A microscopic model of speech recognition for listeners with normal and impaired hearing

Von der Fakultät für Mathematik und Naturwissenschaften

der Carl-von-Ossietzky-Universität Oldenburg

zur Erlangung des Grades und Titels eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

angenommene Dissertation

Dipl.-Phys. Tim Jürgens

geboren am 25. Mai 1979

in Wilhelmshaven

Erstgutachter: Prof. Dr. Dr. Birger Kollmeier

Zweitgutachter: PD Dr. Volker Hohmann

Tag der Disputation: 25. November 2010

für Andreas

Abstract

Degraded speech intelligibility is one of the most frequent complaints of sensorineural hearing-impaired listeners, both in noisy and quiet situations. An understanding of the effect of hearing impairment on speech intelligibility is therefore of large interest particularly in order to develop new hearing-aid algorithms for rehabilitation. However, sensorineural hearing impairment is often found to be very individual in terms of the functional deficits of the inner ear and the entire auditory system. Important individual factors to be considered when modeling the effect of sensorineural hearing impairment on speech intelligibility are the audibility of the speech signal, different compressive properties, or different active processes in the inner ear. The latter two can be termed *supra-threshold* factors, since they affect the processing of speech well above the individual absolute threshold. It is not possible to directly (i.e. invasively) measure and study the influence of these supra-threshold factors on human speech recognition (HSR) for ethical reasons. However, computer models on HSR can provide an insight in how these factors may influence speech recognition performance.

This dissertation presents a microscopic model of human speech recognition, microscopic in a sense that first, the recognition of single phonemes rather than the recognition of whole sentences is modeled. Second, the particular spectro-*temporal* structure of speech is processed in a way that is presumably very similar to the processing that takes place in the human auditory system. This contrasts with other models of HSR, which usually use the *spectral* structure only. This microscopic model is capable of predicting phoneme recognition in normal-hearing listeners in noise (Chapter 2) along with important aspects of consonant recognition in normal-hearing and hearing-impaired listeners in quiet condition (Chapter 5). Furthermore, an extension of this model for the prediction of word recognition rates in whole German sentences is capable of predicting speech reception thresholds of normal-hearing and hearing-impaired listeners as accurately as a standard speech intelligibility model (Chapter 3). Parameters reflecting the supra-threshold auditory processing are assessed in normal-hearing and hearing-impaired listeners using indirect psychoacoustical measurement techniques such as a forward masking experiment and categorical loudness scaling (Chapter 4). Finally, the influence of including supra-threshold auditory processing deficits (assessed using the aforementioned measurement techniques) in modeling speech recognition is investigated (Chapter 5) primarily realized as a loss in cochlear compression. The results show that implementing supra-threshold processing deficits

(as found in hearing-impaired listeners) in a microscopic model of human speech recognition improves prediction accuracy. However, the advantage of taking these additional suprathreshold processing parameters into account is marginal in comparison to predicting speech intelligibility directly from audiometric data.

Zusammenfassung

Eins der Hauptprobleme von Leuten mit einer Schallempfindungsschwerhörigkeit ist eine verschlechterte Sprachverständlichkeit sowohl in Ruhe, als auch in Umgebungen mit Störgeräusch. Ein Verständnis davon zu gewinnen, wie Schwerhörigkeit Sprachverständlichkeit beeinflusst, ist daher von großer Wichtigkeit für die Rehabilitation Schwerhörender, z.B. in Form der Entwicklung neuer Hörgeräte-algorithmen. Schallempfindungsschwerhörigkeit kann allerdings sehr individuell sein, wenn man die Art und Anzahl der geschädigten Komponenten des Innenohres und des gesamten auditorischen Systems betrachtet. Wichtige individuelle Faktoren der Schallempfindungsschwerhörigkeit, welche Sprachverständlichkeit beeinflussen, können zum Beispiel sein: die Hörbarkeit des Sprachsignals, unterschiedliche kompressive Eigenschaften in der Verarbeitung des Innenohres oder unterschiedlich starke aktive Prozesse im Innenohr. Die letzteren beiden können als *überschwellige* Faktoren bezeichnet werden, da sie die Verarbeitung von Sprache oberhalb der Hörschwelle beeinflussen. Es ist aus ethischen Gründen nicht möglich den Einfluss dieser überschwelligen Faktoren auf die menschliche Spracherkennung direkt (also invasiv) zu messen und zu studieren. Allerdings können Computermodelle der menschlichen Spracherkennung einen Einblick geben, wie diese Faktoren die Sprachverständlichkeitsleistung beeinflussen können.

Diese Dissertation präsentiert ein mikroskopisches Modell der menschlichen Spracherkennung, mikroskopisch in dem Sinne, dass erstens die Erkennung von einzelnen Phonemen anstelle der Erkennung von ganzen Wörtern oder Sätzen modelliert wird. Zweitens wird die genaue spektro-*temporale* Struktur von Sprache auf eine Art und Weise verarbeitet, die sehr ähnlich zu der Verarbeitung ist, wie sie auch im menschlichen auditorischen System stattfindet. Andere gängige Modelle der menschlichen Spracherkennung nutzen im Gegensatz dazu nur die *spektrale* Struktur von Sprache und einem optionalen Störgeräusch aus. Dieses mikroskopische Modell ist dazu in der Lage Phonemerkennungsraten für Normalhörende unter Einfluss von Hintergrundrauschen (Kapitel 2) und wichtige Aspekte der Konsonanterkennung für Normal- und Schwerhörende in Ruhe (Kapitel 5) vorherzusagen. Außerdem kann eine Erweiterung dieses Modells auf die Erkennung von Wörtern (eingebettet in ganzen deutschen Sätzen) die Sprachverständlichkeitsschwellen von Normal- und Schwerhörenden mit ebenso großer Genauigkeit vorhersagen wie ein anderes gängiges Sprachverständlichkeitsmodell (Kapitel 3). Parameter, die die überschwellige

auditorische Verarbeitung in Normal- und Schwerhörenden quantifizieren, wurden mit Hilfe von indirekten psychoakustischen Messungen, nämlich einem Nachverdeckungsexperiment und der kategorialen Lautheitsskalierung geschätzt (Kapitel 4). In Kapitel 5 wurde dann schlussendlich untersucht, welchen Einfluss eine Veränderung der überschwelligen Verarbeitung (geschätzt aus den Messungen aus Kapitel 4) auf die modellierte Sprachverständlichkeit hat. Die Ergebnisse zeigen, dass der Einbau einer überschwelligen Verarbeitung, so wie sie in Schwerhörenden beobachtet wird, die Vorhersage der Sprachverständlichkeit verbessert. Allerdings ist der Vorteil, der durch den Einbau der genauen überschwelligen Verarbeitung (geschätzt durch überschwellige psychoakustische Messungen) erreicht wird, marginal im Gegensatz zu einer alleinigen Schätzung dieser überschwelligen Verarbeitung durch das Audiogramm.

List of publications associated with this thesis

Peer-reviewed articles:

Jürgens, T., Brand, T., Kollmeier, B. **(2007)**, "Modelling the human-machine gap in speech reception: microscopic speech intelligibility prediction for normal-hearing subjects with an auditory model," Proceedings of the 8th annual conference of the International Speech Communication Association (Interspeech, Antwerp, Belgium), pp. 410-413.

Jürgens, T., Brand, T. **(2009)**, "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," J. Acoust. Soc. Am. 126, pp. 2635-2648.

Jürgens, T., Fredelake, S., Meyer, R. M., Kollmeier, B., Brand, T. **(2010)**, "Challenging the Speech Intelligibility Index: macroscopic vs. microscopic prediction of sentence recognition in normal and hearing-impaired listeners," Proceedings of the 11th annual conference of the International Speech Communication Association (Interspeech, Makuhari, Japan), pp. 2478-2481.

Jürgens, T., Kollmeier, B., Brand, T., Ewert, S.D. **(2010)**, "Assessment of auditory nonlinearity for listeners with different hearing losses using temporal masking and categorical loudness scaling," submitted to Hear. Res.

Non-peer-reviewed articles:

Jürgens, T., Brand, T., Kollmeier, B. **(2007)**, "Modellierung der Sprachverständlichkeit mit einem auditorischen Perzeptionsmodell," Tagungsband der 33. Jahrestagung für Akustik (DAGA, Stuttgart, Germany), pp. 717-718.

Jürgens, T., Brand, T., Kollmeier, B. **(2008)**, "Sprachverständlichkeitsvorhersage für Normalhörende mit einem auditorischen Modell," Tagungsband der 11. Jahrestagung der Deutschen Gesellschaft für Audiologie (DGA, Kiel, Germany).

Jürgens, T., Brand, T., Kollmeier, B. **(2008)**, "Phonemerkennung in Ruhe und im Störgeräusch, Vergleich von Messung und Modellierung," Tagungsband der 39. Jahrestagung der Deutschen Gesellschaft für Medizinische Physik (DGMP, Oldenburg, Germany).

Jürgens, T., Brand, T., Kollmeier, B. **(2009)**, "Consonant recognition of listeners with hearing impairment and comparison to predictions using an auditory model," Proceedings of the NAG/DAGA International Conference on Acoustics (Rotterdam, The Netherlands), pp. 1663-1666.

Jürgens, T., Brand, T., Ewert, S. D., Kollmeier, B. **(2010)**, “Schätzung der Nichtlinearität der auditorischen Verarbeitung bei Normal- und Schwerhörenden durch kategoriale Lautheitsskalierung,“ Tagungsband der 36. Jahrestagung für Akustik (DAGA, Berlin, Germany), pp. 467-468.

Published abstracts:

Jürgens, T., Brand, T., Kollmeier, B. **(2009)**, “Predicting consonant recognition in quiet for listeners with normal hearing and hearing impairment using an auditory model," J. Acoust. Soc. Am. **125**, p. 2533 (157th meeting of the Acoustical Society of America, Portland, Oregon).

Contents

1 General Introduction

A large proportion of the population of industrialized countries shows a significant hearing impairment (in Germany, for instance, hearing impairment among the population amounts to about 19%; Sohn, 2001). This hearing impairment affects the life of these people in various ways. For example, many people with hearing impairment complain about insensitivity to soft sounds, a degradation of their ability to localize the direction of a sound, and, most importantly, a degradation of their ability to understand speech, especially in noisy conditions. The assessment of speech intelligibility has therefore become an instrument for the diagnosis of hearing impairment and an instrument for the evaluation of rehabilitative strategies, e.g. in hearing-aids. Consequently, an understanding of the effect of hearing impairment on speech intelligibility (i.e. finding appropriate models of the function and dysfunction of human speech recognition) is of large interest.

The first predictions of speech intelligibility using quantitative models were done in the Bell Telephone Laboratories by H. Fletcher, N.R. French and J.C. Steinberg in the 1920s to 1940s. Their research focused on understanding the impact of various distortions on speech intelligibility, especially distortions typical of telephone transmission. The result of their research, the Articulation Index (AI) (French and Steinberg, 1947; Fletcher and Galt, 1950), is a measure for speech intelligibility based on four independent factors: (1) audibility of speech, (2) speech-to-noise-ratio, (3) sensitivity of the auditory system, and (4) a frequency distortion factor. The AI can be termed as the first “macroscopic” model of speech recognition. The term “macroscopic” in relation to speech intelligibility models can be defined twofold. First, a macroscopic model provides a prediction of the average speech recognition performance measured using a complete speech test. This contrasts with predicting the recognition of single words, syllables or phonemes, which is termed “microscopic” in the context of this dissertation. Second, a macroscopic model such as the AI is based on the audibility of parts of the speech signal primarily in the frequency domain, i.e. those parts of the long-term spectrum that can be heard by the subject. A microscopic approach, on the contrary, bases its computation on those spectro-*temporal* features of speech that a listener perceives. The work of French and Steinberg (1947) later on became a standard of the American National Standards Institute (ANSI, 1969). A further improvement

(regarding the type of speech material used) resulted in the Speech Intelligibility Index (SII) (ANSI, 1997). AI and SII work well for normal-hearing (NH) listeners in various stationary noise conditions and extensions have been done to model speech intelligibility in fluctuating noise (Rhebergen and Versfeld, 2005) and in different room acoustics (Beutelmann and Brand, 2006).

Another model of human speech recognition that evaluates slow (i.e. < 16 Hz) speech modulations is the Speech Transmission Index (STI) (Houtgast and Steeneken, 1984). The STI is capable of predicting the distortion of speech intelligibility in room acoustics, but needs long (> 60 s) speech waveforms to reliably estimate the speech modulations. Therefore, the STI can also be termed a macroscopic model of human speech recognition. Although modifications of the STI to model the recognition of smaller speech segments have been investigated (Kollmeier, 1990), however, from a physiological point of view, all these macroscopic models can only be a rough approximation to the human speech recognition process, because of the following reasons.

1. No stage is involved that models the matching of speech to be recognized with the listener's speech knowledge (i.e. the speech memory). Such a stage is assumed to represent the pattern recognition in the human cortex.
2. Many details about the auditory periphery are not included in such a macroscopic model. If a model of human speech recognition shall be applied to model the consequences of hearing impairment on speech intelligibility these details of the auditory periphery may particularly be crucial.
3. The recognition of speech consisting of, e.g., sentences, is not split up into the recognition of smaller speech items, such as words or phonemes. It is very likely that human speech recognition includes analyzing and evaluating single phonemes for the recognition of words and sentences.

Another research field associated with human speech recognition is the application of speech feature extraction for digital transcription, i.e. Automatic Speech Recognition (ASR). Although both, a large commercial interest exists in obtaining a reliable ASR system that matches human performance, and, in addition, a long history of ASR research exists, there still is a large performance gap between human and automatic speech recognition (for an overview see Scharenborg (2007) or Meyer *et al.*

(2007a)). Common ASR systems consist of a feature extraction part, i.e. the transformation of a speech waveform to a set of numbers that represents this speech signal, and a recognizing part that uses statistical models about previously processed speech utterances. Such ASR systems work well if very limited speech response alternatives with high redundancy are used. However, they show surprisingly poor performance compared to humans if less redundancy is associated with the speech material, for example if single phonemes have to be recognized (Meyer, 2009). Furthermore, the robustness of ASR systems against external distortions such as background noise is by far poorer than the robustness of the human auditory system against these distortions (Stern *et al.*, 1996). The strategies usually used by ASR systems resemble human speech recognition in important aspects, some of these strategies implement many research findings about the human auditory system (e.g., Hermansky, 1990; Tchorz and Kollmeier, 1999). For instance, ASR systems use a memory of speech, mostly realized as a statistical model of previously processed speech recordings, and they subdivide the speech recognition process to smaller temporal speech objects, such as syllables and phonemes. Although two out of the three aforementioned drawbacks of macroscopic models are not present in ASR systems, it is however difficult to apply ASR to the prediction of human speech intelligibility, because the aim of ASR systems is not to provide a good model of human speech recognition, but to yield maximum speech recognition rates for applications. Furthermore, the development of ASR systems aims at robustness against disturbances such as background noise or reverberation. As these ASR systems do not even provide a good model of normal-hearing listeners' speech recognition, it is also difficult to implement hearing impairment in these models.

Holube and Kollmeier (1996) were the first to implement a more microscopic model of speech recognition by using an auditory model that extracts an 'internal representation' from a speech signal to be recognized, and a speech recognizer as a pattern matching backend. This model overcomes all the three drawbacks identified with macroscopic speech intelligibility models mentioned above and was used for modeling speech recognition results of a rhyme test, i.e. recognition of single meaningful words. Furthermore, it allows for adjusting the auditory model due to the dysfunction of model stages as observed in HI listeners. However, in the study of Holube and Kollmeier (1996), only average intelligibility prediction results were reported and compared to average measured speech intelligibility, i.e., no detailed assessment of the observed and predicted recognition of *single phonemes* was

performed. The current thesis therefore analyzes speech intelligibility in a more detailed way by using the Oldenburg Logatom speech corpus (OLLO) (Wesker *et al.*, 2005) to test the performance of an advanced auditory processing model on the level of single phoneme recognition.

Due to the variety and complexity of factors that contribute to hearing impairment, the way to implement (particularly sensorineural) hearing impairment into a speech recognition model is not completely clear. The contributing factors are partially difficult to assess in individual listeners and much more difficult to model (for an overview cf. Moore (1998) or Kollmeier (1999)). Audibility seems to be the most important part contributing to reduced speech intelligibility, but even people with similar audiograms may show different performances of speech recognition. Kollmeier (1999) therefore proposed four factors accounting for sensorineural hearing impairment that should be implemented within an auditory model: (1) loss of audibility, (2) loss of dynamic range, (3) increase of an 'internal noise', and (4) one factor that detriments binaural functions. The three latter factors affect the processing of sound well above the individual hearing threshold and can thus be termed "supra-threshold" factors. A supra-threshold processing different from normal might contribute to differences in speech recognition of hearing-impaired listeners with the same audiometric thresholds. Such a supra-threshold processing deficit may be associated with a pathological loudness perception that can be assessed using adaptive categorical loudness scaling (ACALOS) (Brand and Hohmann, 2001). A supra-threshold processing deficit may also manifest in a different input-output (I/O) function of the basilar membrane, which can be estimated using psychoacoustic masking experiments (e.g., Plack *et al.*, 2004). Although both of these measurement methods have frequently been used to characterize an individual's hearing deficit beyond the audiogram, no systematic model-driven investigation has yet been done on whether these supra-threshold processing deficits (estimated using the aforementioned measurement methods) affect speech recognition. The objectives of this dissertation therefore are:

(1) To develop a microscopic model of human phoneme and sentence recognition, which incorporates both a model about the auditory periphery and a speech recognizer. The analysis of the recognition and confusion of single phonemes is used to compare both model and human phoneme recognition thoroughly to get a better understanding about similarities and differences between model and human speech recognition.

(2) To find out in a systematic way how different factors of sensorineural hearing impairment, such as audibility of the speech and an altered peripheral compression, affect modeled speech recognition.

Chapter 2 introduces the microscopic model for the prediction of phoneme recognition in normal-hearing listeners in noise. Different model configurations are used to quantify the performance gap between human and automatic speech recognition. The impact of different perceptive distance measures used within the recognizing stage on predicted speech recognition is analyzed. The microscopic model is evaluated using nonsense speech material, and phoneme confusion matrices of normal-hearing listeners are compared with that of the model. As a predecessor to the model approach and to the complete results described in Chapter 2, the difference between human speech recognition and automatic speech recognition was already assessed within initial work using only one perceptive distance measure. The paper describing this initial work is reprinted in the appendix of this dissertation (Chapter 7).

Chapter 3 extends the model of Chapter 2 by implementing hearing impairment into the auditory model and by modeling single-word recognition in whole sentences rather than phoneme recognition as in Chapter 2. Furthermore, a comparison of the predictive power of this extended microscopic model of speech recognition to the Speech Intelligibility Index (SII) is done. In this chapter hearing impairment is accounted for only by the audibility (i.e. the absolute hearing threshold) of the speech quantified by the pure-tone audiogram. Supra-threshold factors that might influence individual speech intelligibility results are not regarded. Therefore, this model approach resembles the approach standardized within the SII that also regards only the individual audibility of hearing-impaired listeners.

A method of assessing supra-threshold factors in normal-hearing and hearing-impaired listeners is described in Chapter 4. In addition to assessing the audibility using the pure-tone audiogram only, parameters of the supra-threshold processing, such as outer hair cell loss and inner hair cell loss, are assessed using adaptive categorical loudness scaling (ACALOS). ACALOS has the advantage of being a fast and efficient measurement method that has the potential of being used widely in clinical practice. The results are compared to results from temporal masking curves (TMCs), a forward-masking experiment that is widely accepted in the literature for inferring I/O function of the auditory system, but requires much more measurement time compared to ACALOS.

In Chapter 5 different model versions of the auditory periphery are realized within the microscopic model of speech recognition. Some of these versions incorporate parameters inferred from the method introduced in Chapter 4. Consonant recognition of normal-hearing and hearing-impaired listeners in quiet condition is predicted and the impact of adjusting supra-threshold parameters on predicted consonant recognition is investigated.

At large, this dissertation covers a wide range of topics from psychoacoustics to human speech recognition and automatic speech recognition in order to obtain a better understanding of the normal and impaired human auditory system.

2 Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model[1]

Abstract

This study compares the phoneme recognition performance in speech-shaped noise of a microscopic model for speech recognition with the performance of normal-hearing listeners. “Microscopic” is defined in terms of this model twofold. First, the speech recognition rate is predicted on a phoneme-by-phoneme basis. Second, microscopic modeling means that the signal waveforms to be recognized are processed by mimicking elementary parts of human’s auditory processing. The model is based on an approach by Holube and Kollmeier [J. Acoust. Soc. Am. **100**, 1703–1716 (1996)] and consists of a psychoacoustically and physiologically motivated preprocessing and a simple dynamic-time-warp speech recognizer. The model is evaluated while presenting nonsense speech in a closed-set paradigm. Averaged phoneme recognition rates, specific phoneme recognition rates, and phoneme confusions are analyzed. The influence of different perceptual distance measures and of the model’s *a-priori* knowledge is investigated. The results show that human performance can be predicted by this model using an optimal detector, i.e., identical speech waveforms for both training of the recognizer and testing. The best model performance is yielded by distance measures which focus mainly on small perceptual distances and neglect outliers.

[1] This chapter was published as Jürgens and Brand (2009), reprinted with permission from Jürgens T., Brand T., J. Acoust. Soc. Am., Vol. 126, Pages 2635-2648, (2009).

2.1 Introduction

The methods usually used for speech intelligibility prediction are index-based approaches, for instance, the articulation index (AI) (ANSI, 1969), the speech transmission index (STI) (Steeneken and Houtgast, 1980), and the speech intelligibility index (SII) (ANSI, 1997). AI and SII use the long-term average frequency spectra of speech and noise separately and calculate an index that can be transformed into an intelligibility score. The parameters used for the calculation are tabulated and mainly fitted to empirical data. These indices have been found to successfully predict speech intelligibility for normal-hearing subjects within various noise conditions and in silence (e.g., Kryter, 1962; Pavlovic, 1987). The STI is also index-based and uses the modulation transfer function to predict the degradation of speech intelligibility by a transmission system. All of these approaches work "macroscopically", which means that macroscopic features of the signal like the long-term frequency *spectrum* or the *signal-to-noise ratios* (SNRs) in different frequency bands are used for the calculation. Detailed *temporal* aspects of speech processing that are assumed to play a major role within our auditory speech perception are neglected. Some recent modifications to the SII improved predictions of the intelligibility in fluctuating noise (Rhebergen and Versfeld, 2005; Rhebergen *et al.*, 2006; Meyer *et al.*, 2007b) and included aspects of temporal processing by calculating the SII based on short-term frequency spectra of speech and noise. However, even these approaches do not mimic all details of auditory preprocessing that are most likely involved in extracting the relevant speech information. Furthermore, the model approaches mentioned above are "macroscopic" in a second sense as they usually predict average recognition rates of whole sets of several words or sentences and not the recognition rates and confusions of single phonemes.

The goal of this study is to evaluate a "microscopic" speech recognition model for normal-hearing listeners. We define microscopic modeling twofold. First, the *particular stages* involved in the speech recognition of normal-hearing human listeners are modeled in a typical way of psychophysics based on a detailed "internal representation" (IR) of the speech signals. Second, the recognition rates and confusions of *single phonemes* are compared to those of human listeners. This definition is in line with Barker and Cooke (2007), for instance. In our study, this kind of modeling is aimed at understanding the factors contributing to the perception of speech in normal-hearing listeners and may be extended to other acoustical signals or to understanding the implications of hearing impairment on speech perception (for an overview see, e.g.,

Moore (2003)). Toward this goal we use an auditory preprocessing based on the model of Dau *et al.* (1996a) that processes the signal waveform. This processed signal is then recognized by a dynamic-time-warp (DTW) speech recognizer (Sakoe and Chiba, 1978). This is an approach proposed by Holube and Kollmeier (1996). The novel aspect of this study compared to Holube and Kollmeier (1996) is that the influence of different perceptual distance measures used to distinguish between phonemes within the speech recognizer is investigated in terms of the resulting phoneme recognition scores. Furthermore, we evaluate the predictions of this model on a phoneme scale, which means that we compare confusion matrices as well as overall speech intelligibility scores. This is a method commonly used in automatic speech recognition (ASR) research.

2.1.1 Microscopic modeling of speech recognition

There are different ways to predict speech intelligibility using auditory models. Stadler *et al.* (2007) used an information-theory approach in order to evaluate preprocessed speech information. This approach predicts the speech reception threshold (SRT) very well for subjects with normal hearing for a Swedish sentence test. Another way was presented by Holube and Kollmeier (1996) who used a DTW speech recognizer as a back-end to the auditory model proposed by Dau *et al.* (1996a). They were able to predict speech recognition scores of a rhyme test for listeners with normal hearing and with hearing impairment with an accuracy comparable to that of AI and STI. Both Stadler *et al.* (2007) and Holube and Kollmeier (1996) used auditory models that were originally fitted to other psychoacoustical experiments, such as masking experiments of non-speech stimuli, for instance.

