1 and the top of the bars represents one or more confusions each contributing less than 15%. Surprisingly, there is a large variation in the confusions between the female and male talkers. For many of the presented consonants, the listeners make, on average, different confusions for the female talkers compared to the male talkers. In general, for the female talkers, there is a tendency of utterances to be confused with two other consonants, while for the male talkers, the utterances are more consistently confused with only one other consonant.

The following describes some of the major differences in the confusion patterns between utterances of the female and male talkers. (i) /sa/ is mainly confused with /fa/ for the female talker, while it is confused with /za/ for the male. (ii) /Sa/ is mainly confused with /sa/ with the female talker, while it is confused with /za/ for the male. (iii) /ba/ is mainly confused with /da/ in the case of the female talker, while it is confused with /va/ in the case of the male. (iv) /ma/ and /na/ are strongly confused with each other, but they are not confused with any other in case of the male talker, while they are also considerably confused with /va/ in case of the female talker. Thus, overall, the results clearly show that the confusions of a particular consonant strongly depend on the talker that produced the utterance. However, it is unclear whether all the HI listeners make similar confusions.

In order to investigate this, the left panels of the figures 5.4 through 5.7 show the major confusions of the individual listeners for the utterances *f103ka*, *f103ma*, *f105Za* and *m120sa*, respectively. The rows show the confusions of the different listeners while the columns indicate the SNR at which the confusions were obtained. For each SNR, the dotted lines indicate 0% and 100% error and the bars show the proportion of the error constituted by each of the major confusions, where minor confusions (<30%) are indicated by black bars. The right panel of the figures show the audiogram of the listeners making considerable confusions, where the gray-shades of the audiograms indicate which confusion group they correspond to.

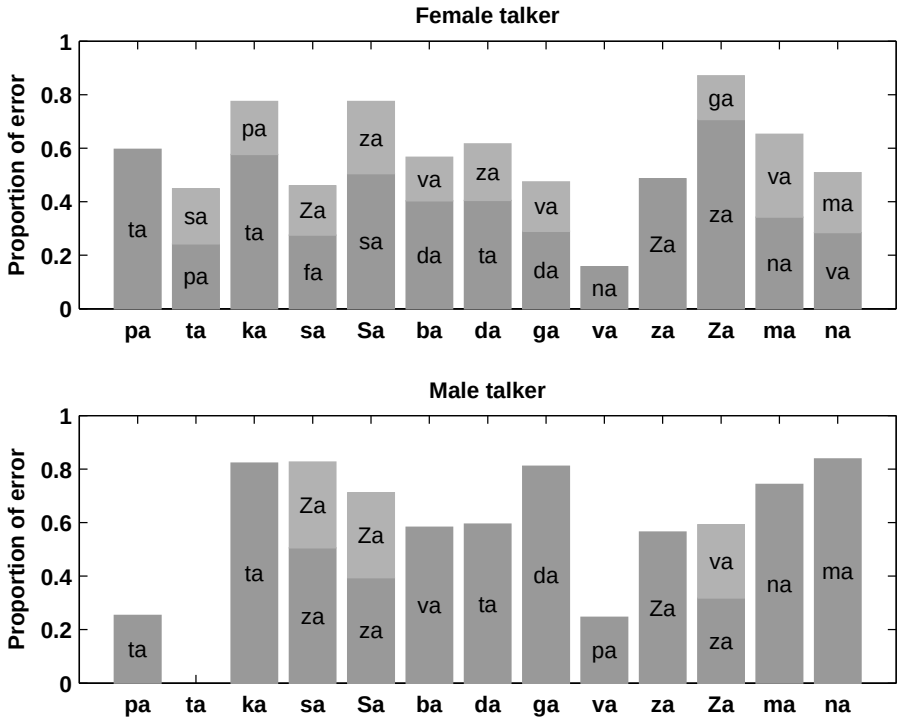


Figure 5.3: Major consonant confusions for each of the female (top panel) and male (bottom panel) utterances shown as the proportion of the total error for each utterance. For each of the presented utterances the different shades of gray indicate the proportion of, the error, that the corresponding labeled confusions constitute. Only the major confusions, those contributing more than 15% to the total error, are shown in the figure. The white space between the top of the bars and a proportion of 1 represents one or more minor confusions each contributing less than 15% to the total error.

Utterance /ka/ (female)

For the female /ka/ utterance (Fig. 5.4), listeners HI1, HI4 and HI8 consistently confuse /ka/ with /ta/, while listeners HI2 and HI3 consistently confuse /ka/ with /pa/. Thus, interestingly, the confusion with /ta/ and /pa/, found when the results were averaged across listeners, does not imply that all listeners confuse /ka/ with both /ta/ and /pa/. Instead, each listener consistently confuses /ka/ with only one other consonant. Thus, the overall confusion group shown in figure 5.3 arises because the HI listeners form two groups each making one specific confusion. Most of the errors occur at the lowest SNR, except for listener HI1 who makes almost 100% errors across all four SNRs. Interestingly, for this listener, the confusions are very consistent for all SNRs, except for the lowest SNR where they are randomly distributed indicating that, for listener HI1, the speech in this condition probably was masked by noise. Comparing the confusions shown in the left panel to the audiograms in the right panel, there is no clear difference between the pattern of the the audiograms of HI listeners that confuse /ka/ with /ta/ versus those that confuse /ka/ with /pa/. Thus, the specific pattern of confusions appears to be unrelated to the audiogram.

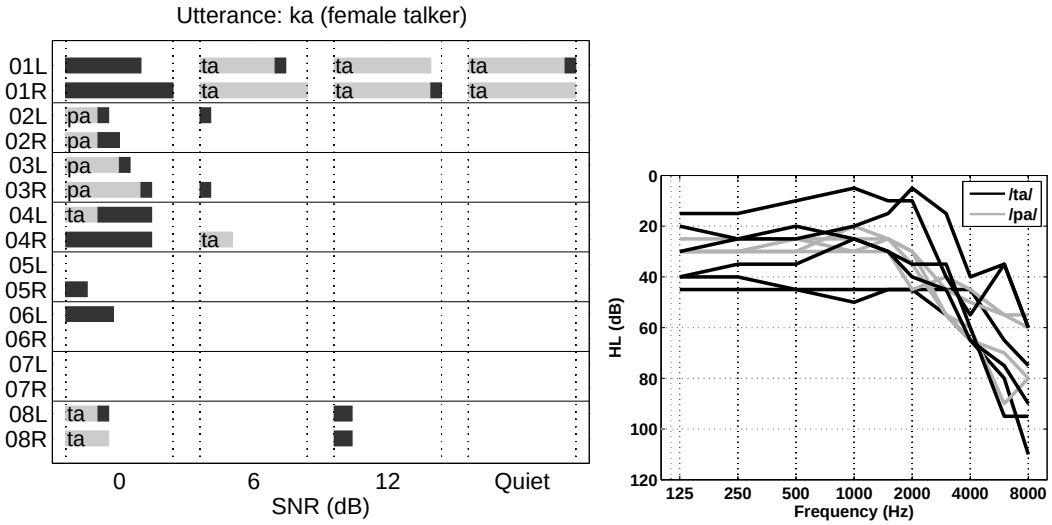


Figure 5.4: The left panel shows the the major confusions obtained with each of the HI listeners for the female /ka/ utterance. The rows show the confusions for the eight HI listeners, where the columns show the confusions for the four different SNR. In each column the left dotted line indicates 0% error while the right dotted line indicates 100% errors. For each vertical bar the different shades of gray indicate the proportion of error constituted by the confusions corresponding to the different labels, while black unlabelled bars indicates one or more minor confusions each contributing less than 30% to the total error. The right panel show the audiograms of all the HI listeners demonstrating a major confusion with /ta/ (black) and the audiograms of all the HI listeners demonstrating a major confusion with /pa/ (gray).

