Assessing speech intelligibility in hearing impaired listeners

Scheidiger, Christoph

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

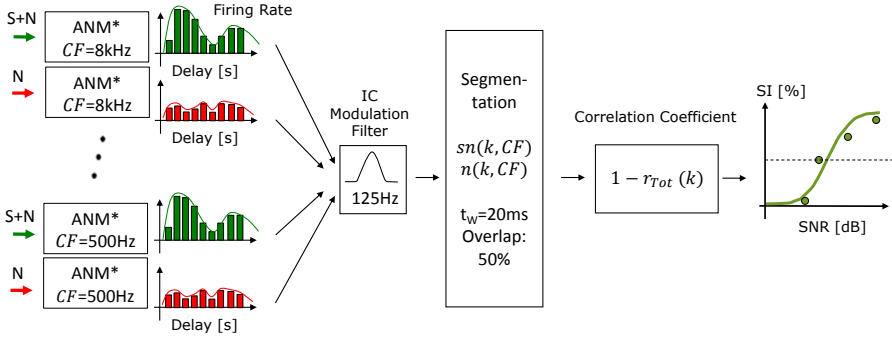
Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 31

Christoph Scheidiger

Assessing speech intelligibility in hearing impaired listeners



Hearing Systems
Department of Electrical Engineering

Assessing speech intelligibility in hearing impaired listeners

PhD thesis by
Christoph Scheidiger

Preliminary version: October 23, 2017

Technical University of Denmark

2017

This PhD dissertation is the result of a research project carried out at the Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark.

The project was partly financed by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement No. PITN-GA-2012-317521 (2/3) and by the Technical University of Denmark (1/3).

Supervisors

Prof. Torsten Dau
Dr. Johannes Zaar
Hearing Systems Group
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Abstract

Quantitatively assessing the speech intelligibility deficits observed in hearing-impaired (HI) listeners is a basic component for a better understanding of these deficits and a crucial component for the development of successful compensation strategies. This dissertation describes two main streams of work aiming at a better quantitative understanding: (i) Chapter 2 focused on describing a new analysis framework based on a confusion entropy and a distance metric to analyze consonant-vowel (CV) perception in HI listeners across different listening conditions; (ii) Chapters 3, 4, and 5 focused on developing a speech intelligibility (SI) model to account for observed deficits in HI listeners. In Chapter 2, HI listeners were provided with two different amplification schemes to help them recognize CVs. In the first experiment, a frequency- independent amplification (flat-gain) was provided. In the second experiment, a frequency-dependent prescriptive gain was provided. An entropy measure and an angular distance measure were proposed to assess the highly individual effects of the frequency-dependent gain on the consonant confusions in the HI listeners. These measures along with a standard analysis of the recognition scores suggested that, while the average recognition error score obtained with the frequency-dependent amplification was lower than that obtained with the flat-gain, the main confusions made by the listeners on a token basis remained the same in a majority of the cases. Chapter 3 describes the introduction of the HI deficits of reduced audibility and decreased frequency selectivity into a speech-intelligibility model for normal-hearing (NH) listeners. The NH model is based on a signal-to-noise ratio measure in the envelope domain (SNR_{env}), as presented in the framework of the speech-based envelope power spectrum model (sEPSM, Jørgensen and Dau, 2011; Jørgensen et al., 2013). The predictions of the model were compared to data in three different noise maskers. While the model was able to account for the relative difference of the HI listeners performance in these different noise interferers; it faild to account for the absolute performance in the noise interferers. Chapter 4 replaced the linear peripheral model, i.e. the gammatone filterbank, by a nonlinear auditory nerve model. The SI predictions showed good agreement with human data when the model operated at an overall presentation level (OAL) of 50 dB sound pressure level (SPL) and with only medium-spontaneous-rate fibers. However, when all fiber types and a realistic OAL of 65 dB SPL were considered, the model overestimated SI in conditions with modulated noise interferers. In Chapter 5,

the front-end processing of an auditory-nerve (AN) model was combined with a correlation-based back end inspired by the vowel-coding hypothesis of stable rate patterns in the inferior colliculus. The proposed model assesses the correlation between the noisy speech and the noise alone, as represented by the AN model's bandpass-filtered instantaneous firing rates, assuming an inverse relationship with SI. The NH listeners' SI data were accurately predicted for all noise types, additionally demonstrating reasonable changes across presentation levels. Furthermore, the SI for 13 HI subjects was predicted by adjusting the front end parameters specifying the inner and outer hair-cell loss based on the audiogram of the listeners. The predictions showed good agreement with the measured data for four out of the thirteen subjects and reasonable agreement for a total of eight subjects. The work provides a foundation for quantitatively modeling individual effects of inner and outer hair-cell loss on speech intelligibility.

Resumé

Kvantitativ evaluering af problemer med taleforståelighed observeret i hørehæmmede personer er vigtig for at kunne opnå en bedre forståelse af de underliggende mekanismer og en væsentlig komponent for udviklingen af succesfulde kompenseringsstrategier. Arbejdet der beskrives i denne afhandling kan opdeles i to overordnede retninger, der har til formål at skabe en øget kvantitativ forståelse: (i) Kapitel 2 fokuserer på beskrivelsen af en ny analysestruktur baseret på et mål for "forvekslings entropi" og et distancemål, der bruges til analyse af resultater fra lytteforsøg med konsonant-vokal kombinationer under forskellige forsøgsbetingelser; (ii) Kapitel 3, 4, og 5 fokuserer på udviklingen af en taleforståelsesmodel, der kan forudsige taleforståelse i hørehæmmede lyttere. I kapitel 2 bliver hørehæmmede lyttere testet med to typer forstærkning, der har til formål at hjælpe dem med at genkende konsonant-vokal kombinationer. I det første eksperiment testes en frekvensuafhængig forstærkning. I det andet eksperiment bruges frekvensafhængig forstærkning (NAL-R). Et entropimål og et vinkelafstandsmål til at bedømme den meget individuelle påvirkning af forstærkning på konsonant-vokal forvekslinger i disse lyttere bliver foreslået. Disse mål antyder, at de vigtigste forvekslinger mellem konsonant-vokal kombinationer forbliver de samme, selvom der i gennemsnit forekommer færre fejl ved brug af NAL-R end ved frekvensuafhængig forstærkning. Kapitel 3 inkorporerer høretab, dvs. reduceret hørbarhed og en formindsket frekvensselektivitet, i en eksisterende taleforståelsesmodel for normalt-hørende. Denne model er baseret på signal-til-støj forhold i "envelope"-domænet (SNRenv), som blev præsenteret i den tale-baserede "envelope power spectrum" model (sEPSM, Jørgensen and Dau, 2011; Jørgensen et al., 2013). Modellens prædiktioner bliver sammelignet med data fra forsøgsbetingelser med tre forskellige former for støjbaggrund. Selvom modellen er i stand til at redegøre for forskellene i de hørehæmmedes lytteres data mellem de forskellige forsøgsbetingelser kan den ikke forudsige de korrekte værdier for taleforståelighed. Kapitel 4 omhandler en mere realistisk model af cochlear processering, der erstatter sEPSM's lineære processering med en ikke-lineær model af hørenerven. Taleforståeligheds-forudsigelser er i overensstemmelse med data ved lavt til medium lydtryksniveau og når kun medium-spontane- hørenerve fibre tages i betragtning. Dog overvurderes taleforståelighed for medium til høje lydtryksniveauer. I kapitel 5, bliver den perifere processering af den ikke-lineære hørenerve model kombineret med en korrelationsbaseret "back end". Den foreslåede model, evaluerer korrelationen mellem den støjfulde tale og talen alene ved output af hørenerve-processeseringen

og et efterfølgende modulationsfilter (ved 125 Hz). Der kan redegøres for de hørehæmmede lytteres taleforståeligheds data for alle støjtyper der er taget i betragtning og for alle de forskellige lydtryksniveauer. For de hørehæmmede lyttere, kan taleforståelighedsdata blive redegjort for ved at justere de parametre i hørenervemodellen, der specificere tabet af indre og ydre hårceller, estimeret i henhold til de enkelte lytteres hørekurve. Prediktionerne for de hørehæmmede lyttere er i god overensstemmelse med de målte data for fire ud af 13 lyttere og er i rimelig overensstemmelse for otte af disse lyttere. Generelt set giver dette arbejde et grundlag for en bedre kvantitativ forståelse af konsekvenserne af tab af indre og ydre hårceller hos individuelle lyttere og især de perceptuelle konsekvenser af sådanne tab i forbindelse med taleforståelighed.

Acknowledgments

I am thankful for Professor Torsten Dau's guidance as a mentor and advisor during my time as a PhD student at the Tehnical University of Denmark (DTU). I have learned a lot during the 3+ years and was blessed to get to work with someone with as much experience as Professer Dau. My professional and personal life will always be influenced by what I learned during our many meetings. I had the remarkable opportunity to work with smart, nice, and helpful colleagues at the Hearing Systems Group (HEA) at DTU. I am especially thankful to Dr. Johannes Zaar, who has been an invaluable help with his insights on technical as well as tactical issues during my PhD. This thesis would not be what it is without him. Furthermore, the speech intelligibility modeling group, consisting of Johannes, Alex, and Helia, was an ideal environment to get feedback and ideas. It was a great privilege to work with such wonderful and bright people. Also invaluable were the discussions with my office mates in office 111, they provided me with a broader perspective on auditory science and on my work. I will always cherish the memories of these discussions about work, politics, and life. Even though the distance made tangible support more difficult, the biggest emotional support came from my family. Knowing of their unconditional love and support helped me to manage difficult times. Most of all, I am very grateful to my wife Erin Olson for believing in me on days I didn't believe in myself.

Related publications

Journal papers

- Relaño-Iborra, H., May, T., Zaar, J., Scheidiger, C., and Dau, T. (**2016**). "Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain," J. Acoust. Soc. Am. **140**, 2670–2679
- Scheidiger, C., Allen, J. B., and Dau, T. (**2017**). "Assessing the efficacy of hearing-aid amplification using a phoneme test," J. Acoust. Soc. Am. **141**, 1739–1748

Conference papers

- Scheidiger, C. and Allen, J. B. (**2013**). "Effects of NALR on consonant-vowel perception," Proceedings of the ISAAR **4**, 373-380
- Scheidiger, C., Jørgensen, S., and Dau, T. (**2014**). "Modeling Speech Intelligibility in Hearing Impaired Listeners," Proceedings of Forum Acusticum, 243-250
- Scheidiger, C., Jørgensen, S., and Dau, T. (**2015**). "Modeling speech intelligibility in hearing impaired listeners," Proceedings of DAGA 2015, 205–210
- Scheidiger, C. and Dau, T. (**2016**). "Modeling speech intelligibility based on neural envelopes derived from auditory spike trains," Proceedings of DAGA, 152–156

Contents

1

General introduction

Speech communication allows humans to transmit complex information through acoustic waves. The complexity and effectiveness of this acoustic communication is one of the main distinguishing factors that sets humans apart from other species. Impaired hearing negatively affects speech communication and can thus have a severe impact on a person's quality of life and the chances of success in life. Understanding how hearing impairment affects our speech understanding is closely tied to our understanding of the physiology of the hearing system. A better understanding of the impaired system through effective hearing tests and models of the auditory system may help develop effective compensation strategies in hearing devices.

In general terms, the human auditory system is composed of three stages; (i) the auditory periphery, (ii) the auditory midbrain, and (iii) the auditory cortex. The auditory periphery can be further subdivided into the outer, middle, and inner ear. The outer ear captures and filters the acoustic stimuli which makes the tympanic membrane vibrate. In the middle ear, a chain of small bones transmits these vibrations to the inner ear. In the inner ear, the vibrations travel as a wave through a fluid filled duct called cochlea. Its mechanical properties allow the cochlea to function as a frequency analyzer, as changes in the stiffness along the basilar membrane in the cochlea cause the resonance frequency of the basilar membrane to change, making each location along the basilar membrane most sensitive to a certain frequency. Hair cells along the cochlea pick up these

membrane resonances and transform them into tonotopically organized neural discharge patterns. These patterns encode sound information via the timing of spikes (Pickles, 2008). This elicited encoding is robust against background noise and highly redundant. Through the auditory nerve, a neural message is conveyed to the cochlear nucleus located in the brainstem (Young and Oertel, 2003). The diversity of cell types and neural circuitry located in the cochlear nucleus in the auditory brainstem produces different spectro-temporal representations and enhances different aspects of sound information. The tonotopically organized auditory cortex assembles the information from all the auditory features of the sound into auditory objects that have perceptual relevance to the listener (Pickles, 2008). Acute or chronic degeneration of the auditory system, particularly in the cochlea, results in degraded representations along the auditory pathway and in measurable performance gaps of hearing-impaired (HI) listeners compared to normal-hearing (NH) listeners on behavioral tests.

The auditory system and its functionality can be probed with objective and behavioral tests. Objective tests measure the internal representation of acoustic stimuli at different stages of the system. This can either be done intrusively in animals or non-intrusively in humans. For example, single unit recordings of the auditory nerve in animals allow to observe how firing patterns of one auditory nerve fiber change in response to different stimuli (e.g., Young and Sachs, 1979). Such measures build a valuable basis for understanding how information is coded at different stages of the system. Comparing the firing patterns of a healthy system to the patterns of an impaired system further allow to quantify the loss of information due to the impairment.

While understanding the neural coding of information along the pathway is crucial for an understanding of the system, the ultimate goal is to understand how the perceived sounds depend on this code. To measure perceptual con-

sequences of the elicited representations, behavioral tests are essential. These tests can either use simplified artificial stimuli or natural complex stimuli. Much progress has been made with respect to measuring and understanding the perception of simplified artificial stimuli, such as pure tones, modulated tones, and modulated noises in NH listeners (e.g. Dau et al., 1999). Computational models based on physiological recordings and psychoacoustic data have been developed to study such data (e.g., Carney, 1993; Dau et al., 1996).

Speech is a complex and natural stimulus and has been investigated in much detail. Speech intelligibility is usually measured as the proportion of correctly identified speech units in controlled listening conditions. Recognition performance depends on the listener, the speech presented and the listening condition. In most studies, interfering background noise is used to test listeners at their recognition limits. The recognition rate decreases as the background noise level increases. The background noise level at which a listener recognizes 50% of the presented speech units can serve as a single number to quantify a listener's performance and is usually referred to as speech-reception threshold (SRT). The slope at which the performance decreases with increasing noise level indicates the robustness of the listener to the background noise.

Different noise types have been used as background noises. They can be categorized in stationary and fluctuating noises according to their temporal properties. White noise and white noise spectrally shaped to have a long-term average speech spectrum, also referred to as speech-shaped noise (SSN), represent examples of stationary noises. Fluctuating noises describe background noises for which the level exhibits significant temporal fluctuations. Most real-world background noises are fluctuating. For example, the "noise" caused by other talkers in the background will fluctuate with the speech rhythm of those talkers. Recognition performance in NH listeners in fluctuating noise

exceeds the performance in stationary noise, as listeners are capable of extracting speech information in the dips of the background noise fluctuations (e.g., Festen and Plomp, 1990). The performance benefit a listener obtains from fluctuating noises as compared to stationary noises is typically referred to as masking release (MR). HI listeners typically show worse performance on SI tests than NH listeners. Their SRTs are higher in both stationary and fluctuating background noises. Furthermore, they typically exhibit lower amounts of MR than NH listeners.

Depending on the speech stimuli used in a study, studies can be divided into "macroscopic" and "microscopic" speech intelligibility (SI) studies. Microscopic intelligibility focuses on the perception of short speech units without a meaning or context, i.e., phonemes (e.g., Allen, 1996a). In contrast, macroscopic SI tests utilize whole words or sentences to probe speech intelligibility in listeners.

Measuring and understanding the consequences of hearing impairment on behavioral measures, especially measures of natural stimuli like speech, has posed a challenge due to the heterogeneity observed in such data (Trevino and Allen, 2013). In order to simplify the problem, recent studies have focused on understanding the implications of hearing impairment on the perception of small meaningless speech units, such as monosyllabic consonant-vowels (CVs). This microscopic approach allows to exclude factors like context processing and between-subject differences in cognitive abilities from the behavioral results. Microscopic speech studies have been able to quantify the impact of stimulus variations on perception (Singh and Allen, 2012; Toscano and Allen, 2014; Zaar and Dau, 2015). For example, Trevino and Allen (2013) demonstrated that the heterogeneity of the HI listeners decreases when the stimuli variations are reduced.

Testing human listeners can be costly or ineffective for certain scientific

questions or technical problems. Models of speech intelligibility aim at predicting a listener's performance based on signal properties of the target stimuli and the background noise or interferers. Such models typically use a linear model of the auditory periphery and a signal-to-noise ratio (SNR) or correlation-based decision metric in their backends (e.g., Fletcher and Steinberg, 1929; Fletcher and Galt, 1950; Kryter, 1962; Houtgast and Steeneken, 1971; Payton and Braida, 1999; Jørgensen and Dau, 2011; Taal et al., 2011). SNR decision metrics assume that if the speech energy exceeds the noise energy in a certain band, then this band positively contributes to SI. In contrast, if the noise energy exceeds the speech energy, the band adversily affects overall SI. Correlation-based models assume that the more correlated a noisy speech signal is to the clean speech signal (i.e., the template), the easier it should be to understand the noisy signal.

Modeling the consequences of individual impairment on speech understanding has been a focus of several recent modeling studies (Bruce et al., 2013; Hossain et al., 2016; Wirtzfeld, 2016; Moncada-Torres et al., 2017). However, predicting individual results based on audiometric data of a listener's hearing loss remains a challenge.

This project focused on advancing SI tests and models to better characterize SI deficits observed in HI listeners. In a first part, new methods to analyze and interpret microscopic SI data are presented. Specifically, *Chapter 2* describes a method to asses and compare the results of microscopic speech perception on an individual level across different conditions. The measures described allow to isolate the individual problems of listeners out of a large data set. The results in this chapter show the potential benefits of using natural meaningless stimuli for hearing aid testing. In a second part, a model to predict SI in HI listeners is proposed. Such a model that can predict the speech intelligibility scores for an individual listener with hearing loss could be an important tool for hearing-aid

fitting or the development of hearing-aid algorithms and could provide insights into the auditory processing of speech in NH and HI listeners.

In *Chapter 3*, consequences of hearing impairment are introduced into an existing model of macroscopic speech intelligibility (Jørgensen and Dau, 2011; Jørgensen et al., 2013). The newly introduced model modifications are verified with data from past studies. The scope and limitations of such models are discussed.

Chapter 4 describes a new modeling approach where a detailed peripheral model (e.g., Carney, 1993; Bruce et al., 2003; Zilany et al., 2009; Zilany et al., 2014) is combined with the framework of predicting speech intelligibility based on the signal-to-noise ratio in the envelope domain.

Chapter 5 presents a promising model, which again uses the auditory-nerve model as a front end, but combines it with a correlation metric in the back end instead of a SNR_{env} decision metric. Inspired by a recent study on vowel coding in the midbrain Carney et al. (2015), the model uses a single modulation filter centered at a frequency close to the fundamental frequency of the male target speaker.

Finally, *Chapter 6* summarizes the main findings and discusses the limitations and perspectives of the proposed models.

2

Assessing the efficacy of hearing-aid amplification using a phoneme test[a]

Abstract

Consonant-vowel (CV) perception experiments provide valuable insights into how humans process speech. Here, two CV identification experiments were conducted in a group of hearing-impaired (HI) listeners, using 14 consonants followed by the vowel /a/. The CVs were presented in quiet and with added speech-shaped noise at signal-to-noise ratios of 0, 6, and 12 dB. The HI listeners were provided with two different amplification schemes for the CVs. In the first experiment, a frequency- independent amplification (flat-gain) was provided and the CVs were presented at the most- comfortable loudness level. In the second experiment, a frequency-dependent prescriptive gain was provided. The CV identification results showed that, while the average recognition error score obtained with the frequency-dependent amplification was lower than that obtained with the flat-gain, the main confusions made by the listeners on a token basis remained the same in a majority of the cases. An entropy measure and an angular distance measure were proposed to assess the highly individual effects of the frequency-dependent

[a] This chapter is based on Scheidiger et al. (2017) JASA

gain on the consonant confusions in the HI listeners. The results suggest that the proposed measures, in combination with a well-controlled phoneme speech test, may be used to assess the impact of hearing-aid signal processing on speech intelligibility.

2.1 Introduction

Most day-to-day communication between humans is based on speech. Deficits in speech communication, e.g., as a result of a hearing impairment, can have strong effects on a person's quality of life and personal success. Hearing aids can help to regain the ability to hear speech, e.g., by compensating for the audibility loss. However, aided hearing- impaired (HI) listeners typically perform worse in speech understanding tasks than normal-hearing (NH) listeners. In particular, hearing-aid users commonly experience difficulties in challenging acoustical environments, such as noisy and/or reverberant spaces. In contrast, speech communication over a noisy transmission channel in NH listeners is typically robust.

Speech recognition can be limited by internal noise and external noise. External noise describes interfering acoustical signals that may mask or distract from the target signal. Internal noise characterizes the limitation and probabilistic nature of a listener's auditory system. A hearing loss may be viewed as an increase of internal noise. According to Plomp (1986), the internal noise can be further divided into an audibility component and a distortion component. The typical measure of the audibility component is an audiogram or a speech reception threshold in quiet (SRTq). The SRTq is defined as the speech level at which the recognition score equals the error score ($p_c = p_e$). While the SRTq is linked to the speech reception threshold in noise in Plomp's model, the audiogram and

speech intelligibility in noise are not directly linked. Several studies have tried to link pure-tone thresholds to speech intelligibility of both NH and HI listeners (Humes et al., 1986; Zurek and Delhorne, 1987; Pavlovic, 1986; Mueller, 1990). Mueller (1990) proposed the "Count-the-dots" method to calculate the articulation index, which can be transformed to a speech intelligibility score. Their method assesses how much of the long-term average speech spectrum (LTASS) is audible, i.e., above the pure-tone thresholds. This has become a widely used method to numerically quantify the benefit of a hearing instrument.

Speech intelligibility in noise may be measured with different speech materials. Phonemes (e.g., consonant-vowels, CVs) represent one class of speech materials. Phoneme identification experiments record which phoneme out of the phoneme set used in the experiment was chosen by a listener in response to a presented stimulus. The recorded responses are often presented in the form of a confusion matrix (CM), wherein each cell corresponds to one of the stimulus- response pairs. The stimuli are usually denoted as rows and the responses as columns. The diagonal of the matrix represents the counts of the correct responses and the row sum equals the total number of presentations for a given stimulus.

Phoneme perception research has a long history and started with the classical studies by French and Steinberg (1947) and Miller (1955). French and Steinberg (1947) based their analysis on recognition scores only, i.e., the CM diagonal, and proposed a model to predict the percent correct value of phoneme pairs or triplets based on the individual phone scores. Later, Miller (1955) applied an information theoretical analysis to their recorded CMs. Their entropy measure, which quantifies the random- ness of responses, represents an approach to describe the communication process beyond pure recognition scores. In the case of a phoneme which is always misclassified (i.e., 100% error), this

phoneme could be always confused with one specific other phoneme, which would correspond to an entropy of 0 bits. Alternatively, the phoneme could be con- fused with many other phonemes (instead of only one specific phoneme), in which case the entropy would be close to its maximum $log_2(J)$ bits with J representing the number of possible response choices.

Entropy is powerful in quantifying the randomness of responses but is insensitive to the kind of confusions. Two different phonemes might produce the same randomness in terms of observed responses but the individual confusions can be very different. Allen (2005) used confusion patterns (CPs) to visualize the individual confusions along with the recognition score. CPs show the response probabilities for all response alternatives as a function of the signal-to-noise ratio (SNR) for a given stimulus, i.e., they depict normalized CM rows as a function of SNR and thereby illustrate at which SNRs the recognition score drops and which confusion(s) was/were chosen instead of the correct response. If the response probabilities are shown on a logarithmic scale, confusions with low probabilities are clearly represented.