Several studies indicate that temporal information is essential for speech recognition. Chi *et al.* (1999) and Elhilali *et al.* (2003), for instance, compared the predictions of a spectro-temporal modulation index to the predictions of the STI and showed that spectro-temporal modulations are crucial for speech intelligibility. They concluded that information within speech is not separable into a temporal-only and a spectral-only part but that also joint spectro-temporal dimensions contribute to overall performance. Christiansen *et al.* (2006) showed that temporal modulations of speech play a crucial role in consonant identification. For these reasons, this study uses a slightly modified version of the approach by Holube and Kollmeier (1996). The modification is a modulation filter bank (Dau *et al.*, 1997) extending the perception model of Dau *et al.* (1996a), which gives the input for the speech recognition stage. It

provides the recognizer with information about the modulations in the different frequency bands. The whole auditory model is based on psychoacoustical and physiological findings and was successful in describing various masking experiments (Dau *et al.*, 1996b), modulation detection (Dau *et al.*, 1997), speech quality prediction (Huber and Kollmeier, 2006), and aspects of timbre perception (Emiroğlu and Kollmeier, 2008). Using a speech recognizer subsequently to the auditory model, as proposed by Holube and Kollmeier (1996), allows for predicting the SRT of an entire speech test. This approach can certainly not account for syntax, semantics, and prosody that human listeners take advantage of. To rule out these factors of human listeners' speech recognition, in the experiments of this study nonsense speech material is presented in a closed response format. The use of this speech material provides a fair comparison between the performance of human listeners and the model (cf. Lippmann, 1997). Furthermore, a detailed analysis of recognition rates and confusions of single phonemes is possible. Confusion matrices can be used in order to compare phoneme recognition rates and phoneme confusions between both humans and model results. Confusion matrices, like those used by Miller and Nicely (1955), can also be used to compare recognition rates between different phonemes provided that systematically composed speech material such as logatomes (short sequences of phonemes, e.g., vowel-consonant-vowel-utterances) is used.

The nonsense speech material of the Oldenburg logatome (OLLO) corpus (Wesker *et al.*, 2005), systematically composed from German vowels and consonants, is used for this task. This corpus was used in a former study (cf. Meyer *et al.*, 2007a) to compare human's speech performance with an automatic speech recognizer. The OLLO speech material in the study of Meyer *et al.* (2007a) allowed excluding the effect of language models that are often used in speech recognizers. Language models store plausible possible words and can use this additional information to crucially enhance the performance of a speech recognizer. Nonsense speech material was also used, for instance, in speech and auditory research to evaluate speech recognition performance of hearing-impaired persons (Dubno *et al.*, 1982; Zurek and Delhorne, 1987) and to make a detailed performance comparison between automatic and human speech recognition (HSR) (Sroka and Braida, 2005). Furthermore, nonsense speech material was used, for instance, to evaluate phonetic feature recognition (Turner *et al.*, 1995) and to evaluate consonant and vowel confusions in speech-weighted noise (Phatak and Allen, 2007).

2.1.2 A-priori knowledge

A model for the prediction of speech intelligibility, which uses an internal ASR stage deals with the usual problems of such ASR systems: error rates are much higher than those of normal-hearing human listeners in clean speech (cf. Lippmann, 1997; Meyer and Wesker, 2006) and in noise (Sroka and Braida, 2005; Meyer *et al.*, 2007a). Speech intelligibility models without an ASR stage, e.g., the SII, are provided with more *a-priori* information about the speech signal. The SII "knows" which part of the signal is speech and which part of the signal is noise because it gets them as separate inputs, which is an unrealistic and "unfair" advantage over models using an ASR stage. For modeling of HSR the problem of too high error rates when using a speech recognizer can be avoided using an "optimal detector" (cf. Dau *et al.*, 1996a), which is also used in many psychoacoustical modeling studies. It is assumed that the recognizing stage of the model after the auditory preprocessing has perfect *a-priori* knowledge of the target signal. Limitations of the model performance are assumed to be completely located in the preprocessing stage. This strategy can be applied to a speech recognizer using template waveforms (for the training of the ASR stage) that are identical to the waveforms of the test signals except for a noise component constraining the performance. Holube and Kollmeier (1996) applied an optimal detector in form of a DTW speech recognizer as a part of their speech intelligibility model using identical speech recordings that were added with different noise passages for the model training stage and for recognizing. Hant and Alwan (2003) and Messing *et al.* (2008) also used this "frozen speech" approach to model the discrimination of speech-like stimuli. Assuming perfect *a-priori* knowledge using an optimal detector (i.e., using identical recordings as templates and as test items) is one special case of modeling human's speech perception. Another case is using different waveforms for testing and training, thus assuming only limited knowledge about the target signal. This case corresponds not to an optimal detector but to a limited one. The latter is the standard of ASR; the former is widely used in psychoacoustic modeling. In this study, we use both the optimal detector approach and a typical ASR approach. In this way it is possible to investigate how predictions of these two approaches differ and whether the first or the second method is more appropriate for microscopic modeling of speech recognition.

2.1.3 Measures for perceptual distances

Because the effects of higher processing stages (like word recognition or use of semantic knowledge) have been excluded in this study by the use of nonsense speech

material, it is possible to focus on the sensory part of speech recognition. As a working hypothesis we assume that the central human auditory system optimally utilizes the speech information included in the IR of the speech signal. This information is used to discriminate between the presented speech signal and other possible speech signals. We assume that the auditory system somehow compares the incoming speech information to an internal vocabulary "on a perceptual scale." Therefore, the following questions are of high interest for modeling: What are the mechanisms of comparing speech sounds and what is the best distance measure, on a perceptual scale, for an optimal exploitation of the speech information? For the perception of musical tones Plomp (1976) compared the perceived similarity of tones to their differences within an equivalent rectangular bandwidth (ERB) sound pressure level spectrogram using different distance measures. Using the absolute value metric, he found higher correlations than using the Euclidean metric. For vowel sounds, however, he found a high correlation using the Euclidean metric. Emiroğlu (2007) also found that the Euclidean distance is more appropriate than, e.g., a cross-correlation measure for the comparison of musical tones. The Euclidean distance was also used by Florentine and Buus (1981) to model intensity discrimination and by Ghitza and Sondhi (1997) to derive an optimal perceptual distance between two speech signals. Although the Euclidean distance was preferred by these authors for modeling the perception of sound signals, especially of speech, it still seems to be useful in this study to analyze the differences occurring on the model's "perceptual scale." By using an optimal distance measure, deduced from the empirically found distribution of these differences, the model recognition performance can possibly be optimized.

2.2 Method

2.2.1 Model structure

2.2.1.1 The perception model

Figure 2.1 shows the processing stages of the perception model. The upper part of this sketch represents the training procedure. A template speech signal with optionally added background noise serves as input to the preprocessing stage. The preprocessing consists of a gammatone-filterbank (Hohmann, 2002) to model the peripheral filtering in the cochlea. 27 gammatone filters are equally spaced on an ERB-scale with one filter per

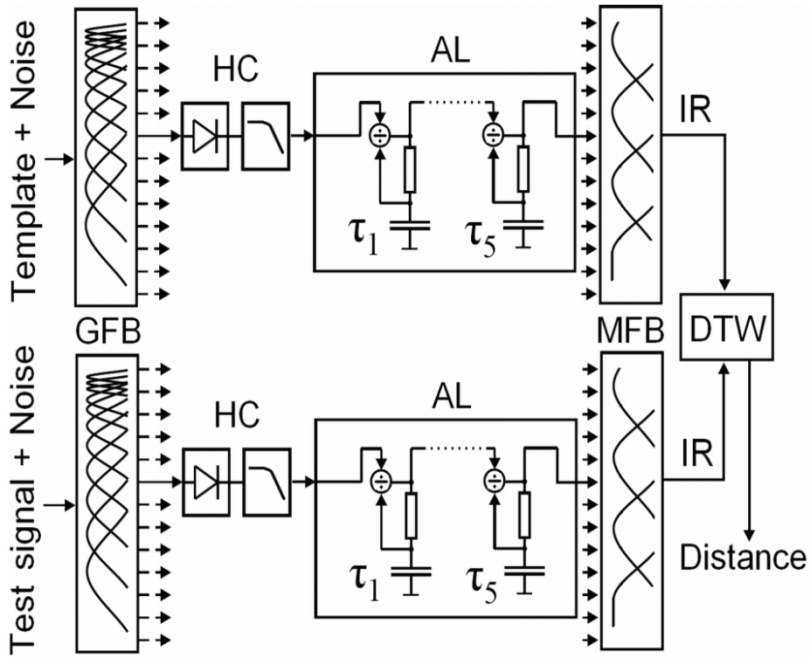

Figure 2.1: Scheme of the perception model. The time signals of the template recording added with running noise and the time signal of the test signal added with running noise are preprocessed in the same effective "auditorylike" way. A gammatone filterbank (GFB), a haircell (HC) model, adaptation loops (ALs), and a modulation filterbank (MFB) are used. The outputs of the modulation filterbank are the internal representations (IRs) of the signals. They serve as inputs to the Dynamic-Time-Warp (DTW) speech recognizer that computes the "perceptual" distance between the IRs of the test logatome and the templates.

ERB covering a range of center frequencies from 236 Hz to 8 kHz. In contrast to Holube and Kollmeier (1996), gammatone filters with center frequencies from 100 to 236 Hz are omitted because these filters are assumed not to contain information that is necessary to discriminate different phonemes. This is consistent with the frequency channel weighting within the calculation of the SII (ANSI, 1997) and our own preliminary results. A hearing threshold simulating noise that is spectrally shaped to human listeners' audiogram data (according to IEC 60645-1) is added to the signal before it enters the gammatone-filterbank (GFB) (cf. Beutelmann and Brand, 2006). The noise is assumed to be 4 dB above human listeners' hearing threshold for all frequencies, as proposed by Breebaart *et al.* (2001)[1]. Each filter output is half-wave rectified and filtered using a first order low pass filter with a cut-off frequency of 1 kHz mimicking a very simple hair cell (HC) model. The output of this HC model is then compressed using five consecutive adaptation loops (ALs) with time constants as given in Dau *et al.* (1996a) (τ_1=5 ms, τ_2=50 ms, τ_3=129 ms, τ_4 =253 ms, and τ_5=500 ms).

[1]Breebart *et al.* (2001) found out that a 9.4 dB SPL Gaussian noise within one gammatone filter channel just masks a sinusoid with 2 kHz frequency at absolute hearing threshold (5 dB SPL, which is about 4 dB lower). This approach was extrapolated for other audiometric frequencies.

These ALs compress stationary time signals approximately logarithmically and emphasize on- and offsets of non-stationary signals. Furthermore, a modulation filterbank (MFB) according to Dau *et al.* (1997) is used. It contains four modulation channels per frequency channel: one low pass with a cut-off frequency of 2.5 Hz and three band passes with center frequencies of 5, 10, and 16.7 Hz. The bandwidths of the band pass filters are 5 Hz for center frequencies of 5 and 10 Hz, and 8.3 Hz for the band pass with center frequency of 16.7 Hz. The output of this model is an IR that is downsampled to a sampling frequency of 100 Hz. The IR thus contains a twodimensional feature-matrix at each 10 ms time step consisting of 27 frequency channels and four modulation frequency channels. The elements of this matrix are given in arbitrary model units (MU). Without the MFB 1 MU corresponds to 1 dB sound pressure level (SPL).

2.2.1.2 The DTW speech recognizer

The IR is passed to a DTW speech recognizer (Sakoe and Chiba, 1978) to "recognize" a speech sample. This DTW can be used either as an optimal detector by using a configuration that contains perfect *a-priori* knowledge or as a limited detector by withholding this knowledge (for details about these configurations see below). The DTW searches for an optimal time-transformation between the IRs of the template and the test signal by locally stretching and compressing the time axes.

The optimal time-transformation between two IRs is computed by first creating a distance matrix *D*. Each element *D*(*i*, *j*) of this matrix is given by the distance between the feature-matrices of the template's IR (IR_{templ}) at time index *i* and the feature-matrix of the test item's IR (IR_{test}) at time index *j*. Different distance measures are possible in this procedure (see below). As a next step a continuous "warp path" through this distance matrix is computed (Sakoe and Chiba, 1978). This warp path has the property that averaging the matrix elements along the warp path results in a minimal overall distance. The output of the DTW is this overall distance and thus is a distance between these IRs. From an assortment of possible templates the template with the smallest distance is chosen as the recognized one.

2.2.1.3 Distance measures

In a first approach the Euclidean distance

$$D_{Euclid}(i,j) = \sqrt{\sum_{f_{\mathrm{mod}}}\sum_{f}\left(IR_{templ}(i,f,f_{\mathrm{mod}}) - IR_{test}(j,f,f_{\mathrm{mod}})\right)^2} \qquad (2.1)$$

between the feature-vectors IR_{templ} and IR_{test} was used with f denoting the frequency channel and f_{mod} denoting the modulation-frequency channel of the IRs (Jürgens *et al.*, 2007). In many studies this Euclidean distance is used when comparing perceptual differences (e.g., Plomp, 1976; Holube and Kollmeier, 1996). The Euclidean distance measure implies a Gaussian distribution of the differences between template and test IR.

As an example, Figure 2.2 panel 1 shows the normalized histogram of differences Δd between the IRs (IR_{templ} and IR_{test}) of two different recordings of the logatome /ada:/:

$$\Delta d(f, f_{\mathrm{mod}}, i, j) = IR_{templ}(i, f, f_{\mathrm{mod}}) - IR_{test}(j, f, f_{\mathrm{mod}}) \qquad (2.2).$$

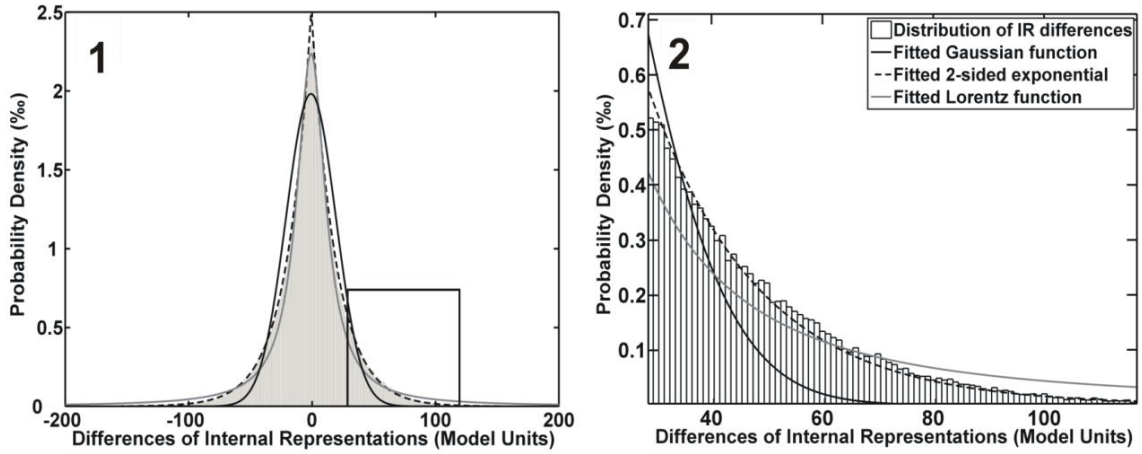

Figure 2.2: Distribution of differences (in MU) between IRs of two different recordings of the logatome /ada:/. The recordings were spoken by the same male German speaker with "normal" speech articulation style and mixed with ICRA1-noise at 0 dB SNR. A Gaussian, a two-sided exponential, and a Lorentz-function were fitted to the data, respectively. Panel 1: complete distribution; panel 2: detail (marked rectangular) of panel 1.

In this example, the logatomes were spoken by the same male German speaker and mixed with two passages of uncorrelated ICRA1-noise (Dreschler *et al.*, 2001) at 0 dB SNR. The ICRA1-noise is a steady-state noise with speech-shaped long-term spectrum. Note that Δd corresponds to all differences occurring within a distance matrix, even those that are not part of the final warp path. However, the shape of the histogram is typical of almost all speakers and all SNRs. To investigate the shape of the histogram of differences Δd between these two IRs a Gaussian probability density function (PDF)

$$PDF_{Gauss}(\Delta d) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{1}{2}\left(\frac{\Delta d_{\max} - \Delta d}{\sigma} \right)^2 \right) \qquad (2.3)$$

is fitted to the distribution which corresponds to the Euclidean metric (Eq. (2.1)) and a two-sided exponential PDF

$$PDF_{Exp}(\Delta d) = \frac{1}{2\sigma} \cdot \exp\left(-\left| \frac{\Delta d_{\max} - \Delta d}{\sigma} \right| \right) \quad (2.4),$$

and a Lorentzian probability density function

$$PDF_{Lorentz}(\Delta d) = \frac{1}{\sqrt{2}\pi\sigma} \cdot \frac{1}{1 + \frac{1}{2}\left(\frac{\Delta d_{\max} - \Delta d}{\sigma} \right)^2} \quad (2.5)$$

are also fitted to the distribution, respectively. Two fitting parameters, the width of the fitted curve given by σ and the position of the maximum Δd_{max}, must be set. The fits in Figure 2.2 panel 1 show that the distribution is almost symmetrical with $\Delta d_{max} = 0$ and that high distances of about 50 MU or more are very much more frequent than expected when assuming Gaussian distributed data. Especially, very high distances of about 80 MU or more (cf. Figure 2.2 panel 2) are present in the tail of outliers. The Lorentzian PDF provides a better fit than the Gaussian function. However, it slightly overestimates the amount of outliers. The two-sided exponential function provides the best fit to the data. The two-sided exponential function is capable of reproducing the shape of the mean peak at 0 MU as well as the shape of the tail of outliers.

By taking the negative logarithm of a PDF (Eqs. (2.3)–(2.5)) and summing up the distances across all frequency channels and modulation frequency channels, a distance measure is obtained (cf. Press *et al.*, 1992) that can be used within the speech recognition process. This gives the Euclidean distance metric (Eq. (2.1)) (for Gaussian distributed data), the absolute value distance metric

$$D_{abs}(i,j) = \sum_{f_{\mathrm{mod}}} \sum_{f} \left(\left| IR_{templ}(i, f, f_{\mathrm{mod}}) - IR_{test}(j, f, f_{\mathrm{mod}}) \right| \right) \quad (2.6),$$

and the Lorentzian distance metric

$$D_{Lorentz}(i,j) = \sum_{f_{\mathrm{mod}}} \sum_{f} \log\left[1 + \frac{1}{2} \left(IR_{templ}(i, f, f_{\mathrm{mod}}) - IR_{test}(j, f, f_{\mathrm{mod}}) \right)^2 \right] \quad (2.7),$$

Note that the prefactors that normalize the PDFs are not included within Eqs. (2.1), (2.6), and (2.7) because they represent a constant offset in the distance metric which has no effect on the position of the minimum of the overall distance. The parameter σ is set to 1 MU for simplicity. For Eqs. (2.1) and (2.6) the value of σ is not relevant to finding the best warp path through the distance matrix (i.e., solving a constrained minimizing problem). However, in Eq. (2.7), σ is relevant to finding the best warp path because it

cannot be factored out as it can for the Euclidean and the absolute value metric. Choosing σ equal to 1 MU results in a very flat behavior of the distance metric for middle and high distances. Other values of σ in the range from 60 to 0.1 MU showed only minor influence to the performance results in preliminary experiments.

A hypothesis for the present study is that using either Eq. (2.6) or Eq. (2.7) instead of the Euclidean distance (Eq. (2.1)) within the DTW speech recognition process may better account for the characteristic differences of the IRs and may improve matching.

2.2.2 Speech corpus

Speech material taken from the OLLO speech corpus (Wesker *et al.*, 2005)[1] is used in this study. The corpus consists of 70 different vowel-consonant-vowel (VCV) and 80 consonant-vowel-consonant (CVC) logatomes composed of German phonemes. The first and the last phoneme of one logatome are the same. The middle phonemes of the logatomes are either vowels or consonants which are listed below (represented with the International Phonetic Alphabet, (IPA, 1999)).

• Consonants:

/p/, /t/, /k/, /b/, /d/, /g/, /s/, /f/, /v/, /n/, /m/, /ʃ/, /ts/, /l/

• Vowels:

/a/, /a:/, /ɛ/, /e/, /ɪ/, /i/, /ɔ/, /o/, /ʊ/, /u/

Consonants are embedded in the vowels /a/, /ɛ/, /ɪ/, /ɔ/, and /ʊ/, respectively, and vowel phonemes are embedded in the consonants /b/, /d/, /f/, /g/, /k/, /p/, /s/, and /t/, respectively. Most of these logatomes are nonsense in German[2]. The logatomes are spoken by 40 different speakers from four different dialect regions in Germany and by ten speakers from France. The speech material covers several speech variabilities such as speaking rate, speaking effort, different German dialects, accent, and speaking style (statement and question). In the present study, only speech material of one male German speaker with no dialect and with “normal” speech articulation style is used.

[1] The OLLO corpus is freely available at http://sirius.physik.uni-oldenburg.de.
[2] Even if very few logatoms in this corpus are forenames or may have a meaning in certain dialect regions in Germany these logatoms are not excluded in this study to preserve the very systematic composition of this speech corpus.

2.2.3 Test conditions

Calculations with the perception model as well as measurements with human listeners were performed under highly similar conditions. The same recordings from the logatome corpus were used. The logatomes were arranged into groups in which only the middle phoneme varied. With this group of alternatives a closed testing procedure was performed. This means that both the model and the subject had to choose from identical groups of logatomes. This allowed for a fair comparison of human and modeled speech intelligibility because the humans' semantic and linguistic knowledge had no appreciable influence. Furthermore, it allowed the recognition rates and confusions of phonemes to be analyzed. The speech waveforms were set to 60 dB SPL. Stationary noise with speechlike long-term spectrum (ICRA1-noise, Dreschler *et al.*, 2001) downsampled to a sampling frequency of 16 kHz was added to the recordings and 400 ms prior to the recording. The whole signal was faded in and out using 100 ms Hanning-ramps. After computing the IR of the speech signals as described in Section 2.2.1.1, the part of it corresponding to the 400 ms noise prior to the speech signal was deleted. This was done in order to give only the information required for discriminating phonemes to the speech recognizer and not the preceding IR of the preceding background noise.

2.2.4 Modeling of a-priori knowledge

Two configurations of *a-priori* knowledge of the speech recognizer were realized.

- In configuration A five IRs per logatome calculated from five different waveforms were used as templates. The waveforms were randomly chosen from the recordings of one single male speaker with normal speech articulation style. None of the five waveforms underlying these IRs (the vocabulary) was identical to the tested waveform. The logatome yielding the minimum average distance between the IR of the test sample and all five IRs of the templates was chosen as the recognized one. This limited detector approach mimics a realistic task of automatic speech recognizers because the exact acoustic waveform to be recognized was unknown.
- Model configuration B used a single IR per logatome as template. The waveform of the correct response alternative was identical to the waveform of the test signal. Thus, the resulting IRs of test signal and the correct response alternative differed only in the added background noise and hearing threshold simulating noise that were uncorrelated in time. In contrast to configuration A, this configuration disregards the natural variability of speech. Thus, it assumes perfect knowledge of

the speech template to be matched using the DTW algorithm and corresponds to an optimal detector approach.

The calculation was performed ten times using different passages of background noise and hearing threshold simulating noise according to the individual audiograms of listeners participating in the experiments. The whole calculation took 100 h for configuration A (ten times for 150 logatomes at nine SNR values) and 13 h for configuration B on an up to date standard PC.

2.2.5 Subjects

Ten listeners with normal hearing (seven male, three female) aged between 19 and 37 years were employed. Their absolute hearing threshold for pure tones in standard audiometry did not exceed 10 dB hearing level (HL) between 250 Hz and 8 kHz. Only one threshold hearing loss of 20 dB at one audiometric frequency was accepted.

2.2.6 Speech tests

The recognition rates of 150 different logatomes were assessed using Sennheiser HDA 200 headphones in a sound-insulated booth. The calibration was performed using a Brüel&Kjaer (B&K) measuring amplifier (Type 2610), a B&K artificial ear (Type 4153), and a B&K microphone (Type 4192). All stimuli were free-field-equalized using an FIR-filter with 801 coefficients and were presented diotically. SNRs of 0, −5, −10, −15, and −20 dB were used for the presentation to human listeners. For each SNR a different presentation order of the 150 logatomes was randomly chosen. For this purpose, the 150 recordings were split into two lists, and the order of presentation of the recordings within the two lists was shuffled. Then all ten resulting lists of all SNRs were randomly interleaved for presentation. Response alternatives for a single logatome had the same preceding and subsequent phoneme (closed test); hence, the subject had to choose either from 10 (CVC) or 14 (VCV) alternatives. The subject was asked to choose the recognized logatome from the list and was asked to guess if nothing was understood. The order of response alternatives shown to the subject was shuffled as well. Before the main measurement all subjects were trained with a list of 50 logatomes.