Utterance /ma/ (female)

The confusions averaged across listeners showed a confusion group consisting of /na/ and /va/ with the /na/ confusion being slightly larger. However, in contrast to the expectation that /ma/ is equally confused with /na/ and /va/, the individual results (Fig. 5.5) show that most listeners confuse /ma/ with /va/ and that the major /na/ confusion is dominated by listener HI1 making a large amount of errors. Again, the confusions are consistent across all the four SNRs. Thus, while noise increases the rate at which confusions occur, the pattern of confusions remain the same. Comparing the confusions shown in the left panel to the audiograms in the right panel reveals that the audiograms of the ears making /na/ confusions (listener HI1 only) show somewhat higher thresholds in the frequency range below 2 kHz than the audiograms of the ears making /va/ confusions. Thus, the confusion with /na/ instead of /va/ could be explained by /na/ having perceptually important information above 2 kHz while /va/ has perceptual important information below 2 kHz which is not audible for listener HI1 due to higher thresholds in the low frequencies.

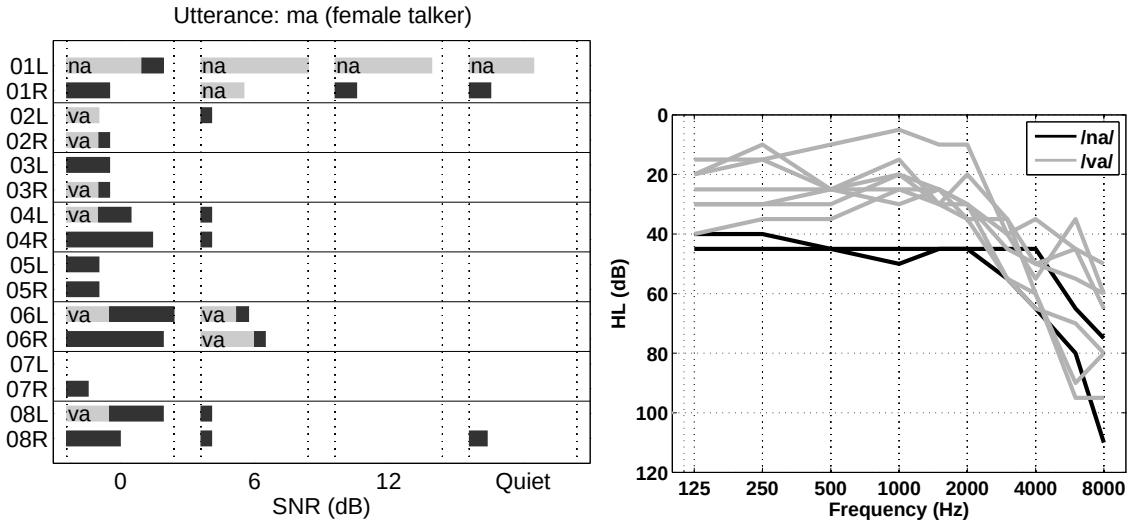


Figure 5.5: The left panel shows the the major confusions obtained with each of the HI listeners for the female /ma/ utterance. The rows show the confusions for the eight HI listeners, where the columns show the confusions for the four different SNR. In each column the left dotted line indicates 0% error while the right dotted line indicates 100% errors. For each vertical bar the different shades of gray indicate the proportion of error constituted by the confusions corresponding to the different labels, while black unlabelled bars indicates one or more minor confusions each contributing less than 30% to the total error. The right panel show the audiograms of all the HI listeners demonstrating a major confusion with /na/ (black) and the audiograms of all the HI listeners demonstrating a major confusion with /va/ (gray).

Utterance /Za/ (female)

The female /Za/ utterance (Fig. 5.6) is a very interesting example that shows that listeners HI1 and HI2 consistently confused /Za/ with /za/. However, while listener HI8 in a similarly way confused /Za/ with /za/ when presented to the right ear, /Za/ was consistently confused with /ga/ when presented to the left ear. For this utterance, all three listeners (HI1, HI2 and HI8) showed a high percentage of errors across all SNRs (including quiet). This is particular interesting as listener HI8 has a relatively mild hearing loss and highest overall percentage of correctly identified consonants when averaged across all consonants. While the percentage of error is high for three of the listeners, the confusions are, once again, consistent across all SNRs. Again, the overall groups of confusion observed in figure 5.3 does not emerge because each listener confuses the presented consonant with several other consonants. The /ga/ confusion that appeared in the overall average is due to the confusions of a single listener for stimuli presented to one particular ear. This is similar to the results from /ma/ utterance produced by the female talker. In the right panel, the left ear of listener HI8 (gray line) shows somewhat lower thresholds up to 1 kHz than all the other ears. Thus, the /ga/ confusion of the left ear of listener HI8 might be explained in terms of the audiogram. This would mean that /ga/ is triggered by energy in the low frequencies and that this energy is just audible for the left ear of listener HI8, but not audible in the right ear or in the ears of the two other listeners.

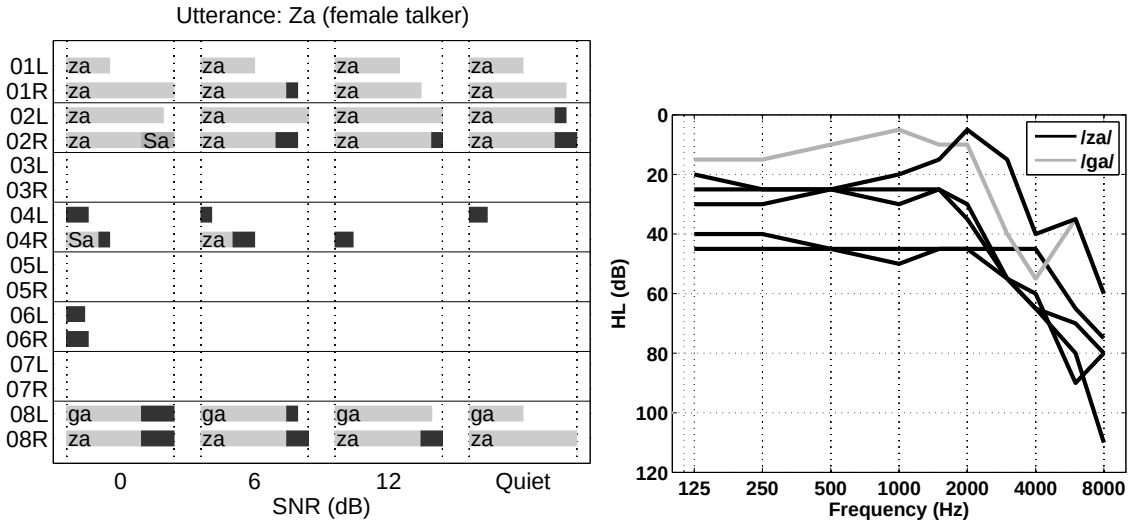


Figure 5.6: The left panel shows the the major confusions obtained with each of the HI listeners for the female /Za/ utterance. The rows show the confusions for the eight HI listeners, where the columns show the confusions for the four different SNR. In each column the left dotted line indicates 0% error while the right dotted line indicates 100% errors. For each vertical bar the different shades of gray indicate the proportion of error constituted by the confusions corresponding to the different labels, while black unlabelled bars indicates one or more minor confusions each contributing less than 30% to the total error. The right panel show the audiograms of all the HI listeners demonstrating a major confusion with /za/ (black) and the audiograms of all the HI listeners demonstrating a major confusion with /ga/ (gray).