However, in order to use CV experiments to assess a HI listener, the perceptually relevant factors that underlie consonant perception need to be known and the CV experiments need to be designed accordingly. Despite the extensive research and elaborate analysis methods, only a few studies have revealed the effect of acoustic stimulus variability on consonant perception in individual listeners (Li and Allen, 2011; Phatak and Allen, 2007; Kapoor and Allen, 2012; Singh and Allen, 2012; Toscano and Allen, 2014; Zaar and Dau, 2015). This variability may be particularly relevant in studies with HI listeners Trevino and Allen (2013). Consonant perception has been demonstrated to be strongly affected by a high-frequency sensorineural hearing loss (e.g. Owens, 1978), reflecting the importance of high-frequency information contained in consonants (Li

et al., 2010; Li and Allen, 2011). Several studies thus proposed to control for the variance in the stimuli (e.g. Bilger and Wang, 1976; Boothroyd, 1984) as well as the variability across the HI listeners to reduce the variability in the CM data (e.g. Owens, 1978; Dubno et al., 1984; Zurek and Delhorne, 1987; Trevino and Allen, 2013).

Miller (1955) found that only a few of the possible response alternatives were chosen for a specific consonant, i.e., CM rows were sparse and the entropy thus small. Owens (1978) discussed a dependency of consonant perception on the specific selection of a consonant-vowel-consonant token, whereby a token represented a single phoneme recording. It was argued that the robustness and confusions obtained for individual tokens were specific to these tokens. The token dependency was later confirmed by Trevino and Allen (2013) who showed that the confusions in CV experiments became more consistent when the token variability was controlled for. Trevino and Allen (2013) analyzed confusions in HI listeners on a token basis and found that listeners with different audiograms showed similar confusions at the token level. This suggested that responses for a given CV token obtained across listeners can be more homogeneous than previously assumed. Furthermore, the authors found that different tokens of the same CV can result in different confusions in the same listener group. For example, the main confusion for a specific /ba/ token was /va/, whereas it was /da/ for another /ba/ token (Table II in Trevino and Allen, 2013). These results demonstrated the importance of considering consonant perception at the token level.

Dubno et al. (1984) reported a degraded CV recognition performance in HI listeners in the presence of noise, even in conditions when the speech was presented at high sound pressure levels, indicating that audibility alone was not sufficient to restore correct recognition. Furthermore, it was found that age

had a detrimental effect on CV recognition in listeners with the same average hearing loss in terms of the audiogram. Zurek and Delhorne (1987) tested average consonant recognition scores both in HI and NH listeners. For the NH listeners, the phonemes were presented together with spectrally-shaped masking noise to simulate the sensitivity- related hearing loss of a matched HI listener. In contrast to the results from Dubno et al. (1984), Zurek and Delhorne (1987) found that matching NH ears to HI audiometric measures can result in a similar performance in terms of their average recognition errors. However, Zurek and Delhorne's conclusions were based on average recognition scores of their listeners and did not compare the confusions between the two listener groups, i.e., the off-diagonal elements of the CM, nor did they take the strong token dependence effect into account.

Trevino and Allen (2013) presented their stimuli to 16 HI ears at a comfortable overall loudness without a frequency- dependent gain to compensate for the audibility loss. They presented the CVs in quiet and at SNRs of 0, 6, and 12 dB in speech-shaped noise (SSN). It remained open if their observed consistency of the main confusions across listeners would also be observed if an individual frequency-dependent amplification was provided. For example, it is possible that the main confusion of /va/ observed in one token of /ba/ and the main confusion of /da/ observed in the other token of /ba/ would change if a frequency-dependent gain were provided.

The present study investigated phoneme perception on a token level in the same HI listeners as the Trevino and Allen (2013) study. In contrast to Trevino and Allen (2013), the listeners were provided with an individual frequency-dependent amplification to compensate for their audibility loss. It was tested how much the listeners improved in CV recognition as a result of the high-frequency amplification as compared to the earlier results obtained with flat

(i.e., frequency-independent) amplification. The results were analyzed on a token basis using a response entropy measure to quantify the distribution of confusions as well as a vector space angular distance to evaluate how the specific nature of confusions changed between the two amplification conditions. It is argued that the two metrics together reveal a detailed picture of the relative efficacy of different amplification schemes and could be used to assess strategies to improve speech intelligibility in general.

2.2 Method

2.2.1 Listeners

Eight HI listeners (16 HI ears) with a mean age of 74 years participated in the two experiments. All listeners reported American English as their first language and were regular users of hearing aids. They were paid to participate in the IRB-approved experiments. Tympanometric measures obtained before the start of the experiments showed no middle-ear pathologies (type A tympanogram). All 16 ears had a mild-to-moderate sensorineural hearing loss. Figure 2.1. shows the fitted pure tone threshold (PTT) functions of the individual listeners (Trevino and Allen, 2013). The audio- grams were modeled as two piece-wise linear functions. These fittings were characterized by three parameters: the breakpoint f_0 , the low-frequency loss h_0 , and the slope of the high-frequency loss s_0 . The break-point f_0 between the two linear functions indicates the frequency at which the sloping loss begins. At frequencies below f_0 , the hearing loss was assumed to be constant over frequency (h_0). At frequencies above f_0 , the audiogram was modeled by a linear function with a negative slope (s_0). The average root-mean-square error of the fitted curves over all audiogram frequencies ($f = [125, 250, 500, 1000, 1500, 2000, 3000, 4000, 6000, 8000]$ Hz) was

5 dB (see the Appendix).

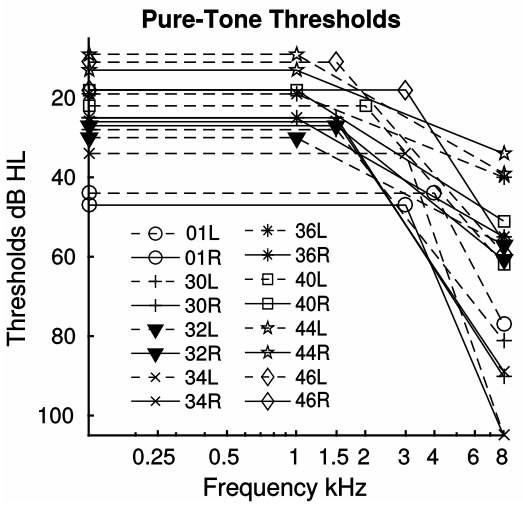


Figure 2.1: Fitted pure-tone thresholds for all the listeners that participated in the study. All listeners had a steeply sloping hearing loss at high frequencies. The average root-mean-square error of the fitting was 5 dB (see the Appendix).

2.2.2 Stimuli

The CV syllables consisted of 14 consonants (six stops /p, t, k, b, d, g/, six fricatives /f, s, ʃ, v, z, ʒ/, and two nasals /m, n/) followed by /ɑ/. Two tokens (one recording of a male talker and one of a female talker) were selected per consonant from the Linguistic Data Consortium Database (LDC-2005S22; Fousek et al., 2000). The tokens were chosen from those for which the recognition error was below 3% at a SNR of >2 dB in earlier experiments with NH listeners (Singh and Allen, 2012; Toscano and Allen, 2014). They were presented at 12, 6, and 0 dB SNR in SSN; a range in which NH listeners would not make any recognition errors. The CV tokens had previously been investigated using the three-dimensional-deep-search method (Li et al., 2012) to identify perceptually relevant spectro-temporal cues in the stimuli. Furthermore, NH reference data with the same CV tokens had been collected in white noise as well as SSN

(Phatak and Allen, 2007). Four of the male tokens (/f, n, s, ʒ/ + /a/) had to be excluded from the analysis, as they were found to have been adversely affected by a stimulus pre-processing algorithm. The algorithm was intended to truncate all stimuli to the same duration by removing silent periods before and after the target token. Unfortunately, it truncated the weak bursts of these CV male tokens. The remaining 24 CV tokens were presented in two amplification conditions which were analyzed in the present study. The stimuli were presented to the listeners over an Etymotic Research (Elk Grove Village, IL) in-ear speaker (ER-2) in a single-walled sound booth in a room with the outer door closed.

2.2.3 Amplification schemes

The stimuli were presented in two different amplification conditions. These conditions were tested on separate days after verifying that the audiometric thresholds of the listeners had not changed since the last session. The listeners completed a 20-min long training session per amplification condition with separate tokens before starting the testing. In the first amplification condition (FG), a frequency-independent gain was provided. The gain was chosen by the listeners in a calibration run before the training session. The levels chosen by the listeners are indicated in the Appendix. The listeners were able to adjust the gain during the experiment. However, only listener 40L made use of this option (2 dB change). For the second amplification condition (NAL-R), the CV stimuli were amplified with an NAL-R gain adjusted for each listener according to their audiogram (Byrne and Dillon, 1986). The goal of the NAL-R amplification scheme is to provide equal loudness in all frequency bands. The insertion gain prescription is based on the PTTs at the frequencies f =0.25, 0.5, 1, 2, 3, 4, and 6 kHz. Also in this condition, the listeners were allowed to adjust the overall gain of the amplification. The corresponding chosen levels are represented in Table

I.

2.2.4 Experimental procedure

A token could be repeated as many times as required to select one of the 14 response alternatives displayed on a computer screen. The display presented symbols from the International Phonetic Alphabet (as well as a common English word that started with the respective consonant. For each condition, SNR, and listener, a token was presented between 5 and 10 times. The data collection for each amplification condition was split into two sessions in which the stimuli were presented in a fully randomized order. The number of stimulus presentations per SNR, ear, and token was four in the first session. In the second session, the number of presentations per SNR, ear, and token depended on the number of confusions in the first session. Zero or one confusion in the first session led to two more presentations in the second session. Two confusions led to five more presentations and more than two confusions led to six additional presentations. This resulted in 800–1000 trials per listener, with more presentations allocated to the CVs that were confused by the individual listeners. This helped in identifying specific problems of individual listeners at realistic SNRs with CV tokens that were known to be robustly recognized by NH listeners at the given SNRs.

2.2.5 Analysis

In the experiments, one CM per ear (16 ears), amplification condition (2 conditions), SNR (4 SNRs), and token (2 tokens) was obtained, resulting in a total of 256 CMs. In addition to the recognition scores (i.e., diagonal CM values), two measures were considered to analyze the data.

Entropy

In information theory, entropy describes the randomness of a communication process. In phoneme experiments, it can be used to quantify the randomness of responses. The CM cell $CM(i,j)$ contains the counts of the listeners' responses with the response alternative $j = 1,...,J$ when the stimulus $i = 1,...,I$ was presented. The value $CM(i,j)$ of the CM, normalized by the respective row sum $RS(i) = \sum_j CM(i,j)$, represents the response probability $p_{ij} = CM(i,j)/RS(i)$, whereby the P overall sum of response probabilities for a row is one ($\sum_j p_{ij} = 1$). In terms of information theory, the observation of a listener responding with j when presented with stimulus i contains the information $log_2(1/p_{ij})$, implying that a more likely response (e.g., the correct response $j = i$) carries less information than a rarely observed response. The response entropy $\mathscr{H}(i)$ is defined as the expected information from observing all responses to a stimulus

$$\mathscr{H} = \sum_j p_{ij} log_2(1/p_{ij}) \tag{2.1}$$

Entropy as defined with the log base 2 is measured in bits. If a listener were to only use one of the response alternatives, the entropy would be 0 bit, irrespective of whether or not the response used by the listener is correct. In contrast, if all 14 possible response alternatives were to occur equally likely ($p_{ij} = 1/14$ for all j), the response entropy would reach its maximum value, $\mathscr{H}_{max} = log_2(J = 14) = 3.81$ bits. The higher the entropy, the more uncertain is the listener regarding his/her responses.

The entropy, as defined above, strongly depends on the recognition score (p_{ii}) as well as the distribution of the confusions. To use the entropy as a complementary measure to the recognition score, a measure independent of the

recognition score is needed. The confusion entropy $\mathscr{H}_{Conf}$ used in this study is obtained by replacing the normalized response vector p_{ij} by the normalized confusion vector p_{Conf} in Eq. 2.1. To obtain p_{conf} the count of correct responses is excluded from a CM row before normalizing it by the row sum, i.e., the vector only consists of counts representing confusions. The values in p_{conf} therefore express the probability of a confusion occurring given an error occurs.

Hellinger Distance

A metric that is sensitive to changes in confusion probabilities was considered. Each CM defines a vector space, with each row CM(i) representing a vector in that space. The vector space is defined by the basis vectors (e_j), where each basis vector represents a possible confusion. In order to find the distance between two rows (e.g., two CVs or two tokens), a norm must be defined. Here, the Hellinger Distance was used (Scheidiger and Allen, 2013), which utilizes the square roots of the probability vectors $p_i = [p_{i1}, ..., p_i J]$. All vectors defined by the square roots of the probabilities yield the same norm and there- fore have the same length. Thus, the distance between two vectors can be expressed by the angle between the vectors. Via the Schwartz inequality, it is possible to calculate an angle θ_{kl} between any two response vectors p_k and p_l in the vector space

$$cos(\theta) = \sum_j \sqrt{p_{kj}} \sqrt{p_{lj}} \tag{2.2}$$

The angle is a measure of how different the two vectors are. In addition to ensuring unit length of all vectors, the square- root transformation emphasizes less likely confusions and makes the metric more sensitive to small changes

in the response vectors than correlation-based metrics. This angular distance measure was used in the present study to represent the difference between two confusion vectors obtained in the condition with frequency-dependent gain (NAL-R) and the flat-gain (reference) condition. A Hellinger distance of 0° between the normalized confusion vector (p_{conf}) of the flat-gain and the NAL-R condition implies that the same confusions were equally likely in the two conditions. In contrast, a Hellinger distance of 90° represents cases in which the confusions in one condition (e.g., flat-gain) were not present in the other condition (e.g., NAL-R). The Hellinger distance between confusion vectors is not defined and thus yields NaN (not a number), if one of the conditions does not exhibit any errors.

2.3 Results

Figure 2 shows the CPs of four listeners (30R, 32L, 36L, 40L) for the /ba/ token #1. The flat-gain condition is shown in the left panels, whereas the results obtained with NAL-R are shown on the right. The recognition score for /ba/ (black solid line), in general, dropped as the SNR decreased from the quiet condition (Q) to lower SNRs, i.e., at 12, 6, and 0 dB. For example, in the flat-gain condition, listener 30R (upper left panel) showed a recognition score for /ba/ of 63% in the quiet condition. At 12 dB SNR, the recognition score was 13% while the response probabilities for the /va/ and /fa/ confusions increased from 0% in the quiet condition to 73% and 13%, respectively. At 6 dB SNR, listener 30R always indicated to have perceived /va/. At 0 dB SNR, the confusion /va/ still represented the dominating response, showing a probability of 60%, whereas the remaining responses were equally distributed over the correct response /ba/ and the two confusions /fa/ and /da/.

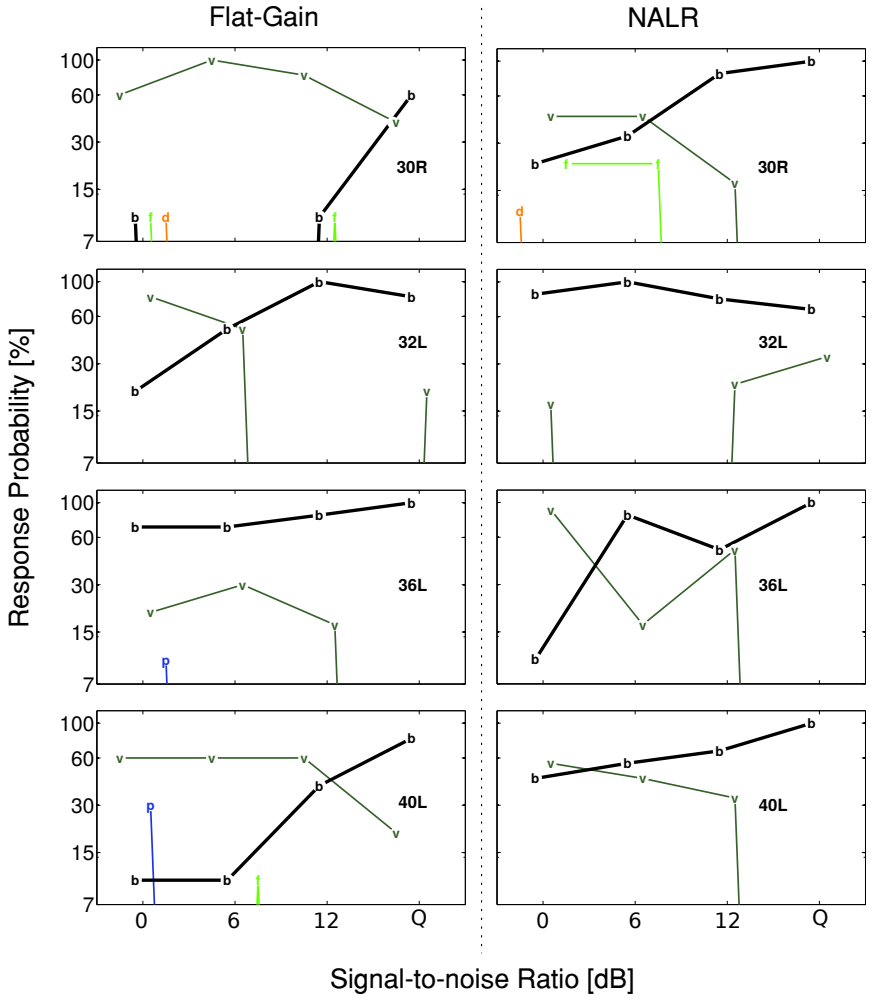


Figure 2.2: Confusion patterns for four of the subjects showing the response probabilities as a function of SNR for the token #1 of the CV /ba/. The left column shows the data with the flat-gain as also presented in Trevino and Allen (2013). The right column presents the data for the same listeners but with NAL-R gain. The main confusion with both gains is /va/. A slight horizontal jitter was introduced to the data for better readability.

When a frequency-dependent gain was provided using the NAL-R scheme (right column of Fig. 2.2), the obtained CPs differed. For example, in the case of listener 30 R, the recognition score became more robust to noise; the recognition score for /ba/ was at 100% in quiet, decreased to 85% at an SNR of 12 dB, and dropped to 30% at 0 dB SNR. However, despite the more robust recognition score than in the flat-gain condition, the /va/ confusion was still also dominant

in the NAL-R condition. With decreasing SNR, the response probability for /va/ increased to 15%, 50%, and 50% at SNRs of 12, 6, and 0 dB SNR, respectively. For all four listeners shown in Fig. 2.2, the main confusion /va/ observed in the flat-gain condition also represented the main confusion in the NAL-R condition. Less likely responses, such as /pa/ and /da/, disappeared in the NAL-R condition. Despite the different audiograms and, therefore, different gains applied to the individual listeners in the NAL-R condition, the main confusions among the listeners remained the same. This finding is consistent with the observations reported in Trevino and Allen (2013), regarding their token- specific confusions.

Figure 2.3 shows the CPs obtained with the same listeners but for the other /ba/ token. As in Fig. 2.2, the recognition scores dropped as the SNR decreased. For listeners 30R, 36L, and 40L, the recognition scores with NAL-R gain were found to be more robust to noise than those obtained with flat-gain. The main confusions in the flat-gain condition for the second token were /ga/ and /da/, in contrast to /va/ in the case of the first token (Fig. 2.1). With the NAL-R gain (right panel), the /ga/ and /da/ error patterns for /ba/ token #2 remained dominating. For example, for listener 30R (top panel), the recognition score of /ba/ became more robust to noise in the NAL-R condition and never dropped below 60%, but the main confusion, /ga/, also became more robust. For listener 32L, the NAL-R gain produced more prominent /ga/ confusions even at high SNRs, i.e., the presence of noise morphed the /ba/ into a /ga/.

When considering all results across all listeners, averaged across SNRs and the 24 tokens, the error rate (i.e., 1- recognition score) decreased from 20.1% in the flat-gain condition to 16.3% in the NAL-R condition. There was a significant relationship between the type of amplification and the correct recognition of the 14 phonemes [$\chi^2(1)$=56.1, p<0.00001]. The odds of a correct response

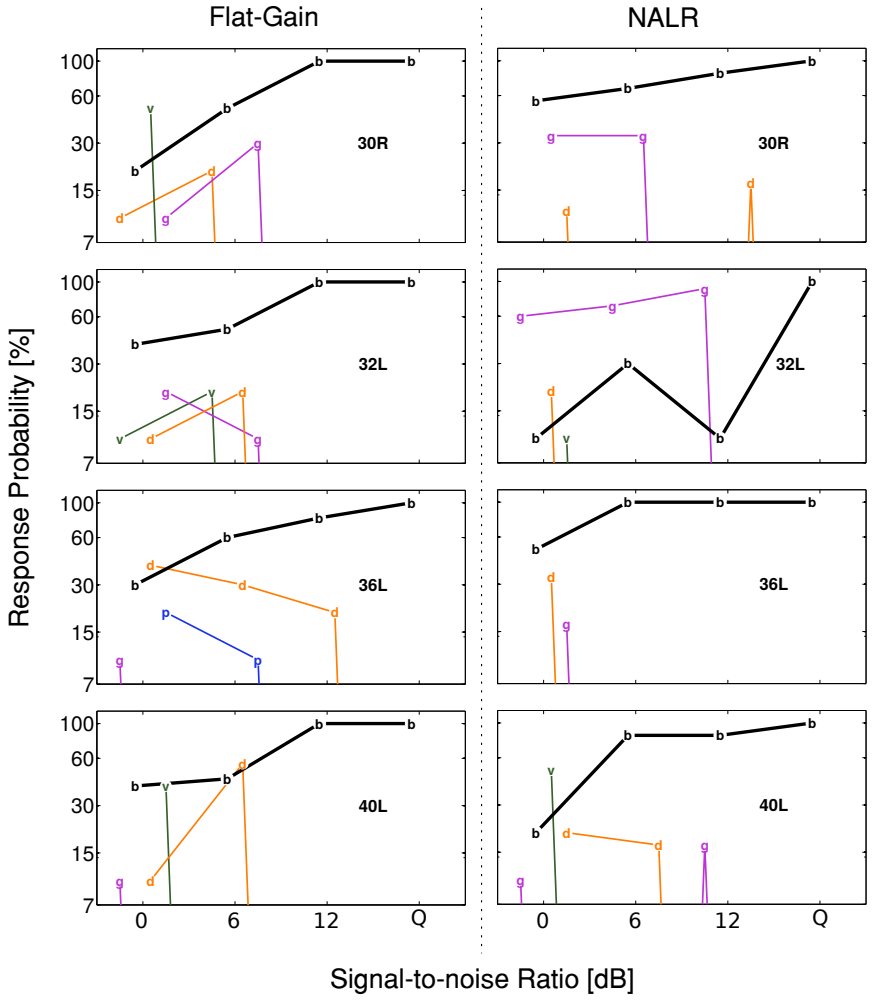


Figure 2.3: Confusion patterns for four of the subjects showing the response probabilities as a function of SNR for token #2 of the CV /bɑ/. The left column shows the data with flat-gain as also presented in Trevino and Allen (2013). The right column presents the data for the same listeners but with NAL-R gain. The main confusion with both gains is /dɑ/. A slight horizontal jitter was introduced to the data for better readability.

with the NAL-R amplification were 1.25 (1.18 1.33) times higher than with the flat-gain amplification. The average normalized confusion entropy ($\mathcal{H}_{Conf}$) decreased from 0.5 ($s = 0.1$) in the flat-gain condition to 0.3 ($s = 0.1$) in the NAL-R condition.