For characterizing the mean intelligibility scores across all logatomes the model function

$$\Psi(L) = \frac{1-g}{1+\exp(4\cdot s\cdot(SRT-L))} + g \qquad (2.8)$$

was fitted to the mean recognition rate (combined for CVCs and VCVs) for each SNR by varying the free parameters SRT and *s* (slope of the psychometric function at the SRT). The SRT is the SNR at approximately 55% recognition rate (averaged across all CVCs and VCVs), which is the midpoint between the guessing probability and 100%. *L* corresponds to the given SNR and *g* is the guessing probability averaged across all CVCs and VCVs (g = 8.9%). The fit is performed by maximizing the likelihood assuming that the recognition of each logatome is a Bernoulli trial (cf. Brand and Kollmeier, 2002). Note that this fitting function assumes that 100% recognition rate is reached at high SNRs. This is feasible for listeners with normal hearing and for speech recognition modeling using an optimal detector, but is not necessarily the case for a real ASR system as such an ASR system will still show high error rates on speech material with a low redundancy even when the SNR is very high (Lippmann, 1997). For model configuration A the fitting curve is therefore fixed at the highest recognition rate that occurred in the ASR test.

2.3 Results and discussion

2.3.1 Average recognition rates

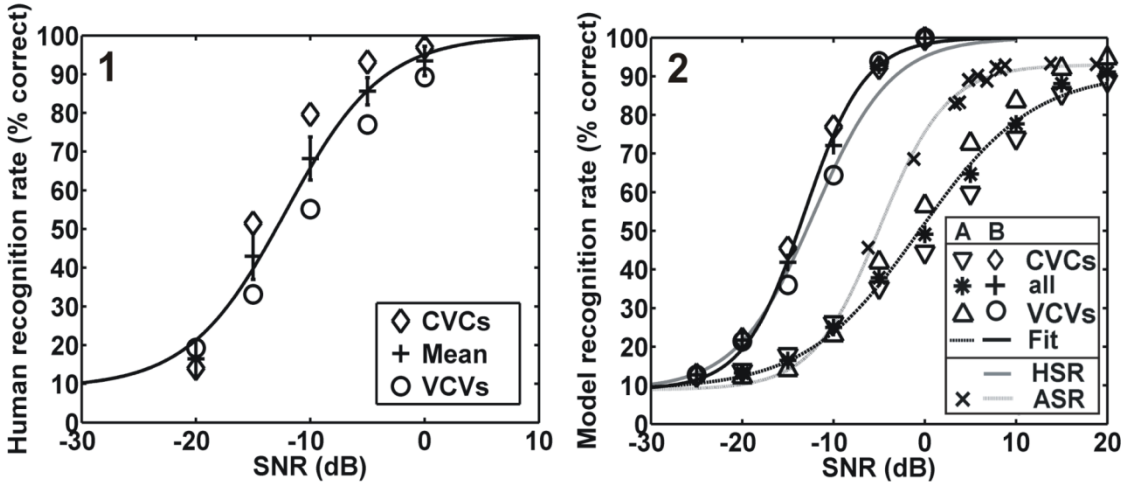

Figure 2.3: Panel 1: Psychometric function (recognition rate versus SNR) of ten normal-hearing listeners using logatomes in ICRA1-noise. Error bars correspond to the inter-individual standard deviations across subjects. Lines show the fit by Eq. (2.8). Panel 2: Psychometric function of the perception model with configurations A and B derived with the same utterances of the OLLO speech corpus as for the measurement. The measured psychometric function (taken from panel 1) is additionally shown for comparison as gray line (HSR). For a further comparison, data of Meyer *et al.* (2007a) are plotted (ASR).

Figure 2.3 panel 1 shows the mean phoneme recognition rates in percent correct versus SNR across all phonemes. Error bars denote the inter-individual standard deviations of the ten normal-hearing subjects. Furthermore, the recognition rates of CVCs and VCVs are plotted separately. The recognition rates for CVCs are higher than for VCVs except for −20 dB SNR. The fitting of the psychometric function to the data yields a slope of 5.4 ± 0.6%/dB and a SRT of −12.2 ± 1.1 dB. Note that even the recognition rate at −20 dB SNR is significantly above chance and therefore included in the fitting procedure.

Table 2.1: List of fitted parameters characterising observed and predicted psychometric functions for the discrimination of logatomes in ICRA1 noise. Rows denote different distance measures used by the Dynamic-Time-Warp speech recognizer and different model configurations (see Section 2.2.1 and 2.2.4 for details) as well as values of human listeners. Pearson's rank correlation coefficients (last column) were calculated using the observed data of individual human listeners. * denotes significant ($p < 0.05$) and ** highly significant ($p < 0.01$) correlations.

	SRT / dB SNR	Difference to observed SRT / dB	Slope / (%/dB)	Pearson's r^2
Human listeners	-12.2	0†	5.4	1†
Euclidian, Conf. A	-0.4	11.8	5.7	0.64**
Euclidian, Conf. B	-8.1	4.1	10.0	0.83**
2-sided exp., Conf. A	-0.4	11.8	5.8	0.65**
2-sided exp., Conf. B	-10.6	1.6	8.4	0.92**
Lorentzian, Conf. A	-0.6	11.6	3.5	0.83**
Lorentzian, Conf. B	-13.2	-1.0	6.8	0.97**

†: by definition

The observed and the predicted results calculated with different distance measures and model configurations are shown in Table 2.1. The smallest differences from the observed SRT values are found for configuration B. Using this configuration, the slope of the predicted psychometric function is slightly overestimated. However, model configuration A, which performs a typical task of speech recognizers, shows a large gap of about 12 dB between predicted and observed SRTs, which is typical of ASR (see below). This gap is nearly independent of the type of distance measure, while the slope is slightly underestimated. The last column of Table 2.1 shows Pearson's squared rank correlation coefficient r^2 between the individual observed and predicted speech recognition scores. The Lorentzian distance measure using model configuration B shows the highest r^2 of 0.97 ($p<0.01$), whereas the two-sided exponential and the Euclidean distance measure show somewhat lower correlation coefficients and higher

differences between observed and predicted SRTs. Different distance measures do not substantially affect the prediction of the SRT using model configuration A.

The predicted psychometric function of this best fitting model realization (configuration B with Lorentzian distance measure) is displayed in Figure 2.3 panel 2. In addition, the fitted psychometric function of Figure 2.3 panel 1 is replotted (HSR), and the predicted psychometric function of model configuration A with Lorentzian distance measure is shown. Furthermore, ASR-data of Meyer *et al.* (2007a) were included for comparison (see Section 2.4.1). For model configuration B the resulting SRT using the Lorentzian distance measure is −13.2 dB SNR and thus within the interval of the subjects' inter-individual standard deviation. The ranking of the recognition of vowels and consonants (i.e., that CVCs are better understood than VCVs) is predicted correctly except for −20 dB SNR. Model configuration A, which performs a typical task of speech recognizers, shows a SRT of −0.6 dB and a slope of 3.5%/dB using the Lorentzian distance measure. With this configuration the ranking of the recognition of vowels and consonants could not be predicted, i.e., the model shows higher recognition rates for consonants than for vowels.

2.3.2 Phoneme recognition rates at different SNRs

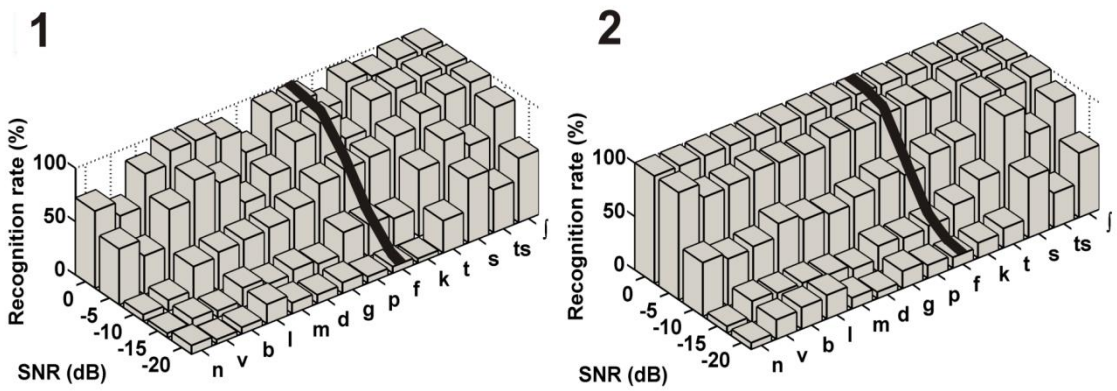

Figure 2.4: Recognition rates of consonants, separately, as a function of SNR for ten normal-hearing listeners (panel 1) and for model configuration B with Lorentzian distance measure (panel 2). As an example the psychometric function for the discrimination of /f/ in noise is shown (solid line).

Figure 2.4 shows the recognition rates of single consonants embedded in logatomes as a function of SNR for normal-hearing listeners (panel 1) and for model configuration B using the Lorentzian distance measure (panel 2). Picking out one phoneme, the psychometric function for this specific phoneme can be seen. The solid lines in panels 1 and 2 show these psychometric functions for the phoneme /f/ as an example. Normal-

hearing listeners show quite poor recognition rates for the phonemes /n/, /v/, or /g/ at the SNRs chosen for measurement. However, there are also some phonemes like /s/, /ts/, and /ʃ/ that show very high recognition rates at these SNRs. The predicted recognition rates for the latter phonemes (see panel 2) fit the observed recognition rates quite well. This is also the case for /l/, /m/, /p/, /f/, and /t/. For the other phonemes there is a discrepancy between observed and predicted recognition rates especially at high SNRs. For instance, at 0 dB SNR the predicted recognition rate is almost 100% for all phonemes, but normal-hearing listeners actually show poor recognition rates of 58% for /v/ or 70% for /g/. The recognition rates for vowels across SNR are shown in Figure 2.5. Normal-hearing listeners show quite a steep psychometric function for the phonemes /e/, /ɛ/, /a:/, and /i/ but a shallower psychometric function for the other phonemes. The predicted recognition rates for /o/ and /u/ fit the observed recognition rates quite well across all SNRs investigated in this study. However, for /e/, /ɛ/, /a:/, and /i/the predicted psychometric functions are too shallow. Note that for vowels, contrary to consonants, at 0 dB SNR almost 100% recognition rates are reached by both normal-hearing listeners and model configuration B.

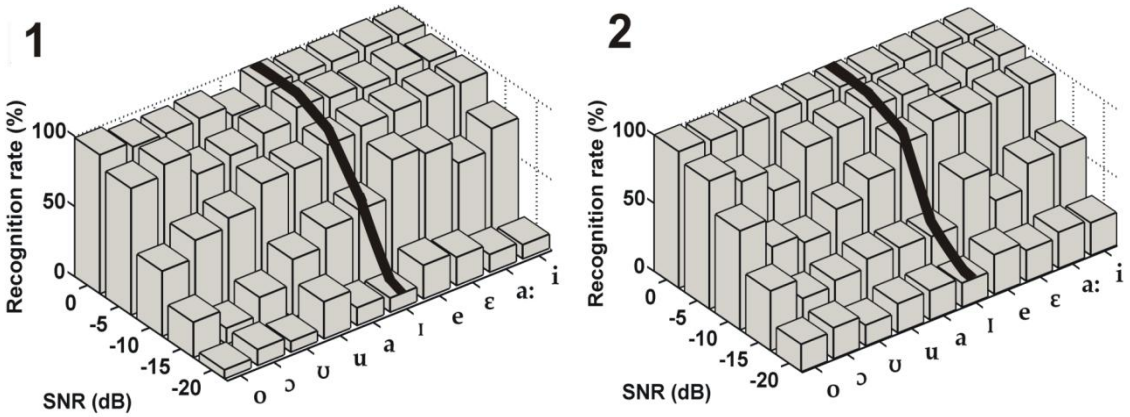

Figure 2.5: Recognition rates of vowels. The display is the same as in Figure 2.4.

2.3.3 Phoneme confusion matrices

Confusion matrices are calculated for all SNRs, which were used in the experiment. In the following section the confusion matrices at -15 dB SNR are analyzed. The recognition rates at this SNR are the least influenced by ceiling effects (see Figure 2.4

and Figure 2.5) and show the largest variation across phonemes. Therefore, at this SNR, the patterns of recognition are most characteristic. Figure 2.6 panel 1 shows the observed confusion matrices of the VCV discrimination task and panel 2 the corresponding predictions using the Lorentzian distance measure with model configuration B. Each row of the confusion matrix corresponds to a specific presented phoneme and each column corresponds to a recognized phoneme. The diagonal elements denote the rates of correct recognized phonemes and the non-diagonal elements denote confusion rates of phonemes. All numbers are given in percentages.

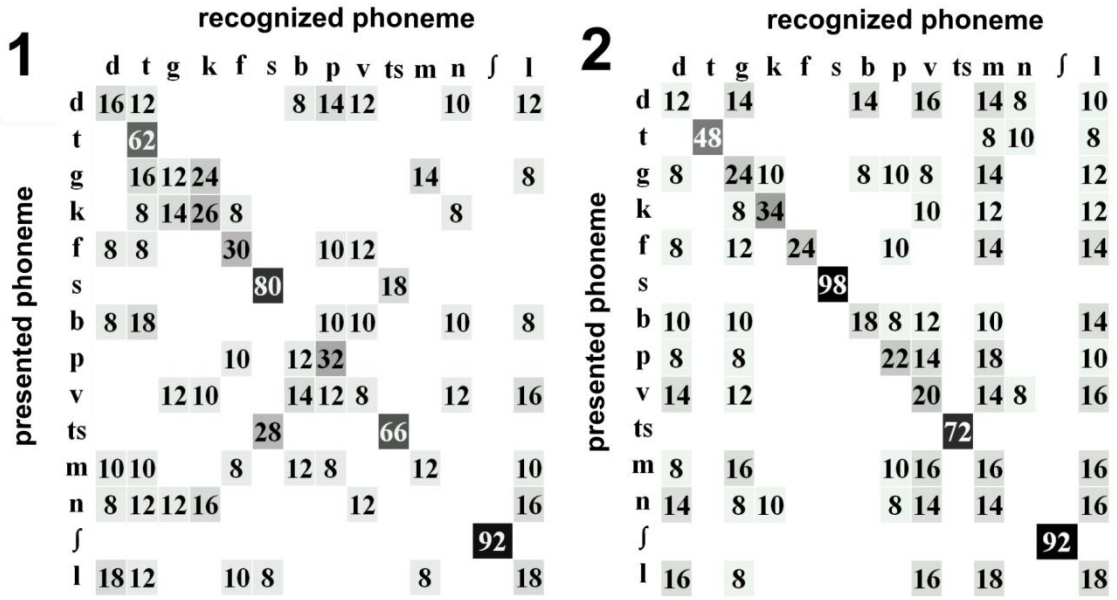

Figure 2.6: Confusion matrices (response rates in percent) for consonants at −15 dB SNR for normal-hearing subjects (panel 1) and for model configuration B with Lorentzian distance measure (panel 2). Row: presented phoneme; column: recognized phoneme. For better clarity, the values in the cells are highlighted using gray shadings with dark corresponding to high and light corresponding to low response rates. Response rates below 8% are not shown.

At -15 dB SNR the average recognition rate for all consonants is 33% (human) and 36% (model configuration B, see also Figure 2.3). In the following text the comparison of the two matrices will be described element-wise. Two elements differ significantly if the two-sided 95% confidence intervals surrounding the respective elements do not overlap (cf. Section 2.7). The observed and the predicted correct consonant recognition rates do not differ significantly, except for the phonemes /s/, and /b/, and /v/. Rates below 17% do not differ significantly from the guessing probability of 7% (cf. Section 2.7). Hence, almost all non-diagonal elements of the model confusion matrix do not differ significantly from the corresponding elements of the human listeners' confusion matrix. One exception is the confusion 'presented /ts/ - recognized /s/', found in the

observed confusion matrix, which cannot be found in the predicted confusion matrix. Other exceptions like ‘presented /p/ - recognized /m/’ differ just significantly and shall not be discussed in detail in this section. Unfortunately, the size of confidence intervals of the matrix elements decreases very slowly with an increasing amount of data. Therefore, it is not possible to find many significant differences between predicted and observed matrix elements although the amount of data is already relatively large. However, if we compare the correct recognition rates within one matrix many phonemes can be found that differ significantly in recognition rate. Note that within one single matrix only matrix elements from different rows should be compared (cf. Section 2.7).

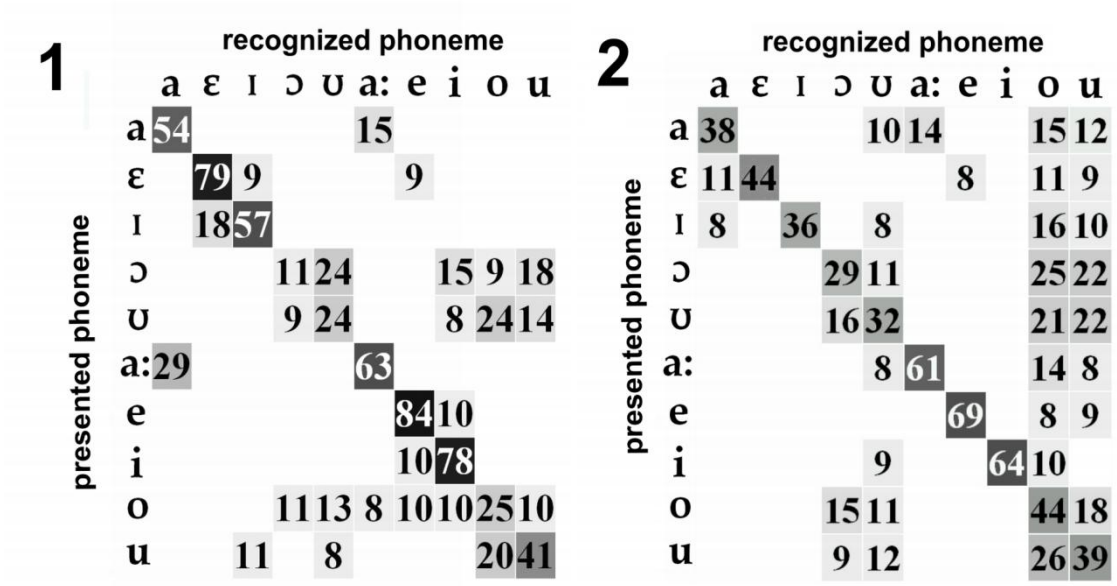

Figure 2.7: Confusion matrices (response rates in percent) for vowels at −15 dB SNR for normal-hearing subjects (panel 1) and of model configuration B (panel 2). The display is the same as in Figure 2.6.

Figure 2.7 panel 1 shows the observed confusion matrices of the CVC discrimination task and panel B the corresponding predictions using the Lorentzian distance measure with model configuration B. At -15 dB SNR the average recognition rate for all vowels is 52% (human) and 46% (model configuration B, see also Figure 2.3) panel 2. The ranking of the best recognized phonemes /e/ and /i/ as well as the ranking of the worst recognized phonemes /o/ and /u/ is predicted correctly. However, the overall “contrast” (i.e. the difference between best and worst recognized phonemes) of the predicted matrix is much less pronounced than in the observed matrix. The largest number of confusions occurred between the phonemes /ʊ/, /ɔ/, /o/, and /u/ for

both predictions and observations. However, the significant observed confusion 'presented /a:/ - recognized /a/' cannot be found in the predicted confusion matrix. Furthermore, the phonemes /o/ and /u/ are recognized with a bias, i.e., no matter what phoneme is presented, the model shows a slight preference for these phonemes.

Pearson's φ^2 (Lancaster, 1958) index was used for comparing the similarity between measured and modeled confusion matrix data. This index is based on the chi-square test of equality for two sets of frequencies and provides a normalized measure for the dissimilarity of two sets of frequencies. A value $\varphi^2 = 1$ is related to complete dissimilarity, whereas a value of $\varphi^2 = 0$ is related to equality. Table 2.2 shows φ^2 values for comparing the confusion patterns, i.e. each φ^2 value is a measure for the dissimilarity of the x-th row of the observed confusion matrix and the x-th row of the predicted confusion matrix of Figure 2.6 and Figure 2.7 respectively. For the consonant confusion matrices highest similarity is found for the confusion patterns of /t/, /s/, and /ʃ/. This very high similarity is mainly due to the high correct response, i.e. the diagonal element.

Table 2.2: Pearson's φ^2 index, a measure of dissimilarity, for comparing the confusion patterns, i.e. one row of a confusion matrix, of observed and predicted phoneme recognition from Figure 2.6 and Figure 2.7, respectively.

Presented consonant	φ^2	Presented vowel	φ^2
/d/	0.21	/a/	0.10
/t/	0.12	/ɛ/	0.24
/g/	0.24	/ɪ/	0.19
/k/	0.20	/ɔ/	0.21
/f/	0.16	/ʊ/	0.11
/s/	0.12	/a:/	0.24
/b/	0.15	/e/	0.14
/p/	0.16	/i/	0.15
/v/	0.14	/o/	0.14
/ts/	0.25	/u/	0.10
/m/	0.21		
/n/	0.14		
/ʃ/	0.08		
/l/	0.18		

Generally, many observed and predicted confusion patterns show high similarity due to low φ^2–values. However, the observed and predicted confusion patterns of /ts/ show

the lowest similarity. This is mainly due to the significant confusion of ‘presented /ts/ - recognized /s/’, which was not predicted by the model. The confusion patterns of the phonemes /f/, /l/, and /p/ show moderate similarity. These phonemes also show a poor recognition rate at -15 dB SNR and thus higher percentages in the non-diagonal-elements. This gives support to the supposition that the model is not able to predict the consonant *confusions* of normal-hearing listeners. For comparing the patterns of recognition, i.e. the diagonal of the confusion matrix, the correlation coefficients between observed and predicted data are shown in Table 2.3 as a function of SNR. For an SNR of -15 dB this correlation coefficient amounts to $r^2 = 0.91$ ($p < 0.01$). This strong correlation means that the model is quite good in modeling the *correct responses*. For observed and predicted consonants there are also highly significant correlations found at -10 dB and -20 dB SNR. The correlation decreases rapidly for higher SNR mainly due to ceiling effects, i.e. many phoneme recognition scores are in the range of 100%. Note that at 0 dB SNR a correlation coefficient for consonants could not be assigned due to the fact that at this SNR all consonants are predicted at a recognition rate of 100%, whereas some were observed at lower recognition rates.

Table 2.3: Correlation coefficients r^2 for comparing observed and predicted recognition scores from Figure 2.4 and Figure 2.5, i.e. the diagonals of confusion matrices, as a function of SNR. * denotes significant ($p < 0.05$) and ** highly significant ($p < 0.01$) correlations.

SNR (dB)	r^2 for consonants	r^2 for vowels
0	not assigned	0.09
-5	0.34*	0.52*
-10	0.78**	0.56**
-15	0.91**	0.57*
-20	0.86**	0.26

For the vowel confusion matrices highest similarity is found for the observed and predicted confusion patterns of /a/, /ʊ/, and /u/. Many confusion patterns show a high similarity except for those of /ɛ/, /ɔ/, and /a:/ which show only modest similarity. The high similarity for the former phonemes is mainly due to the correct modelling of confusions ‘presented /a/- recognized /a:/’, ‘presented /ʊ/- recognized /u/’, and ‘presented /u/- recognized /o/’ and the correct responses, respectively. The modest similarity for /ɛ/, /ɔ/, and /a:/ is mainly due to the high discrepancy in predicting the correct diagonal element score. Correlating the diagonals at this SNR (cf.

also Table 2.3) shows that the patterns of recognition are significantly (r^2 = 0.57, p < 0.05) correlated but not as high as for the consonant recognition patterns. This also holds for -10 dB and -20 dB SNR. For higher SNRs, i.e. higher average recognition scores, the correlation of predicted and observed vowels is higher than the correlation of consonants. This leads to the assumption that the model can better predict the confusion patterns for vowels than for consonants at low recognition rates as, e.g. for /ʊ/ and /u/. In predicting the correct responses, however, the model is not as good for the vowels as for the consonants.

The fact that the model is not able to predict confusion patterns correctly, especially for consonants, may be due to two reasons: The first reason may be that the model is partly not able to exploit similarities between the internal representations of phonemes that might in fact be similar to one another for normal-hearing listeners. This is supported by a confusion that is not predicted (‘presented /ts/- recognized /s/’), but not e.g. by the confusions between /u/ and /o/ that are almost correctly predicted. The second reason may be simply due to the high ranges of confidence intervals (see Section 2.7) due to the inherent binomial statistics of this speech test.