Utterance /sa/ (male)

The final example is the /sa/ utterance produced by the male talker (Fig. 5.7). Here, listeners HI1 and HI3 consistently confused /sa/ with /za/, while listener HI2 mostly confused /sa/ with /Za/. Once again, the main confusions are consistent across the different SNRs and the overall confusion group shown in figure 5.3 arises because the listeners form two groups each making one specific confusion and not because each listener confuses the presented consonant with several other consonants. Thus, overall, the results from figures 5.4 to 5.7 demonstrate that although some HI listeners make many errors they are very consistent in their responses. The right panel reveals that the audiograms of the ears making /za/ and /Za/ confusions overlap each other. Thus, again, the specific pattern of confusions across the HI listeners seem to be unrelated to the audiograms.

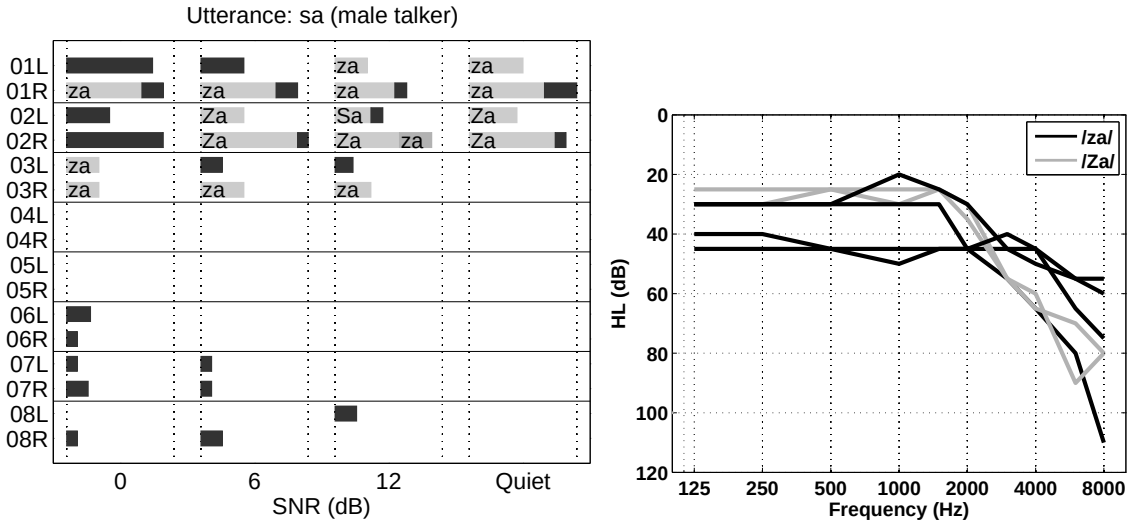


Figure 5.7: The left panel shows the the major confusions obtained with each of the HI listeners for the female /sa/ utterance. The rows show the confusions for the eight HI listeners, where the columns show the confusions for the four different SNR. In each column the left dotted line indicates 0% error while the right dotted line indicates 100% errors. For each vertical bar the different shades of gray indicate the proportion of error constituted by the confusions corresponding to the different labels, while black unlabelled bars indicates one or more minor confusions each contributing less than 30% to the total error. The right panel show the audiograms of all the HI listeners demonstrating a major confusion with /za/ (black) and the audiograms of all the HI listeners demonstrating a major confusion with /Za/ (gray).

5.4 Discussion

5.4.1 Summary of main results

The results from the present study support the observations from Phatak and Allen (2007) that, based on the average score across listeners, the consonants can be divided into roughly three groups. A high scoring group, a low scoring group and a noise-sensitive group, where the score is high in quiet conditions but drops rapidly with an increasing amount of noise. Furthermore, the results showed a large variation in the number of errors across HI listeners and indicated that some listeners are more sensitive to noise than others.

An interesting observation of the present study was that HI listeners rarely show an evenly distributed number of errors across the different consonants. Instead, they often show large errors with a few specific consonants and, even more interesting, they frequently have problems with utterances from one talker and not the other. A spectro-temporal analysis of these specific consonants, and especially the differences between the utterances, might provide valuable information about the hearing loss of the listeners.

The presented consonants were, on average, confused mainly with one or two other consonants. However, there was a large difference in the confusions obtained from the female and male talker, consistent with Phatak et al. (2008).

An important result is that for many of the presented consonants the confusion groups that emerge from data averaged across listeners are not present in the individual results. Instead, the results show

that each listener consistently confuses the presented consonant with one specific other consonant and that the confusion groups emerge because the individual listeners make different confusions.

Compared to previous studies that showed very different confusion patterns across HI listeners, the present study demonstrated that, when analyzed on an utterance-by-utterance basis, most of the HI listeners make the same confusions. However, the confusions depend strongly on the individual utterances. Thus, the reason why earlier studies with HI listeners have shown large variation is probably due to averaging across utterances. It is also likely that investigating confusions of individual NH listeners would show the same large variation of confusions across listeners, since the main source of variation seems to be the different utterances.

Another important observation is that for individual HI listeners the rate of errors varies across SNR while the pattern of errors are consistent across SNRs. Finally, there seems to be very little correlation between the consonant confusions and the audiogram of the individual HI listeners.

5.4.2 A possible explanation for different consonant confusions

Based on the results of Phatak and Allen (2007), Phatak et al. (2008), Li et al. (2010) and Kapoor and Allen (2012), the perception of consonants seems to be governed by primary cues, secondary cues and conflicting cues and the difference in the consonant confusions between different utterances of the same consonant seems to be caused by the utterances having different conflicting cues. It is also possible be that the utterances contain the same conflicting cues but that their strengths are different.

The reason why individual HI listeners in many cases show different confusions for the same utterance might be explained by most utterances having more than one conflicting cue. Impaired hearing can render certain cues inaudible due to reduced sensitivity or noise masking and different hearing losses could make different cues inaudible. Thus, if a given consonant contains two conflicting cues, one type of hearing loss might render the first conflicting cue inaudible, while another type of hearing loss might render the second conflicting cue inaudible. Even if both conflicting cues are still audible, different types of hearing loss could change the relative strength of the different conflicting cues and thereby promote different confusions. Thus, knowing exactly which spectro-temporal energy patterns promote the perception of which consonants for a given utterance can be very valuable in explaining the confusions of the HI listeners.

5.4.3 Consonant confusions and auditory functions

Consonant confusion patterns in individual HI listeners obtained from individual utterances showed very little correlation with the audiograms, which is in agreement with previous studies. This indicates that auditory functions apart from sensitivity play an important role for the recognition of consonants.