Figure 2.4 shows a more granular analysis of how error rates and normalized confusion entropies were affected by the two amplification conditions

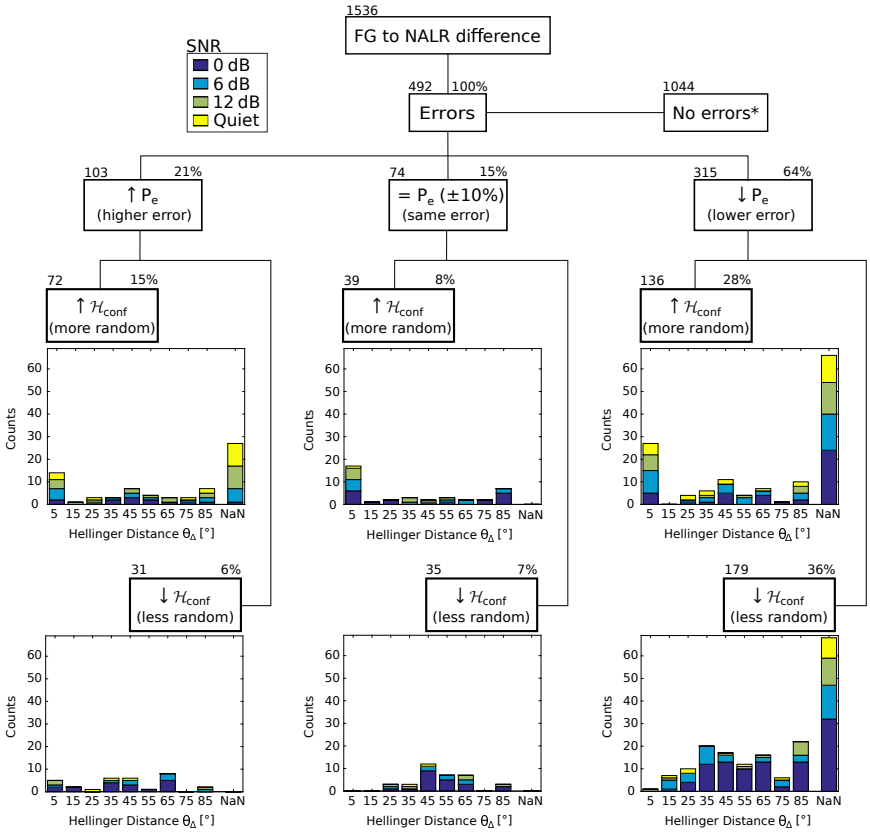


Figure 2.4: Categorization of the CV perception data for the 24 tokens, 16 listeners, and 4 SNRs. The category "No errors*," contains cases with just one or zero errors out of all trials. The response patterns with at least two errors in one of the conditions were divided into three categories according to how the error rate changed from the flat-gain condition to the NAL-R condition. Twenty-one percent of the erroneous response patterns had an increased error, 15% showed the same error (±10%), and the remaining 64% showed at least 10% fewer errors. These three categories were each further divided into two sub-categories depending on how $\mathcal{H}_{Conf}$ changed in the NAL-R condition as compared to the flat-gain condition. For each subcategory, a count histogram of the Hellinger angles θ_Δ is shown on the bottom, the bins are 10° wide, and labeled by their center.

in an individual HI listener responding to a given token at a given SNR. For each listener-token pair, the error rate and confusion entropy ($\mathcal{H}_{Conf}$) at each SNR were calculated in the flat-gain condition and in the NAL-R condition. To compare the results obtained in the two different amplification conditions, the values in the flat-gain condition were considered as refer- ence. The responses of the 16 HI ears to the 24 tokens at 4 SNRs resulted in 1536 response patterns for each condition.

The response patterns were divided into two categories: (i) $P_e = 0$, containing all 1044 (68%) patterns that showed maximally one erroneous response in either condition and (ii) P e > 0, comprising the remaining 492 (32%) patterns which had more than one error in at least one condition. As consonant recognition was at ceiling for the $P_e = 0$ category, these response patterns were not considered in the subsequent analysis. In contrast, the $P_e > 0$ response patterns, which represent the critical/interesting cases, were further divided into three subcategories according to their error rates.

For 103 (21%) of the 492 considered token-listener pairs, P e in the NAL-R condition increased by more than 10% as compared to the flat-gain condition (left branch in Fig. 2.4). In 74 response patterns (15%) the error rate did not change by more than 10% in either the positive or negative direction in the NAL-R amplification condition (middle branch). For the remaining 315 response patterns (64%), the error in the NAL-R condition decreased by at least 10% as compared to the flat-gain condition (right branch). Each of the three categories was in a last step subdivided into two subcategories according to how H_{Conf} changed in the NAL-R condition with respect to the flat-gain condition. The subcategories "more random" and "less random" contain the response patterns in which H_{Conf} in the NAL-R condition increased or decreased, respectively, compared to the flat-gain condition.

This categorization provides a detailed picture of how the NAL-R amplification scheme affected the responses to the considered CVs on a token basis. If NAL-R had improved all listeners' performance, this would have resulted in a decrease in P e along with a decrease or no change in $\mathscr{H}_{Conf}$. However, only 36% of the considered response patterns fell into this category, while 28% showed a decrease in P e along with more random response behavior (right branch in Fig. 2.4). Furthermore, 21% of the considered response patterns showed an

increase in the error rate with the NAL-R amplification (left branch). Rather few response patterns were unaffected by NAL-R (15%; middle branch).

The error rate and normalized confusion entropy do not characterize the nature of confusions. Two response vectors obtained with two different tokens of the same CV might result in the same error rate and normalized confusion entropy; however, one token may show a different main confusion than the other (Trevino and Allen, 2013; cf. Figs. 2.2 and 2.3). To quantify specific confusions, the Hellinger distance was used to measure the angular distance between different response vectors. The two bottom rows of Fig. 2.4 show θ_Δ count histograms for each $\mathscr{H}_{Conf}$-P_e subcategory. The bins of the histogram are 10° wide and are labeled by their center angle. Each response pattern is color-coded according to the SNR at which it was obtained (blue for 0 dB, turquoise for 6 dB, green for 12 dB, yellow for Quiet). It can be seen that response patterns at lower SNRs (blue, turquoise) mostly fall into the decreased error rate category (right branch in Fig. 2.4) and also that the cases in which NAL-R increased both the error and the randomness of the error are dominated by quiet conditions (yellow).

The angular distance is undefined and thus yields NaN if no errors were recorded in one of the conditions. In the upper-left category ($\uparrow P_e$, more random), the 30 response patterns with θ_Δ =NaN did not show any error in the flat-gain condition but showed errors in the NAL-R condition. These error rates were by no means small. The average error rate in the NAL-R condition for these cases was 45%, one- fifth of these cases showed error rates of >90%, indicating significant changes in the percept. Those cases can be referred to as “morphs,” as NAL-R morphed them from a perceptually robust correct response into a robust confusion. For the 140 response patterns for which θ_Δ =NaN in the $\downarrow P_e$ -categories (right panel of Fig. 2.4), NAL-R reduced the error rate to zero. These

can be referred to as “optimal” cases.

The two extreme bins of the θ_Δ histograms (centered at 5° and 85°) indicate listener-token pairs with the same or entirely different confusions in the two conditions, respectively. The 5° bin contains the cases for which the confusions and their proportions remain virtually unchanged irrespective of the amplification. In the case of the $\uparrow P_e$-categories (left panel in Fig. 2.4) they represent cases in which the flat-gain main confusions were chosen even more frequently in the NAL-R condition. In the $\downarrow P_e$ category (right panel in Fig. 2.4), they represent cases for which the error rate decreased but the main confusion remained the most likely confusion. A low θ_Δ indicates that the confusions in the flat-gain condition also dominated the response pattern in the NAL-R condition. The 5° -bin reflects the most prominent examples for this behavior but the same trend can also be observed for bins where $\theta_\Delta = 45°$. Considering a threshold of $\theta_\Delta = 45°$ to indicate whether the main confusion remained the same (<45°), the analysis reveals that in 63% of the cases the main confusions remained unchanged.

$\theta_\Delta = 90°$ —contained in the bin centered at 85° —indicates that the confusions were different and that the response vector for the NAL-R condition did not contain the confusions in the flat-gain condition and vice versa. Thus, in these cases, NAL-R introduced new confusions that were not present in the flat-gain responses (morphs). In all but two θ_Δ-histograms, the 5°-bins exhibited larger counts than the 85°-bins, indicating that the main confusions in these patterns were unchanged.

2.4 Discussion

The results from the present study support the findings of Trevino and Allen (2013) that the confusions in CV experiments are token specific, even if a frequency-dependent gain (NAL-R) is provided. While NAL-R, on average, decreased the error rate in the listeners' responses, the occurrence of the main confusions often remained the same (Figs. 2.2 and 2.3 and <45 ° in Fig. 2.4), indicating that NAL-R alone does not effectively compensate for the deficits that cause the main confusion. The observation of small values for the normalized confusion entropy in both amplification conditions (0.5 bit in the flat-gain condition as com- pared to 0.3 bit in the NAL-R condition) suggests that the main confusion is a robust and consistent phenomenon caused by token-specific cues and deficits in the individual auditory system. The different main confusions for the two /ba/ tokens that are robust across the two amplification conditions, suggest that they are caused by the acoustic properties of the stimulus, i.e., by conflicting consonant cues (Kapoor and Allen, 2012). A stimulus that evokes responses with low entropy but a high error rate must have been chosen based on a robust auditory percept. This percept must therefore result from some distorted internal auditory representation of the stimulus which could be considered as reflecting a "supra-threshold" distortion (such as, e.g., a temporal and/or spectral auditory processing deficit). Such a distortion could affect the primary consonant cue and increase the perceptual salience of a secondary cue that then causes the main confusion. In the case of the 30 morphs observed in the results, the robust confusions resulted from supra-threshold deficits in the HI listeners' auditory processing in combination with the high-frequency amplification. An understanding of which specific cues were used by the HI listeners would require a closer analysis of the individual audiometric configuration, the applied amplification, and the specific cues of the confused tokens (Li et al.,

2010, 2012) which were not undertaken in the present study. In contrast to the conditions with low-entropy response patterns, conditions where the confusion entropy was large are not based on a robust percept and should be assessed differently. The high entropy in these responses indicates that the listener did not respond based on a robust cue, but instead selected the response randomly. Such randomness may be caused by the effect of "internal" noise or attention deficits of the listener.

To define the entropy threshold for a robust percept, the average size of the Miller and Nicely (1955) confusion groups (/p, t, k, b, d, g/; /f, t, s, ʃ/; /v, ð, z, ʒ/; /m, n/) may be used. A listener is most likely guessing and therefore not responding based on a robust percept when confusions outside of the known confusion groups appear. The average size of confusion groups is three; thus, if more than three confusions occur, a decision-threshold for a robust percept could be defined in terms of the normalized confusion entropy which would be $H_{Conf} = 0.43$ bit (3 equally likely confusions out of the 13 possible confusions). When assessing the flat-gain response vectors with this definition, only 268 out of the 1536 token-listener pairs (17%) would not qualify as robust percepts.

A robust auditory percept might also be more appropriate than the traditional PTT and LTASS (i.e., count-the-dots method) to assess the audibility of CV signals. In experiments such as the ones from the present study, the differentiating perceptual cues of CVs may be manifested as local energy bursts or spectral edges in the signal (Li et al., 2010; Li et al., 2012). These cues can be more intense than the LTASS in a critical band over several 10 ms (Wright, 2004), but have a negligible contribution to the LTASS which is dominated by the vowel energy. It has been shown that CV recognition on a token level in NH listeners can drop from 100% correct to chance level if the energy of the noise masker is increased by less than 6 dB (Singh and Allen, 2012; Toscano and Allen,

2014). This “binary”-like recognition supports the importance of specific acoustic speech cues. These cues are either detectable, in which case the CV can be recognized despite the presence of noise, or are masked by the noise, in which case the listener might use a secondary cue or might start guessing. PTTs do not characterize a listeners' sensitivity to recognize these spectro-temporal consonant cues. Furthermore, if a different amplification scheme were chosen instead of NAL-R that aims at restoring audibility, e.g., a scheme as proposed in Reed et al. (2016), the specific confusions that exist after compensating for audibility can be used as an indicator of a supra-threshold distortion loss. To quantify the distortion loss based on CMs, the angular Hellinger distance measure could be used.

The response-patterns where both the error rate and the confusion entropy increased with NAL-R indicate the listener-specific phonemes for which the improvement strategy failed. These specific confusions could not be eliminated by NAL-R alone and should be addressed by alternative compensation strategies. Such strategies should take the token-specific consonant cues into account; the primary consonant cue should be amplified and conflicting secondary cues attenuated (Kapoor and Allen, 2012). For example, individually tuned frequency transposition algorithms may be able to transpose the spectro-temporal cues of the affected CVs to bands that are less affected by the distortion loss. Phoneme tests can help determine sensible limits for such frequency transposition algorithms to avoid further distortions (Schmitt et al., 2016). Such phoneme tests should consist of several well-characterized tokens for each consonant. These tokens should be correctly perceived by NH listeners at the SNRs tested. The recognition results should be analyzed on a token-specific level taking confusions and not only recognition scores into account. Zaar and Dau (2015) emphasized that the additive noise should be frozen noise, i.e., one

noise realization per token, to further decrease the within-listener variance in the responses.

2.5 Summary and Conclusion

CV perception in the same HI listeners as in Trevino and Allen (2013) was analyzed on a token level in two amplification conditions: a condition with frequency-independent amplification (flat-gain) and a condition with frequency-dependent amplification (NAL-R). The response patterns were analyzed in terms of their recognition scores, their confusion entropy, and an angular distance between the confusions in the two amplification conditions. The recognition score in the NAL-R condition was shown to be significantly higher than in the flat- gain condition. In a granular analysis (Fig. 2.4), the response pat- terns showed mixed results for the NAL-R condition, despite the overall increased recognition score.

Two measures were proposed to analyze the efficacy of speech intelligibility improvement strategies using a phoneme test, namely, the confusion entropy and an angular distance. The effect of a frequency-dependent gain was exemplarily investigated. The confusion entropy measure showed robust perception in all but 17% of the token-listener pairs in the flat- gain condition and thus demonstrated the validity of the results obtained at the most comfortable listening level. The proposed angular distance measure revealed that in 63% of the token- listeners pairs, the main confusions remained unchanged despite NAL-R, suggesting these are caused by acoustic properties of the chosen tokens rather than the amplification condition. The results suggest that a compensation strategy different than NAL-R would be needed to eradicate the main confusion. It was also observed that NAL-R in combination with the individ-

ual loss introduced new robust confusions in 30 cases. Phoneme recognition tests and methods that analyze confusions on a token-level, as the ones used in the experiments presented here, may be useful in the evaluation process of hearing-instrument algorithms. The tests could be conducted with selected robust tokens that have been shown to be correctly identified by NH listeners at the SNRs used in the test. Knowing the token-specific consonant cues and using a test that is focused on natural speech without context, a detailed diagnosis of an individual listener's speech loss seems possible and appropriate. A carefully constructed speech test could be used as a diagnostic tool where individual CPs of well characterized tokens may provide detailed information about a listener's hearing loss beyond what PTTs reveal.

2.6 Appendix

Additional information about the listeners who participated in the study can be found in Table 2.1.

Acknowledgments

The authors are grateful for helpful input by Johannes Zaar and the HSR group at the University of Illinois. The work leading to this deliverable and the results described therein has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement No. PITN-GA-2012-317521.

Table 2.1: Information about all the listeners participating in the experiments. The columns contain the following information: (i) label for each listener and the identifier for the left or right ear, (ii) age of the listener, (iii) the pure-tone average of the audiogram of the ear, (iv) the root means square error of the fitted audiogram, (v) the overall presentation level chosen by the listener in the FG experiment, and (vi) the overall presentation level chosen by the listener in the NAL-R experiment.

HI ear	Age	PTA	RSME	FG	NALR
44L	65	10	11	82	77
44R	65	15	7	78	77
46L	67	8.3	9	82	85
46R	67	16.6	7	82	86
40L	79	21.6	5	79,81	80
40R	79	23.3	5	80	80
36L	72	26.6	8	68	75
36R	72	28.3	4	70	75
30L	66	30	3	80	79
30R	66	26.6	5	80	79
32L	74	35	3	79	81
32R	74	26.6	3	77	78
34L	84	31.6	6	84	85
34R	84	28.3	4	82	85
02L	82	45	2	83	88
02R	82	46.6	4	82	89
(m,s)e	(74,4)	(29,15)	(5,2)	(79,4)	81,5

3

Modeling Hearing Impairment within the ESPM Framework[a]

Abstract

Models of speech intelligibility (SI) have a long history, starting with the articulation index (AI, ANSI S3.5, 1969), followed by the speech intelligibility index (SII, ANSI S3.5, 1997) and the speech transmission index (STI, IEC 60268-16, 2003), to only name a few. However, these models fail to accurately predict SI obtained with nonlinearly processed noisy speech, e.g., phase jitter or spectral subtraction. Recent studies predict SI for normal-hearing (NH) listeners based on a signal-to-noise ratio measure in the envelope domain (SNR_{env}), in the framework of the speech-based envelope power spectrum model (sEPSM, Jørgensen and Dau, 2011; Jørgensen et al., 2013). These models have shown good agreement with measured data in various conditions, including stationary and fluctuating interferers, reverberation, and spectral subtraction. Despite the advances in modeling intelligibility in NH listeners, a broadly applicable model that can predict SI in hearing-impaired (HI) listeners is not yet available. As a first step towards such a model, this study investigates to what extent effects of hearing impairment on SI can be predicted

[a] This chapter is based on Scheidiger, C., Jørgensen, S., Dau, T. (2014).

using the sEPSM framework. Our results indicate that, by only modeling the loss of audibility, the model cannot account for the increased speech reception thresholds (SRT) of HI listeners in stationary noise compared to NH listeners. However, this approach can, to some extent, account for the reduced ability of HI listeners "to listen in the dips" of fluctuating noise as compared to stationary noise. The results further indicate that effects of an outer hair-cell (OHC) loss (e.g., broader filters) on SI cannot easily be accounted for by this model. These limitations are discussed and alternative solutions are sketched out.

3.1 Introduction

Most communication between humans is based on speech. A loss of one's ability to communicate using speech can have a detrimental effect on one's social life. Hearing aids can help regain the ability to hear speech; however, compared to the normal-hearing (NH) system, their performance is suboptimal, especially in challenging acoustical environments, such as noisy rooms. In order to resolve this problem, a better understanding of how humans perceive speech is necessary. Modeling speech perception in HI listeners may represent one way to achieve a more detailed understanding as well as a powerful tool for the development and evaluation of hearing aids.

3.1.1 Speech-based envelope power spectrum model (sEPSM)

SI depends on the acoustic properties of the sound entering the ears and the auditory system's processing of the acoustic waveform. Macroscopic models of speech recognition predict average intelligibility scores obtained for a large

amount of speech material. They differ from microscopic models which consider details of the speech signal (e.g., onsets and transitions) and predict the intelligibility of small units of speech, such as phonemes. Macroscopic attempts to model SI have a long history. A first such model, called the Articulation Index (AI, ANSI S3.5, 1969), was developed in the early 1920s by researchers at the Bell Labs who tried to improve SI in the American telephone network (French and Steinberg, 1947; Allen, 1996b). The AI only considers effects of energetic masking in crude auditory bands and its internal representation relies on signal-to-noise ratios (SNRs) in 1/3 octave bands. It has been shown that the predicitive power of the AI is limited in conditions with temporal distortions (e.g., reverberation). Subsequent models, such as the SI index (SII, ANSI S3.5, 1997) and the speech transmission index (STI, IEC 60268-16, 2003), accounted for a larger range of conditions as compared to the AI. Whereas the AI and SII are based on the long-term spectrum of the stimulus and the noise, an extended version of the SII (ESII, Rhebergen et al., 2006) reflects short-term effects and can thus account for SI in fluctuating noise. In contrast to the AI, SII, and ESII models, the STI utilizes a modulation-frequency selective analysis to consider the change in the amplitude modulation of a processed (e.g., noisy) speech signal with respect to the clean speech signal. The STI is therefore able to account for reverberant conditions. However, all these models fail in conditions with non-linear stimulus processing (e.g., spectral subtraction).

Recent studies have predicted SI for NH listeners in a broader range of conditions by using an SNR measure in the envelope domain of the acoustical signal (SNR_{env}, Jørgensen and Dau, 2011; Jørgensen et al., 2013). Instead of measuring the reduction of the envelope modulations with respect to the clean speech (as used in the STI), Jørgensen and Dau (2011) estimated the power ratio of the envelope modulations in the noisy speech and the envelope modulations

of the corrupting noise. The model obtains the envelope modulation power by filtering the Hilbert envelopes of the time domain output of the peripheral filters with a set of modulation filters ranging from 1-64 Hz (Jørgensen and Dau, 2011). A high ratio indicates a high potential of a particular speech segment to contribute to intelligibility. This potential of providing intelligibility is averaged over a whole speech signal and integrated over all modulation channels as well as peripheral channels in order to predict the average recognition score. Jørgensen and Dau (2011) showed that their results were consistent with the STI predictions in "STI-friendly" conditions (e.g., reverberation and additive noise). Furthermore, the SNR_{env} metric correctly predicted SI in the condition of spectral subtraction (noise reduction). In this condition, the AI and STI predict improvements from the noise reduction algorithm, whereas listeners do not experience any benefit.

The sEPSM is based on a long-term integration of the envelope power and therefore fails to predict increased intelligibility in the case of fluctuating maskers. In order to compensate for this limitation, Jørgensen et al. (2013) extended the model to a multi-resolution version, which analyses the stimuli in short time frames, the duration of which is inversely related to the cut-off frequency of the corresponding filter. Furthermore, the modulation filterbank was extended by two filters centered at 128 and 256 Hz, respectively. The extended model has been shown to accurately predict SRTs in various types of background noise, in particular stationary speech-shaped noise (SSN) as well as various fluctuating interferers, including the fluctuating background noises used in the present study (see Sec. 3.2). Despite the advances in modeling intelligibility in NH listeners, a broadly-applicable model of SI in HI listeners has not yet been presented. The present study presents a framework to model SI in HI listeners based on the above described SNR_{env} measure.

3.1.2 Hearing impairment

An individual hearing loss is often described by an attenuation component as well as a supra-threshold distortion component (Plomp, 1978). The attenuation, represented by increased pure-tone detection thresholds, is able to account for much of the between-listener variance of SI in quiet. For speech perception in noise, the distortion component is thought to be responsible for the difficulties experienced by HI listeners. The distortion component has usually been attributed to reduced temporal and spectral resolution or to a deficit in temporal fine structure (TFS) processing. The impact of these components on SI has been highly debated.

A reduced spectral resolution in HI listeners compared to NH listeners is strongly supported by psychoacoustic data (e.g., Glasberg and Moore, 1986). It may be thought of as a broadening of the auditory filters, which could smear the spectral details of the internal representation of speech in the auditory system. In addition, broader auditory filters may increase the noise power falling into an auditory filter, resulting in more masking of salient speech cues. Even though NH studies have shown a decreased SI for spectrally smeared speech (Baer and Moore, 1993), data to support a similar relationship between SI and spectral resolution in HI listeners have been ambiguous (Buss et al., 2004; Summers et al., 2013; Strelcyk and Dau, 2009).

The above described research has established a wealth of knowledge about how a hearing loss affects psychoacoustic measures. However, the mechanisms underlying the individual listener's deficit to perceive speech in a given situation are still not clear.

3.1.3 Release from masking in speech percpetion

Various studies demonstrated a reduced MR for HI listeners (Festen and Plomp, 1990; George et al., 2006; Lorenzi et al., 2006; Bernstein and Grant, 2009; Strelcyk and Dau, 2009). It has been argued that the reduced frequency selectivity may partly be responsible for the reduced, or absent, MR in HI listeners. However, a correlation between MR and frequency selectivity could not be established (George et al., 2006; Strelcyk and Dau, 2009).

Reduced MR has also been linked to a decreased temporal resolution. A correlation between temporal resolution and SI in fluctuating background has been established in several studies (Hou and Pavlovic, 1994; Dubno et al., 2003; George et al., 2006). The degree to which reduced temporal resolution is responsible for the reduced MR remains unclear.