2.4 General discussion

2.4.1 Microscopic prediction of speech intelligibility

This study compares the recognition performance in noise of a microscopic speech intelligibility prediction model to the phoneme recognition performance of human listeners. The model was also used with the same approach as in this study to predict speech intelligibility of a rhyme test (Holube and Kollmeier, 1996). Our results, as well as the results of Holube and Kollmeier (1996), show that this combination of perception model and DTW speech recognizer is able to discriminate noisy speech signals in a closed-set testing procedure. The model used here is also similar to the microscopic model used by Barker and Cooke (2007). Their model is inspired by ASR techniques and evaluates speech parts that ’glimpse‘ the spectro-temporal pattern of the signal to be recognized out of background noise. One main novelty of this study is that the use of the speech database of Wesker *et al.* (2005), which provides many recordings of the same logatome, allows the investigation of the influence of a-priori-knowledge about the speech. This investigation is possible because the speech recognizer is realized with two model configurations. In model configuration B templates are used, which are

identical to the test items; this corresponds to maximum a-priori-knowledge. In model configuration A the recognizer used templates, which are not identical with the test items corresponding to less a-priori-knowledge.

Assuming limited a-priori-knowledge within model configuration A results in a much poorer performance than observed in the results of human listeners. This reflects the gap between human and machine speech reception (cf. Jürgens *et al.*, 2007) because configuration A is the standard case for ASR. The gap of about 11 dB to 12 dB SNR is consistent with findings of other studies employing common speech recognition systems like Hidden-Markov-Models (HMM). Meyer *et al.* (2007a) found a gap of about 10 dB SNR (averaged across different speakers) between human listeners' SRT and the SRT of a speech recognizer using Mel-Frequency-Cepstral-Coefficients and an HMM using the same OLLO speech corpus and very similar listening experiments. As a direct comparison, a subset of the ASR-data of Meyer *et al.* (2007a) is plotted as an additional psychometric function in Figure 2.3. The subset of speech material to be tested is limited to the same speech material that was used in the present study. For this speech material the gap in SRT between ASR and normal-hearing listeners' performance extends to about 8 dB. The difference of 3 to 4 dB from our results might be due to different speech recognizers used. Meyer *et al.* (2007a) used a speech recognizer that benefitted from decades of research. Also the amount of training material in their study was much larger (49 speakers with different articulation styles) than in the present study.

Speech intelligibility can be predicted with greater accuracy using model configuration B in which the amount of information about the speech signal prior to the recognizing process is assumed to be perfect. It has to be stated that in this point the model differs from human listeners' speech processing because human listeners have not stored the exact internal representation of the signal to be recognized. Human listeners are able to generalize their internal representation of a speech utterance to different speech waveforms, even if different articulation styles or speakers are involved. However, our speech recognition model includes a pattern recognizer that has to find a speech pattern among different alternatives, which is closer to human speech processing than, for example, the Speech Intelligibility Index (ANSI, 1997). This optimal detector concept is a standard in psychoacoustic modeling and predicts, e.g., forward, backward and simultaneous masking thresholds (Dau *et al.*, 1996b), modulation detection thresholds (Dau *et al.*, 1997), and the time resolution of the binaural system (Breebaart *et al.*, 2002). As this speech recognition study is in line with

other psychoacoustic experiment studies, because of the closed test paradigm and the nonsense speech material used here, such an approach seems to be appropriate. The very accurate agreement of observed and predicted phoneme recognition rates using model configuration B does not mean that human listeners have a perfect decision device. Humans' limitations in discriminating speech in noise are certainly due to energetic masking of the speech signal by background noises and also due to errors in the inherent processing in the subsequent word recognition stage. However, the speech discrimination performance of the model is very similar to that of human listeners if all limitations of performance are assumed entirely in the preprocessing stage of the model. For the experiments presented here this may be interpreted as that lifelong training of humans in speech makes the pattern recognizing part of human speech recognition perform as well as the model's optimal detector.

With configuration B the model is capable of predicting the SRT of this speech test with an accuracy of about 1 dB. The SII (ANSI, 1997) predicts the SRT within the same accuracy range: For -15 dB SNR the SII-value is found to be 0.045, for instance, and for -10 dB the SII is 0.18. Transformed to intelligibility scores by using the SII transfer function for Hagerman's sentences in noise (Magnusson, 1996), the resulting SRT is -11.2 dB SNR. The main advantage of the microscopic modelling approach compared to the SII is that, whereas the SII is able to predict only average recognition scores, this approach is able to predict the recognition scores for each phoneme separately. Furthermore, this approach draws out some characteristic phoneme confusions that are commonly seen.

2.4.2 Distance measures

The type of distance measure crucially influences the performance of the speech recognizer when using model configuration B. The Euclidian distance used by e.g. Plomp (1976), Holube and Kollmeier (1996), and Jürgens *et al.* (2007) shows the poorest performance among the distance measures investigated here. In this study, there is a gap of more than 4 dB between the SRT of model configuration B and human listeners' SRT. Using the Euclidian distance, outlying passages are strongly weighted and consequently the DTW algorithm tries to minimize the occurrence of outlying passages as far as possible. This may cause the warp path, i.e., the temporal matching function between two internal representations, to be fitted more to the passages containing different speech or noise. Passages with low distances are disregarded. By applying a distance measure that is less sensitive to outliers in the matching procedure

of two internal representations (i.e., using the two-sided exponential measure or the Lorentzian measure) this gap is substantially decreased or vanishes. Using the two-sided exponential distance measure, all distances are weighted with their usual occurance probability (cf. Figure 2.2). Therefore, this can be called a ’natural‘ distance measure for speech in noise. Although no substantial influence of the type of distance measure was found on the performance of model configuration A, it was found for model configuration B. One could argue, since configuration A is typical of an ASR system, that other ASR systems may not benefit from an optimization of the distance measure they use. However, as this approach uses a speech recognizer that does not require a large amount of training material as common ASR systems do, this is speculative. Nevertheless, for further optimizing of ASR systems it may be useful to study the influence of different distance measures on the ASR systems’ performance.

Using the Lorentzian distance measure, all outlying passages get approximately the same constant weight because of the flatness of the logarithm for large input values. Therefore, the overall distance between two internal representations is mainly dominated by the smallest elements of the distance matrix. In other words, the steepness of the logarithm at low values causes similar passages of the internal representations to be matched as closely as possible. This may particularly be an advantage for discriminating noisy speech samples because the speech recognizer is dominated by matched (i.e., similar speech) patterns and neglects unmatched (i.e., noise or different speech) patterns. Hence, the detector can separate the objects ‘matched speech’ passages from ‘unmatched speech’ or ‘noise only’ passages more appropriately. If we conceive of noise and speech as different acoustical objects this mechanism may have some similarities to the mechanism of acoustical object separation within the human auditory system. Neglecting passages that do not match passages of stored response alternatives is a candidate for modeling human’s mechanism of object separation. In that way the distinction between a ‘matchable speech object’ and a ‘not matchable speech object’ or ‘noise-only object’ may be enhanced. Using model configuration B, the Lorentzian distance measure performs best and results in a high agreement in phoneme recognition. Therefore, this set-up was chosen for the prediction of speech recognition in noise of listeners with normal hearing.

2.4.3 Phoneme recognition rates and confusions

In this study both human listeners and the model show the highest performance at the same consonants /t/, /s/, /ʃ/, and /ts/as in the study of Phatak and Allen (2007),

who investigated consonant recognition rates in speech weighted noise. The results obtained in this study are in line with those of Phatak and Allen (2007), although they used speakers and listeners of a different native language and different speech material. Furthermore, the amount of alternatives that could be recognized was completely different from our measurements. However, the separation of consonants into a low scoring and a mid scoring group with the same phonemes as in Phatak and Allen (2007) could not be observed in this study. They concluded that differences in recognition rates can mainly be explained by differences in the long-term spectra of speech and noise. However, this may not account for consonants with characteristics that are mainly determined by the temporal structure as, e.g., for plosives like /p/, /t/, or /k/. Our approach regards this temporal structure by the temporal matching performed in the DTW speech recognizer.

By and large, the confusion matrices of human listeners and of model configuration B with Lorentzian distance measure are very similar. Except for a small number of elements, the consonant confusion matrices do not differ significantly element-wise regarding the binomial statistics valid for these discrimination tasks (see Section 2.7). The correlation between predicted and observed recognition rates of single phonemes is very high. This is promising and it may indicate that for all phonemes speech information is conserved or emphasized during the modeled ’effective‘ auditory preprocessing in a way similar to human listeners.

The vowel confusion matrix of the model shows a slight preference, i.e. a bias, concerning the vowels /ʊ/, /ɔ/, /o/, and /u/ independent of the presented vowel. This is one main difference between the predicted and observed vowel confusion matrices. Meyer *et al.* (2007a) found that the phonemes /o/ and /u/ within this speech corpus have the least distinctive average spectrum compared to speech-shaped noise. Consequently, these phonemes are the phonemes best masked in the background noise at low SNRs. If the speech recognizer is not able to match a presented phoneme, it is very probable that it matches the internal representation that is the most similar to the internal representation of the background noise. These are the internal representations of logatomes with /o/ and /u/ as middle phonemes. In some cases the procedure probably matches mainly the background noise characteristics of the internal representations and is not able to focus on the speech characteristics any more. One reason why the prediction of vowel recognition rates is poorer than for consonants, while the prediction of vowel confusions is better than for the consonants, may be the

spectrotemporal structure of these two phoneme groups. Generally, vowels are more stationary signals than consonants. Furthermore, there is no clear separation between different vowels but a continuous transition in the frequency range. Therefore, it seems reasonable to assume that two different vowels are ’perceptually‘ more close to one another than are two different consonants. This may explain why confusions occur more frequently in both normal-hearing listeners’ and modeled data.

2.4.4 Variability in the data

Data obtained by speech tests using human listeners always show both intra-individual and inter-individual variability. One factor for the inter-individual variability is the variability of the hearing threshold across listeners. Preliminary simulations, however, showed that adapting only the hearing threshold simulating noise results in less variability than found in normal-hearing listeners’ speech recognition data. This can be explained by the low RMS level of the hearing threshold simulating noise, which is masked by the much higher level of the background noise. For this reason a much more effective way to include variability was to use running background noise. In other words the variability in the simulations originates almost exclusively from the statistics of the background noise. However, this is somewhat unrealistic, because in the measurements the background noise stimuli were identical for every participant whereas, in reality the auditory processing varied. It still remains an open question how to obtain a comparable variability by modifying the auditory processing without using this workaround. For speech intelligibility modelling in silence, e.g., Holube and Kollmeier (1996) achieved some variability using a fluctuating absolute threshold of hearing which improved their predictions in silence. Due to the small influence of the exact form of the absolute hearing threshold in our study, this procedure was not applied here.

2.4.5 Practical relevance

There are at least two different applications that may benefit from this modelling approach. Firstly, this approach may be used to model sensorineural hearing loss by appropriate manipulation of the auditory preprocessing. Hence, consequences of the auditory preprocessing on speech recognition for listeners with impaired hearing can be investigated. As a long-term aim the model may serve as a tool for distinguishing between reduced speech recognition caused by impaired preprocessing or by further problems in the patient’s central processing. A further long-term aim is to find

processing strategies that substantially enhance the recognition performance of certain phonemes and that can be used in hearing-aids. Secondly, automatic speech recognizers may be improved especially for functioning in noise if they focus on passages fitting well to their vocabulary and if they neglect outlying passages in a manner similar to that used in the weighting of the perceptual distance in this study.

2.5 Conclusions

(1) The microscopic approach for predicting speech intelligibility by using an auditory model as a pre-processor to a DTW speech recognizer is capable of discriminating CVC and VCV logatomes in noise.

(2) If the detector stage is assumed to be optimal by using identical templates for test signal and vocabulary, the speech discrimination performance of the model is very similar to that of human listeners. This means that the recognition of logatomes by humans can be modelled effectively by assuming a limited auditory-like preprocessing stage and a perfect speech matching process. In other words: The prediction of normal-hearing listeners' speech recognition is only possible if exactly the same stimulus is available as a-priori-knowledge.

(3) No substantial improvement in performance of the model with *imperfect* knowledge about the speech signal was found when changing the distance measure.

(4) For the model with *perfect* knowledge about the speech signal, the Lorentzian measure is the best distance measure where outlying passages have the smallest weight compared to the other distance measures such as the Euclidian or the two-sided-exponential.

(5) Predicted recognition rates of each single phoneme are very similar to observed recognition rates but some of the observed characteristic patterns of human confusions did not occur within the predictions.

2.6 Acknowledgements

We thank Birger Kollmeier for his substantial support and contribution to this work and Bernd Meyer for making available the ASR data. Thanks to Mitchell Sommers, Amy Beeston, and one anonymous reviewer who helped to greatly improve the manuscript. We also like to thank the EU HearCom Project, the 'Förderung wissenschaftlichen

Nachwuchses des Landes Niedersachsen' (FwN), and SFB/TR 31 'Das aktive Gehör' (URL: http://www.uni-oldenburg.de/sfbtr31) for funding the research reported in this paper.

2.7 Appendix: Significance of confusion matrices elements

For deciding whether or not two matrix elements differ significantly, a statistical analysis has to be made. One element of a confusion matrix is given by $p = x/n$, with x denoting the number of recognitions of the phoneme specified by the column and n denoting the number of presentations specified by the row of the matrix. There are n = 50 (VCV) and n = 80 (CVC) presentations respectively of each phoneme at each SNR (i.e. each confusion matrix). Each single presentation is followed by a subjects' decision for one response alternative given in the list. Therefore, each decision is a Bernoulli-trial with an unknown underlying probability π for the correct item and $(1-\pi)$ for all other items. Note that p is just an estimate of π. By estimating π using p, both-sided 95%-confidence intervals can be calculated based on binomial statistics (Sachs, 1999). The upper boundary is given by

$$\pi_{upper} = \frac{(x+1)F_{upper}}{n-x+(x+1)F_{upper}} \qquad (2.9),$$

with $F_{upper} = F_{\{2(x+1),2(n-x)\}}$ taken from Fisher's F-distribution. The lower boundary is given by

$$\pi_{lower} = \frac{x}{x+(n-x+1)F_{lower}} \qquad (2.10),$$

with

$$F_{lower} = F_{\{2(n-x+1),2x\}} \qquad (2.11).$$

The range of confidence intervals for an observed percentage p in the speech test, i.e. $(\pi_{upper} - \pi_{lower})$, results in 4% to 22% for n = 80 (CVC presentation) and 6% to 29% for n = 50 (VCV presentation), whereas the wider range can be found at p = 50% and the smaller range at p = 0% and p = 100%. These confidence intervals contain the underlying probability π with a confidence of 95%. Furthermore, they offer a criterion to decide if two percentages, that are statistically independent of each other, differ significantly (i.e., their confidence intervals must not overlap). The precondition, statistical independence within one confusion matrix, is warranted only for two matrix elements that are not part of the same row because in this case completely different

phonemes were presented to obtain the two percentages. Two elements of the same row are not independent of each other because the recognition of one phoneme affects the percentages for the other phonemes of that row. A comparison of two elements being part of the same row requires a different statistical analysis that is not discussed here. Therefore, only elements of different rows (or different confusion matrices) can be tested for difference using the methods described in this section. When comparing two different confusion matrices (e.g., observed with predicted) this problem does not occur.

3 Challenging the Speech Intelligibility Index: Macroscopic vs. microscopic prediction of sentence recognition in normal and hearing-impaired listeners[1]

Abstract

A “microscopic” model of phoneme recognition, which includes an auditory model and a simple speech recognizer, is adapted to model the recognition of single words within whole German sentences. “Microscopic” in terms of this model is defined twofold, first, as analyzing the particular spectro-temporal structure of the speech waveforms, and second, as basing the recognition of whole sentences on the recognition of single words. This approach is evaluated on a large database of speech recognition results from normal-hearing and sensorineural hearing-impaired listeners. Individual audiometric thresholds are accounted for by implementing a spectrally-shaped hearing threshold simulating noise. Furthermore, a comparative challenge between the microscopic model and the “macroscopic” Speech Intelligibility Index (SII) is performed using the same listeners’ data. The results are that both models show similar correlations of modeled Speech Reception Thresholds (SRTs) to observed SRTs.

[1] This chapter was published as Jürgens *et al.* (2010). The paper was presented at the 11th annual conference of the International Speech Communication Association (Interspeech, Makuhari, Japan).

3.1 Introduction

The Speech Intelligibility Index (SII) (ANSI, 1997) is widely used to predict human speech recognition (HSR) in different noise conditions or for subjects with different audiometric hearing losses. The SII can be called a "macroscopic" model, as it uses only the long-term spectra of speech and noise separately, whereas the particular temporal structure of speech and noise is disregarded. Speech intelligibility is predicted using a weighted sum over the Signal-to-Noise-Ratios (SNRs) in different frequency bands, resulting in an SII value between 0 and 1. The weighting factors are tabulated and depend on the context or the articulation style of the speech material used (ANSI, 1997). Subsequently, the SII value is transformed to a speech recognition rate in percent using a nonlinear function that depends on the speech material.

A psychoacoustically-driven, "microscopic" model of HSR, on the other hand, models the recognition of single phonemes (Jürgens and Brand, 2009) by analyzing the particular spectro-temporal structure of speech and noise. An "internal representation" (IR) is computed from the waveform of the speech/noise-mixture using an auditory model and employing a simple speech recognizer. Thus, it mimics the individual auditory signal processing in a much more realistic way than the SII.

The main goal of this study is first, to adapt this microscopic model from phonemes to sentences and second, to compare the predictive power of this modeling approach with that of the SII. For the comparative challenge of the two models, an ambitious speech recognition data set is used with perceptually similar (rather than physically equal) acoustic measurement conditions for all listeners. This means that signals with higher levels were used for hearing-impaired listeners to ensure equal loudness perception of these signals.

3.2 Measurements

3.2.1 Subjects

15 normal-hearing (NH) listeners aged from 24 to 34 years and 48 sensorineural hearing-impaired (HI) listeners aged from 17 to 82 years participated in this study. In 51 listeners both ears were tested separately, resulting in a total of 114 investigated ears. NH listeners showed pure-tone thresholds of not more than 15 dB Hearing Level (HL) using standard audiometry (IEC60645-1). Figure 3.1 displays averaged audiogram data

of the NH group and two groups of HI listeners (black lines), including the ranges between the 5th and 95th percentiles. The first group of HI listeners showed nearly normal hearing at low frequencies (≤ 30 dB HL between 125 Hz and 1 kHz) and hearing loss at higher frequencies (HI-H). The second group showed a hearing loss both at low and high frequencies (HI-LH). Listeners were paid for their participation in the experiments.

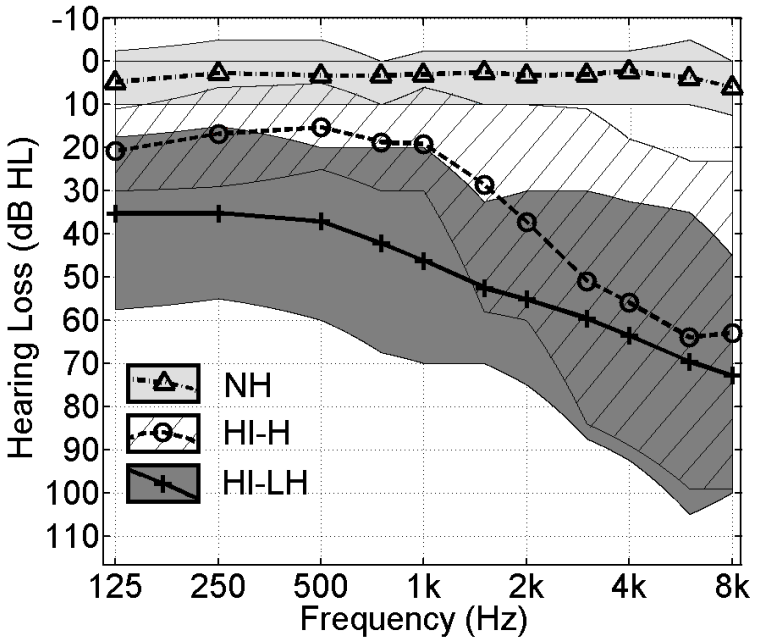

Figure 3.1: Average audiometric thresholds and ranges between the 5th and 95th percentiles for normal-hearing listeners (NH, light gray) and two groups of hearing-impaired listeners (HI-H, hatched; HI-LH, dark gray).

3.2.2 Apparatus

All stimuli were presented monaurally via Sennheiser HDA 200 headphones that were free-field equalized using an FIR-filter with 801 coefficients, while the listeners were seated in a sound-insulated booth. The headphones were connected to a computer-controlled audiometry workstation that was developed within a German joint research project on speech audiometry (Kollmeier *et al.*, 1992).

3.2.3 Speech intelligibility measurements

Speech intelligibility in stationary ICRA1 noise (Dreschler *et al.*, 2001) was measured using the Oldenburg sentence test (Wagener *et al.*, 1999a) that is part of the Oldenburg Measurement Applications (OMA) software by HörTech gGmbH. The Oldenburg sentence test consists of German sentences with the fixed syntactic structure *name-verb-*

number-adjective-object, e.g. *'Peter gets five wet cars'*, spoken by a male speaker. Each word of the sentence was chosen from ten alternatives, respectively. Such sentences were combined in lists consisting of 30 sentences each that were optimized with respect to equal speech intelligibility (Wagener *et al.*, 1999b). Within one measurement run, one list of sentences was presented. An adaptive procedure (Brand and Kollmeier, 2002) was used to measure the Speech Reception Threshold (SRT), i.e. the SNR at 50% speech recognition rate for the sentences of this list as follows. After the presentation of each sentence, the level of the speech was adaptively varied in two randomly interleaved tracks. One track converged at 80% and the other track converged at 20% speech recognition rate. Both tracks started with an SNR of 0 dB. After each run the SRT was calculated by fitting a logistic function with the parameters SRT and slope to all collected data using a maximum likelihood estimator (Brand and Kollmeier, 2002). During the measurements the level of the noise was fixed at a level that individually corresponded to medium loudness. This means that all listeners were tested under perceptually similar, rather than physically equal conditions. At least two test lists were measured as training in advance. Subjects were asked to repeat each presumably understood word after presenting the whole sentence (open test). An investigator marked the correctly recognized words using a touch screen response box.

3.3 Modeling

3.3.1 Speech Intelligibility Index

The SII was calculated according to the ANSI (1997) standard using the long-term spectrum of the ICRA1 background noise and the long-term spectrum of the complete speech material of the Oldenburg sentence test (Wagener *et al.*, 1999a). The critical frequency band method was used and the standard speech spectrum level for stated vocal effort was chosen according to 'normal' speech articulation. Individual audiogram data, interpolated at the center frequencies of the critical frequency bands, was used to calculate the equivalent hearing threshold level. As critical band importance function the values for SPeech In Noise (SPIN) were chosen. The modeled psychometric function, i.e. SII-values for each listener as a function of SNR, was calculated using the same fixed noise level as in the measurements and speech levels in the range of 40 to 100 dB SPL in 2.5 dB steps. An SII-value of 0.24 was defined as the value corresponding to 50% speech intelligibility. Individually for each listener, the modeled SRT was obtained by an interpolation of the psychometric function at that SII-value.

3.3.2 Microscopic model

The microscopic model of speech recognition was implemented very similar to the approach of Jürgens and Brand (2009) for NH listeners and was extended to HI listeners and to a sentence test in the present study. A word from the Oldenburg sentence test, mixed with ICRA1 background noise with an SNR ranging from –15 to 15 dB in 3 dB steps is added to a hearing threshold simulating noise that is spectrally shaped to the individual audiogram data of the listener's ear (cf. Figure 3.2). Subsequently, the Perception Model (PeMo) (Dau *et al.*, 1997) computes an IR from this signal.

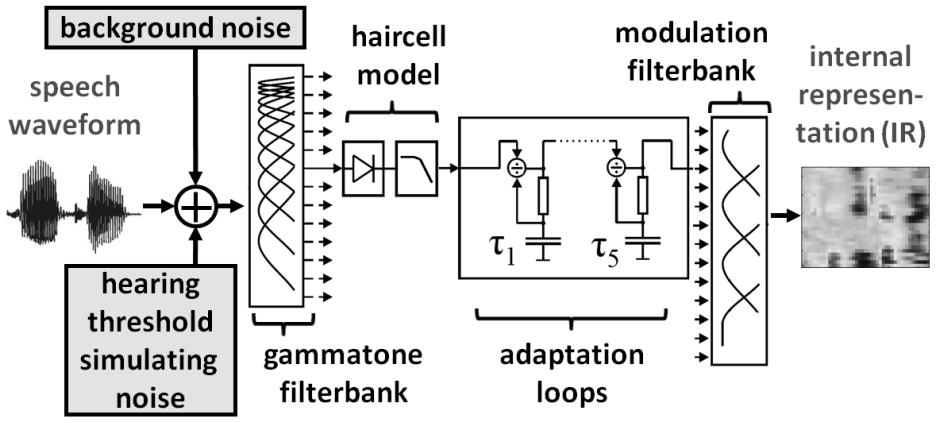

Figure 3.2: Block diagram of the auditory model (white blocks). Background noise and hearing threshold simulating noise are added to the speech waveform in advance. The auditory model computes an internal representation from the speech/noise mixture.