One of the known consequences of impaired hearing is cochlear dead regions (Moore and

Glasberg, 1997). For many HI listeners, the audiogram indicates reduced sensitivity at high frequencies. However, reduced sensitivity may conceal one contiguous or several smaller dead region (e.g., Moore, 2001; Halpin and Rauch, 2009a; Halpin, 2011). Halpin and Rauch (2009b) demonstrated how patients with gradually sloping high-frequency hearing losses, examined post-mortem, showed a sharp transition from a normal population of inner and outer hair cells to completely dead regions above a certain frequency. In these cases, the gradually sloping audiograms are caused by cells at lower frequencies responding to high frequency basilar membrane vibration. Li et al. (2010) identified the perceptually relevant cues for /p/, /k/ and /t/ as low-, mid-, and high-frequency bursts occurring approximately 50 ms before the onset of the vowel and /b/, /g/ and /d/ as low-, mid-, and high-frequency bursts occurring roughly at the same time as the vowel. Furthermore, they also found that the burst of energy for each consonant was relatively narrowband and varied across talkers producing the same consonant. Thus, relatively small dead regions at very specific frequencies may explain why some listeners only have problems in recognizing one specific utterance of a given consonant. Thus, knowing exactly which spectro-temporal energy segments are important for correct recognition of each utterance can, combined with consonant recognition measurements, potentially be used to precisely identify dead regions. Furthermore, assuming that many utterances contain more than one conflicting cue, dead regions corresponding to specific frequency bands could lead to very specific confusions for the individual listeners which would not be captured by the audiogram.

Frequency selectivity has been shown to be correlated with speech intelligibility in noise (e.g., Festen and Plomp, 1983; Dreschler and Plomp, 1985; Horst, 1987). Broader auditory filters are considered to lead to increased noise masking. Thus, differences in frequency selectivity could probably explain why some HI listeners were more sensitive to noise than others. It might also explain why some had problems recognizing specific consonants, such as the noise-sensitive ones.

Tests of temporal fine structure (TFS) processing, such as frequency discrimination and frequency modulation detection, have been shown to correlate with reduced recognition of speech in stationary noise (Horst, 1987; Glasberg and Moore, 1989; Noordhoek et al., 2001; Buss et al., 2004). Furthermore, Lorenzi et al. (2006) showed that HI listeners performed very poorly compared to NH listeners, in understanding so-called TFS speech. The TFS is considered to be important for the extraction of F0 information, which is important for estimating the voice-onset-time between consonant burst and vowel-onset in a CV task. Thus, differences in TFS processing could probably explain differences in the confusion patterns between individual listeners.

5.5 Summary and conclusions

Consonant confusion experiments with NH listeners have shown that different utterances of the same consonant often lead to different confusions (Phatak and Allen, 2007; Phatak et al., 2008). Thus, in order to get a better understanding of the confusions made by HI listeners, the present study investigated the consonant confusions of individual HI listeners on an utterance-by-utterance

basis, with a relatively large number of trials (5-10) for each utterance and SNR.

The main results of the study were as follows.

(1) Apart from a variation in the general performance of the listeners, some listeners were found to be much more affected by noise than others. However, none of these results were related to the audiograms.

(2) A small number of errors, averaged across consonants does not imply a small number of errors evenly distributed across all consonants, but often implies a large number of errors with a few very specific consonants. Interestingly, in several cases, the errors do not even correspond to a general problem with a specific consonant, but to a problem with a specific utterance of the consonant.

(3) The HI listeners are often consistent in their response behaviour and often confuse the presented consonants with only one other consonant, independent of the SNR.

(4) For each utterance, the HI listeners are typically divided into two groups, where the listeners in each group make the exact same confusion. The confusion groups for each utterance seem to emerge because the HI listener constitute one or two confusion groups and not because each listener makes several confusions for the same utterance.

(5) The large variation of confusions between individual listeners seems to be caused by the individual listener's problems with a specific utterance and that each of these utterances gives rise to different confusions. Thus, the variation across HI listeners is more a variation across utterances than an actual difference in the confusions between the listeners.

(6) Even at an utterance-by-utterance basis, there is very little or no correlation between the audiograms and the consonant confusions made by the individual listeners.

Overall, the present study shows that speech decoding is a very complicated process and that the perception of consonants is affected by other auditory processing deficits than just reduced audibility. An important result was that most utterances are only confused with one or two other consonants across all HI listeners and that the large variability in the confusions is due to individual listeners making errors with different utterances. This indicates that, instead of investigating the confusions made by individual listeners, investigating which specific utterances cause confusions for each listener might provide more information about the impairments of the individual listeners. This should be combined with measures of different auditory functions, since the results in the present study were unrelated to the audiometric thresholds.

6

Summary and final thoughts

The work presented in this thesis was motivated by the great difficulties of HI listeners to comprehend speech in adverse listening conditions. The overall objective was to obtain a better understanding of speech perception in different types of interferers and how this is affected by hearing loss. This was done through modeling of speech intelligibility, measurements of MR in NH and HI listeners and an investigation of consonant confusions of HI listeners.

The modeling framework established in chapter 2 revealed that, by using a psychoacoustically validated auditory preprocessing model, it was possible to predict speech intelligibility in various adverse conditions with a very simple central processing stage. A model analysis showed that the individual stages of the auditory preprocessing were essential to achieve accurate predictions throughout the tested conditions. This suggested that the internal representation generated by the auditory model might represent similar information as that available to the listeners in the experiments. The model provided the best results when the predicted speech intelligibility was solely based on the high-energy segments of the speech. This indicated a large redundancy of the speech signal and suggested that NH listeners can understand speech based on information stemming only from these high-energy segments. An advantage of the proposed model is that it can be used as a framework to study speech intelligibility of HI listeners by modifying the preprocessing of the model according to a given hearing impairment.

Focusing on one of the major problems experienced by HI listeners, chapter 3 investigated the reduced MR of HI listeners. When measuring speech intelligibility in stationary noise, HI listeners often exhibit higher SRTs. It has been suggested that this increase of the stationary-noise SRT causes the reduction of MR observed for HI listeners. In the study presented in chapter 3, using different types of processing, the stationary-noise SRTs of NH listeners were increased to the same level as found in HI listeners. In these processed conditions, NH listeners still showed a considerably larger MR than HI listeners, indicating that the stationary-noise SRT only partly accounts for the difference between NH and HI listeners. In modulated noise, audibility seemed to be an important factor for the MR of HI listeners. Interestingly, strongly reducing TFS and F0 information by noise-vocoding reduced the MR of NH listeners in the presence of a competing talker to the same level as that obtained with the HI listeners, suggesting that HI listeners might have difficulties utilizing F0 information.