Christiansen and Dau (2012) compared HI listeners' MR with NH listeners' MR measured with vocoded stimuli (i.e., stimuli with degraded TFS) and found a good correspondence between the two. This finding supports the salient role of TFS for MR. However, it is questionable if TFS processing also affects speech reception in other (non-speech) fluctuating backgrounds. Strelcyk and Dau (2009) argued that TFS provides MR in condition where streaming (i.e., separating multiple talkers) is important but does not help in separating the speech from random noise.

Oxenham and Simonson (2009) reported a decreasing MR as a function of the stationary SRT of NH listeners, in an experiment where SRTs were decreased by high- and low-pass filtering the speech. Inspired by Oxenham and Simonson (2009), Bernstein and Grant (2009) argued that the decreased MR in HI listeners might only be due to the increased stationary SRTs (see Fig. 3.1), i.e., that MR in HI listeners was equivalent to MR in high- or low-passed conditions in NH listeners. They supported their argument by calculating MR for NH and HI

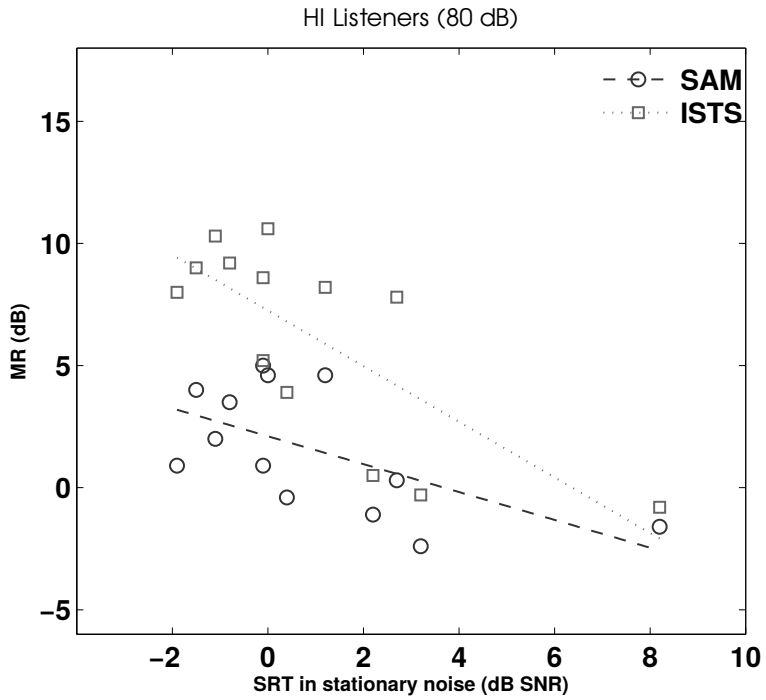


Figure 3.1: Masking relaease of HI listeners as function of their speech reception thresholds in stationary noise as measured by Christiansen and Dau (2012).

listeners at the same SNR (i.e., at different points on the psychometric function: high percentage correct for NH listeners and low percentage correct for HI listeners) and found that the HI listeners exhibited only slightly decreased MRs compared to NH listeners. Christiansen and Dau (2012) found the MR of NH listeners for differently processed stimuli (i.e., vocoding, high and low-pass filtering) to strongly depend on the type of processing, which contradicts the view that MR is only depended on the stationary SRT.

Thus, there still exist different hypotheses as to why HI listeners have a decreased MR. In conclusion, the findings suggest that both impaired TFS processing and the loss of audibility may be important to account for the reduced MR in HI listeners (Christiansen and Dau, 2012). A modeling approach to test the different hypotheses may shed light on which deficits of the impaired system are responsible for the reduced MR and to what degree.

3.2 Methods

The MR data from Christiansen and Dau (2012) were simulated with different versions of a modified sEPSM model. The simulations were run for three different noise types. One of them was a stationary speech-shaped noise (SSN), the other two were fluctuating interferers: An 8-Hz sinusoidally amplitude-modulated speech-shaped noise (SAM) and the international speech test signal (ISTS; Holube et al., 2010b). The ISTS consists of recordings of female talkers speaking six different languages that are segmented into short segments and recombined in random order.

3.2.1 Model A: Audibility Loss

The models used in the present study represent extensions of the sEPSM model as proposed by Jørgensen et al. (2013). The original model only processes auditory bands whose root mean square (RMS) power exceeds the diffuse-field hearing threshold in quiet (ISO 389-7, 2005). In the current study, an audibility loss of the model was incorporated by adding the hearing thresholds of individual HI listeners to the internal threshold of the model, such that the number of bands processed by the model decreased.

3.2.2 Model F: Reduced frequency selectivity

Model F incorporates broader auditory filters, simulating the decreased spectral-resolution typically observed in HI listeners. Auditory filter bandwidths can, for example, be estimated by time-consuming psychoacoustic tests or through otoacoustic measurements (Glasberg and Moore, 1990; Shera et al., 2002). For the current study, no such data were available. Filter bandwidths were instead estimated from a fitted function obtained from Fig. 3.23 in Moore (2007), which

depicts filter-bandwidths as a function of hearing loss in dB. This fit served as a starting point for the simulation and is not meant to accurately represent filter bandwidths in individual listeners. The broadening factor b was obtained by:

$$b(f) = HL(f) * (3.5-1)/60 + 1. \tag{3.1}$$

It represents the ratio between the filter bandwidth of a NH listener and the bandwidth of a listener with a hearing loss of HL at frequency f. With this formula, a HL of 70 dB leads to four times broader filters as compared to NH filters.

3.2.3 Model AF: Audibility loss & reduced frequency selectivity

Model AF combines the models A and F and simulates increased audibility thresholds in combination with broader filters.

3.3 Results

In the simulations, the masking release obtained with the two fluctuating interferers (i.e., SAM and ISTS) was calculated by taking the difference between SRT_{SSN} and SRT_{SAM} or SRT_{ISTS}, respectively. A positive value indicates that the listener was able to take advantage of the dips in the background. Three psychometric functions, as seen in Fig. 3.2 (example for Model A), were simulated for all 26 ears (i.e., 13 HI listeners). The two ears of a listener were combined by selecting the better ear (i.e., the ear with the lower SRT). It can be seen that the predictions for ear HI01L show a small benefit from “listening in the dips” in the case of the SAM interferer and a larger benefit in the case of the ISTS interferer.

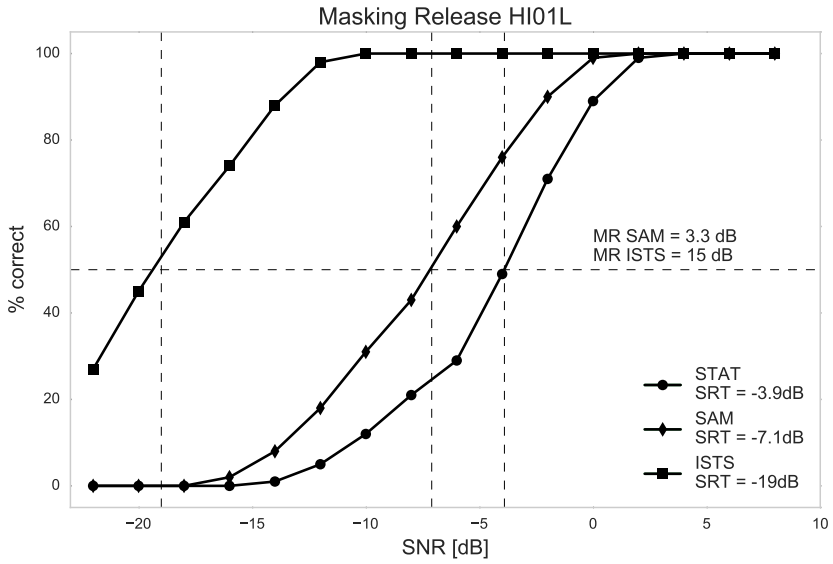


Figure 3.2: Word recognition score predictions as function of SNR for the left ear of HI listener 01 in SSN, SAM and ISTS noise obtained from Model A.

3.3.1 Model A: Audibility

In order to compare the model simulations to the measured data of Christiansen and Dau (2012), the model predictions are shown as a function of the measured data. Figure 3.3 depicts four scatter plots: Three of them show SRTs in SSN (upper left panel), SAM (upper right panel), and ISTS noise (lower left panel), respectively. The lower right panel shows the MR, i.e., the difference between the SAM/ISTS SRTs and the SSN SRT. The scatter plot for the SRTs in SSN noise (upper left panel) shows that this audibility-loss based model fails to account for the difficulties that HI listeners experience in the stationary masker: All but one predicted SRTs were too low (i.e., the predicted SI was too good). The SRTs for SAM noise (upper right panel) exhibited a similar trend, as all predictions were lower than the measured SRTs. However, despite the model's global underestimation of the SRTs, the trends across listeners for SSN and SAM noise were quite well predicted (Pearson's correlation of 0.27 and 0.74, respectively).

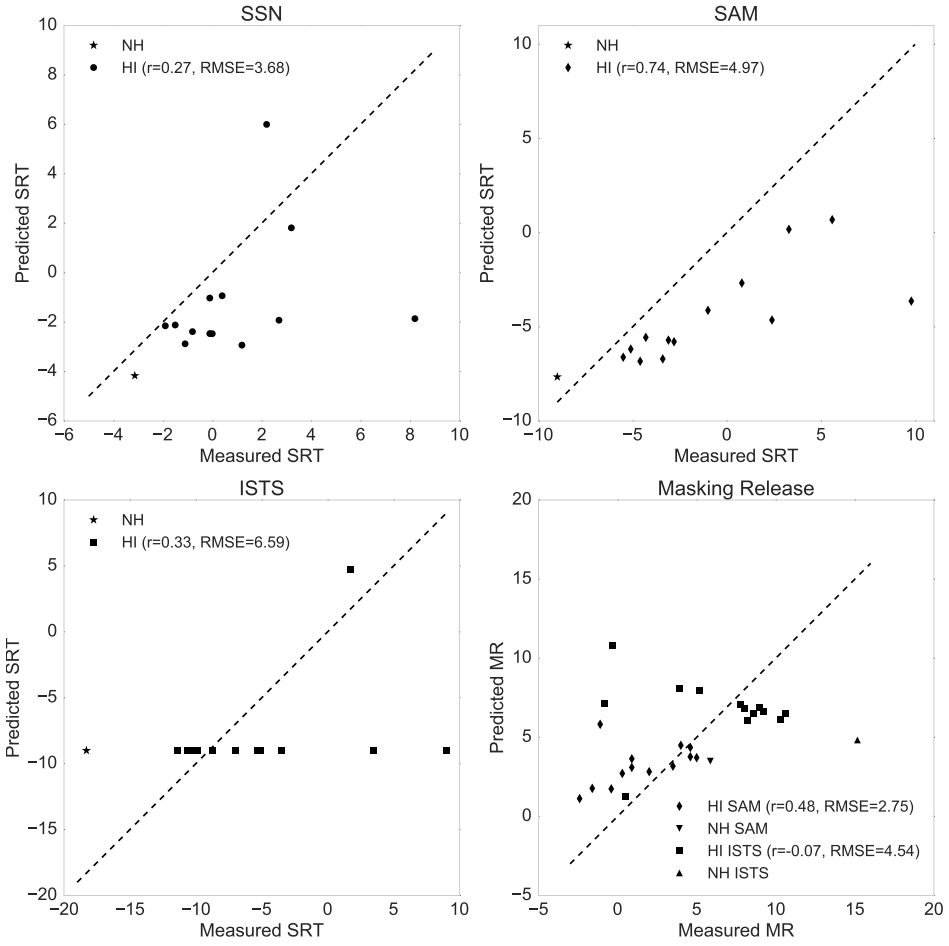


Figure 3.3: Measured versus predicted SRTs for individual HI listeners in SSN (top left), SAM noise (top right), and ISTS noise (bottom left), along with measured versus predicted MR for SAM and ISTS noise (bottom right). Predictions obtained using the mr-sEPSM with audibility loss, Model A. If the model were to predict the data correctly, all points would lay on the diagonal dashed line.

In the case of the ISTS SRTs (lower left panel), it can be seen that the model failed completely, as it predicted NH performance for all (but one) HI listeners. Interestingly, the model is able to account for the MR measured in the SAM noise (diamonds in lower right panel). This is surprising, since MR data are by definition, based on the SSN SRTs. However, the model overestimates the MR of the HI listeners in the case of the ISTS noise (squares in the lower right panel). Figure 3.4 shows the predicted MRs as a function of the SSN SRTs for SAM and

ISTS noise. The simulations (Fig. 3.4) show a decrease of MR with increasing SSN SRT similar to the data in Fig. 3.1. However, the predicted SSN SRTs are higher than the coresponding data.

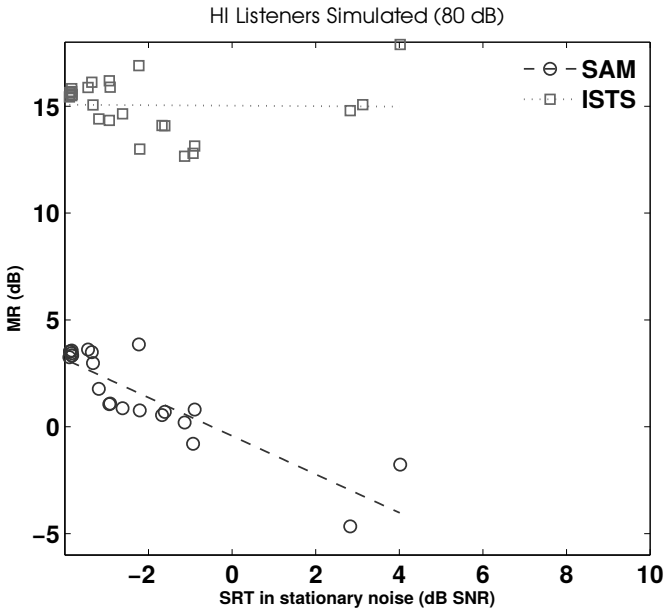


Figure 3.4: Masking release of HI listeners as function of their stationary speech reception threshold as predicted by the mr-sEPSM with audibility loss, Model A (*cf.* Fig. 3.1).

3.3.2 Model F: Reduced frequency selectivity

Figure 3.5 depicts the simulation results for Model F. The four panels show the predictions as a function of the measured data for the 13 HI listeners. The upper left panel shows the scatter plot for the SRTs in stationary noise. In the upper right panel, the SRTs for SAM noise are shown. The SRTs for the ISTS noise interferer are depicted in the lower left panel. Lastly, the differences between SRTs in stationary noise and fluctuating noise, i.e., the MRs, are shown in the lower right panel. The measured SRTs for SSN, SAM, ISTS show a large across-subject variance. The predicted SRTs do not cover this range; in fact, they do not change as a function of the bandwidths factors $b(f)$ applied for the individual

listeners, implying that the SRTs for the HI listeners remain unchanged when compared to the NH predictions. This also explains the simulated MRs observed for this model, which amount to about 4 dB, irrespective of the noise type, and thus show no correlation with the measured values.

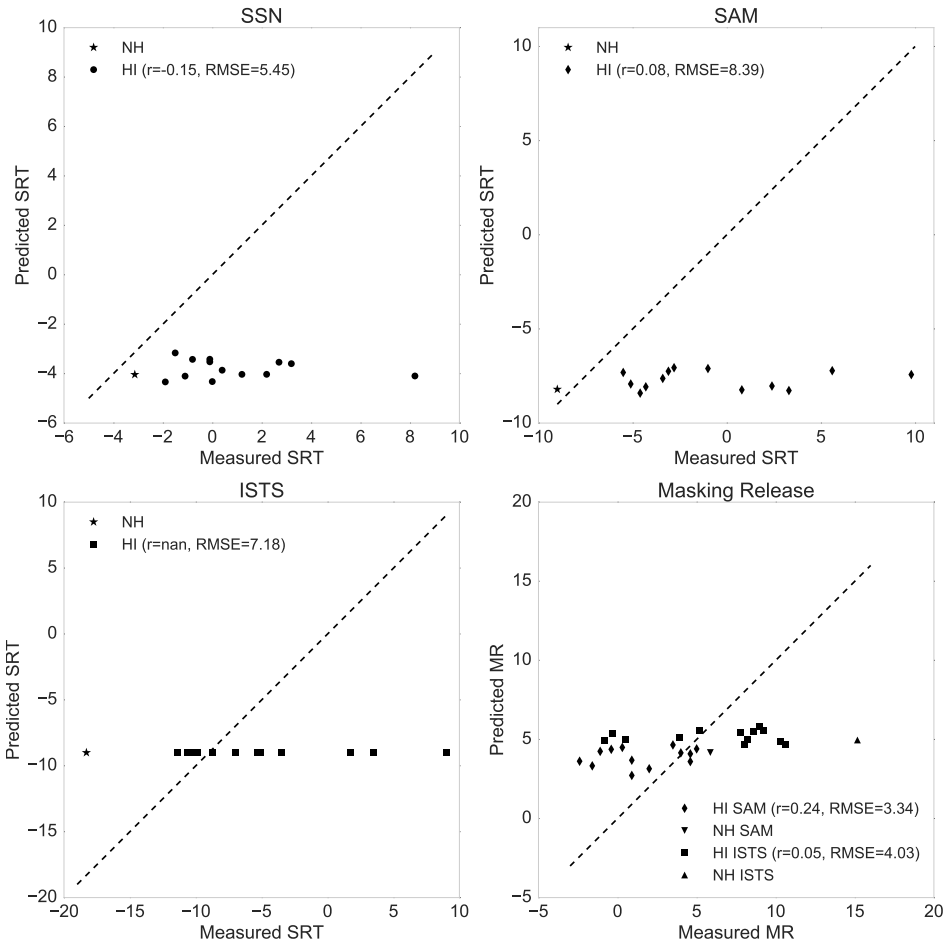


Figure 3.5: Measured versus predicted SRTs for individual HI listeners in SSN (top left), SAM noise (top right), and ISTS noise (bottom left), along with measured versus predicted MR for SAM and ISTS noise (bottom right). Predictions were obtained using the mr-sEPSM with broader filters, Model F.

3.3.3 Model AF: Audibility & frequency selectivity

Figure 3.6 depicts the simulation results for Model AF. The measured data are identical to those shown in Fig. 3.5. Compared to Fig. 3.5 it can be seen that the spread of the predictions is greater. The audibility component clearly dominates the predictions (cf. Fig. 3.3). Even though broader filters, considered in isolation, did not have a noticable effect on the MR predictions (*cf* Fig. 3.5), they decreased the accuracy of the MR predictions when applied in combination with an audibility threshold (lower right panel Fig. 3.6).

3.4 Discussion

3.4.1 Predictive power of different model versions

Increasing the filter bandwidths in Model F did not account for any performance deficits observed in the HI listeners; instead, Model F predicted SRTs in the range of the NH SRTs, irrespective of the filter bandwidths (*cf* third row in Table 3.1). Model AF, which incorporated broader filters as well as a loss of audibility, predicted the SRTs less accurately than Model A, i.e., the broader filters decreased the predictive power of the model. Overall, the influence of the broader filters seems to be small, despite the assumption of up to four times broader filters. Other modeling studies have also argued that broader auditory filters do not represent a strong factor affecting SI in HI listeners (Kollmeier et al., 2016).

The predictions of the ISTS SRTs are generally insensitve to the hearing-loss simulations incorporated into the models, i.e., the predicted SRTs for HI listeners were similar to the SRTs for NH listeners, except for one outlier in the models A and AF (*cf* fourth column in Table 3.1). The predicted SRT values are also generally too high. The ISTS signal consists of different female talkers

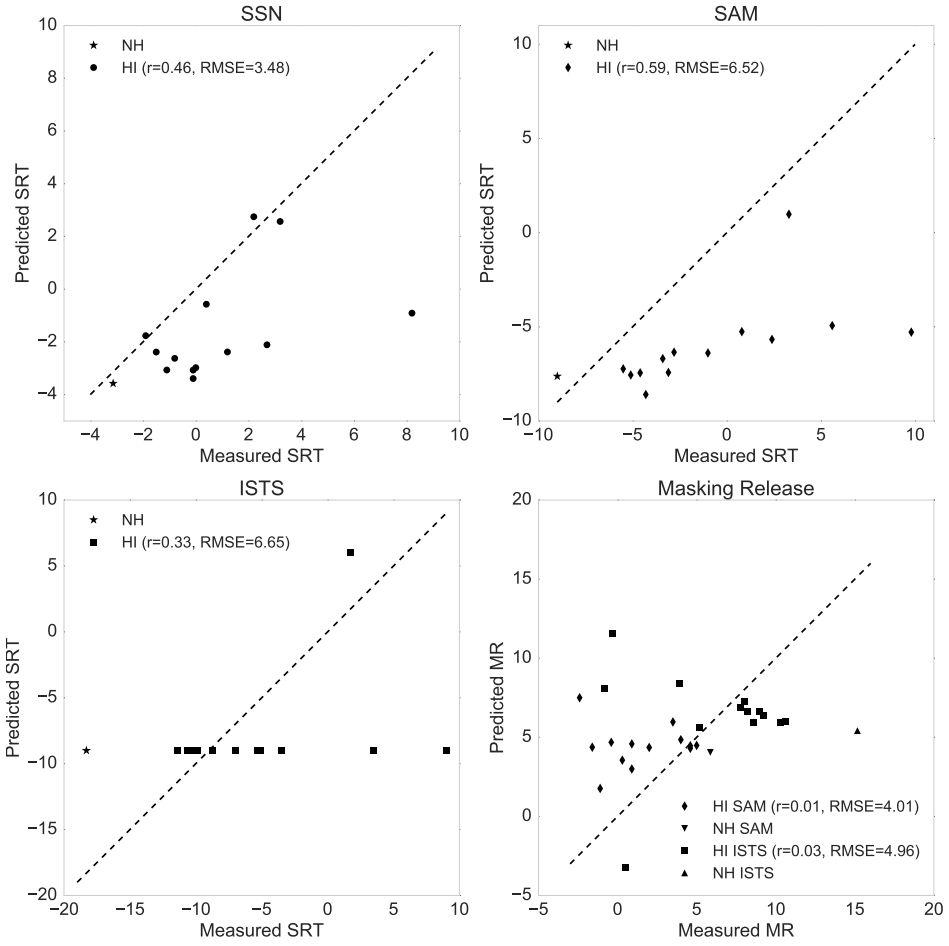


Figure 3.6: Measured versus predicted SRTs for individual HI listeners in SSN (top left), SAM noise (top right), and ISTS noise (bottom left), along with measured versus predicted MR for SAM and ISTS noise (bottom right). Predictions were obtained using the mr-sEPSM with audibility loss and broader filters, Model AF.

speaking six different languages. The average fundamental frequency (F0) of the six talkers is 214 Hz (Holube et al., 2010b), whereas the F0 of the target speech is 119 Hz (Christiansen and Dau, 2012). These differences represent important cues to separate the different sources into streams. However, a simple HI model solely based on envelope cues, as the one presented in this study, is blind to streaming effects caused by F0 differences between talkers. Such streaming effects may rely on TFS information which is not represented in the model.

The models used in this study are therefore likely to predict higher SRTs than measured in listeners.

Table 3.1: The summary of Pearson correlation coefficents shows that correlations are low in general. Model A predicts the SRTs in SAM noise the best. Model AF exhibits a higher correlation for the SSN noise than Model A does.

r, RMSE	SSN	SAM	ISTS	MR SAM	MR ISTS
Model A	0.27, 3.68	**0.74, 4.97**	0.33, 6.59	**0.48, 2.75**	-0.07, 4.54
Model F	0.15, 5.45	0.08, 8.39	nan, 7.18	0.24, 3.34	0.05, 4.03
Model AF	**0.46, 3.48**	0.59, 6.52	0.33, 6.65	0.01, 4.01	0.03, 4.96

3.4.2 Prediction of masking release in HI listeners

As seen in Fig. 3.3 (lower right panel), the MRs in SAM could be accounted for by the audibility-loss based model (A), except for one outlier, despite the inaccurate predictions of SSN SRTs on which the SAM MR values are based. This suggests that both the SSN SRTs and SAM SRTs predictions show an offset of about the same value. It also suggests that audibility alone might be able to account for the reduced MR in SAM noise. This is consistent with results from other studies that observed a restored masking release in HI subjects when audibility was restored in short time windows rather than on a long-term basis (Desloge et al., 2010; Reed et al., 2016).