The PeMo implementation used in the present study consists of a gammatone filter bank with 27 frequency channels ranging from 236 Hz to 7469 Hz center frequency. The gammatone filterbank models the peripheral filtering in the cochlea. A haircell-model computes the temporal envelope in each frequency channel and adaptation loops emphasize on- and offsets of the signal. A modulation filterbank with four modulation channels evaluates low speech modulations up to about 20 Hz. Consecutively, the IR is downsampled to a sampling frequency of 100 Hz and thus contains a feature-matrix of 27 frequency channels and four modulation frequency channels at each 10 ms time step. PeMo is capable of modeling psychoacoustical data, e.g. of forward and backward masking experiments, and modulation detection in normal-hearing listeners (Dau *et al.*, 1997).

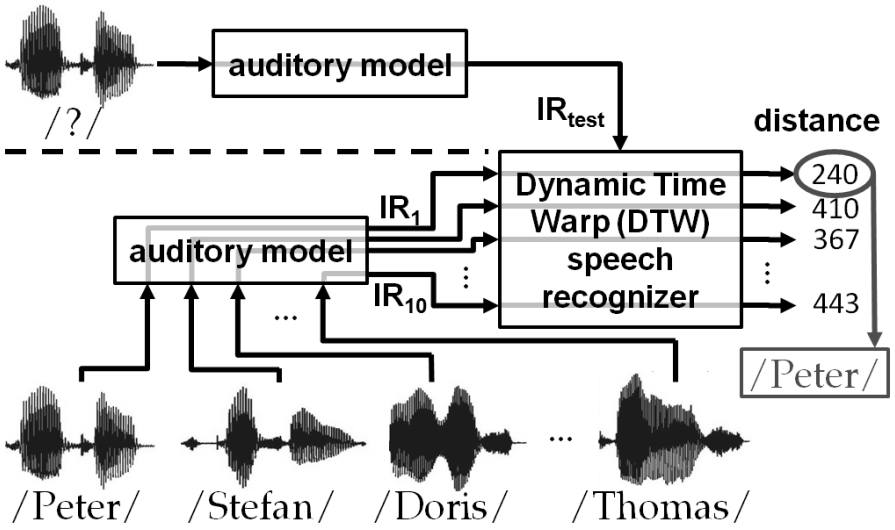

Figure 3.3: Microscopic modeling approach: An internal representation (IR_{test}) of the speech waveform to be recognized mixed with noise (top left) is computed by the auditory model. The IRs of ten different response alternatives (vocabulary, bottom) are also computed by the same auditory model. Both, IR_{test} and one of IR_1 to IR_{10} are given pair wise to the DTW speech recognizer that computes a "perceptive" distance of each pair. The word with the smallest distance is recognized.

Figure 3.3 shows the approach for the recognition of one unknown word (in this example: /Peter/). For a given listener's ear and SNR the IR_{test} of the unknown word (same speech waveform as in the measurements) is computed (upper part). To initialize the adaptation loops of PeMo 0.4 s of preceding noise with the same level as the background noise are added to the waveform. The corresponding passage in the IR was deleted before entering the recognition stage. For each one of the ten possible words of the same clause, an IR was randomly chosen from the pool of all IRs calculated from the speech waveforms presented during the measurements, mixed with noise at the same SNR as the unknown word (vocabulary, lower part of Figure 3.3). A Dynamic-Time-Warp (DTW) speech recognizer computes pair wise the Lorentzian distance ("perceptive" distance, cf. Jürgens and Brand (2009)) between IR_{test} and the IRs in the vocabulary by locally stretching and compressing the time axes. That word from the vocabulary with the smallest perceptive distance to the test word is taken as the recognized one. Note that the exact speech waveform to recognize is always also contained in the vocabulary, i.e. the detector stage is assumed to be optimal (cf. Jürgens and Brand (2009)). However, the waveforms of the speech/noise mixtures that enter PeMo are always different due to different temporal passages of the background noise and the hearing threshold simulating noise. For each test word the recognition procedure was conducted nine times using different temporal passages of background noise and

hearing threshold simulating noise. The speech recognition rate for a given listener and SNR was then calculated as the average over the nine repetitions, different parts of the sentence, and different sentences. The whole calculation was performed on a computer cluster of the University of Oldenburg.

3.4 Results and comparison

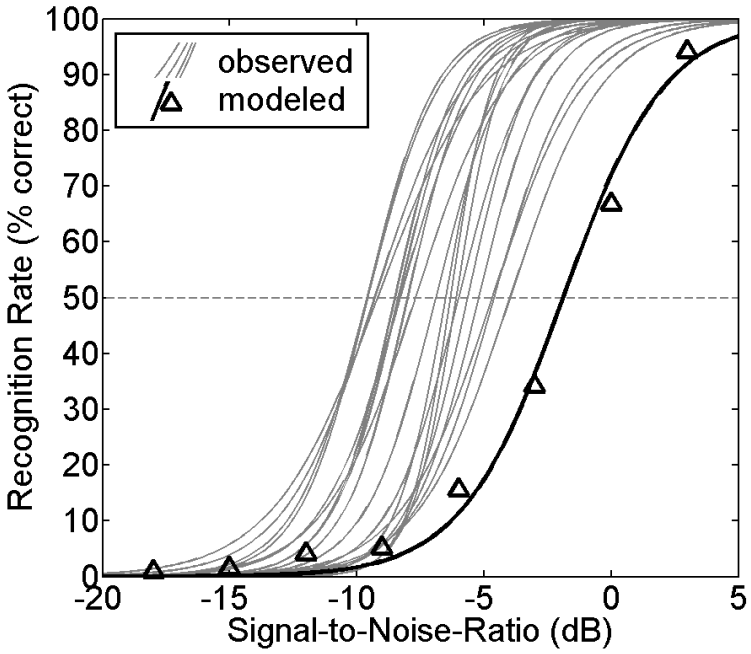

Figure 3.4: 20 observed (gray solid lines) and one modeled psychometric function (triangles and black solid line) of speech intelligibility of NH listeners using the microscopic model.

Figure 3.4 shows the modeled recognition rates (triangles) using the microscopic model for a NH listener with 0 dB HL at all audiometric frequencies. Note that the modeled recognition rates were corrected for the random hit rate of 10% that is inherent in this modeling approach, but not inherent in the open-set speech intelligibility measurements. A fit to the modeled recognition rates using a logistic function (psychometric function, black solid line, cf. Jürgens and Brand (2009)) results in the optimal fit parameters SRT = -1.9 dB SNR and slope = 11.3 %/dB. Additionally, the psychometric functions of the 20 NH ears are plotted as gray solid lines. Substantial inter-individual differences (about 5 dB) in the SRT between NH ears can be observed. The modeled psychometric function shows an SRT that is about 5 dB higher than the average SRT of the NH listeners. Furthermore, the slope of the modeled psychometric function is slightly shallower than the slopes of the psychometric functions of the NH ears.

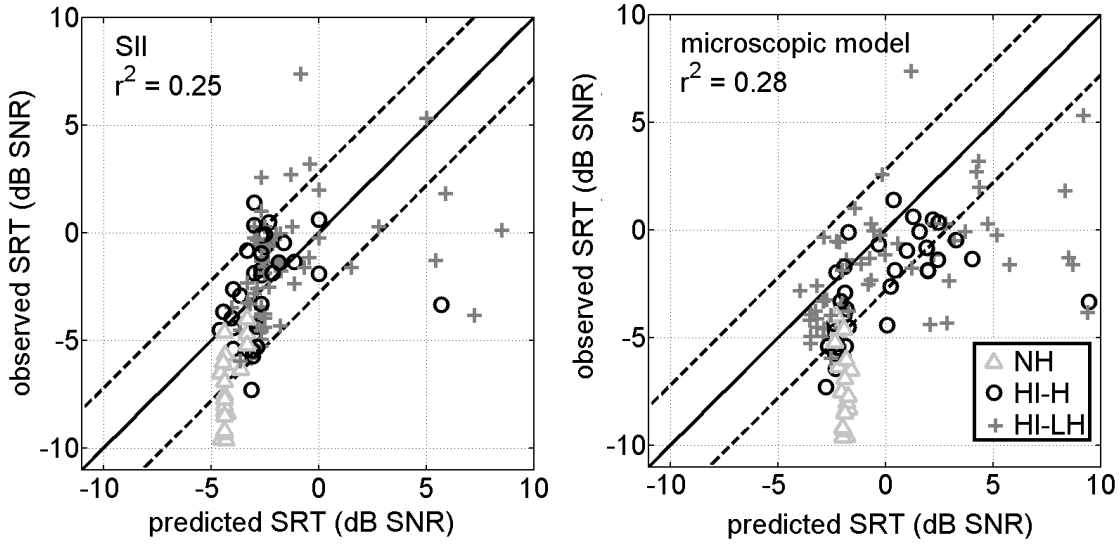

Figure 3.5: Observed vs. predicted SRTs using the SII (panel on the left) and the microscopic model (panel on the right). Different groups of listeners are denoted with different symbols and gray scales. The dashed lines show the confidence interval of the measurement procedure.

Figure 3.5 presents the predicted SRTs using the SII (panel on the left) and the microscopic model (panel on the right) versus the observed SRTs for NH and HI listeners. Dashed lines indicate the 95% confidence boundaries that were calculated as ± twice the standard deviation of the test-retest SRT-difference (1.4 dB) measured in a subset of the listeners. The SII shows a correlation of r^2 = 0.25 (p < 0.001) using Pearson's correlation coefficient r. 82% of the predicted SRTs fall within the confidence boundaries of the measurement procedure. HI-LH data (gray crosses) show the largest inter-individual variation in both, observation and prediction. NH data (light gray triangles) show only inter-individual variation in the observations, but almost no variation in the predictions. Most of the data are clustered in a region that covers about 3 dB in the predictions and about 10 dB in the observations. The microscopic model shows a correlation of r^2 = 0.28 (p < 0.001). 57% of the data points fall within confidence boundaries. The microscopic model, as well as the SII, is not able to predict the variation in NH listener's data according to different audiometric thresholds. However, using the microscopic model, HI data points show less clustering than observed using the SII, which indicates that the microscopic model predicts larger differences of SRTs due to individual audiometric thresholds and testing conditions (i.e. background noise levels) of the HI listeners. Concerning the individual slopes of the psychometric functions, the microscopic model shows only a poor correlation of r^2 = 0.09 (p < 0.01). On average, the modeled psychometric functions (average slope 10%/dB) are shallower than the observed psychometric functions (16%/dB).

3.5 Discussion

SII and microscopic model show similar correlations between predicted and observed SRTs. This indicates that it is possible to achieve the same performance as the SII, concerning individual differences of HI listeners, using a psychoacoustically-driven microscopic model. However, there is a difference between the models regarding the number of SRT values in confidence intervals. The fact that over 80% of the SII-predicted SRT values fall within confidence boundaries was achieved by assuming an SII-value of 0.24 at the SRT. This value is adjustable and was set in order to reproduce the *average* SRT of all listeners. With the microscopic model such an optimization was not necessary or rational. However, both models are not able to model the remarkable inter-individual differences of listeners with nearly the same audiometric thresholds (e.g., of NH listeners), which indicates the existence of an important, not adequately modeled individual factor on speech intelligibility. Some of the following factors might explain parts of the differences between measurements and predictions.

Semantic context effects might be responsible for the predicted SRTs of NH listeners being higher and for the predicted psychometric functions being shallower than the respective observed value. Although the sentences of the Oldenburg sentence test contain low semantic context, listeners might still benefit in their recognition performance due to co-articulation between subsequent words and due to the prosody of the sentence, which cannot be used by the model. The amount of this benefit might be subject-dependent and thus might explain parts of the remaining variance in the data. Too shallow psychometric functions are also reported in Stadler *et al.* (2007) when modeling human sentence recognition using an auditory model and an information-theoretic framework. Within the scope of their framework, Stadler *et al.* (2007) attributed the too shallow modeled psychometric function to a non-optimal probabilistic speech model they used. However, since an “optimal detector” approach is used in the present study, this reason does not hold here.

The microscopic model assumes the Oldenburg sentence test to be a closed test, although the measurement was performed as an open test. Modeled psychometric functions that show a random hit probability of 10% at low SNRs are scaled to cover the whole range of possible recognition rates (cf. Figure 3.4). A closed test approach was also used in Jürgens and Brand (2009) and has the advantage that only limited speech material is needed as possible response alternatives for the speech recognizer.

Using this approach for the open speech test used here seems to be feasible, since a study comparing the results of the open and closed version of the Oldenburg sentence test for NH listeners revealed no significant differences, as long as the listeners have been trained prior to the test (Brand *et al.*, 2004).

The individual measurement conditions might be a factor responsible for the prediction of better speech recognition performance (i.e., lower SRTs) for some of the HI listeners than for NH listeners by the microscopic model (red crosses in the lower panel of Figure 5 with predicted SRTs between -4 and -3 dB SNR). These HI listeners show a flat hearing loss across all frequencies and were tested at very high background noise levels. Hence, in the model, the individual hearing threshold simulating noise vanishes in the background noise in all frequency channels and thus has only little effect. However, the also higher speech level compared to NH listeners might have resulted in the predicted SRT being slightly lower than the SRT observed in NH listeners.

The present study is a first modeling approach that explicitly mimics the effective signal processing of the auditory system for the prediction of speech intelligibility of a sentence test. Concerning the particular model blocks, the microscopic model is much closer to mimicking human speech processing than the SII. Furthermore, one important difference between the two models is that in contrast to the SII, the microscopic model does not need speech and noise as separate signals. In the future, the predictions of this microscopic approach might be improved by a more realistic implementation of the individual hearing threshold and other aspects of hearing impairment like a reduced dynamic compression or temporal resolution. An advantage of this microscopic model compared to the SII is the possibility to investigate how these individual aspects of hearing impairment affect speech recognition performance by implementing them in the signal processing of the auditory model. Furthermore, this approach could be extended from acoustical to electrical hearing by modeling the individual signal processing of cochlear implant users.

3.6 Conclusions

The microscopic model of human sentence recognition applied to speech recognition data of normal-hearing and sensorineural hearing-impaired listeners shows similar performance as the standard SII. However, the different modeling blocks of the microscopic model aim at mimicking human speech processing much more closely than

the SII. Furthermore, the microscopic model has the potential to be extended to model the effects of context of the speech material on speech recognition and to investigate how different individual aspects of hearing impairment affect sentence recognition.

3.7 Acknowledgements

Thanks to Sven Kissner and the Hörzentrum Oldenburg GmbH for the execution of the measurements. This work is supported by SFB TRR 31 “The active auditory system” and the “Audiologie-Initiative Niedersachsen”.

4 Assessment of auditory nonlinearity for listeners with different hearing losses using temporal masking and categorical loudness scaling[1]

Abstract

A dysfunction or loss of outer hair cells (OHC) and inner hair cells (IHC), assumed to be present in sensorineural hearing-impaired listeners, affects the processing of sound both at and above the listeners' hearing threshold. A loss of OHC may be responsible for a reduction of cochlear gain, apparent in the input/output function of the basilar membrane and steeper-than-normal growth of loudness with level (recruitment). IHC loss is typically assumed to cause a level-independent loss of sensitivity. In the current study, parameters reflecting individual auditory processing were estimated using two psychoacoustic measurement techniques. Hearing loss presumably attributable to IHC damage and low-level (cochlear) gain were estimated using temporal masking curves (TMC). Hearing loss attributable to OHC (HL_{OHC}) was estimated using adaptive categorical loudness scaling (ACALOS) and by fitting a loudness model to measured loudness functions. In a group of listeners with thresholds ranging from normal to mild-to-moderately impaired, the loss in low-level gain derived from TMC was found to be equivalent with HL_{OHC} estimates inferred from ACALOS. Furthermore, HL_{OHC} estimates obtained using both measurement techniques were highly consistent. Overall, the two methods provide consistent measures of auditory nonlinearity in individual listeners, with ACALOS offering better time efficiency.

[1] This chapter is submitted as a manuscript for publication in 'Hearing Research'. Reprint permitted by Elsevier. Please note that this chapter is not the peer-reviewed, final version of the article to be published in Hearing Research.

4.1 Introduction

A large part of the hearing deficit that sensorineural hearing-impaired (HI) listeners suffer from can be attributed to audibility as defined by the pure-tone audiogram (for an overview see, e.g., Moore, 1998). However, additional factors exist in these HI listeners, which cannot be attributed to the individual audiometric thresholds and which affect their sound perception as well. These factors are referred to as supra-threshold factors in the following, and could reflect a different input/output (I/O) function of the nonlinear basilar membrane (BM) processing (Plack *et al.*, 2004), a different temporal and spectral resolution of the auditory system (Moore, 1998; Derleth, 1999), or a different loudness perception (Launer, 1995; Appell, 2002). From a physiological point of view, one supra-threshold factor is the variation of the individual nonlinear I/O function of the BM as a consequence of sensorineural hearing impairment. Psychophysically estimated reduction of the cochlear gain at low levels, accompanied by a loss or reduction of the compressive region has been shown to account for broadening of tuning curves, and thus reduced frequency selectivity (cf. Moore, 1998) in listeners with sensorineural hearing loss. Evidence exists that less compression also may explain reduced temporal resolution in HI listeners (Glasberg and Moore, 1992; Derleth, 1999). From a phenomenological point of view, a steepening of loudness perception as a function of signal level is often observed in sensorineural HI listeners. The steepening of the loudness function compared to normal-hearing (NH) listeners' loudness function is clinically described as the "recruitment phenomenon" (Fowler, 1950) and results in a smaller level range acceptable to the listener from absolute hearing threshold to the uncomfortable level. In contrast to the modified I/O function, loudness perception can be directly assessed by ''ratio scale'' procedures (Steinberg and Gardner, 1937; Stevens, 1957) and categorical loudness scale procedures (Allen, 1990; Kiessling *et al.*, 1993; Kollmeier, 1997; Brand and Hohmann, 2002). Although both I/O functions and loudness perception involve different stages of auditory processing, with the former describing the behavior of a single auditory filter and the latter integrating across auditory filters and involving (additional) more central structures, it may still be reasonable to expect a certain relation between them. On the one hand, some relation between both is expected based on the fact that all neural excitation entering the auditory pathway is driven by the response of the BM. On the other hand, successful loudness models for hearing impairment already include two components that can be associated with properties of BM processing: one component is attributed to a reduction

of cochlear gain caused by dysfunction of the outer hair cells (OHC), while the other component is attributed to a reduction of "signal transmission" related to damaged inner hair cells (IHC) (Launer *et al.*, 1997; Moore *et al.*, 1999; Chalupper and Fastl, 2002). However, a clear connection between loudness perception and recruitment as well as estimates for cochlear gain, gain loss, overall hearing loss, and hearing loss attributed to inner hair cell damage is not yet established on an individual basis in NH and HI listeners.

The aim of the current study is therefore to estimate and to compare parameters that characterize the nonlinear I/O function and parameters that characterize loudness perception in a mixed group of NH and HI listeners. This group covers a wide range of individual I/O functions and loudness perception curves, which is required for the establishment of a hypothetical relationship between both measures. To estimate the nonlinear BM I/O function, the time-consuming method of temporal masking curves (Nelson *et al.*, 2001) was used; the method adaptive categorical loudness scaling (ACALOS; Brand and Hohmann, 2002) that was used to asses loudness perception can be executed within a time that would be acceptable in clinical practice.

In order to measure the nonlinear I/O function, psychoacoustic methods like growth-of-masking (GOM, Oxenham and Plack, 1997) and temporal masking curves (TMCs, Nelson *et al.*, 2001) have been developed, which use different variations of forward-masking experiments. These methods have also been used to investigate nonlinear I/O functions in HI listeners (Plack *et al.*, 2004; Rosengard *et al.*, 2005a; Lopez-Poveda *et al.*, 2005). A direct comparison of these methods to physiological measures with invasive techniques is not possible in humans; however, a very plausible hypothesis is that the I/O function inferred using these methods is largely of cochlear origin. Arguments supporting this hypothesis are the consistency of the observed psychoacoustic data with physiological data of the BM I/O function in animals and the fact that masker and signal are transduced by the same population of inner hair cells at the best place along the BM so that all subsequent processing will affect them likewise. In the present study, the TMC method was employed and I/O functions were derived by comparing two forward masking conditions, one for a masker frequency well below the probe frequency (off-frequency condition) and one for a masker frequency equal to the probe frequency (on-frequency condition) preceding a probe tone. The main assumption of the method is that the off-frequency masker is linearly processed at the BM site where the probe has its best representation, while the on-frequency masker is subject to the nonlinear BM processing at that site. As the probe is held constant in level, it can be

used to evoke a defined excitation at this BM site. One further assumption is that the decay of forward masking with increasing gap between masker offset and signal onset is independent of level and masker frequency. Recently, both assumptions have been challenged, questioning the validity of compression ratios of those studies that use a too-close off-frequency masker as linear reference (Lopez-Poveda and Alves-Pinto, 2008) or too high masker levels (Wojtczak and Oxenham, 2009). While it might be possible to compensate for these violations by model assumptions and corrections to the estimated compression ratios, the current study primarily focuses on the estimation of the low-level gain, which is not affected by violations to the above assumptions.

In the present study, a method referred to as ACALOS (Brand and Hohmann, 2002) was used in the same group of listeners. This method has been optimized with respect to the number of loudness categories, execution of the loudness scaling procedure, and reliability of the results. It has been standardized (ISO 16832, 2006) and is frequently used for the assessment of loudness recruitment in HI listeners. ACALOS provides judgments of loudness in categorical units (CU) as a function of the sound pressure level (SPL) of a stimulus. The aim of ACALOS is not to produce loudness functions that resemble or equal classical loudness functions (e.g., Hellman and Zwislocki, 1961), but to measure the listener's loudness perception in an efficient way within a reasonable amount of time. However, the same perceptual quantity is measured using both classical loudness methods and categorical loudness procedures (Allen *et al.*, 1990). One parameter that can be derived from categorical loudness curves resulting from ACALOS is the slope of the lower portion of the loudness function, which shows loudness recruitment as increased steepness (relative to the steepness observed in NH listeners), and which can be measured very reliably (Al-Salim *et al.*, 2010). Another way to further assess the ACALOS data is to directly associate categorical loudness with the output activity of an auditory model (e.g., Derleth *et al.*, 2001). Further links to underlying physiological mechanisms are also provided by loudness models, e.g., Zwicker (1977), Launer *et al.* (1997), Appell (2002), Chalupper and Fastl (2002), and Moore and Glasberg (2004). To model loudness perception in NH listeners, a compressive power function with an exponent less than 1 is used. Moore (1998) attributed the compressive part of the loudness model to the effects of the BM I/O function and to the nonlinear transformation of the BM response into neural activity. Sensorineural hearing impairment is then accounted for in loudness models typically by a two-component approach (Launer, 1995). One component effectively steepens the loudness function (related to a loss of OHC) and the other component accounts for a

level-independent reduction of loudness (related to a loss of IHC) (Moore *et al.*, 1999). The current study uses the dynamic loudness model (DLM; Chalupper and Fastl, 2002) to estimate the proportion of hearing loss attributed to loss of OHC and IHC from the ACALOS data, and compares the results to the parameters estimated from the TMC data.

The current study first assesses the absolute hearing threshold and observed TMCs; these data are presented in Section 4.3.1. Parameters such as cochlear gain, loss of gain, compression ratio, and IHC loss are inferred from the TMC data in Section 4.4.1 and 4.4.2. Parameters such as OHC loss and slope of the loudness function are inferred from the ACALOS data in Section 4.4.3. The parameters of both methods, TMC and ACALOS are compared in a correlational analysis in Section 4.4.4, and statistical and potential sources of systematic errors are discussed in Section 4.5.