Based on these results, chapter 4 investigated the importance of F0 information for MR of NH listeners in more detail. In particular, the study focused on the contribution of high-rate envelope fluctuations that are related to F0. HP filtering the overall stimuli at 500 Hz had no

effect on MR. However, LP filtering the high-rate envelope fluctuations at 30 Hz reduced MR by approximately 1.5 dB. Importantly, when both processing schemes were applied, MR was strongly reduced (approximately 6.5 dB). This suggests that F0 information obtained from high-rate envelope fluctuations are important for MR, in particular when the F0 information that is available from resolved harmonics has been reduced. Given the importance of F0 information for the large MR exhibited by NH, it is likely that HI listeners, who often exhibit a strong reduction in the MR, might have deficits in the processing of F0 information. In particular, instead of a general deficit in extracting F0 information from a single talker in quiet conditions, HI listeners may have deficits in the processing of F0 information from simultaneous sources (e.g., through an interaction of source carriers within an auditory filter) and thereby have difficulties separating the target speech from an interfering talker.

Chapter 5 investigated the effect of hearing impairment on the ability to decode speech information via consonant confusions. As expected, this chapter demonstrated a large variability in the consonant recognition scores across the HI listeners. However, an analysis of the consonant confusions on an utterance-by-utterance basis showed that each listener very consistently confused the target with only one specific consonant. Furthermore, for each utterance, listeners were clustered into one or two groups in which each listener produced the same consonant confusion. The results in this chapter suggest that the variability in the consonant confusions for a specific consonant observed for HI listeners in previous studies emerged because HI listeners only have problems with a specific utterance of a given consonant and that each utterance promotes different confusions.

The results presented in this thesis provide constraints for future models of auditory signal processing and speech perception. The finding in chapter 4, that NH listeners are able to obtain a large MR mainly based on the F0-related information represented in the high-rate envelope fluctuations indicates that the auditory system may utilize envelope frequencies up to 300 Hz. Most speech perception models only consider frequency modulations up to 32 or 64 Hz (e.g., Steeneken and Houtgast, 1980; Elhilali et al., 2003; Kim, 2005; Jørgensen and Dau, 2011). The results also have implications for the characterization of individual hearing loss. The finding that HI listeners are reasonably consistent in their confusions for a specific utterance of a given consonant, but have problems with different utterances, suggests that studying which specific utterances cause confusions in a given listener might provide more information about his/her impairment than investigating consonant confusions. Dead regions at specific frequencies may explain why some listeners only have problems with recognizing one specific utterance of a given consonant. Thus, information about which spectro-temporal energy segments are important for correct recognition of each utterance, combined with consonant recognition results, may be useful for the identification of dead regions. The large contribution of F0 information to MR, suggests that measures of F0 processing could play an important role in characterizing individual hearing loss. This could be done by traditional measures of F0 detection or discrimination. However, the ability to separate two simultaneous complex tones with different F0 might be a more critical measure in relation to speech perception. Finally, if the reduced MR of HI listeners is related to reduced deficits in the ability process F0 information, this could have important implications for hearing-aid processing.

Although it might be impossible to restore F0 processing with a hearing aid, it may be possible to process the signals in such a way that they impose smaller demands on the processing of F0 information in the impaired auditory system. This could be done by enhancing the modulation depth of the high-rate envelope fluctuations; at least the results presented here indicate that high-rate fluctuations should be preserved as much as possible. Another solution could be to modify the F0 of the sources in such a way that the difference in F0 between them becomes larger. A more elaborate approach would be to separate the target talker from the interferer directly in the hearing aid. However, even if it was possible to develop such an algorithm that works in real-life scenarios, it would be a complex and computationally demanding task that, most importantly, must be able to determine which source represents the target.

Overall, this work provides insights into the mechanisms underlying speech perception with various interferers and MR in NH and HI listeners as well as insights into the effect of hearing loss on consonant confusions. The results may have implications for future auditory models, an advanced clinical characterization of individual hearing loss as well as novel hearing-aid strategies compensating for reduced ability to understand speech in the presence of competing sounds.

Bibliography

Alcántara, J. L., Moore, B. C. J., Kühnel, V., and Launer, S. (2003). Evaluation of the noise reduction system in a commercial digital hearing aid. *Int J Audiol*, *42*(1), 34–42.

Allen, J. B. and Han, W. (2011). Sources of decoding errors of the perceptual cues, in normal and hearing impaired ears. In *3rd International Symposium on Auditory and Audiological Research*.

Amos, N. E. and Humes, L. E. (2007). Contribution of high frequencies to speech recognition in quiet and noise in listeners with varying degrees of high-frequency sensorineural hearing loss. *J Speech Lang Hear Res*, *50*(4), 819–834.

ANSI S3.5 (1997). Methods for calculation of the speech intelligibility index. American National Standards Institute, Inc.

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*(3), 549–563.

Baer, T. and Moore, B. (1993). Effects of spectral smearing on the intelligibility of sentences in noise. *J Acoust Soc Am*, *94*(3), 1229–1241.

Baer, T. and Moore, B. C. (1994). Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech. *J Acoust Soc Am*, *95*(4), 2277–2280.

Beerends, J., Hekstra, A., Rix, A., and Hollier, M. (2002). Perceptual Evaluation of Speech Quality (PESQ) the new ITU standard for end-to-end speech quality assessment: part II - psychoacoustic model. *J Audio Eng Soc*, *50*(10), 765–778.

Beerends, J. and Stemerdink, J. (1992). A perceptual audio quality measure based on a psychoacoustic sound representation. *J Audio Eng Soc*, *40*(12), 963–978.

Bernstein, J. G. and Oxenham, A. J. (2003). Pitch discrimination of diotic and dichotic tone complexes: harmonic resolvability or harmonic number? *J Acoust Soc Am*, *113*(6), 3323–3334.

Bernstein, J. G. W. and Brungart, D. S. (2011). Effects of spectral smearing and temporal fine-structure distortion on the fluctuating-masker benefit for speech at a fixed signal-to-noise ratio. *J Acoust Soc Am*, *130*(1), 473–488.

Bernstein, J. G. W. and Grant, K. W. (2009). Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *J Acoust Soc Am*, *125*(5), 3358–3372.

Bilger, R. C. and Wang, M. D. (1976). Consonant confusions in patients with sensorineural hearing loss. *J Speech Hear Res*, *19*(4), 718–748.

Bird, J. and Darwin, C. J. (1998). *Psychophysical and Physiological Advances in Hearing*, chapter Effects of a difference in fundamental frequency in separating two sentences, (pp. 263 – 269). Grantham, U.K.: Whurr Publishers.

Békésy, G. V. (1960). *Experiments in Hearing*. McGraw Hill.

Brokx, J. and Nooteboom, S. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, *10*(1), 23–36.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am*, *109*(3), 1101–1109.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *J Acoust Soc Am*, *120*(6), 4007–4018.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (2009). Multitalker speech perception with ideal time-frequency segregation: effects of voice characteristics and number of talkers. *J Acoust Soc Am*, *125*(6), 4006–4022.

Buss, E., Hall, J. W., and Grose, J. H. (2004). Spectral integration of synchronous and asynchronous cues to consonant identification. *J Acoust Soc Am*, *115*(5), 2278–2285.

Calandruccio, L., Dhar, S., and Bradlow, A. R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *J Acoust Soc Am*, *128*(2), 860–869.

Carter, G., Knapp, C., and Nuttall, A. (1973). Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing. *IEEE Transactions on Audio and Electroacoustics*, *21*(4), 337–344.

Cherry, E. (1953). Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am*, *25*, 975–979.

Ching, T. Y., 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*(2), 1128–1140.