However, a clear limitation of the presented modeling results is that they do not correctly describe the measured SRTs in HI listeners (*cf* columns two to four in Table 3.1). The MR predictions are by definition a function of the SRTs in SSN; the inaccurate SRT predictions in SSN therefore also question the validity of the MR estimates. Better estimates might be obtained by using more complex hearing loss simulations (i.e., adding temporal smearing, decreased temporal resolution) or by changing the back-end processing, e.g., the integration of peripheral and modulation channels.

3.4.3 Model limitations

Modeling a hearing loss based on the linear gammatone model represents a crude approximation. The human auditory system is highly nonlinear and a hearing loss affects these non-linearities. Modeling a hearing loss without considering non-linearities omits an important functional aspect of the system. Given the physiological impairment, i.e., outer and inner hair-cell loss, it may make sense to use a physiologically realistic model of the periphery instead of incorporating psychoacoustic deficits (i.e., audibility loss or broader filters) seperately. Other studies have shown that the non-linearities are crucial in predicting a change in SI for HI listeners (Hossain et al., 2016). Furthermore, to account for the lower SI of listeners at higher sound pressure levels, known as roll-over effect, the non-linearities also play a crucial role (Studebaker et al., 1999).

Outer hair cells (OHC) are the active parts in the cochlea responsible for level compression. Studies indicate that outer hair-cell loss is associated with broader auditory filters and a loss of compression (e.g., Ruggero and Rich, 1991; Strelcyk and Dau, 2009). Modeling these two aspects separately may thus not be reasonable; using a front end that links these two factors might yield more realistic SI predictions.

3.5 Summary and conclusion

The effects of a sensorineural hearing loss on speech intelligibility are still being investigated. Different hypotheses about how psychoacoustic measures, such as temporal and spectral resolution as well as a deficit in TFS processing, are related to SI have been presented in the literature. The modeling results described in the present study suggest that the reduced MR in HI listeners in SSN noise might

be accounted for by loss of audibility alone, whereas the MR obtained in the presence of a competing talker cannot be described by reduced audibility alone, but is mainly a consequence of a distortion loss.

Acknowledgments

The work leading to this study and the results described therein has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement number PITN-GA-2012-317521.

4

Estimating SNRenv based on Auditory Nerve Firing Rates[a]

Abstract

Speech intelligibility (SI) models aim to predict the human ability to understand speech in adverse listening conditions. However, most current speech intelligibility models are based on a strongly simplified simulation of the auditory periphery, which limits their ability to predict effects of hearing impairment on SI. The goal of the present study was to combine an established speech intelligibility model with the auditory signal processing of an auditory-nerve (AN) model. Specifically, the back-end processing of the multi-resolution speech-based envelope power spectrum model (mr-sEPSM; Jørgensen et al., 2013) was combined with the AN model by Zilany et al. (2014). Signal-to-noise-ratios in the envelope domain (SNR_{env}) were calculated for normal-hearing listeners based on envelope representations derived from the AN model's instantaneous firing rates. The SI predictions showed good agreement with human data when the model was operated at a sound pressure level of 50 dB assuming only medium-spontaneous-rate fibers. However, when all fiber types and presentation level of 65 dB SPL were considered,

[a] This chapter is based on Scheidiger et al., (in preparation).

the model overestimated SI in conditions with modulated noise interferers. A modulation-frequency range analysis showed that these prediction errors mostly resulted from high-frequency modulation channels, indicating that a reduction of the modulation-frequency range considered in the model may be advantageous.

4.1 Introduction

Speech intelligibility (SI) is the measure of how well speech is understood as a function of adversity of a given listening situation. It is quantified as a percentage of correctly identified speech units, e.g., number of correctly identified words in a sentence. SI can be measured by listening tests with human listeners. Models of SI aim to predict the results of listening tests by means of a computer model, typically using the speech stimulus (e.g., speech in noise) as well as an additional reference signal (e.g., the clean speech or the noise alone).

SI models based on the signal-to-noise ratio in the envelope domain (SNR_{env}) have been shown to yield accurate predictions in a wide range of listening conditions, e.g., stationary and fluctuating background noises (Jørgensen and Dau, 2011; Jørgensen et al., 2013). The SNR_{env} framework utilizes Hilbert envelopes derived from the output of a bank of gammatone filters, analyzing the temporal fluctuations of these envelopes using a modulation filterbank. In each modulation filter, the envelope power of the noisy speech, i.e., the mixture of target speech signal and interfering noise signal, is compared to the power of the noise signal by the means of a signal-to-noise ratio (i.e., SNR_{env}). The higher the SNR_{env}, the greater the potential contribution of the corresponding channel to SI. While this approach yields a powerful decision metric for predicting SI measured at medium levels in normal-hearing (NH) listeners, it is limited in

accounting for effects of presentation level and hearing impairment (HI). Consequently, attempts to incorporate a hearing impairment have led to mixed results (see Chapter 3). These limitations are, at least partially, related to simulating the cochlear processing by means of a linear gammatone filterbank, which represents a simplification of the highly complex and non-linear functionality of the cochlea. To overcome this functional limitation, the present study explores the SNR_{env} concept in combination with a non-linear, physiologically inspired model of the auditory periphery using envelopes derived from firing rates of an auditory nerve (AN) model.

Other studies have used AN models to predict speech intelligibility. Zilany and Bruce (2007) assessed the spectrogram-like firing patterns, i.e., neurograms, from multiple AN fibers along the basilar membrane with the spectro-temporal modulation index (STMI), a model that assesses SI based on how the auditory cortex processes spectro-temporal ripples (Elhilali et al., 2003). The clean speech signal at a level of 65 dB serves as a reference template. The deviation, i.e., the distance as assessed by a L2-norm, of a noisy signal from this reference is inversely correlated to SI. Zilany and Bruce (2007) predicted SI for both NH and HI listeners by adjusting the inner and outer hair-cell loss factors of the AN model for individual subjects. This AN-model based STMI qualitatively matched the word recognition scores of NH and HI subjects for low- and high-pass filtered speech, as well as presentation-level effects at three different signal-to-noise ratios (SNRs) in stationary noise. However, the study did not address the influence of fluctuating maskers on SI and only tested a limited range of SNRs. Furthermore, all predictions were averaged over a group of listeners with similar audiograms while the accuracy of individual predictions was not assessed.

Bruce et al. (2013) extended the AN-based SI model approach to also account for fluctuating noise maskers. Instead of using the STMI model to process

the neurograms, they used the Neurogram Similarity (NSIM) metric (Hines and Harte, 2010; Hines and Harte, 2012). The NSIM metric compares a neurogram for a given condition to a template neurogram obtained for the NH configuration of the AN model at 65 dB presentation level in quiet. It compares the two neurograms, which it treats as images, based on the mean (luminance), the variance (contrast) and the correlation (structure) of pixel values in time-frequency segments. The model predictions were compared to perceptual data from Léger et al. (2012), where NH and HI listeners were presented with low-pass and band-pass filtered VCV stimuli in noise. The NSIM model yielded better predictions than the extended speech intelligibility index (ESII, Rhebergen and Versfeld, 2005; Rhebergen et al., 2006).

Both the STMI and NSIM models use reference signals to asses the degradation of a speech signal. In contrast to these approaches, Hossain et al. (2016) proposed a reference-free model based on neurograms derived from an AN model. Third-order statistics obtained from neurograms were used to predict SI. These statistics, also referred to as bispectrum, capture the extent of phase coupling between frequency components. Compared to second-order statistics of the signal (e.g., the power spectrum or the autocorrelation), bispectra take phase information into account. This allows to account for any changes to the nonlinearities in the periphery due to a hearing loss. Phoneme and word recognition test data from Studebaker et al. (1999) were used for verification. The model successfully accounted for effects of presentation level, hearing loss, audibility, and additive stationary noise. The proposed metric could account for SI of phonemes and words, but failed to predict sentence recognition. The reference-free approach is also bound to fail in conditions with speech-like maskers, as the model has no means to distinguish between masker and target. Furthermore, while the bispectrum is a powerful signal processing approach,

there is no physiological or behavioral evidence that the brain uses bispectra-like features.

The AN based SI models described above are all limited to predicting speech intelligibility of phonemes or words as opposed to whole sentences. Furthermore, the mentioned studies considered only stationary masking noise, except for Bruce et al. (2013), where SI was predicted in both stationary and fluctuating noise. However, Bruce et al. (2013) compared only two noise types and neither of them was speech-like. In the present study, the same AN model as in Hossain et al. (2016) was used to predict SI of sentences mixed with stationary noise, sinusoidally amplitude-modulated noise, and a speech-like interferer in NH listeners. The proposed metric in this study extends the SNR_{env} framework used in the speech-based envelope power spectrum models (sEPSM, Jørgensen and Dau, 2011; Jørgensen et al., 2013) towards accounting for effects of presentation level and hearing impairment.

4.2 Model description

The model proposed in the present study consists of two main stages, each subdivided into further stages. These two main stages are a peripheral front-end stage, represented by the AN model, and a decision stage, which converts the firing patterns at the output of the AN model to an SNR_{env} value and subsequently to a SI score (see Fig. 5.1). The noisy speech signal (SN) and the noise alone (N) serve as inputs to the model.

4.2.1 Front end: AN Model

The front end of the model consists of an auditory nerve model (Zilany et al., 2014). The responses of the model have been validated against a large dataset

of AN recordings from the literature (e.g., Carney, 1993; Bruce et al., 2003; Zilany et al., 2009; Zilany et al., 2014). The AN model consists of several stages, each providing a phenomenological description of a major part of the auditory periphery, starting at the middle ear and ending at a specific AN synapse. The input to the model is an acoustic waveform (in pascals) which is first processed by the middle-ear (ME) filter. After the ME filter, the signal is further passed through a basilar membrane (BM) filter. A feed-forward control path of the BM filter controls the gain and bandwidth to account for level-dependent properties in the cochlea, e.g., less amplification and broader filters at higher levels. The inner-hair-cell (IHC) stage converts the mechanical BM response to an electrical potential, which is further low-pass filtered before it is applied to the IHC-AN synapse model. The synapse model determines the spontaneous rate, adaptation properties, and rate-level behavior of the AN model. In a last stage, the spike times are generated by a non-homogeneous Poisson process that includes refractory effects. The working of the IHCs and the outer hair cells (OHCs) in the model may be adjusted by specifying the amount of IHC and OHC damage.

4.2.2 Back end: sEPSM

The back-end is inspired by the multi-resolution speech-based envelope power spectrum model (mr-sEPSM; Jørgensen et al., 2013), which predicts SI based on the SNR_{env} metric. The model is not meant to be an accurate functional representation of the physiological stages of the auditory pathway. It rather models the behavioral modulation sensitivity of the human auditory system with a modulation filterbank (Dau et al., 1997; Dau et al., 1999). The average firing rate of each simulated characteristic frequency (CF) is filtered by this modulation filterbank. In each modulation band, specified by the center fre-

quency of the filter f_m, the envelope power $\left(P(CF, f_m)\right)$ of the input signals, i.e., the noisy speech mixture (SN) and the noise signal (N), is estimated in time windows. The duration of these time windows is defined as the inverse of the respective modulation filter's center frequency (f_m). Thus, short time windows are applied for high-frequency and long time windows for low-frequency modulation bands. The envelope powers $P_{SN}(CF, f_m)$ and $P_N(CF, f_m)$ are compared to each other by calculating the $\text{SNR}_{env}(CF, f_m)$:

$$SNR_{env}(CF, f_m) = \frac{P_{SN} - P_N}{P_N}. \tag{4.1}$$

The greater the ratio, the higher the contribution of the considered time window to SI. In a further step, the SNR_{env} values are averaged across time windows and integrated across all CFs and modulation bands using the root of a sum of squares. The resulting integrated SNR_{envTot} is lastly converted to a percentage correct using a fitting condition.

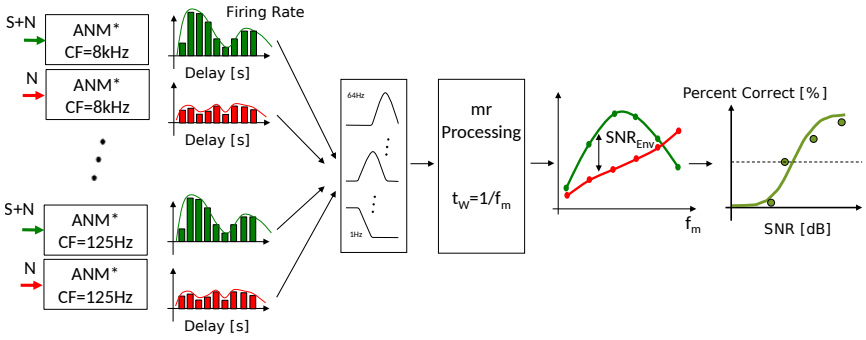


Figure 4.1: Structure of the proposed model: The auditory-nerve model estimates average firing rates for 21 CFs along the basilar membrane for both the noisy speech signal (SN) and the noise signal (N). The firing rates are further processed by a modulation filterbank. At the output of each filter the envelope powers P_{SN} and P_N are estimated in time windows and compared by means of a signal-to-noise ratio (SNR_{env}). The ratio is averaged across time and integrated across both CF and modulation filters. In a last stage, the integrated SNR_{envTot} is transformed to a percentage correct.

4.3 Methods

4.3.1 Speech and noise material

The stimuli used for the modeling work were similar to the speech stimuli used in Christiansen and Dau (2012). The stimuli consisted of natural meaningful Danish five-word sentences from the CLUE corpus (Nielsen and Dau, 2009). Ten sentences of the speech corpus were chosen for the simulations. All simulation results were averaged over these ten sentences in order to obtain stable predictions. The sentences were recorded from a male talker with an average fundamental frequency (F0) of 119 Hz. Behavioral data were available for three different interferers: (i) speech-shaped noise (SSN), a stationary masker with a long-term spectrum identical to the average of all sentences in the study, (ii) an 8-Hz sinusoidally amplitude-modulated (SAM) speech-shaped noise, and (iii) the international speech test signal (ISTS; Holube et al., 2010a) which consists of natural speech from six female talkers speaking different languages. The ISTS signal was created by truncating recorded sentences into segments and randomly remixing the segments, yielding a largely unintelligible signal with natural speech properties in terms of periodicity and modulation (average F0: 207 Hz). The measured speech reception thresholds (SRTs) in NH listeners, which were obtained directly using an adaptive procedure, were made available by Christiansen and Dau (2012). For the model simulations, the simulated input SNRs ranged from -21 dB in 3-dB steps up to 12 dB. This range covers the SRTs for all three noise types, which are at -18.3 dB, -9 dB, and -3.1 dB for the ISTS, SAM, and SSN interferers, respectively.

4.3.2 Model configurations

Different configurations of the front end and the back end of the model were tested in this study. The model front end always consisted of 21 CFs (125, 160, 200, 250, 315, 400, 500, 630, 800, 1000, 1250, 1600, 2000, 2500, 3150, 4000, 5000, 6300, 8000 Hz). If not mentioned otherwise, the back end's modulation filterbank consisted of the nine modulation filters used in the mr-sEPSM (Jørgensen et al., 2013), namely a third-order butterworth low-pass filter with cut-off frequency of 1 Hz, and band-pass filters with center-frequencies of 2, 4, 8, 16, 32, 64, 128, and 256 Hz. For all configurations, the SNR_{envTot} values obtained in the SSN condition were used for fitting. A non-linear fitting algorithm was applied to iteratively estimate the two parameters a_1 and a_2 of the following logistic function in order to convert SNR_{envTot} to SI:

$$SI(SNR) = \frac{100}{1 + e^{a_1 SNR_{envTot}(SNR) + a_2}}, \tag{4.2}$$

where SI is the SI expressed as percentage correct and SNR_{envTot} is the total SNR_{env} integrated across all modulation channels and all CFs. In order to quantify the accuracy of the model predictions, the predicted SRT was obtained as the 50% point on the predicted psychometric function and the prediction error E was calculated as $E = SRT_{predicted} - SRT_{measured}$.

"Linear" operation mode

As a first step, the front end of the model was configured such that it behaves as similarly as possible to the linear gammatone filterbank used in the original mr-sEPSM model. This configuration was used to verify the successful coupling of the front end and the back end. In order to obtain a linear behavior in the AN model, only medium spontaneous-rate fibers were selected. Out of the three

available fiber types (LSR: low spontaneous rate; MSR: medium spontaneous rate, and HSR: high spontaneous rate), the MSR fibers show the shallowest rate-level curves, i.e., their average firing rate does not change as much as a function of level as for the other two types. Furthermore, the model was operated at an overall presentation level (OAL) of 50 dB sound pressure level (SPL), at which the BM filter does not exhibit any broadening due to the forward-feeding feedback loop described above. The SNR_{envTot} values obtained in the SSN condition were used for fitting according to Eq. 4.2.

"Realistic" operation mode

The human auditory system consists of HSR, MSR, and LSR fibers. For the realistic model configuration it was assumed that 60% of all fibers are HSR, 20% MSR, and 20% LSR (Zilany and Bruce, 2007). The OAL was set to 65 dB SPL, the same level at which the NH SI data were collected (Christiansen and Dau, 2012). To further evaluate the model behavior for lower presentation levels, an OAL of 50 dB SPL was additionally considered in the model simulations. The SNR_{envTot} values obtained in the SSN condition at an OAL of 65 dB SPL were used for fitting according to Eq. 4.2.

4.4 Results

"Linear" operation mode

Figure 4.2 shows the SI predictions (dots) as a function of the input SNR. The SI predictions decrease with decreasing SNRs. A psychometric function as defined by Eq. 4.2 was fitted to the predictions (dashed lines). From this function, the SRT (SI= 50%) was estimated. Note that the SSN condition (red) was used for fitting, such that the model predictions fit the data by definition for this

condition. However, it can also be observed for the other conditions (SAM, green; ISTS, blue) that the predicted SRTs line up with the SRTs measured in human listeners (squares). The SRT error E did not exceed 0.8 dB for any of the three interferers.

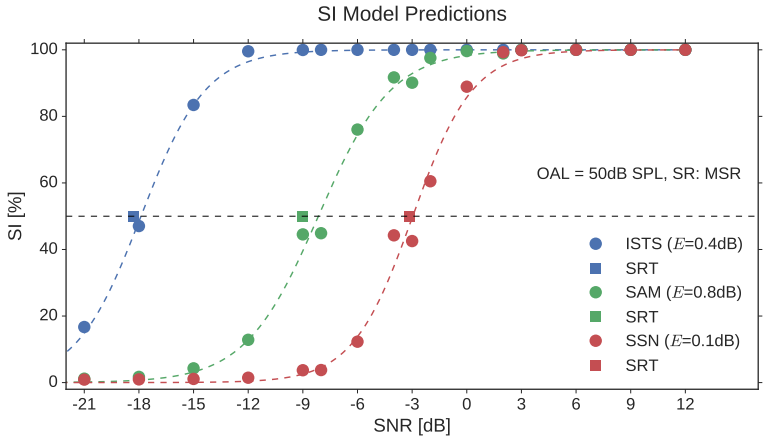


Figure 4.2: Simulated SI depicted as a function of the input SNR for the SSN (red), SAM (green), and ISTS (blue) interferers obtained with the model's "linear" operation mode (only MSR, 50 dB SPL). The round dots represent the simulated SI predictions for each input SNR. The dashed lines depict psychometric functions (Eq. 4.2) fitted to the predictions. The squares indicate the SRTs measured in NH listeners (at 65 dB SPL), as provided by Christiansen and Dau, 2012.

"Realistic" operation mode

Figure 4.3 shows the SI predictions as a function of the input SNR for the model's "realistic" operation mode (60% HSR, 20% MSR, 20% LSR, 65 dB). The measured SRT lines up with the predictions for the SSN interferer (red), which was to be expected as this is the fitting condition. However, the model strongly overestimated SI for the SAM (green) and ISTS (blue) interferers, leading to very low predicted SRTs. Compared to the measured SRTs, there was an offset of the predicted SRTs for SAM and ISTS of -8.7 dB and -5.4 dB, respectively.

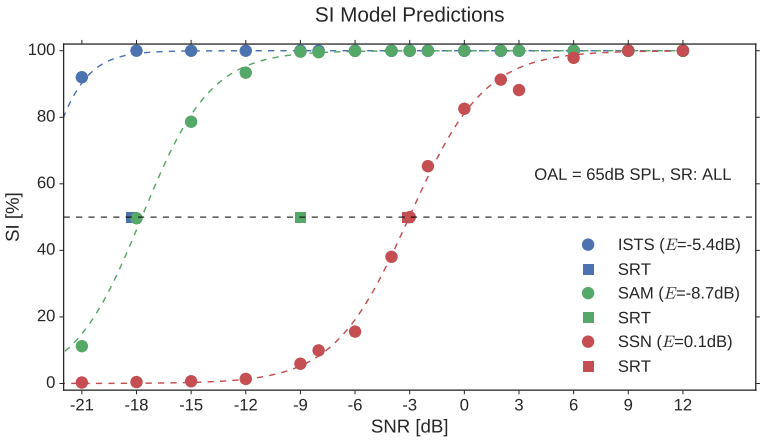


Figure 4.3: Simulated SI is depicted as a function of the input SNR for SSN (red), SAM (blue), and ISTS (green) obtained with the model's "realistic" operation mode (ALL SR, 65 dB SPL). The round dots represent the simulated SI predictions for each input SNR. The dashed lines depict psychometric functions (Eq. 4.2) fitted to the predictions. The squares indicate the SRTs measured in NH listeners, as provided by Christiansen and Dau, 2012.

Analysis of modulation-frequency range

Figure 4.4 shows the SRT error (E) as function of the number of modulation channels included in the analysis of the back end. For example, for the data point reflecting 8 Hz as the highest modulation channel, all modulation channels up to and including 8 Hz (i.e., 1, 2, 4, and 8 Hz) were used in the analysis. Panel A shows the performance of the different model configurations at an OAL of 50 dB SPL, while panel B depicts the same at an OAL of 65 dB SPL. It should be noted that the E for SSN at OAL = 65 dB SPL (red dots in panel B) stays quasi-constant at a value close to zero since this was the fitting condition. In contrast, the E for SSN at OAL = 50 dB SPL (red dots in panel A) shows deviations from 0 as the fitting was not adapted to the OAL, i.e., it was also based on the SSN results obtained at an OAL of 65 dB SPL. An $E < 0$, as depicted for SSN in panel A (red dots), indicates that the predicted SRTs were lower (i.e., SI better) than the measured SRTs, which is in contrast to what one might expect for a lower levels, such as 50 dB SPL.

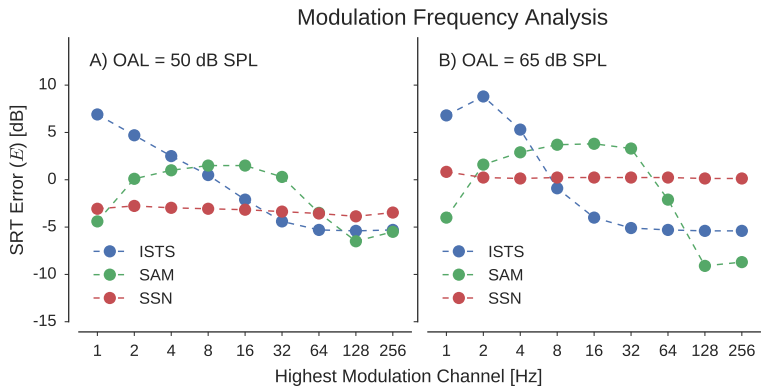


Figure 4.4: SRT error for the three noise interferers as a function of the highest modulation channel that was used in the analysis. Panel A) shows the error for an OAL of 50 dB, Panel B) for an OAL of 65 dB. The SSN interferer in Panel B) was the fitting condition.