4.2 Method

4.2.1 Subjects

Five NH listeners aged from 24 to 40 years (3 female and 2 male) and 12 HI listeners aged from 64 to 75 years (5 female and 7 male) participated in the study. Listeners were recruited using the database of the Hörzentrum Oldenburg GmbH, which contains volunteers with normal hearing and with various types of hearing losses. Pure-tone thresholds were measured using standard audiometry (IEC 60645-1, 2002) with continuously presented sinusoids using step sizes of 5 dB. These measurements were done at 11 audiometric frequencies between 125 Hz and 8 kHz (see Table 4.1). For each participant in the study, the ear with better average pure-tone thresholds was selected for testing with the TMC and ACALOS measurements. NH listeners had thresholds at or below 15 dB hearing level (HL) for all frequencies tested. The HI listeners showed mild-to-moderate symmetric hearing loss, i.e., threshold differences between the right and left ears did not exceed 20 dB for any tested frequency. Audiometric thresholds for the tested ears of HI listeners are given in Table 4.1. Examination using an otoscope revealed no abnormalities. The air-bone gap in the ears of the 11 HI listeners did not exceed 10 dB for all frequencies between 500 Hz and 4 kHz, indicating sensorineural origin of the hearing loss. One listener (GF, 74 years of age) showed an air-bone gap of 15 dB at 4 kHz. GF participated in all measurements presented in the present study and the results are shown in the Appendix as an example of the effect of an additional

conductive hearing loss component on the results. All listeners received a compensation for their participation in the experiments on an hourly basis.

Table 4.1: Absolute thresholds of hearing-impaired listeners measured using pure-tone audiometry (IEC 60645-1, 2002) ordered alphabetically by listener label.

	side	0.125 kHz	0.25 kHz	0.5 kHz	0.75 kHz	1 kHz	1.5 kHz	2 kHz	3 kHz	4 kHz	6 kHz	8 kHz
AM	right	20	20	15	15	15	15	15	30	30	40	35
BG	right	5	15	20	30	30	45	40	45	60	70	70
GF	left	10	15	10	10	20	25	30	40	50	45	65
MC	right	20	20	20	20	20	25	30	30	30	55	45
MH	right	35	40	30	25	20	25	40	50	50	80	75
NB	right	35	40	55	55	55	55	55	50	50	70	65
QH	right	5	10	15	20	20	25	35	35	50	60	60
RM	right	20	10	5	5	10	10	20	40	40	55	60
SB	left	0	5	20	30	25	35	25	30	25	50	50
SG	left	10	5	10	10	5	10	10	30	45	60	65
SS	left	20	10	5	10	10	20	30	35	40	60	50
WH	left	5	5	5	5	5	5	25	30	35	30	55

4.2.2 Apparatus and calibration

All stimuli were presented via Sennheiser HDA 200 headphones to the listeners seated in a sound-insulated booth. The headphones were connected to a Tucker Davis HB7 headphone amplifier linked to a digital-to-analog converter (RME TDIF 1) that received the digital stimuli generated by a standard PC. All electronic equipment was placed outside the sound-insulated booth. The calibration was performed using a Brüel&Kjaer (B&K) measuring amplifier (Type 2610), a B&K artificial ear (Type 4153), and a B&K microphone (Type 4192). In all measurements, the frequency-dependent attenuation of the specific headphones was equalized by a frequency-dependent amplification of the sinusoidal and narrow-band noise stimuli.

4.2.3 Procedure and stimuli

4.2.3.1 Absolute audiometric threshold

In addition to standard audiometry (with level step sizes of 5 dB), the absolute threshold at 4 kHz was assessed more precisely using a three-interval, three-alternative forced-choice (3I-3AFC) procedure with adaptive tracking. This precise absolute threshold was used for the analysis of the loudness curves in section 4.3. The AFC software package

for MATLAB (The MathWorks™) developed at the Universität Oldenburg was used. The 4-kHz tone had a duration of 1000 ms including 50-ms raised-cosine ramps. The level of the sinusoid was decreased after two consecutive correct responses and increased after one incorrect response from the listener (1up-2down procedure), converging at a threshold value that corresponds to 70.7% correct on the psychometric function (Levitt, 1971). Feedback indicating the correct response was provided after each trial. The starting level was set to 20 dB above the individual audiometric threshold, thereby ensuring that the sinusoid was clearly audible in the first trial. A run consisted of trials with 6 dB level steps up to the second reversal, 3 dB up to the fourth and 1 dB up to the final tenth reversal. The threshold estimate of a run was defined as the average over the levels of the sinusoid at the last six reversals. Each listener finished five runs, including one training run. The absolute threshold was then defined as the average across the threshold estimates of the last four runs. The total testing time for this measurement was about 20 minutes.

4.2.3.2 Temporal masking curves (TMCs)

Prior to measuring the TMCs, the absolute threshold of the sinusoidal probe tone to be detected in the forward masking experiment was measured. The probe tone had a frequency f_p = 4 kHz, a steady-state portion of 5 ms and was gated with 2.5-ms raised-cosine ramps, resulting in a total duration of 10 ms. The procedure for measuring the absolute threshold of the probe tone was exactly the same as in Section 4.2.3.1, except that the starting level of the probe tone was set to 40 dB above the individual audiometric threshold (for long-duration tones).

TMCs were measured using the same sinusoidal probe tone fixed at 10 dB sensation level (SL). The probe tone followed an on-frequency forward masker ($f_m = f_p$) or an off-frequency forward masker ($f_m = 0.55 \cdot f_p$), which was varied in level during the measurement procedure. The masker had a steady-state portion of 105 ms and was gated using 2.5-ms raised-cosine ramps, resulting in a total duration of 110 ms. The temporal gap between masker and probe was defined as the interval between the zero-points of the envelopes of masker and probe. No background noise was presented during the presentation of the probe tone; the level of the probe tone was always so close to the individual threshold that neither off-frequency listening (considering a flat hearing loss profile in the hearing-impaired subject, cf. Table 4.1) nor detection in the contralateral ear was deemed likely. The TMC measurement was performed in blocks consisting of one run with on-frequency maskers and one run with off-frequency maskers in

randomized order (with a fixed temporal gap per block). Temporal gaps ranged from 0 ms to 75 ms. The masker level required to just mask the probe tone was measured using the same 3I-3AFC procedure with adaptive tracking as in Section 4.2.3.1, except for the following differences. A 2up-1down procedure was used to vary the level of the forward masker. The step sizes were 8, 4, and 2 dB for NH listeners. In pilot runs using these step sizes with HI listeners, the adaptive procedure showed much slower convergence than observed in NH listeners. Therefore, step sizes of 9, 6, and 3 dB were chosen for the HI listeners. If one listener chose the incorrect interval in the first trial of a run, two additional reversals with step sizes of 9 dB were inserted, resulting in a total of twelve reversals in that run. All listeners were explicitly advised to pay attention to the soft "click"-like tone after the sinusoid and were asked to select the interval that contained the probe tone. The maximum level of the masker tone was set to 102 dB SPL. If this maximum level would have been exceeded, the masker level was set to this maximum level. A run was skipped if this maximum level would have been exceeded more than four times within one run and then no data was collected for that run. The starting level L_{start} of the masker was chosen according to

$$L_{start} = \max(\Delta t \frac{dB}{ms}, L_{\text{target}}) \qquad (4.1),$$

where Δt is the temporal gap (in ms) and L_{probe} is the level of the probe tone (in dB SPL). Eq. (4.1) was found by analyzing TMC data of Rosengard *et al.* (2005a); Although Rosengard *et al.* (2005a) used a probe tone of 5 ms duration in their study rather than a 10 ms probe tone as in the present study, a clearly audible masker and probe tone for each listener in the first trial were obtained using Eq. (4.1). Furthermore, in the current experiment, this choice of L_{start} resulted in a reasonable amount of trials within one run for the listener to get used to the task before reaching the threshold region. At least five different temporal gaps were measured for each listener. The assortment of temporal gaps to be measured was selected on an individual basis in order to cover as wide of a range of masker levels as possible, while keeping the measurement time as short as possible. The minimal gap step was 5 ms. Each temporal gap was measured in five to seven blocks consisting of two runs each. Data collection did not begin until a listener had a minimum of half an hour practice (NH listeners) or an hour practice (HI listeners). At least four blocks of each temporal gap were obtained for the subsequent data analysis. The whole procedure lasted four to five hours that were distributed over five sessions and interleaved by the ACALOS measurements (see Section 4.2.3.3). One session lasted a maximum of two hours including an obligatory

pause of about 15 minutes. Listeners were asked to make an additional break whenever they felt tired or could not concentrate on the task any more.

4.2.3.3 Adaptive categorical loudness scaling (ACALOS)

ACALOS was measured using both one-third-octave band noises with center frequencies of 0.5, 1, 2, and 4 kHz, and sinusoids with frequencies of 0.5, 1, 2, and 4 kHz. The reason for using both stimuli was that one-third-octave band noises are typically used in clinical examinations as a tool for hearing-aid adjustment, while sinusoids resemble more closely the sinusoidal probe tone used in the TMC experiment. All stimuli were 1000 ms in duration including 50-ms raised-cosine ramps. ACALOS was measured using an adaptive procedure according to ISO 16832 (2006) and using the software OMA (Oldenburg Measurement Applications) by HörTech gGmbH. Listeners were asked to rate the presented stimuli in loudness using eleven categories ranging from ‘inaudible’ to ‘too loud’ that were mapped to categorical units (CU) from 0 to 50 in 5 CU-steps: for instance ‘soft’ was mapped to 15 CU, ‘medium’ to 25 CU, and ‘loud’ to 35 CU. In the first measurement phase, the auditory dynamic range was roughly estimated starting with a stimulus at 80 dB HL and both increasing and decreasing the stimulus level in an interleaved way until the listener gave the responses ‘inaudible’ (at lower levels) and ‘too loud’ (at higher levels). If the level evoking the ‘too loud’ rating was below 105 dB HL, it defined the individual maximum level to be used in the experiment. Otherwise, the maximum level was set to 105 dB HL. In the second measurement phase, the estimated dynamic range found in the first phase was re-estimated twice in more detail than in the first measurement phase. More details about the adaptive procedure that was found to be efficient and accurate in measuring individual categorical loudness scaling are given in Brand and Hohmann (2002). Within one run, the loudness functions of the stimuli with the four (center) frequencies were measured in an interleaved way, i.e., the presentation order of the four stimuli trial-by-trial was completely random. The testing time for one run measuring ACALOS with one-third-octave band noises at four center frequencies was about 15 minutes. The same testing time was required for ACALOS using sinusoidal stimuli. Both ACALOS measurements were done three times on different days to average out day-to-day performance differences.

4.3 Experimental results

4.3.1 Temporal masking curves

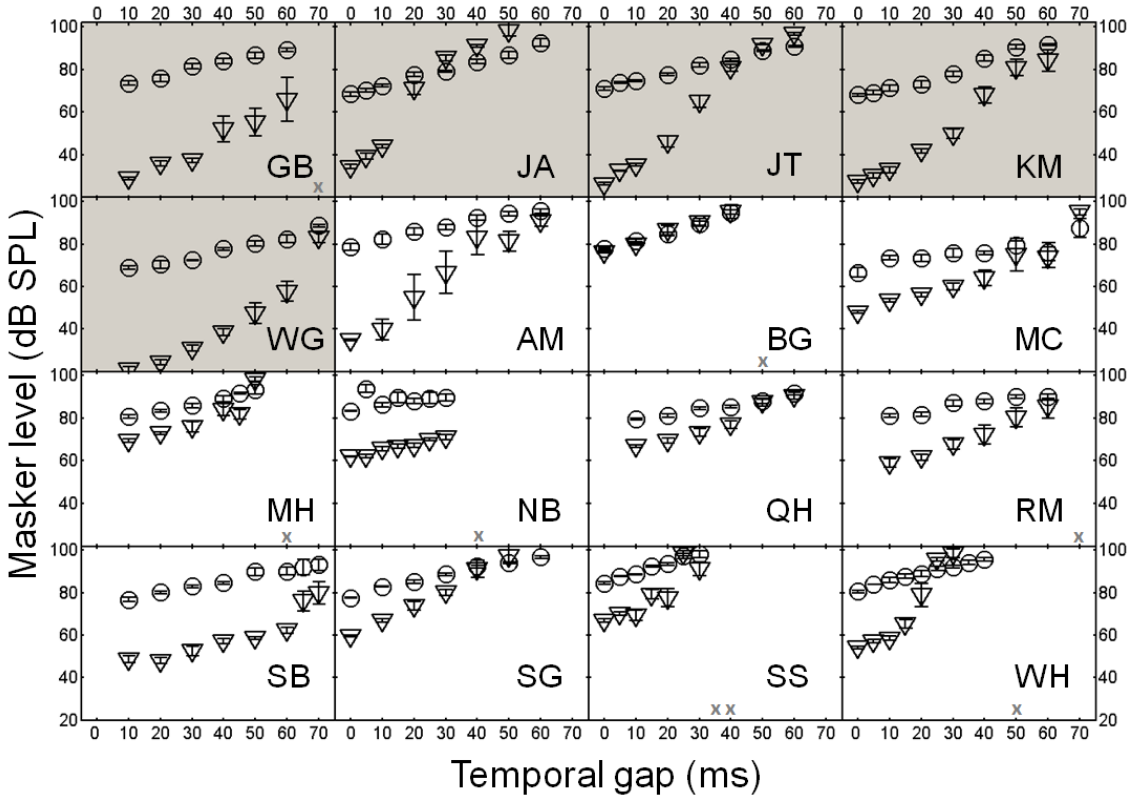

Figure 4.1: Temporal Masking Curves for normal-hearing (gray backgrounds) and hearing-impaired listeners (white backgrounds), plotted as mean masker level at threshold as a function of temporal gap between masker and target for the on-frequency masker (4 kHz, triangles) and the off-frequency masker (2.2 kHz, circles). Error bars denote the standard error. A gray x indicates a temporal gap at which no threshold could be measured.

Figure 4.1 shows individual TMCs, i.e. the average masker level at threshold as a function of the temporal gap between masker and probe, for NH listeners (panels with gray backgrounds) and HI listeners (panels with white backgrounds). Error bars denote the standard error of at least four measurement runs. In some listeners it was not possible to reliably measure either the on-frequency or the off-frequency masker threshold at certain temporal gaps when the maximum allowed masker level was exceeded, resulting in a termination of the measurement run. Only valid thresholds of at least four measurement runs are included in the average data. The gray 'x' at the abscissa denotes temporal gaps at which neither the on-frequency nor the off-frequency masker threshold could be reliably measured. In general, off-frequency TMCs show mostly higher masker levels at threshold than on-frequency TMCs. On-frequency TMCs tend to be steeper than the corresponding off-frequency TMCs. The latter effect is more

pronounced for the NH listeners than for the HI listeners and completely absent in some listeners (e.g., BG and NB). If the BM response to the off-frequency masker is assumed to be linear at the probe frequency place, a steeper on-frequency TMC indicates the response of the on-frequency masker to be compressive at the probe frequency place. Standard errors are in the region of 1 dB to 3 dB for the off-frequency TMCs, and are similar for the shallow portions of the on-frequency TMCs. Considerably larger (up to 10 dB) standard errors are observed for the steep portions of the on-frequency TMCs.

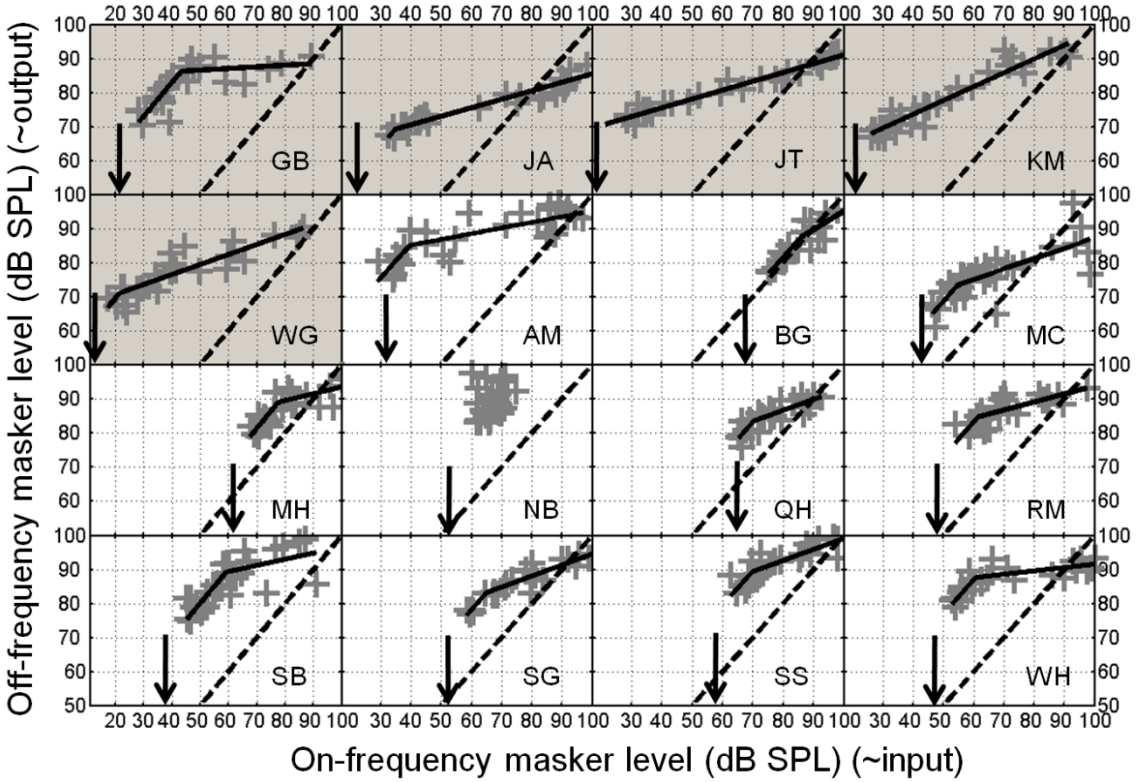

Figure 4.2: Estimated input/output functions for normal-hearing (gray backgrounds) and hearing-impaired (white backgrounds) listeners. The mean masker levels at threshold for the off-frequency masker (2.2 kHz) are plotted versus the respective thresholds of the on-frequency masker (4 kHz) paired according to the temporal gap between masker and target and block number (dark gray crosses). The black dashed line shows the hypothetical response of a passive, linear system. Abscissa intercepts of the black downward arrows indicate the individual absolute threshold of the target. The black solid line shows a fit according to Eq. (4.2) and Eq. (4.3) for characterizing low-level gain and compression ratio.

An input/output (I/O) function can be inferred from TMC data by plotting off-frequency masker threshold versus on-frequency masker threshold (Nelson *et al.*, 2001), paired according to identical temporal gaps. Figure 4.2 presents the resulting individual I/O functions for 4 kHz, inferred from the *non-averaged* TMC data (see below) of Figure 4.1 (dark gray cross symbols). Data were only included in Figure 4.2 if both on- and off-frequency masker threshold were measurable for a given temporal gap. Thus, Figure

4.2 shows the growth of masking for an off-frequency masker relative to the growth of masking for an on-frequency masker. Off-frequency masker levels are identified as output levels (ordinate) and on-frequency masker levels as input levels (abscissa) of the I/O function. The dashed black line shows the assumed response of a passive, linear system and the abscissa intercept of the black vertical arrow shows the absolute threshold of the probe. The gray cross symbols represent the masker thresholds for one temporal gap and one measurement run. This display of I/O functions differs from that of Nelson *et al.* (2001), who plotted *averaged* on- versus off-frequency masking thresholds rather than thresholds of single runs. As the measurements of the 'single-run' off- and on-frequency masker thresholds were carried out in direct succession, single-run thresholds used to infer gray cross symbols are highly comparable and similarly influenced by a possible day-to-day fluctuation of the listener's performance. Overall, in most of the HI listeners, data points at low levels show a slope of almost unity implying a linear I/O-response. At medium-to-high levels a shallower slope is observed implying a compressive I/O-response. In the NH listeners, the compressive portion starts at very low input levels combined with a very small region with slope of almost unity observed at the lowest levels. For listeners JT and KM, the I/O function appears compressive over the whole range (the linear regions at low levels may be located at input levels lower than measurable). The black solid line shows a model fit for characterizing low-level gain and compression ratio, and will be described in Section 4.4.1.

4.3.2 Categorical loudness scaling data

In the following, ACALOS data using one-third-octave band noises are shown (cf. ISO 16832, 2006), although ACALOS was measured both with sinusoids and one-third-octave band noises. Both signal configurations led to comparable results in the present study. The correlations of parameters estimated using ACALOS with sinusoids to parameters obtained using TMCs are also presented in Section 4.4 and discussed further in Section 4.5.

Figure 4.3 presents individual ACALOS data (dark gray circles) of NH (gray backgrounds) and HI listeners (white backgrounds) using a one-third-octave band noise with 4 kHz center frequency (dashed line). In general, all NH listeners show a very similar growth of loudness with level, which consists of a portion with shallower slope at low levels and a portion with steeper slope at high levels, respectively. Also in most of the HI listeners a shallow and a steep portion can be observed. On average, HI listeners rated the category "very soft", i.e. 5 CU, at higher levels than the NH listeners.

Different HI listeners show different slopes in the low-level portion of the loudness functions that tend to be steeper than the slopes observed in NH listeners, thereby indicating loudness recruitment (Kollmeier, 1997). There are some listeners who show low variation of stimulus levels within one loudness category (i.e. one value of categorical loudness), e.g. QH and JA, and some, who show very large variation, e.g. WH and SB.

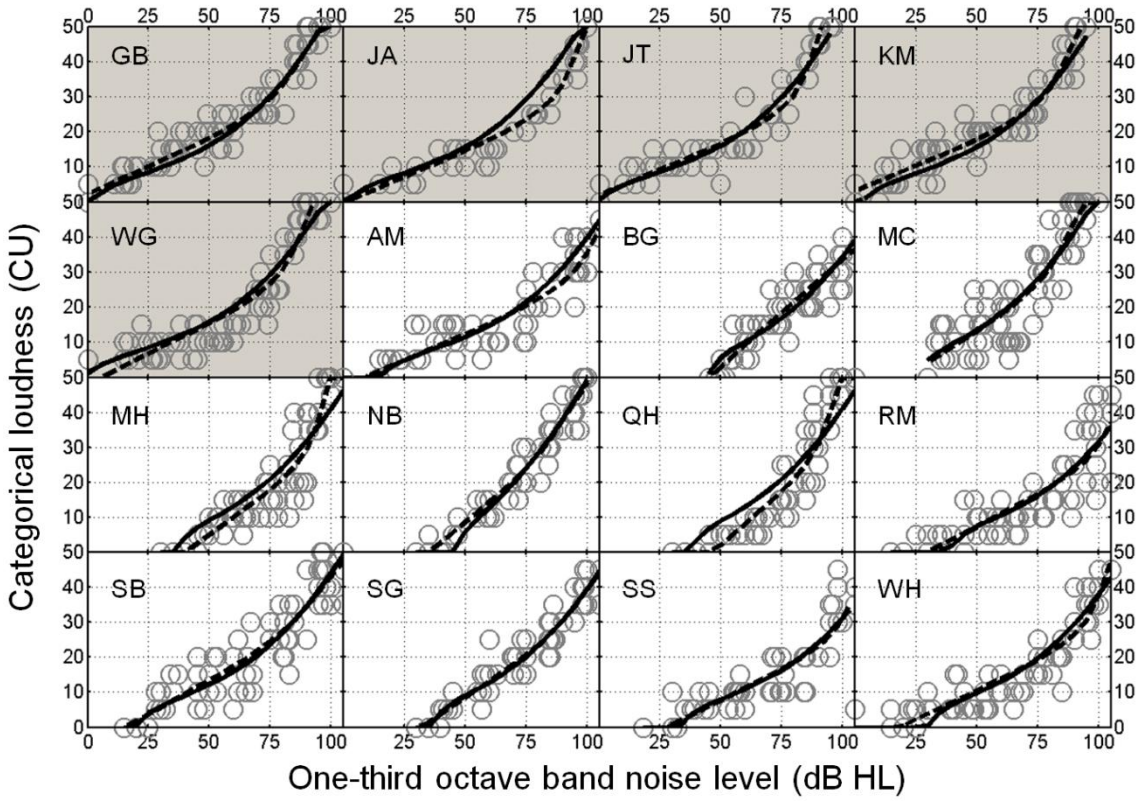

Figure 4.3: Adaptive categorical loudness scaling data of normal-hearing (gray backgrounds) and hearing-impaired (white backgrounds) listeners using one-third octave band noises with a center frequency of 4 kHz. Data (dark gray circles) are pooled across three measurements conducted on three different days. A two-section Bezier-fit was applied to the data (black dashed line). The black solid lines show modeled loudness scaling curves (see text for details) using the parameter k_{fit}, given in Table 4.4.