Christiansen, C. and Dau, T. (2012). Relationship between masking release in fluctuating maskers and speech reception thresholds in stationary noise. J Acoust Soc Am (Accepted).

Christiansen, C., MacDonald, E. N., and Dau, T. (2012). Contribution of high-rate envelope fluctuations to release from speech-on-speech masking. J Acoust Soc Am (Submitted).

Christiansen, C., Pedersen, M. S., and Dau, T. (2010). Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Comm*, *52*(7-8), 678–692.

Christiansen, C., Trevino, A., Allen, J. B., and Dau, T. (2012). Analyzing the variation of consonant confusions in hearing-impaired listeners. J Speech Lang Hear R (Submitted).

Cohen, M. A., Grossberg, S., and Wyse, L. L. (1995). A spectral network model of pitch perception. *J Acoust Soc Am*, *98*(2 Pt 1), 862–879.

Darwin, C. J. (1997). Auditory grouping. *Trends in Cognitive Sciences*, *1*(9), 327–333.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997a). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *J Acoust Soc Am*, *102*(5 Pt 1), 2892–2905.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997b). Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *J Acoust Soc Am*, *102*(5 Pt 1), 2906–2919.

Dau, T., Püschel, D., and Kohlrausch, A. (1996a). A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *J Acoust Soc Am*, *99*(6), 3615–3622.

Dau, T., Püschel, D., and Kohlrausch, A. (1996b). A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements. *J Acoust Soc Am*, *99*(6), 3623–3631.

de Cheveigné, A. (1998). Cancellation model of pitch perception. *J Acoust Soc Am*, *103*(3), 1261–1271.

Derleth, R. P. and Dau, T. (2000). On the role of envelope fluctuation processing in spectral masking. *J Acoust Soc Am*, *108*(1), 285–296.

Desloge, J. G., Reed, C. M., Braida, L. D., Perez, Z. D., and Delhorne, L. A. (2010). Speech reception by listeners with real and simulated hearing impairment: effects of continuous and interrupted noise. *J Acoust Soc Am*, *128*(1), 342–359.

Dreschler, W. A. and Plomp, R. (1985). Relations between psychophysical data and speech perception for hearing-impaired subjects. ii. *J Acoust Soc Am*, *78*(4), 1261–1270.

Drullman, R., Festen, J. M., and Plomp, R. (1994). Effect of reducing slow temporal modulations on speech reception. *J Acoust Soc Am*, *95*(5 Pt 1), 2670–2680.

Dubno, J. R., Dirks, D. D., and Langhofer, L. R. (1982). Evaluation of hearing-impaired listeners using a nonsense-syllable test. ii. syllable recognition and consonant confusion patterns. *J Speech Hear Res*, *25*(1), 141–148.

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*(4 Pt 1), 2084–2094.

Duquesnoy, A. J. and Plomp, R. (1983). The effect of a hearing aid on the speech-reception threshold of hearing-impaired listeners in quiet and in noise. *J Acoust Soc Am*, *73*(6), 2166–2173.

Elhilali, M., Chi, T., and Shamma, S. A. (2003). A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Comm*, *41*(2-3), 331–348.

Ewert, S. D. and Dau, T. (2000). Characterizing frequency selectivity for envelope fluctuations. *J Acoust Soc Am*, *108*(3 Pt 1), 1181–1196.

Festen, J. M. and Plomp, R. (1983). Relations between auditory functions in impaired hearing. *J Acoust Soc Am*, *73*(2), 652–662.

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*(4), 1725–1736.

Füllgrabe, C., Berthommier, F., and Lorenzi, C. (2006). Masking release for consonant features in temporally fluctuating background noise. *Hear Res*, *211*(1-2), 74–84.

Fousek, P., Svojanovsky, P., Grezl, F., and Hermansky, H. (2004). New Nonsense Syllables Database - Analyses and Preliminary ASR Experiments. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*. IDIAP-RR 2004-29.

French, N. and Steinberg, J. (1947). Factors governing the intelligibility of speech sounds. *J Acoust Soc Am*, *19*, 90–119.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *J Acoust Soc Am*, *106*(6), 3578–3588.

George, E. L. J., Festen, J. M., and Houtgast, T. (2006). Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners. *J Acoust Soc Am*, *120*(4), 2295–2311.

Glasberg, B. R. and Moore, B. C. (1989). Psychoacoustic abilities of subjects with unilateral and bilateral cochlear hearing impairments and their relationship to the ability to understand speech. *Scand Audiol Suppl*, *32*, 1–25.

Glasberg, B. R. and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear Res*, *47*(1-2), 103–138.

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

Goldsworthy, R. L. and Greenberg, J. E. (2004). Analysis of speech-based Speech Transmission Index methods with implications for nonlinear operations. *J Acoust Soc Am*, *116*(6), 3679–3689.

Gordon-Salant, S. (1985). Phoneme feature perception in noise by normal-hearing and hearing-impaired subjects. *J Speech Hear Res*, *28*(1), 87–95.

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

Hagerman, B. (1982a). Measurement of speech reception threshold. a comparison between two methods. *Scand Audiol*, *11*(3), 191–193.

Hagerman, B. (1982b). Sentences for testing speech intelligibility in noise. *Scand Audiol*, *11*(2), 79–87.

Hagerman, B. (1984a). Clinical measurements of speech reception threshold in noise. *Scand Audiol*, *13*(1), 57–63.

Hagerman, B. (1984b). Some aspects of methodology in speech audiometry. *Scand Audiol Suppl*, *21*, 1–25.

Hagerman, B. and Kinnefors, C. (1995). Efficient adaptive methods for measuring speech reception threshold in quiet and in noise. *Scand Audiol*, *24*(1), 71–77.

Halpin, C. (2011). Re-focusing on the clinical targets. In *3rd International Symposium on Auditory and Audiological Research*.

Halpin, C. and Rauch, S. D. (2009a). Clinical implications of a damaged cochlea: pure tone thresholds vs information-carrying capacity. *Otolaryngol Head Neck Surg*, *140*(4), 473–476.

Halpin, C. and Rauch, S. D. (2009b). Hearing aids and cochlear damage: the case against fitting the pure tone audiogram. *Otolaryngol Head Neck Surg*, *140*(5), 629–632.

Han, W. (2011). *Methods For Robust characterization of consonant perception in hearing-impaired listeners*. PhD thesis, University of Illinois at Urbana-Champaign.

Heinz, M. G. (2012). Physiological correlates of perceptual deficits following sensorineural hearing loss. *Acoustics Today*, *8*(2), 34–40.

Hällgren, M., Larsby, B., Lyxell, B., and Arlinger, S. (2005). Speech understanding in quiet and noise, with and without hearing aids. *Int J Audiol*, *44*(10), 574–583.

Hogan, C. A. and Turner, C. W. (1998). High-frequency audibility: benefits for hearing-impaired listeners. *J Acoust Soc Am*, *104*(1), 432–441.

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (2010). Development and analysis of an International Speech Test Signal (ISTS). *Int J Audiol*, *49*(12), 891–903.

Holube, I. and Kollmeier, B. (1996). Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *J Acoust Soc Am*, *100*(3), 1703–1716.

Hopkins, K., Moore, B. C. J., and Stone, M. A. (2008). Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. *J Acoust Soc Am*, *123*(2), 1140–1153.