For an OAL of 50 dB SPL, the errors for the SAM and ISTS interferers decreases with increasing number of included modulation channels. The error curves converge to $E \approx -3$ dB, the minimum, when all channels are included. Note that if SSN at an OAL of 50 dB would have been used as a fitting condition, instead of SSN at an OAL of 65 dB, all the errors in panel A would be shifted up by about 3 dB and the predictions would thus almost be perfect ($E \approx 0$ dB) when using all modulation channels. This is consistent with the excellent fit between the model predictions and the data observed in Fig. 4.2 for the "linear" operation mode, where an OAL of 50 dB SPL was used (in combination with only MSR fibers). However, for an OAL of 65 dB SPL, there is no clear convergence, implying that there is no model configuration that works for all three interferers. The smallest mean average SRT error (MAE) (i.e., the best model configuration) was obtained for a model with modulation channels ranging from 1 Hz to 8 Hz (MAE= 1.6 dB). The configuration with modulation channels ranging form 1 Hz to 64 Hz showed the second smallest error (MAE= 2.5 dB). In this configuration, the SRT in SSN and SAM is well accounted for, whereas the SRT for the ISTS interferer is predicted as too low (i.e., too good). Model configurations

with modulation channels of 128 and 256 Hz substantially over-predicted SRT differences between the different interferers, as also seen in Fig. 4.3.

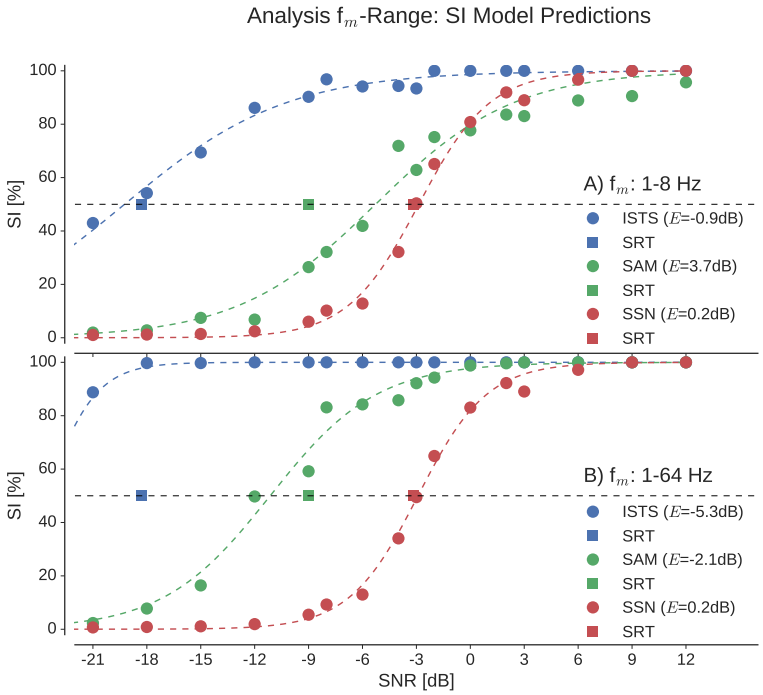


Figure 4.5: Simulated SI is depicted as a function of the input SNR for SSN (red), SAM (blue), and ISTS (green) obtained with the model's "realistic" operation mode (ALL SR, 65 dB SPL) for back ends with modulation channels ranging from 1-8 Hz (panel A) and 1-64 Hz (panel B). The round dots represent the simulated SI predictions for each input SNR. The dashed lines depict psychometric functions (Eq. 4.2) fitted to the predictions. The squares indicate the SRTs measured in NH listeners, as provided by Christiansen and Dau, 2012.

Figure 4.5 illustrates the simulated psychometric functions for the two model configurations that showed the smallest MAE in the "realistic" operation mode (ALL SR, 65 dB SPL). Panel A shows the simulations (dots) and the fitted psychometric function (dashed lines) for the model configurations that only considered the four lowest modulation channels, i.e., 1, 2, 4 and 8 Hz. The SSN condition (red; fitting condition) and the ISTS condition (blue) are well accounted for. The SRT for the SAM interferer (green) is predicted as too high (i.e., too bad). It can also be seen that the slopes of the pyschometric function are shallower than in Fig. 4.2 and 4.3. Panel B shows the predictions for the model with modulation

channels ranging from 1 to 64 Hz. In this configuration, the SSN and SAM condition are well accounted for, whereas the SRT in the ISTS condition is predicted as too low (i.e., too good).

4.5 Discussion

The model presented in this study represents an extension of the SNR_{env} framework. Instead of predicting SI based on the envelope power derived from outputs of a linear gammatone filterbank, the model derives the envelope power from instantaneous firing rates from an AN model. The model worked well for the "linear" operation mode, in which the AN model operates in its most linear way, by only considering the MSR fibers and by scaling down the OAL to 50 dB. This is in agreement with the results obtained with the original mr-sEPSM using the same stimuli. For the "realistic" operation mode in which all fiber types (LSR, MSR, and HSR) were considered at a realistic level of 65 dB, the model underestimated the SRTs for the modulated interferers, thus showing poorer performance.

An analysis of the modulation-frequency range revealed that mostly the two highest modulation channels (128 and 256 Hz) contributed to the deviation of the predicted SRTs from the measured SRTs in the "realistic" operation mode. Reducing the modulation-frequency range to 64 Hz (Fig 4.5, Panel B) yielded a model that was able to account well for the SSN and SAM interferer, however, the model underestimated the ISTS SRT by more than 5 dB. This deviation may partly be accounted for by the perfect streaming assumed in the model, which is represented by the fact that the model has a perfect internal representation of the noise signal. Even though the difference in F0 between the interferer and the target may help a human listener to segregate the two streams, the segregation

will likely not be perfect. This less-than-perfect segregation could lead to higher SRTs in human listeners as compared to the model.

In the "realistic" operation mode, the model configuration with modulation filters up to 64 Hz must be considered the most promising model configuration. This might seem counter-intuitive as it is only the second best model in terms of MAE, second to the model with modulation filters ranging from 1-8 Hz. However, the latter model configuration (1-8 Hz) exhibited unrealistically shallow slopes of the psychometric functions. Further evidence that the modulation-frequency range from 1-64 Hz might be the most appropriate was provided by the original sEPSM (f_m : 1 − 64 Hz; Jørgensen and Dau, 2011) and the STMI models (f_m : 2 − 32 Hz; Zilany and Bruce, 2007) that use the same or similar modulation frequency ranges. A possible explanation as to why the modulation-frequency range needed to be reduced as compared to the mr-sEPSM (f_m : 1 − 256 Hz; Jørgensen et al., 2013) is the adaptation in the front end considered in the present study (i.e., the AN model), which was not represented in the front end of the original model. The adaptation enhances transients/onsets and thereby acts as a modulation enhancer. This modulation enhancement leads to high SNR_{env}, especially in the higher modulation channels where fast-acting transients are represented. Whereas the mr-sEPSM benefitted from these higher modulation channels to account for the difference between steady and fluctuating interferers, these channels respond too strongly to the adaptaion in the front end of the proposed model.

In a further analysis using an OAL of 80 dB (not shown) and modulation filters from 1-64 Hz, the differences between the different SRTs increased even more, i.e., the predictions for the SAM and ISTS SRTs were much lower than their measured counterparts. While one might expect intelligibility to improve slightly with increasing levels, the effects should be within a few dB. Even though

it was originally intended to extend this model approach to HI listeners, the level-limitation effects prevented the authors from pursuing these plans. One major limitation of the model, which might contribute to the difficulties of integrating the AN model into the SNR_{env}-based framework, is the assumption of linearity that is intrinsically embedded in the SNR_{env} decision metric, i.e., it is assumed that the difference between the power of the noisy speech and the power of the noise alone estimates the power of the clean signal (see Eq. 4.1). While the AN model might be quasi-linear when only considering MSR fibers and an OAL of 50 dB, it is highly non-linear if all fibers are considered at realistic presentation levels. Other back-end decision metrics, such as correlation or distance-based metrics, may therefore be more appropriate in combination with the AN model. This is in agreement with Bruce et al. (2013), who used correlation metrics, and with Zilany and Bruce (2007), who used a distance metric.

Despite the difficulties discussed above, the use of an AN model to successfully predict SI is highly desirable as it may bridge the gap between psychophysics and physiology. In other words, model predictions may be compared to predictions that are derived from actual auditory nerve recordings in animals. In addition to the results shown in the present study, different methods of coupling the AN-model front end and the mr-sEPSM back end were investigated. A method that gave favorable results was the use of neural metrics (i.e., SUMCORs, e.g., Louage et al., 2004) to extract envelope information from actual spike trains. The Fourier transform of a SUMCOR represents the power spectrum of the envelope of the time signal, in the same way that a Fourier transform of an autocorrelation function of a time signal is equivalent to the power spectrum of that time signal. By applying the modulation filterbank to the magnitude spectrum of the SUMCOR in the frequency domain, the envelope power in each

modulation band can be estimated. It should be noted that this was not done in multiple time windows, but instead over the whole signal duration. The SNR_{env} can be derived by estimating the envelope power for both the noisy speech and the noise signal and comparing them according to Eq. 4.1. Predictions obtained in this way (not shown) showed promising results for the SSN and SAM interferers. However, the SUMCOR calculations are computationally heavy, as they are based on actual spike times and not instantaneous rates. Furthermore, these spike trians are shuffeled to increase the statistical power. While this bridge is certainly interesting, it may not preferable as a speech intelligibility model.

4.6 Summary and conclusion

Previous studies have demonstrated the predictive power of SI models based on the SNR_{env} framework. The present study investigated the extension of this framework to include the non-linear processing of the auditory periphery up to the auditory-nerve synapse. The rationale behind that was that these non-linearities are important for explaining level effects and especially effects of hearing loss. Instead of predicting SI based on the envelope power derived from outputs of a linear gammatone filterbank, the proposed model derives the envelope power from the instantaneous firing rates of an AN model. The proposed model framework was shown to work well in restricted conditions where only MSR fibers and an OAL of 50 dB were considered. For realistic conditions (OAL = 65 dB, LSR, MSR, and HSR fibers), the model predicted the SRTs of the modulated interferers (SAM and ISTS) as too low (i.e., SI as too good). It was shown that a modulation filter bank with fewer modulation channels (1-64 Hz) resulted in reasonable predictions. It is argued that higher modulation channels are too sensitive to the onsets of the adapted stimuli.

Furthermore, the incompatibility of the non-linear front end with the SNR_{env} decision metric, which implicitly assumes linearity, was discussed. Alternatively, back ends based on a cross-correlation or distance metric may be considered in combination with the AN-model.

Acknowledgments

The work leading to this study and the results described therein has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement number PITN-GA-2012-317521. The author thanks Dr. Johannes Zaar for the immensly helpful discussion and inputs during the work leading to this study.

5

Modeling Speech Intelligibility in Hearing-Impaired Listeners[a]

Abstract

Speech intelligibility (SI) models aim to predict the human ability to understand speech in adverse listening conditions, typically using strongly simplified representations of the auditory periphery. The present study focuses on predicting SI in hearing-impaired listeners by taking the highly non-linear peripheral processing of the auditory system into account. The front-end processing of an auditory-nerve (AN) model was combined with a correlation-based back end inspired by the vowel-coding hypothesis of stable rate patterns in the inferior colliculus (Carney et al., 2015). The proposed model assesses the correlation between the noisy speech and the noise alone, as represented by the AN model's instantaneous firing rates, assuming an inverse relationship with SI. The use of the noise alone as a reference signal is inspired by the speech-based envelope power spectrum model (sEPSM Jørgensen and Dau, 2011; Jørgensen et al., 2013). SI data obtained by Christiansen and Dau (2012) in normal-hearing (NH) and hearing-impaired (HI) listeners in conditions of stationary speech-shaped noise (SSN), sinusoidally

[a] This chapter is based on Scheidiger et al., (in preparation).

amplitude-modulated noise (SAM), and speech-like noise (ISTS) were considered for the simulations. The NH listeners' SI data could be accounted for accurately for all noise types, and across presentation levels. The HI listeners' data were predicted by adjusting the front end parameters estimating the inner and outer hair-cell loss based on the audiogram of the listeners. The predictions showed a good agreement with the measured data for four out the thirteen listeners and a reasonable agreement for eight listeners. The work may provide a valuavle basis for quantitatively modeling individual consequences of inner and outer hair-cell loss on speech intelligibility.

5.1 Introduction

Hearing-impaired (HI) listeners exhibit more difficulties in understanding speech in adverse listening conditions as compared to normal-hearing (NH) listeners. These difficulties can be quantified using speech intelligibility (SI) tests. A better understanding of SI in HI listeners could help improve hearing-aid algorithms and fitting procedures (Kates and Arehart, 2005; Hossain et al., 2016). However, performing listening tests with real subjects is a complex and time-consuming process. In order to better understand the impact of hearing deficits on SI in various conditions and to additionally avoid time-consuming listening tests, SI prediction models may serve as valuable tools.

Various NH SI models have been adapted to account for SI in HI listeners. The speech intelligibility index (SII) for NH listeners uses signal-to-noise ratios (SNRs) in critical bands to asses SI (French and Steinberg, 1947; Kryter, 1962; ANSI, 1997). It further includes a spread of masking function, which adds

fractions of the SNR in one band to neighboring channels. In order to adapt the NH model to HI listeners, an empirically determined desensitization factor based on the pure-tone sensitivity (i.e., audiogram) was introduced (Pavlovic, 1986; Magnusson et al., 2001). This desensitization reduces the estimated SII in each band by multiplying it with a factor decreasing from one to zero for hearing losses ranging from 15 to 94 dB.

Kates and Arehart (2005) proposed a SI model extending the SII framework by replacing the standard SNR calculation of the SII by a signal-to-distortion ratio (SDR) derived from the magnitude squared coherence. The coherence SII (CSII) calculates the SDR between the clean signal and the signal distorted by the speech transmission channel. The model was used to predict HI and NH listeners' SI in conditions in which the transmission channel introduced non-linear distortions, such as peak clipping. For the HI listeners, the transmission channel also consisted of a hearing aid that amplified the distorted signal according to the HI listener's specific NAL-R prescription. The SDR derived from the magnitude squared coherence was able to account for the difference in SI between NH and HI listeners in these distorted conditions.

The critical-band filtering in the CSII and SII models is a crude approximation of the human auditory signal processing since a linear model is employed to simulate a highly non-linear system. Thus, these models do not fully represent the physiology of the system but merely its healthy functioning at average conversation levels, i.e., at about 65 dB sound pressure level (SPL). Detailed physiological impairments of the system or non-linear level effects cannot be simulated using such linear models. Other approaches have used more elaborate models of the auditory periphery to model level effects and the consequences of outer hair-cell loss versus inner hair-cell loss on SI. One such elaborate model is an auditory-nerve (AN) model that has been developed to

describe the temporal properties of auditory nerve spike trains from different studies in cats and other species (e.g., Carney, 1993; Bruce et al., 2003; Zilany et al., 2009; Zilany et al., 2014). The model has also been adapted to the sharper human cochlear tunning (Shera et al., 2002).

Zilany, Bruce, and other colleagues have used this AN model as front-end processing for SI models (Zilany and Bruce, 2007; Bruce et al., 2013; Hossain et al., 2016). The internal representations of these SI models are spectrogram-like neurograms. Two of these models use a clean reference signal (at 65 dB SPL) to assess SI. The assessment was either performed through a correlation-based metric (NSIM Bruce et al., 2013) or a distance metric (STMI Zilany and Bruce, 2007). Hossain et al. (2016) proposed a reference-free model based on bi-spectra of the target signal. These SI models based on the AN model have shown promising results for CV and word recognition test, but do not account for sentences intelligibility.

SI models based on the signal-to-noise ratio in the envelope domain (SNR_{env}) have been shown to yield accurate predictions in a wide range of listening conditions for NH listeners, e.g., stationary and fluctuating background noises (sEPSM Jørgensen and Dau, 2011; Jørgensen et al., 2013). These models use the noise signal as reference, instead of the clean signal. The more the noise masks the speech energy and the speech modulations, the more SI is degraded. This has been shown to be a powerful metric that also accounts for effects of noise reduction despite an increase of the SNR (in the energy domain). However, the SNR_{env} decision metric assumes linearity in the peripheral processing and is therefore limited in its use to predict SI based on a non-linear peripheral model.

The aim of the present study was to adapt the decision metric of the sEPSM model to work with a AN model front end. Inspiration for such a metric comes from a model of vowel coding in the midbrain proposed by Carney et al. (2015).

Their proposed neural code for vowel sounds was shown to be robust over a wide range of sound levels and in background noise. It is derived from the AN model responses (Zilany et al., 2014). The code is based on the phenomenon that speech, especially vowels, induce systematic amplitude fluctuations in AN responses close to the fundamental frequency (F0). These F0-related neural fluctuations create patterns of amplitude contrasts across neurons tuned to different CFs. These patterns are robust across presentation level and at moderate levels of background noise. This study proposes to assess these patterns by using a correlation-based metric comparing the noisy speech signal (SN) to the interfering noise (N), thereby applying the sEPSM idea to quantify the degrading influence of the masking noise on the target speech.

5.2 Model description

The proposed model consists of two main stages, each subdivided into further stages. The first main stage is a peripheral front-end stage, represented by the AN model, the second main stage is a decision stage, which converts the firing patterns at the output of the AN model to a correlation-based decision metric and subsequently to a SI score (see Fig. 5.1). The noisy speech signal (SN) and the noise alone (N) serve as inputs to the model.

5.2.1 Front end: AN Model

The front end of the model consists of an auditory nerve model (Zilany et al., 2014). The responses of the model have been validated against a large dataset of AN recordings from the literature (e.g., Carney, 1993; Bruce et al., 2003; Zilany et al., 2009; Zilany et al., 2014). The AN model consists of several stages, each providing a phenomenological description of a major part of the auditory

periphery, starting at the middle ear and ending at a specific AN synapse. The input to the model is an acoustic waveform (in pascals) which is first processed by the middle-ear (ME) filter. After the ME filtering, the signal is further passed through a basilar membrane (BM) filter. A feed-forward control path of the BM filter controls the gain and bandwidth to account for level-dependent properties in the cochlea, e.g., less amplification and broader filters at higher levels. The inner-hair-cell (IHC) stage converts the mechanical BM response to an electrical potential, which is further low-pass filtered before it is applied to the IHC-AN synapse model. The synapse model determines the spontaneous rate, adaptation properties, and rate-level behavior of the AN model. In a last stage, the spike times are generated by a non-homogeneous Poisson process that includes refractory effects. The working of the IHC and OHC in the model may be adjusted by specifying the amount of IHC and OHC damage. The human auditory system contains nerve fibers with different spontaneous rates fiber types (LSR: low spontaneous rate; MSR: medium spontaneous rate, and HSR: high spontaneous rate). For the simulations in this study it was assumed that 60% of all fibers are HSR, 20% MSR, and 20% LSR (Zilany and Bruce, 2007). Each CF was modeled as 50 fibers. The firing rate of all fibers (all types) were averaged to obtain the firing rate for a specific CF.

5.2.2 Back end: Midbrain Model and cross-correlation

The nonlinearities of the auditory periphery have strong effects on the rate fluctuations of AN fibers in response to speech. The contrast in the amplitude of low-frequency rate fluctuations across the AN population is enhanced in the midbrain by the rate tuning of inferior colliculus (IC) neurons to amplitude modulations, which is described by modulation transfer functions (MTFs). The majority of MTFs in the IC have bandpass (BP) tuning to amplitude modulations

(Carney et al., 2015).

In the back end of the model, the MTFs of the IC neurons were modeled as a BP filter with $Q = 1$ centered at $f_c = 125$, i.e., close to the F0 of the target speaker (see below). The instantaneous firing rate of each CF was filtered by this BP, which represents an IC model. The output of this BP filter was segmented into time frames with a duration of 20 ms and a 50% overlap. For each CF, the instantaneous firing rate was squared and averaged within each time frame. The noisy speech and the noise alone were thus represented as functions of time segment k and CF, $sn(k,CF)$ and $n(k,CF)$, respectively. For each segment, a correlation coefficient $r(k)$ between the noisy speech and the noise alone was obtained by:

$$r(k) = \frac{\sum_{CF}(sn(k,CF) - \overline{sn}(k)) \cdot (n(k,CF) - \overline{n}(k))}{\sqrt[2]{\sum_{CF}(sn(k,CF) - \overline{sn}(k))^2} \cdot \sqrt[2]{\sum_{CF}(n(k,CF) - \overline{n}(k))^2}}, \tag{5.1}$$

where $\overline{sn}(k)$ and $\overline{n}(k)$ represent the across-CF mean of $sn(k,CF)$ and $n(k,CF)$. The smaller the correlation coefficient $r(k)$, the higher the contribution of the considered segment to SI. In a further step, the $r(k)$ values were averaged across time windows by means of an unweighted average to obtain r_{Tot}. Lastly, $1 - r_{Tot}$ was converted to a percentage correct using a fitting condition.

5.3 Methods

5.3.1 Speech and noise material

The stimuli used for the modeling work were similar to the speech stimuli used in Christiansen and Dau (2012). The stimuli consisted of natural meaningful

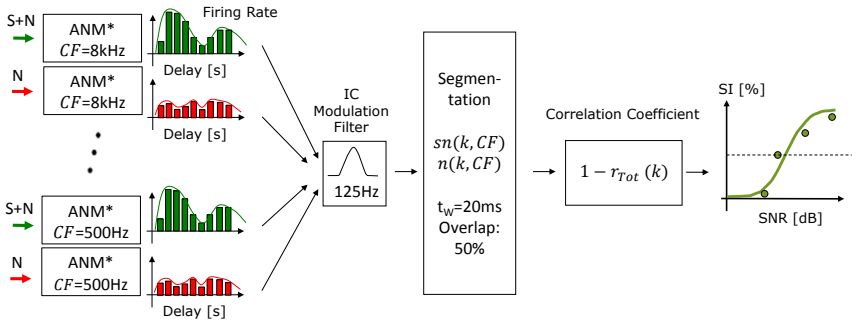


Figure 5.1: Structure of the proposed model: The auditory-nerve model estimates average firing rates for 13 CFs along the basilar membrane for both the noisy speech signal (SN) and the noise signal (N). The firing rates are further processed by a IC filter. At the output of this IC filter, across-CF correlation coefficients $r(k)$ between the noisy mixture $sn(k,CF)$ and the noise alone $n(k,CF)$ are estimated in fixed time windows. The coefficients are averaged across time segments. In a last stage, the integrated r_{Tot} is transformed to a percentage correct.

Danish five-word sentences from the CLUE corpus (Nielsen and Dau, 2009). Ten sentences of the speech corpus were chosen for the simulations. All simulation results were averaged over these ten sentences in order to obtain stable predictions. The sentences were recorded from a male talker with an average fundamental frequency (F0) of 119 Hz. The data from human listeners were available for three different interferers: (i) Speech-shaped noise (SSN), a stationary masker with a long-term spectrum identical to the average of all sentences in the study, (ii) an 8-Hz sinusoidally amplitude-modulated (SAM) speech-shaped noise, and (iii) the international speech test signal (ISTS, Holube et al., 2010a) which consists of natural speech from six female talkers speaking different languages. The ISTS signal was created by truncating recorded sentences into segments and randomly remixing the segments, yielding a largely unintelligible signal with natural speech properties in terms of periodicity and modulation (average F0: 207 Hz). The measured speech reception thresholds (SRTs) in NH listeners, which were obtained directly using an adaptive procedure, were made available by Christiansen and Dau, 2012. For the model simulations, the range

of simulated input SNRs ranged from -21 dB in 3-dB steps up to 12 dB. This range covers the SRTs for all three noise types, which are at -18.3 dB, -9 dB, and -3.1 dB for the ISTS, SAM, and SSN interferers, respectively. The SRTs were measured at overall presentation levels (OALs) of 65 and 80 dB SPL for NH and HI listeners, respectively. The model predictions were obtained at OALs of 50, 65, and 80 dB SPL.