4.4 Data analysis and comparison

4.4.1 Estimates of low-level gain, gain loss, and compression ratio from TMC

A two-section fit was applied to the data points of Figure 4.2, assuming a linear relation of the output levels L_{out} to the low input levels L_{in} and a compressive relation to high L_{in}, which share a joint breakpoint *BP*.

$$L_{out} = L_{in} + (CR - 1) \cdot BP + V_{offset} \quad \text{for } L_{in} \leq BP \tag{4.2}$$

$$L_{out} = CR \cdot L_{in} + V_{offset} \quad \text{for } L_{in} \geq BP \tag{4.3}$$

The free parameters in this fit are the compression ratio *CR* that equals the slope of the fitted I/O function in the compressive region, the breakpoint *BP* between linear and compressive region, and a vertical offset V_{offset}. This fit is similar to the fits used by Yasin and Plack (2003) and Plack *et al.* (2004), except that the additional linear region at very high input levels is omitted, since such a linear region did not appear clearly in any of the listeners of the present study. The best fitting two-section lines are shown in Figure 4.2 as solid black lines. To quantitatively analyze the individual I/O function, a "low-level gain" *G* of the nonlinear system was extracted, which was defined as the vertical difference between the linear region of the I/O function (lower part of the solid black curve) and the unity I/O-function of the passive linear system given by $L_{in} = L_{out}$ (dashed black line). Low-level gain was derived only for listeners who had at least five data points below the breakpoint of the fit, i.e. listener NB had to be excluded and in listeners JA, JT and KM only a lower limit for the low-level gain could be derived. The lower limit is based on the assumption that the low-level linear region is located below the left-most data point in the respective panel of Figure 4.2. This data point corresponded to a temporal gap of 0 ms between masker and target. Furthermore, as an additional parameter, the "gain loss" *GL* was defined as the difference between the individual low-level gain *G* and the average low-level gain $\langle G_{NH} \rangle$ in NH listeners:

$$GL = \langle G_{NH} \rangle - G \tag{4.4}.$$

The gain loss was determined by assuming that the average low-level gain in NH listeners at 4 kHz is $\langle G_{NH} \rangle$ = 43.5 dB. This value was found by analyzing the data of Plack *et al.* (2004). The variability of the low-level gain across NH listeners in the study of Plack *et al.* (2004) was 1.5 dB. Note that the gain loss can have negative values if, e.g., a NH listener shows more low-level gain than the average NH listeners. Table 4.2 shows values of the individual low-level gain, the breakpoint between the linear und the compressive portion of the I/O function, the compression ratio, and the gain loss for NH (upper five) listeners and HI (lower twelve) listeners. Additionally, the individual absolute thresholds of the probe tone and an estimate of the standard error of the low-level gain (see Section 4.4.5.1) are shown. The low-level gain is somewhat decreased in

HI listeners compared to NH listeners and the breakpoint between the linear and the compressive region is shifted to higher levels. No systematic differences in the compression ratios between NH and HI listeners and a high degree of variability in compression ratios across listeners can be observed, consistent with the findings of Plack *et al*. (2004).

Table 4.2: Absolute threshold of the TMC-target and estimated I/O function parameters for normal-hearing (upper five) and hearing-impaired (lower twelve) listeners, ordered alphabetically by listener label. Estimates of the low-level gain and the breakpoint of the I/O function are only included if at least five single data points of the I/O function were below the breakpoint. Low-level gain and compression ratio could not be reasonably estimated for listener NB due to the inconsistency of this listener's TMC-data.

Listener	Absolute threshold of TMC-target (dB SPL)	Low-level gain (dB)	Break-point (dB SPL)	Compression ratio (dB/dB)	Gain loss (dB)	Standard error of gain loss (dB)
GB	21.1	43.0	43	0.05	0.5	1.0
JA	22.4	> 34.3	< 35	0.25	< 9.2	..
JT	18.3	> 46.2	< 24	0.26	< -2.7	..
KM	20.4	> 41.6	< 27	0.41	< 1.9	..
WG	12.5	49.5	21	0.30	-6.0	2.7
AM	30.6	45.4	40	0.17	-1.9	1.4
BG	67.6	1.5	86	0.54	42.0	0.5
GF	58.0	8.8	71	0.47	34.7	0.4
MC	42.9	18.7	55	0.31	24.8	1.1
MH	60.6	11.4	78	0.20	32.1	0.7
NB*	52.7	..	..	..	..	..
QH	65.7	13.5	68	0.38	30.0	1.0
RM	47.4	23.2	61	0.23	20.3	1.2
SB	38.0	30.1	59	0.18	13.4	0.7
SG	52.0	18.1	65	0.32	25.4	0.4
SS	58.0	19.5	70	0.33	24.0	1.2
WH	47.1	26.1	61	0.10	16.6	0.6

*excluded from further analysis of TMC data

4.4.2 Estimates of inner and outer hair cell loss from off-frequency TMCs

A further analysis of the forward masking data was performed to estimate the amount of hearing loss attributable to loss of IHCs and OHCs. The analysis assumes that in a listener with a hearing loss that is solely caused by loss of OHCs, the *off-frequency masker* has an identical masking effect on the probe tone as in a NH listener, and should

thus have the same threshold level. The basis for this assumption is that OHC loss is accompanied by a gain loss at the best frequency (cf. Plack *et al.*, 2004), increasing the absolute threshold of the probe tone by the gain loss relative to the threshold of a NH listener. For the low sensation level used in the TMC measurements (10 dB SL), it is then reasonable to expect linear processing of the probe tone in both NH and HI listeners (as the I/O function is linear at such low levels above absolute threshold, e.g., Plack *et al.*, 2004). This means that the same response of the BM relative to absolute threshold (i.e. the same BM output level) would be present in both NH and pure OHC-loss HI listeners. It can be further assumed that the off-frequency masker is processed linearly at the best frequency of the probe tone, regardless of its level and regardless of the presence of a hearing loss (Lopez-Poveda *et al.*, 2005). If it is then assumed that the off-frequency masker has to evoke the same excursion of the basilar membrane to mask the probe at the probe tone's best place as the on-frequency masker (Nelson *et al.*, 2001), the identical level of the off-frequency masker is required to mask the probe in a NH listener and a HI listener with pure OHC-loss. In turn, any difference in off-frequency masker level to the off-frequency masker levels of NH listeners can be attributed to loss or dysfunction of inner hair cells. To quantify this difference, an average off-frequency-TMC of the five NH listeners (circles in the gray panels of Figure 4.1) was first generated. Second, the average difference of the four lower-most data points of the individual off-frequency-TMCs to the corresponding data points of the average off-frequency TMC of NH listeners was calculated and taken as an estimate of the hearing loss attributable to inner hair cell loss HL_{IHC}. The four lower-most data points were chosen in order to minimize the effect of a potential residual compression that might affect the off-frequency TMCs in some listeners. To determine whether or not the difference of the four lower-most data points of the individual off-frequency-TMCs to the corresponding data points of the average off-frequency TMC of NH listeners is statistically significant in each single listener, Student's t-test was used. In all listeners but four NH listeners (GB, JA, JT, KM) and one HI listener (MC) such a significant difference was found. The use of off-frequency TMC data to estimate IHC loss was suggested by Plack (personal communication), and originally used to estimate IHC loss for listeners with temporary threshold shift (Plack and Howgate, 2010). The amount of outer hair cell loss HL_{OHC} was then obtained by assuming that $HL_{IHC} + HL_{OHC} = HL_{tot}$ (Moore and Glasberg, 1997), with HL_{tot} denoting the total amount of hearing loss. It should be noted that, although the naming of the parameters HL_{IHC} and HL_{OHC} suggests a direct relation to the underlying physiology, HL_{IHC} and HL_{OHC} do not

provide information about the number or proportion of damaged hair cells. They only estimate the amount of hearing loss (in dB) that is *related* to the dysfunction of the IHCs and OHCs, respectively. The estimated HL_{IHC} and the resulting HL_{OHC} values for each listener can be found in Table 4.3.

Table 4.3: Inner hair cell loss and resulting outer hair cell loss estimated from the bias of off-frequency-TMCs

	HL_{IHC} (dB)	HL_{OHC} (dB)
GB	1.6	-2.9
JA	0.4	-1.7
JT	2.4	-7.4
KM	-1.5	3.0
WG	-4.9	-2.4
AM	10.0	2.3
BG	9.6	41.6
GF	6.7	40.0
MC	-1.5	33.0
MH	7.6	35.7
NB	..	..
QH	5.5	39.0
RM	7.3	35.5
SB	3.9	18.1
SG	9.9	32.2
SS	16.7	18.2
WH	13.0	20.7

4.4.3 Estimates of HL_{OHC} from ACALOS

To estimate HL_{OHC} from ACALOS data, first, a two-section Bezier-fit according to Brand and Hohmann (2002) was applied to the pooled categorical loudness data from measurements conducted on three different days and is referred to as "loudness function" in the following. The fits are shown in Figure 4.3 by dashed lines. The loudness function consists of a straight portion at low levels and a straight portion at high levels. Both straight portions are connected using a Bezier curve. This two-section Bezier-fit can be characterized with three parameters: (1) the slope of the low-level portion m_{low}, (2) the slope of the high-level portion m_{high}, and (3) the level that corresponds to medium loudness L_{25} (corresponding to 25 CU). For the current data set, the slope of the low-level portion was the most reliable parameter of the loudness function, because it shows the highest ratio of the *inter-individual* standard deviation

σ_{inter} to the *intra-individual* standard deviation σ_{intra} when examining the ACALOS data of different measurement sessions (ratio for m_{low}: 2.34; ratio for m_{high}: 1.63; ratio for L_{25}: 2.12). A high ratio indicates high variability between subjects compared to a low test-retest reliability. Thus, m_{low} was taken as the most reliable parameter for quantifying the amount of recruitment.

Second, the Dynamic Loudness Model (DLM, Chalupper and Fastl, 2002) was used to estimate the amount of hearing loss due to OHCs, HL_{OHC}, in individual listeners according to their individual loudness functions. The DLM combines properties of the loudness model by Zwicker (1977) for NH listeners and a two-component approach, originally proposed by Launer (1995) and adopted by Moore and Glasberg (1997), to model hearing impairment. It is assumed within the model that the total hearing loss HL_{tot} can be split into two components, one accounting for the attenuation HL_a and one accounting for the linearization of the HI system, often referred to as an expansion HL_{exp} relative to the compressive NH system.

$$HL_{tot} = HL_a + HL_{\exp} \qquad (4.5)$$

Following Moore and Glasberg (1997), these two components can directly be related to the amount of hearing loss attributed to inner hair cell dysfunction $HL_{IHC} = HL_a$ and the amount of hearing loss caused by OHC dysfunction $HL_{OHC} = HL_{exp}$. Note that this two-component approach yields only one degree of freedom as soon as HL_{tot} is determined, since once one of either HL_{IHC} or HL_{OHC} is known the other one can be calculated using Eq. (4.5). In the model, this degree of freedom is implemented as a parameter k, which is the proportion of HL_{OHC} from HL_{tot}:

$$HL_{OHC} = k \cdot HL_{tot} \qquad (4.6).$$

By definition, k can have values ranging from 0 to 1. Figure 4.4 shows schematic loudness functions for a NH listener (gray solid line) and for a HI listener (black lines) with a hearing loss of 40 dB HL and varying proportions k of OHC loss. The steepness of the loudness function for the HI listener increases with increasing parameter k. Details about the DLM can be found in Chalupper and Fastl (2002). Briefly, the DLM was used to compute the loudness $N(t)$ in sone as a function of time t for the one-third octave band of noise signal (as also used in ACALOS). The absolute maximum of $N(t)$ was used to specify the loudness N for the whole signal. The conversion of the loudness in sone N to loudness in categorical units N_{CU}, as measured by ACALOS, was done using the following calculations that were proposed by Appell (2002): The logarithm of

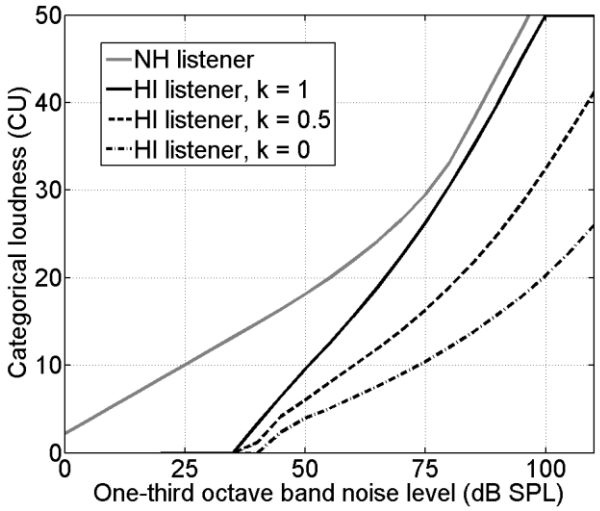

Figure 4.4: Exemplary, schematic loudness functions, i.e. categorical loudness as a function of level of a one-third octave band noise, for a normal-hearing (NH) listener (gray solid line) and for a sensorineural hearing-impaired (HI) listener with 40 dB HL (black lines) and different proportions *k* of outer hair cell loss.

the modeled loudness function in sone for a hypothetical NH listener with 0 dB HL and k = 0 was plotted versus the average categorical loudness function of NH listeners paired according to identical stimulus levels. A polynomial of 3rd degree was fitted to the resulting curve, which yielded the following relationship between the loudness in sone N and the categorical loudness N'_{CU}.

$$N'_{CU} = 2.2793 \cdot (\log_{10}(N))^3 + 2.1426 \cdot (\log_{10}(N))^2 + 8.0289 \cdot \log_{10}(N) + 8.1431 \quad (4.7).$$

This value was limited to strictly positive values of N_{CU} with

$$N_{CU} = \max(N'_{CU}, 0) \quad (4.8).$$

Thus, the model was calibrated to match loudness functions of NH listeners. The coefficients of the polynomial in Eq. (4.7) differ slightly from the coefficients found by Appell (2002) that were also used by Anweiler and Verhey (2006). To find his coefficients, Appell (2002) used loudness functions averaged over both, different listeners and different center frequencies of the stimuli, whereas in the present study, loudness functions are averaged over listeners, but only for the narrow-band stimulus with 4 kHz center frequency. The fitting procedure for finding the best fitting k was applied similar to an approach of Chalupper and Fastl (2002). The two-section Bezier-fit to the individual ACALOS data, i.e. the loudness function, was sampled in 5 dB steps and taken as representative of the data. For a starting value of k = 0.5 a modeled loudness function was calculated. Then k was iteratively changed using Levenberg-Marquardt's algorithm (Marquardt, 1963; Press *et al.*, 1992) until the $k = k_{fit}$ was found

that minimizes the least-squares distance between the representative and the modeled loudness function. The only free parameter in this fitting routine was *k*. The terminating condition to end the iterative loop was fulfilled when the difference between two subsequent *k*-values in the iteration was less than 0.01. In the fitting procedure, *k* was restricted to values ranging from 0 to 1, related to a proportion of OHC loss ranging from 0% to 100%. The resulting modeled loudness functions are shown in Figure 4.3 by solid lines. Due to the restriction to only one parameter (*k*), the modeled and measured loudness functions differ in some listeners, e.g. JA and NB, whereas in most of the listeners a very good agreement between modeled and measured loudness functions is observed.

Table 4.4: Audiometric thresholds measured using a 3-AFC procedure and parameters characterizing individual ACALOS data for normal-hearing (upper five) and hearing-impaired (lower twelve) listeners, ordered alphabetically by listener label.

Listener	Pure-tone audiometric threshold at 4 kHz (dB HL)	Slope of the low-level portion of the loudness function (CU/dB)	k_{fit}	HL_{OHC} (dB)	Standard error of HL_{OHC} (dB)
GB	-1.3	0.32	0.00	0.0	> 0.8
JA	-1.3	0.30	1.00	-1.3	> 0.7
JT	-5.0	0.29	0.86	-4.3	> 1.1
KM	1.5	0.29	1.00	1.5	> 0.5
WG	-7.3	0.35	1.00	-7.3	1.5
AM	12.3	0.30	0.00	0.00	> 1.0
BG	51.2	0.63	0.67	34.3	2.9
GF	46.7	0.43	0.41	19.1	1.9
MC	31.5	0.45	1.00	31.5	> 0.4
MH	43.3	0.51	0.76	32.9	2.9
NB	56.5	0.58	0.98	55.4	0.6
QH	44.5	0.61	0.77	34.3	1.0
RM	42.8	0.37	0.51	21.8	9.6
SB	22.0	0.39	0.64	14.1	3.4
SG	42.1	0.47	0.72	30.3	2.2
SS	34.9	0.35	0.37	12.9	4.2
WH	33.7	0.33	0.58	19.5	2.3

Table 4.4 presents the individual hearing loss at 4 kHz, the slope of the low-level portion of the loudness function, the calculated values k_{fit}, the OHC loss HL_{OHC} according to Eq. (4.6), and the standard error of HL_{OHC} (see Section 4.4.5.1) for NH

(upper five) listeners and HI (lower twelve) listeners. NH listeners show quite similar slopes of the low-level portion of the loudness function of about 0.3 CU/dB. HI listeners show large inter-subject differences ranging from 0.3 CU/dB to about 0.6 CU/dB. Individual values of k_{fit} differ highly between listeners in both the NH and the HI group. Averaged across HI listeners, a mean proportion of HL_{OHC} to HL_{tot} of 64% was found. For the NH listeners, small differences in their total hearing loss lead to small differences in HL_{OHC} nearly independent of the specific k_{fit} value. For the HI listeners large differences in k_{fit} and also large differences in HL_{tot} lead to large differences in HL_{OHC}.

4.4.4 Comparison of parameters derived from TMCs and ACALOS

For the comparison of the parameters derived from TMCs and ACALOS, the data of the NH and HI listeners were treated as stemming from a common subject population with hearing loss ranging from absent to mild-to-moderate. Listener NB had to be excluded from the comparison, as no reliable I/O function could be estimated, resulting in a total of 15 listeners. A correlation analysis of the parameters from the two measurement methods was performed and the results are shown in Figure 4.5. Each panel of Figure 4.5 displays a correlation plot of one parameter derived from TMC measurements (abscissa) and one from ACALOS or the 3-AFC-measurement of the absolute hearing loss (ordinate). Squares denote data of NH listeners and stars denote data of HI listeners. In three panels, the diagonal representing unity is shown as black dashed line. If only a lower limit of a value could be given (as for the parameters *BP* and *GL* in listeners JA, JT, and KM) that value was included in the correlations. The star in brackets denotes the additional listener GF, who was excluded from the Pearson's correlation coefficient *r* (see below), since this listener showed a combined conductive and sensorineural hearing loss.

The upper left panel of Figure 4.5 shows that the gain loss, estimated using TMCs, is highly correlated to the OHC loss, estimated from ACALOS data ($r^2 = 0.88$, $p < 0.001$) and that the data fall close to the unity line (dashed). This result is in line with the assumption that OHC loss causes a reduction of the active process resulting in a loss of the low-level gain.

The lower left panel shows that HL_{OHC} estimated using the off-frequency TMCs (by first estimating HL_{IHC} and then subtracting HL_{IHC} from HL) is very highly correlated to HL_{OHC} estimated using ACALOS ($r^2 = 0.96$, $p < 0.001$). Again, the data closely follow the unity line. It thus appears possible to estimate HL_{OHC} consistently

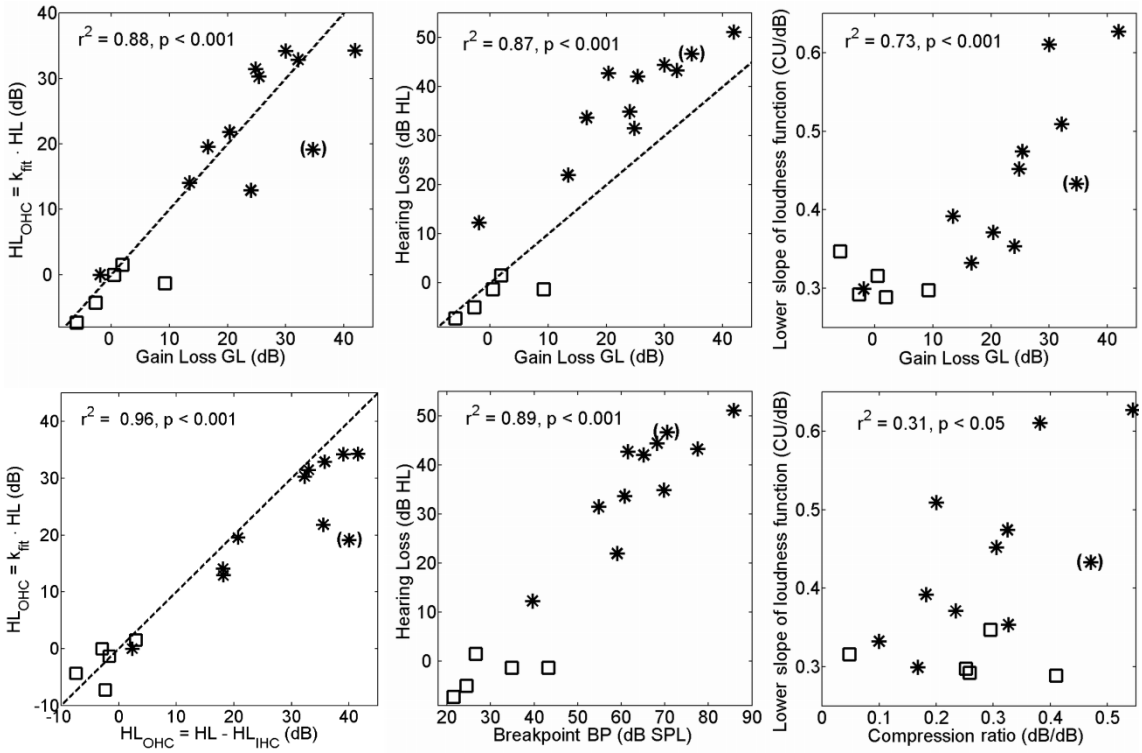

Figure 4.5: Scatter plots of parameters inferred from TMCs (abscissa, data from Table 4.2 and Table 4.3) and ACALOS or the 3-AFC- measurement of the absolute hearing loss (ordinate, data from Table 4.4). Squares: normal-hearing listeners, black stars: hearing-impaired listeners. Upper left panel: gain loss versus HL_{OHC}. Lower left panel: HL_{OHC} estimated from the off-frequency TMCs versus HL_{OHC} estimated from ACALOS data. Upper middle: gain loss versus HL_{tot}. Lower middle panel: input level at the lower breakpoint *BP* of the I/O function versus HL_{tot}. Upper right panel: gain loss versus lower slope of the loudness function. Lower right panel: compression ratios versus lower slope of the loudness function. Squares: normal-hearing listeners, stars: hearing-impaired listeners. The star in brackets denotes listener GF.

using both measurement techniques. Additionally, a high correlation was found between HL_{OHC} estimated using the off-frequency TMCs and gain loss *GL* inferred from TMC ($r^2 = 0.87$, $p < 0.001$, not shown). Also in this case the data closely follow the unity line.

The upper middle panel shows that the total hearing loss is highly correlated to the gain loss, estimated from TMCs ($r^2 = 0.87$, $p < 0.001$). The correlation coefficient is almost as high as in the upper left panel. In this panel it is also obvious that the gain loss in HI listeners is about 10 to 15 dB less than the total hearing loss while showing an otherwise linear relationship. This suggests that very mild sensorineural hearing losses involve a considerable amount of IHC loss, while OHC loss becomes increasingly important with increasing total hearing loss.

The lower middle panel of Figure 4.5 shows that the total hearing loss is highly correlated to the breakpoint between the linear and compressive region of the I/O function, as estimated from TMCs ($r^2 = 0.89$, $p < 0.001$).

The upper right panel shows that the slope of the ACALOS loudness function m_{low} (as a measure of recruitment) is correlated to the gain loss estimated from TMC (r^2 = 0.73, p < 0.001). However, this correlation is not as high as, e.g., the correlation between HL_{OHC} and gain loss (upper left panel). This suggests that it is disadvantageous to *directly* infer the gain loss from ACALOS measurements via the steepness of the loudness function, but that it is necessary to employ a loudness model that extracts a parameter such as HL_{OHC} to characterize the gain loss.

A common hypothesis is that the steepness of the loudness function, i.e. the amount of loudness recruitment, is in some way related to the amount of compression, quantified by the compression ratio (e.g., Launer, 1995; Moore, 1998; Derleth *et al.*, 2001). The lower right panel of Figure 4.5 shows, however, that there is only a poor correlation of these two values (r^2 = 0.31, p = 0.038) using Pearson's correlation coefficient r. NH listeners show a wide range of compression ratios and a small range of slope values, whereas HI listeners show variations in both parameters. Additionally, there was no correlation found between compression ratios and absolute hearing loss (r^2 = 0.10, p = 0.245, not shown). These results are in line with the assumption that hearing loss does not affect the amount of compression but rather reduces the level region where compression is observed in the impaired system as suggested in Plack *et al.* (2004). Furthermore, possible systematic errors in the analysis that could have additionally affected the correlation values are discussed in Section 4.5.