Horst, J. W. (1987). Frequency discrimination of complex signals, frequency selectivity, and speech perception in hearing-impaired subjects. *J Acoust Soc Am*, *82*(3), 874–885.

Hou, Z. and Pavlovic, C. V. (1994). Effects of temporal smearing on temporal resolution, frequency selectivity, and speech intelligibility. *J Acoust Soc Am*, *96*(3), 1325–1340.

Houtgast, T. and Festen, J. M. (2008). On the auditory and cognitive functions that may explain an individual's elevation of the speech reception threshold in noise. *Int J Audiol*, *47*(6), 287–295.

Houtsma, A. J. M. and Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics. *J Acoust Soc Am*, *87*(1), 304–310.

Hu, G. and Wang, D. (2004). Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks*, *15*(5), 1135–1150.

Huber, R. and Kollmeier, B. (2006). PEMO-Q - A new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(6), 1902–1911.

Jepsen, M. L., Ewert, S. D., and Dau, T. (2008). A computational model of human auditory signal processing and perception. *J Acoust Soc Am*, *124*(1), 422–438.

Jørgensen, S. and Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *J Acoust Soc Am*, *130*(3), 1475–1487.

Kapoor, A. and Allen, J. B. (2012). Perceptual effects of plosive feature modification. *J Acoust Soc Am*, *131*(1), 478–491.

Karjalainen, M. (1985). A new auditory model for the evaluation of sound quality of audio systems. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 10, (pp. 608–611).

Kates, J. M. and Arehart, K. H. (2005). Coherence and the speech intelligibility index. *J Acoust Soc Am*, *117*(4 Pt 1), 2224–2237.

Kidd, G., Mason, C. R., Rohtla, T. L., and Deliwala, P. S. (1998). Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. *J Acoust Soc Am*, *104*(1), 422–431.

Kim, D.-S. (2005). Anique: An auditory model for single-ended speech quality estimation. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 821 – 831.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. (2009). Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *J Acoust Soc Am*, *126*(3), 1415–1426.

Koch, R. (1992). *Gehoergerechte Schallanalyse zur Vorhersage und Verbesserung der Sprachverstaendlichkeit*. PhD thesis, Georg August Universität, Physics, Göttingen, Germany.

Li, F., Menon, A., and Allen, J. B. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *J Acoust Soc Am*, *127*(4), 2599–2610.

Licklider, J. C. R. (1951). A duplex theory of pitch perception. *Experientia*, *7*(4), 128–134.

Lin, F., Niparko, J., and Ferrucci, L. (2011). Hearing loss prevalence in the united states. *Archives of Internal Medicine*, *171*(20), 1851–1853.

Lindgreen, T. S. (2009). Measures of temporal fine structure processing in human hearing. Master's thesis, Technical University of Denmark.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proc Natl Acad Sci*, *103*(49), 18866–18869.

Lorenzi, C., Husson, M., Ardoint, M., and Debruille, X. (2006). Speech masking release in listeners with flat hearing loss: effects of masker fluctuation rate on identification scores and phonetic feature reception. *Int J Audiol*, *45*(9), 487–495.

Ludvigsen, C., Elberling, C., Keidser, G., and Poulsen, T. (1990). Prediction of intelligibility of non-linearly processed speech. *Acta Otolaryngol Suppl*, *469*, 190–195.

McClelland, J. L. and Elman, J. L. (1986). The trace model of speech perception. *Cogn Psychol*, *18*(1), 1–86.

Meddis, R. and Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification. *J Acoust Soc Am*, *89*(6), 2865–2882.

Metselaar, M., Maat, B., Krijnen, P., Verschuure, H., Dreschler, W., and Feenstra, L. (2008). Comparison of speech intelligibility in quiet and in noise after hearing aid fitting according to a purely prescriptive and a comparative fitting procedure. *Eur Arch Otorhinolaryngol*, *265*(9), 1113–1120.

Micheyl, C. and Oxenham, A. J. (2007). Across-frequency pitch discrimination interference between complex tones containing resolved harmonics. *J Acoust Soc Am*, *121*(3), 1621–1631.

Miller, G. A. and Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *J Acoust Soc Am*, *22*(2), 167–173.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *J Acoust Soc Am*, *27*(2), 338–352.

Moore, B. C. J. (2001). Dead regions in the cochlea: Diagnosis, perceptual consequences, and implications for the fitting of hearing aids. *Trends in Amplification*, *5*(1), 1–34.

Moore, B. C. J. and Glasberg, B. R. (1997). A model of loudness perception applied to cochlear hearing loss. *Auditory Neuroscience*, *3*(3), 289–311.

Moore, B. C. J., Peters, R. W., and Stone, M. A. (1999). Benefits of linear amplification and multichannel compression for speech comprehension in backgrounds with spectral and temporal dips. *J Acoust Soc Am*, *105*(1), 400–411.

Nelson, P. B. and Jin, S.-H. (2004). Factors affecting speech understanding in gated interference: cochlear implant users and normal-hearing listeners. *J Acoust Soc Am*, *115*(5 Pt 1), 2286–2294.

Nguyen, D. P., Wilson, M. A., Brown, E. N., and Barbieri, R. (2009). Measuring instantaneous frequency of local field potential oscillations using the kalman smoother. *J Neurosci Methods*, *184*(2), 365–374.

Nielsen, J. B. and Dau, T. (2009). Development of a danish speech intelligibility test. *Int J Audiol*, *48*(10), 729–741.

Nielsen, L. B. (1993). *Modeling sound quality for hearing-impaired listeners*. PhD thesis, Technical University of Denmark and Oticon A/S.

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*(2), 1085–1099.

Noordhoek, I. M., Houtgast, T., and Festen, J. M. (2001). Relations between intelligibility of narrow-band speech and auditory functions, both in the 1-khz frequency region. *J Acoust Soc Am*, *109*(3), 1197–1212.

Oxenham, A. J. and Simonson, A. M. (2009). Masking release for low- and high-pass-filtered speech in the presence of noise and single-talker interference. *J Acoust Soc Am*, *125*(1), 457–468.

Palmer, A. and Russell, I. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hear Res*, *24*(1), 1–15.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1987). An efficient auditory filterbank based on the gammatone function. In *Proc. Meeting of the IOC Speech Group on Auditory Modelling at RSRE, December 14-15*.

Payton, K. L. and Braida, L. D. (1999). A method to determine the speech transmission index from speech waveforms. *J Acoust Soc Am*, *106*(6), 3637–3648.

Payton, K. L., Braida, L. D., Chen, S., Rosengard, P., and Goldsworthy, R. (2002). Computing the STI using speech as a probe stimulus. In *Past, Present and Future of the Speech Transmission Index*, (pp. 125–138.)., Soesterberg, The Netherlands. TNO Human Factors.

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*(1), 577–587.

Phatak, S. A. and Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *J Acoust Soc Am*, *121*(4), 2312–2326.

Phatak, S. A., Lovitt, A., and Allen, J. B. (2008). Consonant confusions in white noise. *J Acoust Soc Am*, *124*(2), 1220–1233.

Phatak, S. A., Yoon, Y.-S., Gooler, D. M., and Allen, J. B. (2009). Consonant recognition loss in hearing impaired listeners. *J Acoust Soc Am*, *126*(5), 2683–2694.