5.3.2 Model configurations

Different configurations of the front end and the back end of the model were tested in this study. The model front end always consisted of 13 CFs (500, 630, 800, 1000, 1250, 1600, 2000, 2500, 3150, 4000, 5000, 6300, 8000 Hz). For all model configurations, the r_{Tot} values obtained in the SSN condition in NH operation mode (see below) were used for fitting. A non-linear fitting algorithm was applied to iteratively estimate the two parameters a_1 and a_2 of the following logistic function in order to convert r_{Tot} to SI:

$$SI(SNR) = \frac{100}{1 + e^{a_1 r_{Tot}(SNR) + a_2}}, \tag{5.2}$$

where SI is the SI expressed as a percentage correct and r_{Tot} is the total correlation coefficient integrated across all CFs. In order to quantify the accuracy of the model predictions, the predicted SRT was obtained as the 50% point on the predicted psychometric function and the prediction error E was calculated as $E = SRT_{predicted} - SRT_{measured}$.

NH operation mode

For the normal-hearing operation mode, the OHC and IHC loss factors in the AN model were set to one, i.e., no loss was assumed. Only one ear was simulated for

the NH listeners, assuming that both ears would result in the same predictions. The model was run for three different OALs (50, 65, 80 dB SPL) to investigate the level dependency of the predictions. The preidictions in SSN noise at an OAL of 65 dB SPL served as the fitting condition.

HI operation mode

For the hearing-impaired operation mode, the audiograms of the two ears of the 13 HI listeners were used to determine the OHC and IHC loss factors in the AN model (Table 1, p. 1657; Christiansen and Dau, 2012). It was assumed that one third of the total hearing loss was caused by an IHC loss and two thirds were due to an OHC loss (Zilany and Bruce, 2007; Bruce et al., 2013). For each listener, the two ears were simulated with the model. To obtain one SI score per listener out of the two individual ear scores, the better ear in terms of the predicted SI was chosen for each SNR. The model predictions were obtained for an OAL of 80 dB SPL, the level at which the SRT data were measured (Christiansen and Dau, 2012). The same fitting parameters as in the NH operation mode were used for the HI operation mode.

5.4 Results

5.4.1 NH operation mode

Figure 5.2 shows the SI predictions (dots) as a function of the input SNR. The SI predictions (in terms of percent correct) decrease with decreasing SNRs. A logistic function was fitted to the predictions (dashed line). From this function, the SRT (SI= 50%) was estimated. Note that the SSN condition (red) was used for fitting. It can be seen that the predicted SRTs also in the other two conditions (SAM, green; ISTS, blue) are very close to the measured SRTs of the human

listeners (squares). The SRT error E did not exceed 0.1 dB for any of the three interferers.

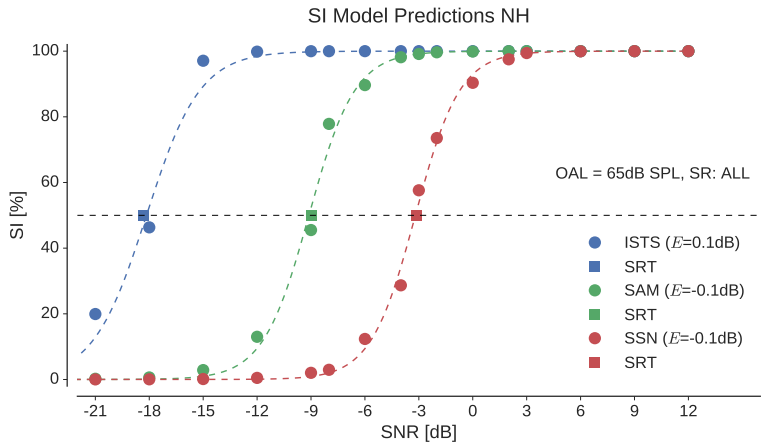


Figure 5.2: Simulated SI depicted as a function of the input SNR for the SSN (red), SAM (green), and ISTS (blue) interferers obtained with the model at an OAL of 65 dB SPL. The round dots represent the simulated SI predictions for each input SNR. The dashed lines depict psychometric functions fitted to the predictions. The squares indicate the SRTs measured in NH listeners at the same OAL, as provided by Christiansen and Dau, 2012.

Figure 5.3 illustrates how the NH predictions change with OAL. Panel A depicts the SI predictions obtained at 50 dB SPL. The SRT errors (E) are positive for all three noise types, implying that these predicted SRTs are higher (i.e., SI worse) than the reference SRTs measured at an OAL of 65 dB SPL. The average overestimation of the SRTs is 2.8 dB. The results for the SAM interferer showed the largest deviation, with $E = 3.8$ dB. Panel B shows the SI predictions obtained at 80 dB SPL. The SRT error (E) was negative for all three noise interferers, implying that the predicted SRTs were lower (i.e., better) than the measured ones at an OAL of 65 dB SPL. The average underestimation of the SRTs was 2.6 dB. The prediction for the ISTS interferer showed the largest deviation, with $E = 5.3$ dB.

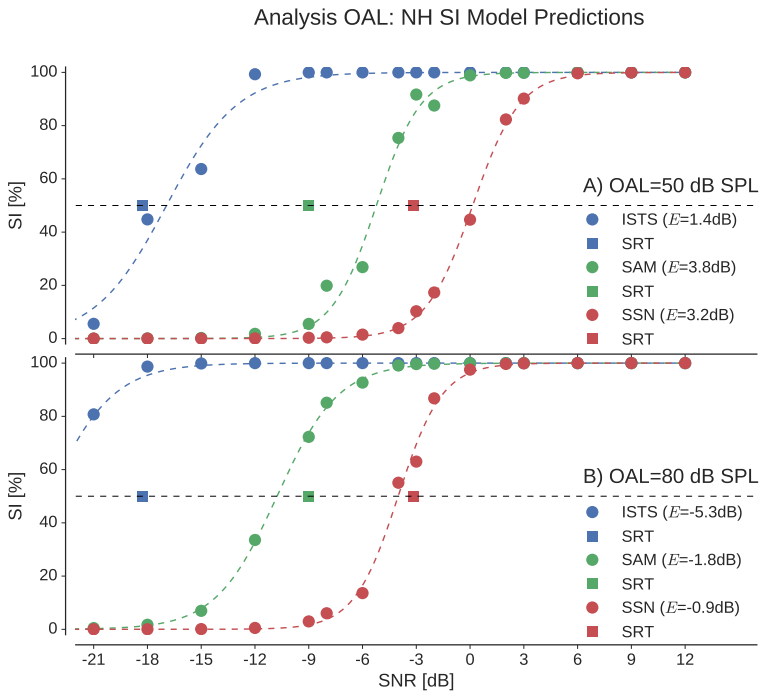


Figure 5.3: Simulated SI depicted as a function of the input SNR for the SSN (red), SAM (green), and ISTS (blue) interferers. The results in panel A were obtained for an OAL of 50 dB SPL, whereas panel B depicts the results for an OAL of 80 dB SPL. The round dots represent the simulated SI predictions for each input SNR. The dashed lines depict psychometric functions fitted to the predictions. The squares indicate the SRTs measured in NH listeners at an OAL of 65 dB SPL, as provided by Christiansen and Dau, 2012.

5.4.2 HI operation mode

Figure 5.4 depicts the SI predictions for two hearing-impaired listeners at an OAL of 80 dB SPL. The predictions are based on the same fitting as described above, i.e., using the SSN condition at an OAL = 65 dB SPL in the NH operation mode. The fitting remained the same for all configurations and the differences between the NH and HI predictions are solely based on changes in the AN-model front end. Panel A shows the predictions for listener HI2, for whom the prediction error E was small for all but the SAM interferer, for which the SRT was predicted as being 2.5 dB too low (i.e., too good). Panel B represents the predictions for listener HI10 and represents a quasi-perfect prediction. Here,

all predictions matched the measured SRTS very well, with an average absolute error of < 0.4 dB. The model can account for the SRT shift to higher signal to noise ratios as compared to the NH SRTs (see Fig. 5.2). Also the decreased difference between the SRTs for the different interferers is well accounted for by the model.

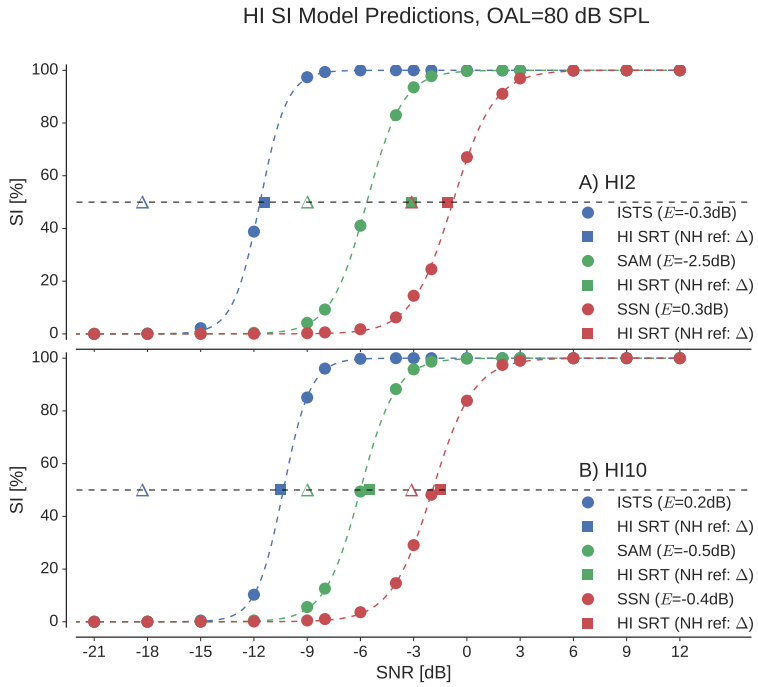


Figure 5.4: Simulated SI depicted as a function of the input SNR for the SSN (red), SAM (green), and ISTS (blue) interferers. All results were obtained at an OAL of 80 dB. Panel A illustrates the results for listener HI2 whereas panel B depicts the results for listener HI10. The round dots represent the simulated SI predictions for each input SNR. The dashed lines depict psychometric functions fitted to the predictions. The squares indicate the SRTs measured in HI listeners at an OAL of 80 dB, as provided by Christiansen and Dau, 2012.

Figure 5.5 shows that not all HI predictions are as good as the ones depicted in Fig. 5.4. The figure illustrates the SRT prediction errors (E) for all listeners. The listeners on the abscissa are sorted in descending order according to their mean absolute error (MAE). Four out of the 13 listeners exhibit small errors ($\bar{E}$=2.6 dB) for all three noise types. For the other remaining listeners the SRTS are generally predicted as too low (i.e., too good). In general, the results for the

the three noise types follow the same trend, indicating that the predictions for some of the HI listeners fail to account for the full deficits experienced by these listeners.

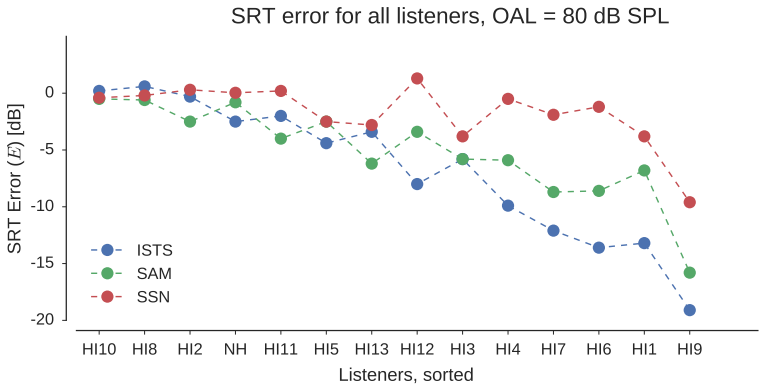


Figure 5.5: The SRT prediction error (E) for all listeners for the SSN (red), SAM (green), and ISTS (blue) interferers. All results were obtained at an OAL of 80 dB. The listeners are sorted in descending order of their mean absolute error.

5.5 Discussion

The SI model presented in this study consists of an AN model as front end and a simplified midbrain model followed by an cross-correlation metric in the back end. The model uses the idea of the vowel-coding patterns across frequency in the IC as discussed in Carney et al. (2015). Inspired by the sEPSM approach (Jørgensen and Dau, 2011; Jørgensen et al., 2013), the decision metric assesses the similarity of the pattern evoked by the noise interferers to the pattern evoked by the noisy-speech signal in short time frames of 20 ms. The lower the correlation coefficient between these two patterns is, the higher the contribution of a particular segment to SI is assumed to be.

The study also investigated the use of the clean speech signal as a reference signal, as used in the STMI and NSIM approaches (not shown here). A correlation to a NH clean speech template proved difficult as the fitting of the back end

was not able to accounht for both NH and HI listeners. A individual fitting to an average HI listener condition could solve this issue, however, such an additional fitting was not congruent with the goal of this study to predict SI solely based on peripheral defficits in HI listeners. A correlation to a HI clean speech template proved difficult as the correlations for the HI listeners increased compared to the NH listeners since the HI representations encoded less information especially at high frequencies. The fact that the decision metric based on the correlation to the noise signal delivered the best results proves the power of the sEPSM concept.

Compared to the model presented in Chapters 3 and 4, the model in this study only considers CF above 500 Hz. This change was motivated by the use of the modulation filter centered at 125 Hz, congruent with the sEPSM assumption that only peripheral channels with a frequency of $>4 f_m$ should be considered for a given modulation filter. If filters below 500 Hz were included into the model, the model predictions were dominated by the strong F0 fluctuations in these filters and were insensitive to loss of information at higher CFs, i.e. the CFs affected by a high-frequency hearing loss. Other center frequencies for the IC modulation filter were tested as well. While a 100-Hz filter or other center frequencies close to the F0 of the traget speaker worked well, lower frequencies (8 and 64 Hz were tested) did poorly compared to the 125 Hz filter.

In the NH operation mode, the model yielded very accurate predictions for both modulated (SAM and ISTS) and steady interferers (SSN) at an OAL of 65 dB SPL, which was also used in the NH experiment. In addition, the model also showed plausible trends as a function of OAL: For a lower OAL of 50 dB SPL, the SI predictions of the model were slightly worse than the predictions at an OAL of 65 dB; although there is no measured reference data available for an OAL of 50 dB SPL, this effect is expected since lower-level parts of the sentences

(e.g., consonants) may become inaudible at this lower OAL. In contrast, the predictions became slightly better for a higher OAL of 80 dB SPL, with more pronounced improvements for the fluctuating noises (SAM, ISTS) than for the steady-state noise (SSN); this could be due to an increased contribution of dip listening, as the speech cues in the masker gaps might become more audible at higher levels.

In the HI operation mode, the model also showed some very promising results. It is important to highlight that the model performed all HI predictions at an OAL of 80 dB SPL (the OAL used in the HI experiment) based on the NH fitting for the SSN interferer at an OAL of 65 dB SPL (the OAL used in the NH experiment). To achieve accurate individual HI predictions at different levels in the presence of different noise types and based on a NH fitting procedure is a challenging test for a SI model. The only difference that the model was "allowed" to take into account were the OHC and IHC loss parameters in the front end, which are based on the audiogram and the assumption that one third of the total loss at a specific CF is due to IHC loss and two thirds are due to OHC loss. The model performed very well for four of the thirteen listeners, and reasonably well for eight listeners.

However, the model shows considerable deviations for five listeners. These listeners are those with SRTs deviating most from the NH SRTs (on average +11 dB). This large deviation might result from deficits not considered in the model (e.g., cognitive deficits). This is further supported by the fact that the audiograms are similar to other listeners audiograms but their SRTs were not. The model in its presented form does not account for any cognitive limitations in the processing.

The assumption of one third IHC and two thirds OHC loss is based on histology results in animals (e.g. Liberman and Dodds, 1984). For individual listeners,

this assumption may not be accurate. Further investigations are required to determine the influence of different IHC versus OHC loss ratios and their influence on SI. The proposed model could be used to systematically test different ratios for each subject. If some of the listeners' predicted SI values improve as a result of different ratios, this might be an indicator that the histology of that subject deviates from the 1/3 versus 2/3 assumption. Furthermore, the knowledge of how different histologies affect speech intelligibility could help tailor compensation strategies to be more effective for a given loss configuration.

The correlation-based metric chosen for the back end could potentially be altered. $1-r_{Tot}$ is strictly speaking not a correlation but a "similarity distance". Other distance measures were also tested (not shown), including a simple euclidean distance and the absolute difference of the envelopes. They delivered reasonable results but were inferior to the correlation-based metric presented here. A combination of distance metrics, such as used in the NSIM study, could potentially outperform the metric proposed in the present study.

Acknowledgments

The work leading to this study and the results described therein has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement number PITN-GA-2012-317521. The author thanks Dr. Johannes Zaar for the immensly helpful discussion and inputs during the work leading to this study.

6

Overall discussion

6.1 Summary of main results

This thesis described two main streams of work: (i) Chapter 2 focused on describing a new analysis framework based on entropy and a distance metric to analyze consonant-vowel (CV) perception in hearing-impaired (HI) listeners across different listening conditions; (ii) Chapters 3, 4, and 5 focused on developing a speech intelligibility (SI) model to account for observed deficits in HI listeners, which could help accelerate the development of hearing-aid algorithms. In particular, the work presented in these three chapters represents an attempt to develop a model that accounts for sentence intelligibility in fluctuating and steady background noises.

6.1.1 Analysis framework for CV experiments

Previous work on CV recognition experiments mostly focused on analyzing recognition scores, thereby missing much of the information that is encoded in the consonant confusions. The study presented in Chapter 2 proposed to utilize the measure of confusion entropy (as used in information theory) to analyze the randomness of confusions as opposed to just the number of confusions. The confusion entropy allows to differentiate between cases where listeners have a low recognition rate because they are guessing, as compared to cases where they have a low recognition rate because the consonant for them has “morphed”

(i.e., changed) into a completely different consonant. If a listener were to always give the same response, the confusion entropy would be 0 bit, irrespective of whether or not the response is correct. In contrast, if a listener were to select all response alternatives the same number of times, which indicates that the listener is randomly guessing, the confusion entropy would reach its maximum value. Thus, the higher the confusion entropy, the more uncertain is the listener regarding his/her responses.

While this response randomness allows to distinguish between cases where the listener is guessing or making a systematic error, it does not take the actual confusions into account. Whether a consonant is confused with one or another response alternative does not affect the confusion entropy measure as long as the confusion probability is equal. However, when comparing results from different recognition experiments, it can be helpful to know if the type of confusions has changed from one condition to another. For example, if a new hearing-aid compensation strategy is tested with CVs, a strong /ga/ confusion for the stimulus /da/ in the unaided condition can turn into an equally strong /ta/ confusion in the aided condition. Such a transformation is undesirable, as it will likely increase the resistance of a listener to accept a new hearing device. The proposed distance measure allows to detect such changes in the confusions.

These two newly proposed metrics were used to analyze CV recognition in the same set of listeners across two experiments: (i) in one experiment, a simple linear gain (flat gain) was provided to the listeners; (ii) in the other condition, the listeners were provided with a frequency-dependent gain (NAL-R) that aims at restoring loudness in all frequency channels. It was shown that NAL-R generally increased the average recognition rate. In a more granular analysis, it was shown that this improvement was multi-faceted when analyzed at a token-level. The confusion entropy measure showed robust perception, i.e., no

guessing, in all but 17% of the token-listener pairs in the flat-gain condition. The proposed angular distance measure revealed that in 63% of the token-listeners pairs, the main confusions remained unchanged despite NAL-R, suggesting that these are caused by acoustic properties of the chosen tokens rather than by the amplification condition. The results suggest that a compensation strategy different than NAL-R would be needed to minimize the main confusions. It was also observed that NAL-R in combination with the individual loss introduced new robust confusions in 30 cases. The analysis framework thus revealed highly relevant information in the data, which was not represented in the consonant recognition scores.

6.1.2 A model accounting for SI in HI listeners

Chapters 3, 4, and 5 presented different stages of the work towards developing a model of SI in HI listeners. The line of work took "baby steps" in altering an existing normal-hearing (NH) SI model, namely the speech-based envelope power spectrum model (sEPSM; Jørgensen and Dau, 2011; Jørgensen et al., 2013), to also account for SI in HI listeners. In the sEPSM framework, SI is predicted based on a signal-to-noise ratio in the envelope domain (SNR_{env}). Chapter 3 describes the attempt to integrate a hearing loss into the linear front end (gammatone filterbank) of the mr-sEPSM model. By simply integrating the audibility thresholds of HI listeners in the model, the predictions agreed reasonably well with the measured relative difference in SI between fluctuating and steady background noises (i.e., masking release). However, the overall decrease of SI measured in the HI listeners was not reflected in the model predictions, which was assumed to be related to its simplistic linear front end. Chapter 4 documents the work to replace the linear gammatone filterbank by a non-linear auditory-nerve model, which allows for the adjustment of inner

(IHC) and outer hair-cell (OHC) loss in the model. The model exhibits accurate predictions of NH listener data when the front end operates at a low input level and when assuming only medium spontaneous-rate fibers, i.e., when the model operates essentially linear. However, when using realistic input levels and a physiologically plausible mixture of low, medium, and high spontaneous-rate fibers, the model overpredicts SI in fluctuating noise. Chapter 5 presents the work conducted to adjust the back end of the sEPSM to work well with the auditory-nerve model as front end. The results show accurate predictions for the NH listeners and promising predictions results for the majority of the HI listeners.

Chapter 3 simulates a hearing loss within the linear mr-sEPSM model whose peripheral processing is represented by a gammatone filterbank. Three model configurations were tested; (i) a gammatone model with an audibility threshold in each filter band according to a listener's audiogram, (ii) a gammatone model with broader filters based on a frequency resolution estimate derived from a listener's audiogram, (iii) a combination of (i) + (ii). The model with configuration (i) performed the best out of the three. It was capable of predicting the difference between a listener's speech reception thresholds (SRT; SNR at which a listener perceives 50% of the presented speech units) in a steady noise and the SRT in a fluctuating noise. However, it did not predict the upward shift of the individual SRTs correctly but only the difference between them. The broader filters of model configuration (ii) showed virtually no effect on the SRT predictions. Model (iii) showed some compounding effect of the two alterations of the front end, but in general the effect of the audibility component was dominating. The chapter concluded that much of the decreased masking release in HI listeners could be accounted for by an audibility loss. Furthermore, it was concluded that the linear front end is limited in its ability to model the important peripheral

effects of a hearing loss.

Chapter 4 uses envelopes derived from the instantaneous firing rate of an auditory-nerve model (Zilany et al., 2014) instead of envelopes derived from a linear gammatone filterbank as in Chapter 3. The auditory-nerve model was operated in two modes: (i) a quasi-linear mode in which medium spontaneous-rate fibers at an overall level (OAL) of 50 dB SPL were considered and (ii) a "realistic" operation mode in which the auditory-nerve model worked at an OAL of 65 dB SPL and the envelopes were derived from all fiber types, i.e., low, medium and high spontaneous-rate fibers. It was shown that the model predictions obtained in the quasi-linear mode agreed well with the NH behavioral data for both steady maskers and fluctuating maskers, whereas the SRT predictions in the "realistic" mode were too low (i.e., too good). It was thus concluded that the model is too sensitive to level variations and originally planned attempts to also predict HI data were therefore not further pursued. A back end analysis revealed that the modulation range of the original mr-sEPSM (1-256 Hz) does not seem to be appropriate for the nonlinear front end. It was shown that the prediction error reaches a minimum for a back end considering a modulation range of 1-64 Hz.