Pickles, J. O. (1988). *An Introduction to the Physiology of Hearing*. Academic Press.

Plack, C. J. (2005). *The Sense Of Hearing*. Routledge.

Plomp, R. (1967). Pitch of complex tones. *J Acoust Soc Am*, *41*(6), 1526–1533.

Plomp, R. (1978). Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J Acoust Soc Am*, *63*(2), 533–549.

Qin, M. K. and Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *J Acoust Soc Am*, *114*(1), 446–454.

Rankovic, C. M. (1991). An application of the articulation index to hearing aid fitting. *J Speech Hear Res*, *34*(2), 391–402.

Régnier, M. S. and Allen, J. B. (2008). A method to identify noise-robust perceptual features: application for consonant /t/. *J Acoust Soc Am*, *123*(5), 2801–2814.

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*(4 Pt 1), 2181–2192.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2006). Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *J Acoust Soc Am*, *120*(6), 3988–3997.

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. (1997). Basilar-membrane responses to tones at the base of the chinchilla cochlea. *J Acoust Soc Am*, *101*(4), 2151–2163.

Santurette, S. and Dau, T. (2011). The role of temporal fine structure information for the low pitch of high-frequency complex tones. *J Acoust Soc Am*, *129*(1), 282.

Santurette, S. and Dau, T. (2012). Relating binaural pitch perception to the individual listener's auditory profile. *J Acoust Soc Am*, *131*(4), 2968–2986.

Shackleton, T. M. and Carlyon, R. P. (1994). The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. *J Acoust Soc Am*, *95*(6), 3529–3540.

Shanks, J. E., Wilson, R. H., Larson, V., and Williams, D. (2002). Speech recognition performance of patients with sensorineural hearing loss under unaided and aided conditions using linear and compression hearing aids. *Ear Hear*, *23*(4), 280–290.

Shield, B. (2006). *Evaluation of the Social and Economic Costs of Hearing Impairment: a report for Hear-it AISBC*. Hear-it.org.

Smith, R. L. (1977). Short-term adaptation in single auditory nerve fibers: some poststimulatory effects. *J Neurophysiol*, *40*(5), 1098–1111.

Steeneken, H. J. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *J Acoust Soc Am*, *67*(1), 318–326.

Stone, M. A., Füllgrabe, C., and Moore, B. C. J. (2008). Benefit of high-rate envelope cues in vocoder processing: effect of number of channels and spectral region. *J Acoust Soc Am*, *124*(4), 2272–2282.

Strelcyk, O. and Dau, T. (2009). Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *J Acoust Soc Am*, *125*(5), 3328–3345.

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*(3 Pt 1), 1307–1314.

Terhardt, E. (1974). Pitch, consonance, and harmony. *J Acoust Soc Am*, *55*(5), 1061–1069.

Thiede, T., Treurniet, W., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J., Colomes, C., Keyhl, M., Stoll, G., Brandenburg, K., and Feiten, B. (2000). Peaq - the itu standard for objective measurement of perceived audio quality. *J. Audio Eng. Soc.*, *48*(1/2), 3–29.

Verhey, J. L., Dau, T., and Kollmeier, B. (1999). Within-channel cues in comodulation masking release (cmr): experiments and model predictions using a modulation-filterbank model. *J Acoust Soc Am*, *106*(5), 2733–2745.

von Helmholtz, H. (1912). *On the sensations of tone as a physiological basis for the theory of music*. Longmans, Green.

Wagener, K., Josvassen, J. L., and Ardenkjaer, R. (2003). Design, optimization and evaluation of a danish sentence test in noise. *Int J Audiol*, *42*(1), 10–17.

Walden, B. E. and Montgomery, A. A. (1975). Dimensions of consonant perception in normal and hearing-impaired listeners. *J Speech Hear Res*, *18*(3), 444–455.

Wang, D. (2004). *Speech Separation by Humans and Machines*, chapter On ideal binary mask as the computational goal of auditory scene analysis, (pp. 181–197). Springer US.

Westerman, L. A. and Smith, R. L. (1984). Rapid and short-term adaptation in auditory nerve responses. *Hear Res*, *15*(3), 249–260.

WHO (2012). World health assembly. *World Health Organization*.

Wightman, F. L. (1973). The pattern-transformation model of pitch. *J Acoust Soc Am*, *54*(2), 407–416.

Xu, L. and Zheng, Y. (2007). Spectral and temporal cues for phoneme recognition in noise. *J Acoust Soc Am*, *122*(3), 1758.

Zilany, M. S. A. and Bruce, I. C. (2006). Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *J Acoust Soc Am*, *120*(3), 1446–1466.

Zilany, M. S. A. and Bruce, I. C. (2007). Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery. In *Proc. 3rd International IEEE/EMBS Conference on Neural Engineering, Vols 1 and 2*, (pp. 481 – 485).

Contributions to Hearing Research

Vol. 1: *Gilles Pigasse*, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.

Vol. 2: *Olaf Strelcyk*, Peripheral auditory processing and speech reception in impaired hearing, 2009.

Vol. 3: *Eric R. Thompson*, Characterizing binaural processing of amplitude-modulated sounds, 2009.

Vol. 4: *Tobias Piechowiak*, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.

Vol. 5: *Jens Bo Nielsen*, Assessment of speech intelligibility in background noise and reverberation, 2009.

Vol. 6: *Helen Connor*, Hearing aid amplification at soft input levels, 2010.

Vol. 7: *Morten Løve Jepsen*, Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.

Vol. 8: *Sarah Verhulst*, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.

Vol. 9: *Sylvain Favrot*, A loudspeaker-based room auralization system for auditory research, 2010.

Vol. 10: *Sébastien Santurette*, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.

Vol. 11: *Iris Arweiler*, Processing of spatial sounds in the impaired auditory system, 2011.

Vol. 12: *Filip Munch Rønne*, Modeling auditory evoked potentials to complex stimuli, 2012.

The end.

To be continued...

The hearing system is very important for development of speech and enables us to communicate with other people in a time where this is more important than ever. Speech communication often takes place in the presence of concurrent talkers, background noise or in a reverberant environment. In such adverse listening conditions, speech intelligibility generally remains high for normal-hearing listeners, whereas hearing-impaired listeners often experience major difficulties. Speech perception is a complex process involving the ability to hear the speech, selectively focus on a specific person talking in the presence of interfering sound sources and the ability to extract meaning from the perceived speech. This volume of “Contributions to hearing research” deals with the different processes of speech perception. The underlying mechanisms enabling normal hearing listeners to understand speech in the presence of interfering sounds and why this ability is more or less reduced in hearing-impaired listeners were studied. Especially the use pitch information in focusing on one out of two talkers was investigated. Finally, the effect of hearing loss on the ability to extract meaning from the perceived speech was also studied. This work provides insights into the auditory mechanisms underlying speech perception in the presence of interfering sound sources, and how this and the decoding of speech are affected by hearing loss. The work presented in this volume, may have implications for future auditory models, clinical characterization of individual hearing loss as well as hearing-aid strategies compensating for reduced ability to understand speech with interfering sounds.

DTU Electrical Engineering
Department of Electrical Engineering

Ørsteds Plads
Building 348
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 38 00
Fax: (+45) 45 93 16 34
www.elektro.dtu.dk

ISBN 978-87-92465-xx-x