Chapter 5 presents a model which again uses the auditory-nerve model as a front end but combines it with a correlation metric in the back end instead of a SNR_{env} decision metric. Inspired by a recent study on vowel coding in the midbrain (Carney et al., 2015), the model uses a single modulation filter centered at a frequency close to the fundamental frequency of the male target speaker. The output of this modulation filter is analyzed in 20-ms segments by means of an across-CF correlation between the noise signal and the noisy mixture. The model was shown to work well across three OALs (50, 65, and 80 dB SPL) for NH listeners. Consecutively, the model that was fitted to the NH

condition was tested for 13 HI listeners. Only the IHC and OHC transduction of the auditory-nerve model were adjusted according to the audiograms of the HI listeners under the assumption that 1/3 of their loss at a specific frequency was caused by a IHC loss and the remaining 2/3 by a OHC loss (Liberman and Dodds, 1984). The model predictions showed good agreement for eight of the thirteen subjects. It is encouraging that the model only uses one NH fitting condition for its back-end, whereas all other conditions are predicted soley based on changes in the periphery.

6.2 Perspectives

The work presented in this thesis could be extended in multiple ways.

6.2.1 Analysis framework for CV experiments

The distance metric that was used in Chapter 2 is a powerful metric to asses how confusions change across different responses. The metric could be used in combination with a clustering algorithm (e.g., k-means) to automatically establish confusion groups. These confusion groups found by an algorithm based on the distance metric could provide a valuable perspective on the grouping of consonant confusions. In the literature, confusion groups have been typically formed *a priori* according to modes of articulation (e.g., voiced vs unvoiced) instead of being based on perceptual attributes. A clustering-algorithm based alternative could test if these groups really exist in the data.

Furthermore, it would be of interest to further investigate the cases in which NAL-R introduced new stable confusions. An analysis that takes both (i) detailed information about the acoustic properties of the specific token that was confused as well as (ii) a detailed assessment of a listener's hearing loss into

account could help explain the cause of such confusions and help to develop compensation strategies to minimize these confusions.

6.2.2 A model accounting for SI in HI listeners

While the modeling approaches presented in Chapters 3 and 4 exhibited different limitations and shortcomings, the model presented in Chapter 5 shows a promising foundation for a model to predict SI in HI listeners. The correlation metric used in the back end was intended as a first step towards a simpler decision metric. Correlation is a powerful metric as it assesses the signals in full detail. However, this may result in the drawback that the metric can react overly sensitively to small changes in the signals. Another metric based on a distance measure, similar to the neurogram similarity index measure (NSIM), could potentially yield equally predictive but more stable (less extreme) predictions for listeners with a severe hearing loss. It could also be interesting to evaluate the proposed back end with a different front end. The AN model could, for example, be replaced with other models of the nonlinear auditory periphery, e.g., the computational model of human auditory signal processing and perception (CASP; Jepsen et al., 2008).

Chapter 4 showed that the high modulation channels (>64 Hz) were detrimental to the prediction accuracy. Interestingly, Chapter 5 showed that the best results were obtained with a modulation filter at 125 Hz, while results with lower modulation filters were not as accurate. These two findings appear to be contradicting and it should be investigated why the "optimal" preprocessing depends so much on the chosen back end, i.e., SNR_{env} as compared to correlation.

So far the model has only been tested under the simplifying assumption that any of the considered hearing losses consists of 1/3 IHC and 2/3 OHC loss. To better capture the hearing deficits of the individual HI listeners, different

IHC-OHC ratios could be systematically tested with respect to their influence on the final SI predictions.

The front end represents a detailed and thus computationally costly model of the auditory periphery. Once the model has been tested with more subjects and different IHC-OHC ratios, it would be of interest to test/evaluate simplified versions of the front-end model. A gradual simplification of the model could reveal which details are actually necessary to model SI in HI listeners in different noise types.

Bibliography

ANSI S3.5 (1969). “ANSI S3.5, 1969”. In: *American National Standard Methods for the Calculation of the Articulation Index*.

ANSI S3.5 (1997). “ANSI S3.5, 1997”. In: *American National Standard Methods for the calculation of the Speech Intelligibility Index*.

ANSI (1997). “Methods for Calculation of the Speech Intelligibility Index”. In: *ANSI* S3.5.

Allen, J. (1996a). “Harvey Fletcher ’ s role in the creation of communication acoustics a)”. In: *Journal of the Acoustical Society of America* 99.June 1995, pp. 1825–1839.

Allen, J. (1996b). “Harvey {F}letcher’s role in the creation of communication acoustics”. In: *The Journal of the Acoustical Society of America* 99.4, pp. 1825–1839.

Baer, T. and B. C. J. Moore (1993). “Effects of spectral smearing on the intelligibility of sentences in noise”. In: *The Journal of the Acoustical Society of America* 94.3, pp. 1229–1241.

Bernstein, J. G. W. and K. W. Grant (2009). “Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners.” In: *The Journal of the Acoustical Society of America* 125.5, pp. 3358–3372.

Bilger, R. C. and M. D. Wang (1976). "Consonant confusions in patients with sensorineural hearing loss." In: *Journal of speech and hearing research* 19.4, pp. 718–48.

Boothroyd, A. (1984). "Auditory Perception of Speech Contrasts by Subjects with Sensorineural Hearing Loss". In: *Journal of Speech Language and Hearing Research* 27.1, p. 134.

Bruce, I. C., M. B. Sachs, and E. D. Young (2003). "An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses". In: *The Journal of the Acoustical Society of America* 113.1, p. 369.

Bruce, I. C., A. C. Leger, B. C. Moore, and C. Lorenzi (2013). "Physiological prediction of masking release for normal-hearing and hearing-impaired listeners". In: *Proceedings of Meetings on Acoustics*. Vol. 19. 1. Acoustical Society of America, pp. 050178–050178.

Buss, E., J. W. Hall III, and J. H. Grose (2004). "Temporal fine-structure cues to speech and pure tone modulation in observers with sensorineural hearing loss". In: *Ear and hearing* 25.3, pp. 242–250.

Byrne, D. and H. Dillon (1986). "The National Acoustic Laboratories' (NAL) New Procedure for Selecting the Gain and Frequency Response of a Hearing Aid". In: *Ear and hearing* 7.4, pp. 257–65.

Carney, L. H. (1993). "A model for the responses of low-frequency auditory-nerve fibers in cat". In: *The Journal of the Acoustical Society of America* 93.1, p. 401.

Carney, L. H., T. Li, and J. M. McDonough (2015). "Speech Coding in the Brain: Representation of Vowel Formants by Midbrain Neurons Tuned to Sound Fluctuations". In: *eNeuro* 2.4, pp. 1–12.

Christiansen, C. and T. Dau (2012). "Relationship between masking release in fluctuating maskers and speech reception thresholds in stationary noise". In: *The Journal of the Acoustical Society of America* 132.3, pp. 1655–1666.

Dau, T, J Verhey, and A Kohlrausch (1999). “Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers.” In: *The Journal of the Acoustical Society of America* 106.5, pp. 2752–2760.

Dau, T., D. Pueschel, and A. Kohlrausch (1996). “A quantitative model of the “effective” signal processing in the auditory system. I. Model structure”. In: *The Journal of the Acoustical Society of America* 99.6, p. 3615.

Dau, T., B. Kollmeier, and A. Kohlrausch (1997). “Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers”. In: *Journal of the Acoustical Society of America* 102.5, pp. 2892–905.

Desloge, J. G., C. M. Reed, L. D. Braida, Z. D. Perez, and L. A. Delhorne (2010). “Speech reception by listeners with real and simulated hearing impairment: effects of continuous and interrupted noise.” In: *The Journal of the Acoustical Society of America* 128.1, pp. 342–59.

Dubno, J. R., D. D. Dirks, and D. E. Morgan (1984). “Effects of age and mild hearing loss on speech recognition in noise”. In: *The Journal of the Acoustical Society of America* 76.1, p. 87.

Dubno, J. R., A. R. Horwitz, and J. B. Ahlstrom (2003). “Recovery from prior stimulation: masking of speech by interrupted noise for younger and older adults with normal hearing”. In: *The Journal of the Acoustical Society of America* 113.4, pp. 2084–2094.

Elhilali, M., T. Chi, and S. A. Shamma (2003). “A spectro-temporal modulation index (STMI) for assessment of speech intelligibility”. In: *Speech Communication* 41.2-3, pp. 331–348.

Festen, J. M. and R. Plomp (1990). “Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing”. In: *The Journal of the Acoustical Society of America* 88.4, p. 1725.

Fletcher, H. and J. C. Steinberg (1929). “Articulation Testing Methods”. In: *Bell System Technical Journal* 8.4, pp. 806–854.

Fletcher, H. and R. H. Galt (1950). “The Perception of Speech and Its Relation to Telephony”. In: *The Journal of the Acoustical Society of America* 22.2, pp. 89–151.

Fousek, P., P. Svojanovsk, R. Simplon, and C. Postale (2000). “New Nonsense Syllables Database – Analyses and Preliminary ASR Experiments”. In: *Analysis*, pp. 1–4.

French, N. R. and J. C. Steinberg (1947). “Factors governing the intelligibility of speech sounds”. In: *The journal of the Acoustical society of America* 19.1, pp. 90–119.

George, E. L. J., J. M. Festen, and T. Houtgast (2006). “Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners”. In: *The Journal of the Acoustical Society of America* 120.4, pp. 2295–2311.

Glasberg, B. R. and B. C. Moore (1990). “Derivation of auditory filter shapes from notched-noise data.” In: *Hearing research* 47.1-2, pp. 103–38.

Glasberg, B. R. and B. C. J. Moore (1986). “Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments”. In: *The Journal of the Acoustical Society of America* 79.4, pp. 1020–1033.

Hines, A. and N. Harte (2010). “Speech intelligibility from image processing”. In: *Speech Communication* 52.9, pp. 736–752.

Hines, A. and N. Harte (2012). “Speech intelligibility prediction using a Neurogram Similarity Index Measure”. In: *Speech Communication* 54.2, pp. 306–320.

Holube, I., S. Fredelake, M. Vlaming, and B. Kollmeier (2010a). “Development and analysis of an International Speech Test Signal (ISTS).” EN. In: *International journal of audiology* 49.12, pp. 891–903.

Holube, I., S. Fredelake, M. Vlaming, and B. Kollmeier (2010b). “Development and analysis of an international speech test signal (ISTS)”. In: *International journal of audiology* 49.12, pp. 891–903.

Hossain, M. E., W. A. Jassim, and M. S. A. Zilany (2016). “Reference-Free Assessment of Speech Intelligibility Using Bispectrum of an Auditory Neurogram”. In: *PLOS ONE* 11.3. Ed. by D. A. Robin, e0150415.

Hou, Z. and C. V. Pavlovic (1994). “Effects of temporal smearing on temporal resolution, frequency selectivity, and speech intelligibility”. In: *The Journal of the Acoustical Society of America* 96.3, pp. 1325–1340.

Houtgast, T. and H. J. M. Steeneken (1971). “Evaluation of Speech Transmission Channels by Using Artificial Signals”. In: *Acta Acustica united with Acustica* 25.6, pp. 355–367.

Humes, L. E., D. D. Dirks, T. S. Bell, C. Ahlstbom, and G. E. Kincaid (1986). “Application of the Articulation Index and the Speech Transmission Index to the Recognition of Speech by Normal-Hearing and Hearing-Impaired Listeners”. In: *Journal of Speech Language and Hearing Research* 29.4, p. 447.

IEC 60268-16 (2003). “IEC 60268-16 (2003)”. In: *Sound system equipment-Part 16: Objective rating of speech intelligibility by speech transmission index.*

ISO 389-7 (2005). “Acoustics - Reference zero for the calibration of audiometric equipment. Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions”. In: *(International Organization for Standardization, Geneva).*

Jepsen, M. L., S. D. Ewert, and T. Dau (2008). “A computational model of human auditory signal processing and perception.” In: *The Journal of the Acoustical Society of America* 124.1, pp. 422–38.

Jørgensen, S. and T. Dau (2011). “Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing.” In: *The Journal of the Acoustical Society of America* 130.3, pp. 1475–87.

Jørgensen, S., S. D. Ewert, and T. Dau (2013). “A multi-resolution envelope-power based model for speech intelligibility.” In: *The Journal of the Acoustical Society of America* 134.1, pp. 436–46.

Kapoor, A. and J. B. Allen (2012). “Perceptual effects of plosive feature modification”. In: *The Journal of the Acoustical Society of America* 131.1, pp. 478–491.

Kates, J. M. and K. H. Arehart (2005). “Coherence and the speech intelligibility index”. In: *The Journal of the Acoustical Society of America* 117.4, p. 2224.

Kollmeier, B., M. R. Schädler, A. Warzybok, B. T. Meyer, and T. Brand (2016). “Sentence Recognition Prediction for Hearing-impaired Listeners in Stationary and Fluctuation Noise With FADE: Empowering the Attenuation and Distortion Concept by Plomp With a Quantitative Processing Model.” In: *Trends in hearing* 20.

Kryter, K. D. (1962). “Methods for the Calculation and Use of the Articulation Index”. In: *The Journal of the Acoustical Society of America* 34.11, pp. 1689–1697.

Léger, A. C., B. C. J. Moore, and C. Lorenzi (2012). “Temporal and spectral masking release in low- and mid-frequency regions for normal-hearing and hearing-impaired listeners.” In: *The Journal of the Acoustical Society of America* 131.2, pp. 1502–14.

Li, F. and J. B. Allen (2011). “Manipulation of Consonants in Natural Speech”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.3, pp. 496–504.

Li, F., A. Menon, and J. B. Allen (2010). “A psychoacoustic method to find the perceptual cues of stop consonants in natural speech”. In: *The Journal of the Acoustical Society of America* 127.4, pp. 2599–2610.

Li, F., A. Trevino, A. Menon, and J. Allen (2012). “A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise”. In: *The Journal of the Acoustical Society of America* 132.4, p. 2663.

Liberman, M. C. and L. W. Dodds (1984). “Single-neuron labeling and chronic cochlear pathology. III. Stereocilia damage and alterations of threshold tuning curves.” In: *Hearing research* 16.1, pp. 55–74.

Lorenzi, C., M. Husson, M. Ardoint, and X. Debruille (2006). “Speech masking release in listeners with flat hearing loss: Effects of masker fluctuation rate on identification scores and phonetic feature reception”. In: *International journal of audiology* 45.9, pp. 487–495.

Louage, D. H. G., M. van der Heijden, and P. X. Joris (2004). “Temporal properties of responses to broadband noise in the auditory nerve.” In: *Journal of neurophysiology* 91.5, pp. 2051–65.

Magnusson, L, M Karlsson, and A Leijon (2001). “Predicted and measured speech recognition performance in noise with linear amplification.” In: *Ear and hearing* 22.1, pp. 46–57.

Miller, G. A. (1955). “An Analysis of Perceptual Confusions Among Some English Consonants”. In: *The Journal of the Acoustical Society of America* 27.2, p. 338.

Miller, G. A. and P. E. Nicely (1955). “An Analysis of Perceptual Confusions Among Some English Consonants”. In: *The Journal of the Acoustical Society of America* 27.2, pp. 338–352.

Moncada-Torres, A., A. van Wieringen, I. C. Bruce, J. Wouters, and T. Francart (2017). “Predicting phoneme and word recognition in noise using a computational model of the auditory periphery”. In: *The Journal of the Acoustical Society of America* 141.1, pp. 300–312.

Moore, B. C. J. (2007). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues (Wiley Series in Human Communication Science)*. Wiley-Interscience, p. 346.

Mueller, H. G.M.C. K. (1990). “An easy method for calculating the articulation index”. In: *The Hearing journal* 43.9.

Nielsen, J. B. and T. Dau (2009). “Development of a Danish speech intelligibility test.” en. In: *International journal of audiology* 48.10, pp. 729–41.

Owens, E. (1978). “Consonant Errors and Remediation in Sensorineural Hearing Loss”. In: *Journal of Speech and Hearing Disorders* 43.3, p. 331.

Oxenham, A. J. and A. M. Simonson (2009). “Masking release for low-and high-pass-filtered speech in the presence of noise and single-talker interference”. In: *The Journal of the Acoustical Society of America* 125.1, pp. 457–468.

Pavlovic, C. V. (1986). “An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals”. In: *The Journal of the Acoustical Society of America* 80.1, p. 50.

Payton, K. L. and L. D. Braida (1999). “A method to determine the speech transmission index from speech waveforms”. In: *http://dx.doi.org/10.1121/1.428216*.

Phatak, S. A. and J. Allen (2007). “Consonant and vowel confusions in speech-weighted noise”. In: *The Journal of the Acoustical Society of America* 121.4, p. 2312.

Pickles, J. O. (2008). *An introduction to the physiology of hearing*. Vol. 3. London: Academic press.

Plomp, R. (1978). "Auditory handicap of hearing impairment and the limited benefit of hearing aids". In: *The Journal of the Acoustical Society of America* 63.2, pp. 533–549.

Plomp, R. (1986). "A Signal-to-Noise Ratio Model for the Speech-Reception Threshold of the Hearing Impaired". In: *Journal of Speech Language and Hearing Research* 29.2, p. 146.

Reed, C. M., J. G. Desloge, L. D. Braida, Z. D. Perez, and A. C. Léger (2016). "Level variations in speech: Effect on masking release in hearing-impaired listeners)". In: *The Journal of the Acoustical Society of America* 140.1, pp. 102–113.

Rhebergen, K. S. and N. J. Versfeld (2005). "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners". In: *The Journal of the Acoustical Society of America* 117.4, p. 2181.

Rhebergen, K. S., N. J. Versfeld, and W. A. Dreschler (2006). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise". In: *The Journal of the Acoustical Society of America* 120.6, pp. 3988–3997.

Ruggero, M. A. and N. C. Rich (1991). "Furosemide alters organ of corti mechanics: evidence for feedback of outer hair cells upon the basilar membrane." In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 11.4, pp. 1057–67.

Scheidiger, C. and J. B. Allen (2013). "Effects of NAL-R on consonant- vowel perception". In: *4th International Symposium on Auditory and Audiological Research*.

Schmitt, N., A. Winkler, M. Boretzki, and I. Holube (2016). “A Phoneme Perception Test Method for High-Frequency Hearing Aid Fitting”. In: *Journal of the American Academy of Audiology* 27.5, pp. 367–379.

Shera, C. A., J. J. Guinan, and A. J. Oxenham (2002). “Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements.” In: *Proceedings of the National Academy of Sciences of the United States of America* 99.5, pp. 3318–23.

Singh, R. and J. Allen (2012). “The influence of stop consonants’ perceptual features on the Articulation Index model.” In: *The Journal of the Acoustical Society of America* 131.4, pp. 3051–68.

Strelcyk, O. and T. Dau (2009). “Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing.” In: *The Journal of the Acoustical Society of America* 125.5, pp. 3328–45.

Studebaker, G. A., R. L. Sherbecoe, D. M. McDaniel, and C. A. Gwaltney (1999). “Monosyllabic word recognition at higher-than-normal speech and noise levels.” In: *The Journal of the Acoustical Society of America* 105.4, pp. 2431–44.

Summers, V., M. J. Makashay, S. M. Theodoroff, and M. R. Leek (2013). “Suprathreshold auditory processing and speech perception in noise: hearing-impaired and normal-hearing listeners”. In: *Journal of the American Academy of Audiology* 24.4, pp. 274–292.

Taal, C. H., R. C. Hendriks, R. Heusdens, and J. Jensen (2011). “An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7, pp. 2125–2136.

Toscano, J. C. and J. Allen (2014). “Across- and Within-Consonant Errors for Isolated Syllables in Noise”. In: *Journal of Speech Language and Hearing Research* 57.6, p. 2293.

Trevino, A. and J. Allen (2013). “Within-consonant perceptual differences in the hearing impaired ear.” In: *The Journal of the Acoustical Society of America* 134.1, pp. 607–17.

Wirtzfeld, M. R. (2016). “Predicting Speech Intelligibility and Quality from Model Auditory Nerve Fiber Mean-rate and Spike-timing Activity”. PhD thesis, p. 121.

Wright, R. (2004). “A review of perceptual cues and cue robustness”. In: *Phonetically Based Phonology*, pp. 34–57.

Young, E. D. and D. Oertel (2003). “The Cochlear Nucleus”. In: *Synaptic Organization of the Brain*. Ed. by G. M. Shepherd. New York, NY: Oxford University Press, pp. 125–163.

Young, E. D. and M. B. Sachs (1979). “Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers”. In: *The Journal of the Acoustical Society of America* 66.5, pp. 1381–1403.

Zaar, J. and T. Dau (2015). “Sources of variability in consonant perception and their auditory correlates”. In: *Journal of the Acoustical Society of America* 137.4, pp. 2306–2306.

Zilany, M. S. A. and I. C. Bruce (2007). “Predictions of Speech Intelligibility with a Model of the Normal and Impaired Auditory-periphery”. In: *2007 3rd International IEEE/EMBS Conference on Neural Engineering*. IEEE, pp. 481–485.

Zilany, M. S. A., I. C. Bruce, P. C. Nelson, and L. H. Carney (2009). “A phenomenological model of the synapse between the inner hair cell and auditory nerve:

long-term adaptation with power-law dynamics." In: *The Journal of the Acoustical Society of America* 126.5, pp. 2390–412.

Zilany, M. S. A., I. C. Bruce, and L. H. Carney (2014). "Updated parameters and expanded simulation options for a model of the auditory periphery." In: *The Journal of the Acoustical Society of America* 135.1, pp. 283–6.

Zurek, P. M. and L. A. Delhorne (1987). "Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment". In: *The Journal of the Acoustical Society of America* 82.5, p. 1548.

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.

Vol. 13: *Claus Forup Corlin Jespersgaard*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.

Vol. 14: *Rémi Decorsière*, Spectrogram inversion and potential applications for hearing research, 2013.

Vol. 15: *Søren Jørgensen*, Modeling speech intelligibility based on the signal-to-noise envelope power ration, 2014.

Vol. 16: *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.

Vol. 17: *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.

Vol. 18: *Márton Marschall*, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.

Vol. 19: *Jasmina Catic*, Human sound externalization in reverberant environments, 2014.

Vol. 20: *Michał Fereczkowski*, Design and evaluation of individualized hearing-aid signal processing and fitting, 2015.

Vol. 21: *Alexandre Chabot-Leclerc*, Computational modeling of speech intelligibility in adverse conditions, 2015.

Vol. 22: *Federica Bianchi*, Pitch representations in the impaired auditory system and implications for music perception, 2016.

Vol. 23: *Johannes Zaar*, Measures and computational models of microscopic speech perception, 2016.

Vol. 24: *Johannes Käsbach*, Characterizing apparent source width perception, 2016.

Vol. 25: *Gusztáv Löcsei*, Lateralized speech perception with normal and impaired hearing, 2016.

Vol. 26: *Suyash Narendra Joshi*, Modelling auditory nerve responses to electrical stimulation, 2017.

Vol. 27: *Henrik Gerd Hassager*, Characterizing perceptual externalization in listeners with normal, impaired and aided-impaired hearing, 2017.

Vol. 28: *Richard Ian McWalter*, Analysis of the auditory system via synthesis of natural sounds, speech and music, 2017.

Vol. 29: *Jens Cubick*, Characterizing the auditory cues for the processing and perception of spatial sounds, 2017.

Vol. 30: *Gerard Encina-Llamas*, Characterizing cochlear hearing impairment using advanced electrophysiological methods, 2017.

The end.

To be continued...

Quantitatively assessing the speech intelligibility deficits observed in hearing-impaired listeners is a basic component for a better understanding of these deficits and a crucial component for the development of successful compensation strategies. This dissertation describes two main streams of work aiming at a better quantitative understanding: Part (i) focuses on describing a new analysis framework based on a confusion entropy and a distance metric to analyze consonant-vowel perception in hearing-impaired listeners. These metrics allow for a speech-token-based comparison of a listener's performance across different listening conditions. Part (ii) focuses on developing a computational speech intelligibility model to account for observed deficits in HI listeners. It presents a model that predicts normal-hearing and hearing-impaired speech intelligibility in stationary and fluctuating back-ground noises.

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