In addition to the data used in the above analysis, ACALOS was also measured with sinusoidal stimuli, unusual for this type of measurement, but more comparable to the stimuli used in the TMC measurements. Qualitatively, the same results were obtained: The corresponding correlations were r^2 = 0.88 (p < 0.001) for the upper left panel, and r^2 = 0.94 (p < 0.001) for the lower left panel (with the data closely following the unity line), and r^2 = 0.87 (p < 0.001) for the upper middle panel, r^2 = 0.89 (p < 0.001) for the lower middle panel, r^2 = 0.73 (p < 0.001) for the upper right panel, and r^2 = 0.32 (p < 0.05) for the lower right panel.

4.4.5 Variability of parameters

Since parameters derived from TMCs and ACALOS measurements were inferred using both nonlinear models like the DLM and assumptions about the auditory processing of signals (common in the literature), it is reasonable to investigate the accuracy of these parameters. Therefore, in this section the standard error, i.e. the *statistical* accuracy, of these parameters is quantitatively investigated. Statements about *systematic* deviations

that originate in the assumptions made in these analyses can be given only qualitatively based on the data that was collected in this study and can be found in the discussion section.

4.4.5.1 Standard error of parameters from TMC measurements

The standard error of the estimated gain loss was calculated as follows. Since the fitting of the linear portion of the I/O function at low levels is mainly based on single data points below the breakpoint *BP*, these data points were analyzed. The standard error of the vertical position of the fit's linear region $\sigma_{lin,fit}$ is defined as the standard error of the (vertical) difference between the fit $y_{lin,fit}(i)$ and these data points $y(i)$

$$\sigma_{lin,\,fit} = \sqrt{\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(y(i) - y_{lin,\,fit}(i))^2} \qquad (4.9),$$

where n is the number of data points below the lower knee point and $y(i)$ and $y_{lin,fit}(i)$ are the vertical coordinates of data point and fit with the same horizontal position i. The last column of Table 4.2 shows the individual standard error of the gain loss, which ranges from 0.4 to 2.7 dB with an average value of 1.0 dB. Note that these values match the values of the standard error of the gain (not shown) if the error in the average low-level gain of normal hearing listeners is assumed to be negligible.

4.4.5.2 Standard error of parameters from ACALOS

The standard error of HL_{OHC} calculated in the analysis of the ACALOS data was estimated by considering the measurement results of the single ACALOS runs and the single runs of the 3-AFC-measurement of the absolute hearing threshold from Section 4.2.3.1. Let ΔHL_{tot} be the standard error of the measurement of the absolute hearing threshold HL_{tot} and let Δk_{fit} be the standard error of k_{fit} if the model fitting was applied to the ACALOS data of each single run. Since HL_{OHC} is the product of HL and k_{fit}, error propagation yields ΔHL_{OHC}, i.e. the standard error of the estimation of HL_{OHC}.

$$\Delta HL_{OHC} = \sqrt{\left|\Delta HL_{tot} \cdot k_{fit}\right|^2 + \left|HL_{tot} \cdot \Delta k_{fit}\right|^2} \qquad (4.10).$$

The resulting values can be found in the last column of Table 4.4, which range from 0.4 to 9.6 dB with an average value of 2.2 dB. Thus, the variability of HL_{OHC} is about 1 dB higher than the standard error of the estimated gain loss. Note that if the value of k_{fit} was found to be equal or very close to 1 or 0, very low values of Δk_{fit} occur, leading to numerical values of $\Delta HL_{OHC} = \Delta HL_{tot}$ for $k_{fit} = 1$ or $\Delta HL_{OHC} = 0$ for $k_{fit} = 0$. In these cases ΔHL_{tot} was taken as a lower limit for ΔHL_{OHC}. Furthermore, it should be noted that

listener RM showed high variations in the ACALOS data of each single run, which results in the high standard error ΔHL_{OHC} for this person.

4.5 Discussion

The present study focused on estimating the amount of hearing loss that can be attributed to the two components inner and outer hair cell loss in a group of 16 listeners with hearing impairment ranging from absent to mild-to-moderate. In the following, the data and findings of the current study are discussed focusing on four specific aspects: (1) the possible systematic deviations of the estimation of parameters of TMC, (2) the possible systematic deviations of parameters from ACALOS, (3) the relation of ACALOS loudness functions to classical loudness curves, and (4) the correlation between ACALOS and TMC parameters.

4.5.1 Possible systematic deviations of parameters derived from TMCs

The standard error, i.e. the statistical accuracy, of the gain loss estimated from TMCs is in the range of 0.4 to 2.7 dB; however, systematic deviations may enlarge the total uncertainty of the gain loss or the compression ratio. These systematic deviations can have origin in the validity of some crucial assumptions used in this data analysis.

The first crucial assumption is that the off-frequency masker with frequency $0.55 \cdot f_p$ used in this study is processed linearly at the place of the BM with a characteristic frequency of f_p = 4 kHz. Lopez-Poveda and Alves-Pinto (2008) suggested that such an off-frequency masker might still show portions of compressive processing and that the amount of compression might be underestimated using such an off-frequency masker as a linear reference. They recommended using a reference with $0.4 \cdot f_p$ as a linear reference. However, using such a reference, the levels of the off-frequency masker required to mask the probe would be much higher than the levels used in the present study. This was avoided in the present study, since it would have reduced the possible range of off-frequency masker levels. Given that Lopez-Poveda and Alves-Pinto (2008) found no evidence for inter-subject differences in the amount of the potential residual compression of the $0.55 \cdot f_t$ masker, the accompanying systematic deviation might be reflected in a slight underestimation of compression ratios constant in all NH listeners. The measured I/O function of HI listeners would potentially be less affected, because BM compression at the probe frequency place is probably reduced in

these listeners. The position of the breakpoint and the low-level gain are deemed not to be influenced by potential residual compression of the $0.55 \cdot f_p$ masker.

The second assumption that may not hold is that the rates of recovery from forward masking are the same for different masker frequencies. Wojtczak and Oxenham (2009) found that this assumption does not hold for high-level on- and off-frequency masker levels. They concluded that compression ratios that are based on high-level data points might be overestimated by as much as a factor of 2. Since compression ratios in HI listeners are mostly based on high-level data points, whereas compression ratios in NH listeners mostly are not, the compression ratios of HI listeners might be estimated as being too compressive. However, a recent study of the same authors in four HI listeners (Wojtczak and Oxenham, 2010) suggests that this problem might not occur in HI listeners. Moreover, since low-level gain and gain loss were determined by data points at somewhat lower levels, these parameters are very likely not to be influenced strongly.

The third assumption that may not hold is that the hypothetical passive, completely linear auditory system shows an I/O function as given by the dashed black line in Figure 4.2 in all listeners uniformly, where the output *equals* the input. In other words: it is not clear if a listener with no remaining low-level gain shows equal threshold levels for on-frequency and off-frequency maskers. This assumption and thus the current definition of estimating OHC loss as gain loss is equivalent with the definition given in Lopez-Poveda *et al.* (2009). Estimated low-level gain and gain loss would be influenced if this assumption does not hold. It appears reasonable that this assumption holds in listeners with either completely normal hearing or with a pure sensorineural hearing loss without any kind of conductive component. Differences in the air-bone gap between the frequencies chosen for the on- and off-frequency masker may lead to a shift of the location of the hypothesized passive, linear system relative to the data measured using TMCs.

An assumption regarding the calculation of the gain loss is that the average low-level gain in NH listeners has an amount of 43.5 dB, as found in the data of Plack *et al.* (2004). The fact that no reliable value for the average low-level gain in NH listeners could be inferred from the data of the present study may be due to the use of slightly different stimuli, or that Plack *et al.* (2004) presented an additional notched-noise with the masker as a cue helping to reduce possible confusion effects. Nevertheless, a different value of the low-level gain in NH listeners would shift all data points of the three upper panels of Figure 4.5 horizontally and would thus have no effect on the correlation values given in the data analysis section.

The estimation of HL_{IHC} from the difference between one individual off-frequency TMC and the average off-frequency TMC of NH listeners is based on the assumption that all off-frequency TMCs carry the same amount of residual compression (at best: no residual compression). This means that all off-frequency TMCs (cf. Figure 4.1) should in theory have the same slope. As this slope varies considerably between both NH and HI listeners, there is no evidence that this assumption is not valid. However, if there would be residual compression in the NH listeners, but not in the HI listeners, HL_{IHC} would have been slightly underestimated, because compression affects those data points with higher temporal gap more than the data points with smaller temporal gap. In turn HL_{OHC} would have been slightly overestimated in HI listeners. This would lead to a slight shift of the black stars in the lower left panel of Figure 4.5 to the left.

Other assumptions like, e.g., the absence of off-frequency listening or inability to detect the probe tone with the contralateral ear are very likely to hold in the present study, because the probe had a very low sensation level and was presented to the better ear. These assumptions are in accordance with many other TMC-studies, e.g., Plack *et al.* (2004), Rosengard *et al.* (2005a) and Wojtczak and Oxenham (2009).

4.5.2 *Relation of ACALOS loudness functions to classical loudness functions*

Classical loudness functions use a ratio scale rather than categories to assess loudness perception (Stevens, 1957; Hellman and Zwislocki, 1961). The ratio scale has a unit of sone, where 1 sone is defined as the loudness of a 1 kHz sinusoidal tone at 40 dB SPL for NH listeners. For NH listeners, a classical loudness function is a power function of the physical intensity of the signal with a compressive exponent of 0.3 for levels above about 40 dB SPL (Stevens, 1957) and is much steeper for lower levels (Hellman and Zwislocki, 1961; Moore, 1998). Thus, classical loudness functions are concave and resemble somewhat the I/O function of NH listeners (cf. Figure 4.2). The relation between this classical loudness function and a loudness function measured with categorical procedures, which usually are convex, was investigated by Allen *et al.* (1990) and Appell (2002). The analysis of Allen *et al.* (1990) supports the view that categorical loudness ratings and the sone scale are both measures of the same perceptual quantity, i.e. the scales can be connected via a monotonic transformation. The exact transformation depends on the number and width of the loudness categories (measured in sone) and should always be investigated in terms of the NH data of the specific stimulus and method, since different loudness scaling procedures relate the perceptual

categories differently to the physical domain (Elberling, 1999). This transformation in terms of the NH data was done, for instance, by Appell (2002) and also in the present study. Using such a transformation, the knee point of classical loudness functions (situated at about 40 dB SPL for NH listeners) is flattened out but the information about the knee point is preserved in the categorical loudness function, as the transformation is monotonic. For these reasons, categorical loudness scaling is as much a valuable tool for examining if an abnormal loudness-growth with level (i.e. recruitment) is present in individual HI listeners, as are classical loudness procedures.

4.5.3 Possible systematic deviations of parameters from ACALOS

Categorical loudness scaling is a relatively simple method and thus easy to understand for inexperienced listeners. It is applied in clinical examinations as a tool for hearing aid adjustment (Kiessling *et al.*, 1993; Kollmeier, 1997). Its results are reliable across sessions both for individual loudness categories and for slopes of loudness functions (Al-Salim *et al.*, 2010). Mostly one-third-octave band noises are used as stimuli, since they are assumed to rule out the influence of the fine-structure of absolute hearing threshold, which was found to affect loudness perception (Mauermann *et al.*, 2004). In this study, one-third-octave band noises and sinusoids were used as stimuli. However, regarding the correlations of the outcomes of the two methods, no significant differences were found using one-third-octave band noises or sinusoids and thus both kinds of stimuli appear equally suited. A possible fine-structure of the hearing threshold does not appear to affect the correlations. This may partly be the case as fine-structure of the absolute hearing threshold would generally be assumed to be present in most of the NH listeners, but is expected to be decreased or absent in the HI listeners.

In the analysis of the ACALOS data presented in the present study, systematic deviations may have the following origins. First, a possible small (< 10 dB) air-bone-gap at 4 kHz attenuates the signal reaching the listener's inner ear and thus shifts the loudness function to higher levels. Fitting the modeled loudness function to the observed loudness function would then result in an overestimation of HL_{IHC} because the loudness model used in this study does not separate between HL_{IHC} and a conductive component of the hearing loss. Thus, HL_{OHC} would be underestimated. Another source of systematic deviations in HL_{OHC} might be the transformation from sone to CU as specified in Eqs. (4.7) and (4.8). Although the loudness model used here was calibrated to match the average NH listeners' data, in some cases modeled loudness functions for individual NH listeners differ from measured loudness functions (e.g. listener JA).

Large differences in NH listeners' loudness functions were also found by Brand and Hohmann (2002). How this finding may qualitatively affect the correlations found in this study is not quite clear. However, a more detailed loudness modeling than available in the loudness models up to now, e.g., by including suprathreshold data, such as individual low-level gain and compression ratios, may improve matching of modeled and measured loudness functions and may help to minimize this effect.

4.5.4 Correlation of parameters derived from TMCs and ACALOS

Although the TMC data can be thought to reflect the processing at a specific site of the BM, while the perception of loudness as used in ACALOS is thought to reflect a process that integrates across frequencies and thus involves different BM sites, the current study showed high correlation between parameters derived from both methods. In particular, the estimates of HL_{OHC} from the off-frequency forward masker level and from ACALOS showed very high correlation (r^2=0.96) and it was shown that *GL* and HL_{OHC} were also highly correlated (r^2=0.88). Furthermore, the lower slope of the loudness function was significantly correlated to *GL*.

The poor correlation between the lower slope of the observed loudness function and the compression ratio is in line with evidence from Plack *et al.* (2004), who found that sensorineural hearing loss reduces the level range where compression is observed rather than affecting the compression ratio. If one neglects the recent discussion on the validity of compression estimates by TMC (Lopez-Poveda and Alves-Pinto, 2008) for the time being, the lack of correlation between compression ratio and the lower slope of the loudness function might have several reasons. First, it can be attributed to the fact that it is not the compression ratio that changes in HI listeners but rather the level range where compression occurs, a notion that is supported by the current data and Plack *et al.* (2004). Second, the lack of correlation could partly be attributed to the fact that different (input) level regions are covered by the two parameters. Especially in moderately HI listeners the compressive portion of the I/O function starts at 65 to 80 dB SPL, but the lower slope of the loudness function mostly covers data points up to not more than 75 to 80 dB. However, an additional analysis comparing the higher slope of the observed loudness function and the compression ratio showed no correlation (r^2 = 0.03, p = 0.55) in addition to the lack of correlation between the compressive exponent of transformed classical loudness functions and the compression ratio. Another reason might be that the aforementioned assumptions for inferring parameters do not hold in all details. Moore *et al.* (1999) investigated the

relationship between loudness matching and a measure of the individual compression ratio, estimated using GOM, and found higher correlation coefficients between this measure of the compression ratio to HL_{OHC} (r^2 = 0.46) than to HL_{tot} (r^2 = 0.32). However, Moore *et al.* (1999) found that the difference between these correlation coefficients failed to reach statistical significance. In their study, individual differences in loudness perception, which might impair the estimation of HL_{OHC}, were accounted for by employing unilaterally HI listeners and using a loudness matching experiment between impaired and normal ear. In the present study, no significant correlations were found between compression ratios and HL_{OHC} (r^2 = 0.35, p = 0.204) and between compression ratios and HL_{tot} (r^2 = 0.10, p = 0.245). The poorer correlation coefficients compared to Moore *et al.* (1999) may partly be due to different measurement methods or due to possible systematic deviations that might occur by inferring parameters from both ACALOS and TMCs.

Regarding loudness, Elberling (1999) concluded that categorical loudness scaling cannot, in general, provide significant information for the fitting process of hearing aids and thus shows no further significant information about individual supra-threshold processing, which cannot be predicted from pure-tone audiometry itself. However, the statement of Elberling (1999) was based on the somewhat unrealistic assumption that loudness functions are linear with the level of the stimulus. Furthermore, he did not take the variability of the estimates of the absolute hearing threshold into account. Concerning the question of further information of ACALOS beyond the pure-tone audiogram, the results of the present study may be interpreted in an ambiguous way: The upper left and upper middle panels of Figure 4.5 suggest that HL_{OHC} estimated using ACALOS is as good a predictor for the gain loss as HL_{tot} itself, i.e., it does not provide additional information. However, the close-to-perfect correlation shown in the lower left panel of Fig. 5 suggests that HL_{OHC} estimated using ACALOS is a surprisingly good predictor for HL_{OHC} estimated using TMCs. The correlation of HL_{tot} to HL_{OHC} estimated using TMCs (r^2 = 0.92, $p < 0.001$) is reasonably lower. To estimate whether HL_{OHC} assessed using ACALOS provides additional information beyond HL_{tot} for estimating the gain loss from TMCs, the partial correlation coefficient (Sachs, 1999) between HL_{OHC} and gain loss with HL_{tot} partialled out was calculated. The squared partial correlation coefficient amounts to r^2 = 0.27, which is just significant on the 5%-level. Although this correlation is not strong, it can be interpreted as indicating that ACALOS provides further significant information.

Further studies should investigate the function or dysfunction of multiple nonlinear components within auditory processing of NH and HI listeners (factors, as proposed by Kollmeier, 1999) rather than assuming the variation of one nonlinear component within the loudness model used in the present study. Adapted to modeling loudness, such models would most likely be better suited to establish the direct connection between psychophysical measures of the nonlinear compression and loudness perception, and may help to better estimate the compressive nonlinearity by fitting modeled loudness functions to observed loudness functions. Adapted to changes in auditory processing due to hearing impairment, such models might serve as tools to investigate and differentiate between peripheral, central, and cognitive components of hearing impairment. The long-term aim of such an approach is an individual tuning of these models using small sets of data or very few individual parameters from HI listeners, in order to make valid predictions of speech intelligibility (Jürgens and Brand, 2009), speech quality, and the benefit of hearing aid algorithms (cf. Rohdenburg *et al.*, 2008; Meddis *et al.*, 2010).

4.6 Conclusions

The amount of hearing loss attributed to IHC and OHC loss or dysfunction was investigated by comparing parameters related to basilar membrane processing that were estimated using temporal masking curves (TMC) and adaptive categorical loudness scaling (ACALOS). The assessment was done by fitting the following parameters to the data: Gain provided by the active mechanism in the auditory system, gain loss relative to normal-hearing listeners, the breakpoint from linear to nonlinear processing, the compression ratio, and inner hair cell loss (HL_{IHC}) were estimated from the TMC data. The slope of the low-level portion of the loudness function and the individual OHC loss (HL_{OHC}) were estimated using ACALOS and fitting modeled loudness functions to measured loudness functions using a loudness model. The following conclusions can be drawn:

(1) HL_{OHC} (estimated using ACALOS and pure-tone audiometric threshold) was highly correlated with the gain loss (estimated using TMCs) and the data fell close to the unity line when plotted against each other, indicating that these two parameters estimate the same property of the (impaired) auditory system. A very high correlation was observed between total hearing loss (HL_{tot}) and gain loss, and between HL_{tot} and the breakpoint from linear to nonlinear processing.

Hence, these parameters presumably carry very limited information in addition to HL_{tot}.

(2) HL_{OHC} estimated using ACALOS showed a very high correlation and a close-to–unity-mapping with HL_{OHC} (derived from the off-frequency masker conditions of the TMC data), which indicates that both methods estimate this supra-threshold parameter consistently.

(3) Gain loss was found to be proportional to HL_{tot} but about 10 to 15 dB lower. This suggests that in the listeners of the present study, very mild sensorineural hearing losses involve a considerable amount of IHC loss with regard to HL_{tot}, while OHC loss or dysfunction becomes increasingly important with increasing total hearing loss.

(4) The gain loss can be estimated using TMCs with slightly higher *statistical* accuracy than that for the estimation of HL_{OHC} using ACALOS. However, some assumptions in estimating parameters from TMCs and ACALOS may not fully be valid and might result in a *systematic* bias of parameters inferred from these methods. This bias can only be estimated qualitatively on the basis of the data used in this study.

(5) To estimate the gain loss, the TMC technique requires about ten times more measurement time than estimating HL_{OHC} using ACALOS combined with the pure-tone audiometric threshold. Given the standard errors of the parameters estimated in this study, the combination of ACALOS and precise measurement of HL_{tot} provides all parameters also found with the TMC technique, except for the compression ratio.

4.7 Acknowledgements

The authors would like to thank Müge Kaya and Kerstin Sommer from Hörzentrum Oldenburg and Angela Josupeit, who performed the measurements. Special thanks go to Christopher Plack for many helpful comments on earlier versions of the manuscript, for very fruitful discussions, and valuable suggestions. We are grateful to Mani Swaminathan and Paul Nelson for proofreading. This study was supported by the Deutsche Forschungsgemeinschaft (DFG, SFB TRR 31 “The active auditory system”) and the BMBF project “Model-based hearing aids”.

4.8 Appendix: Data of a listener with combined conductive and sensorineural hearing loss

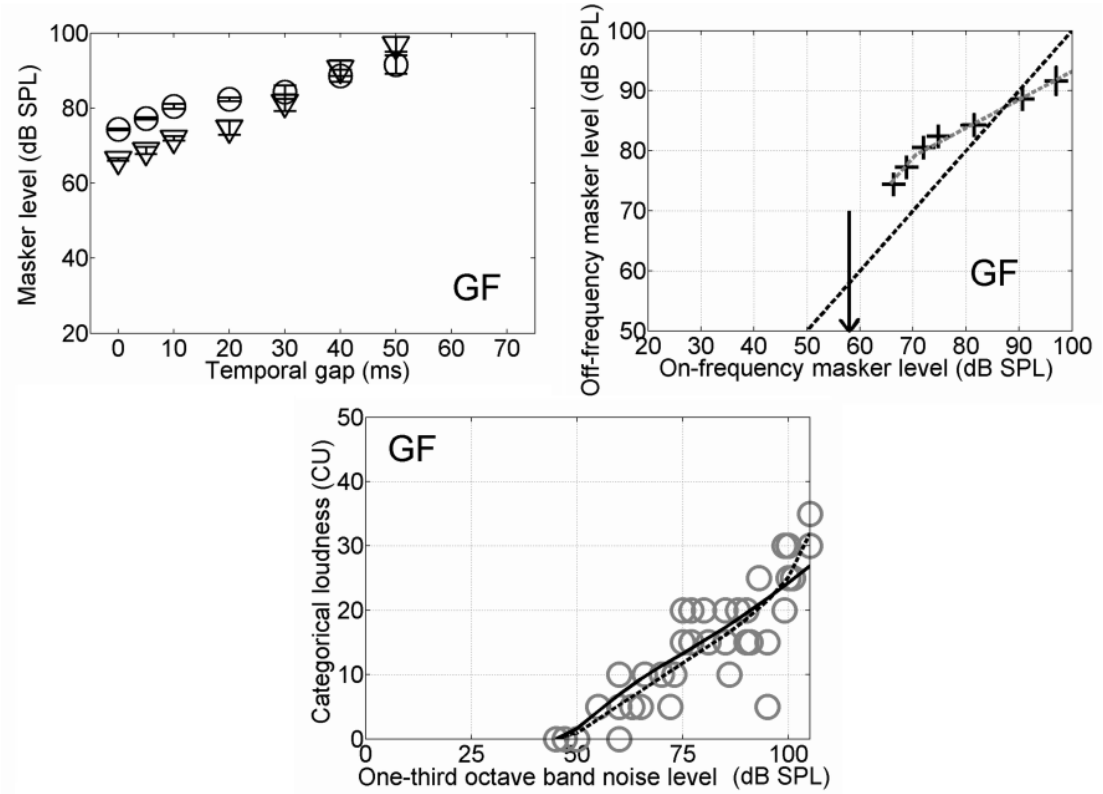

Figure 4.6: TMC-data (upper left panel), inferred I/O function at a best frequency of 4 kHz (upper right panel), and ACALOS data (lower panel) of listener GF, who shows an air-bone-gap of 15 dB at 4 kHz. For details about the display style of the panels see captions of Figure 4.1, Figure 4.2, and Figure 4.3, respectively.

Listener GF shows an air-bone gap of 15 dB at a frequency of 4 kHz and 0 to 5 dB at frequencies 2 and 3 kHz. The effects on the I/O function estimated using TMCs can be qualitatively described as follows. The on-frequency masker and the probe tone experience about 10 to 15 dB more attenuation within the outer- and middle ear than the off-frequency masker. The absolute threshold of the probe tone, as well as the on-frequency masker threshold, was therefore affected in level. For the on-frequency masker, an approximately 10 to 15 dB higher level was required to mask the probe tone. However, the difference between on-frequency masker level and absolute probe level is *not* affected since both masker and probe experience the same attenuation. The off-frequency masker threshold was also not affected; the off-frequency masker reaches the listener's inner ear without substantial attenuation. This frequency-dependent attenuation of the two maskers leads to a horizontal shift of the I/O function to the right

compared to the case where no conductive loss is present at 4 kHz. This can qualitatively be seen in the upper right panel of Figure 4.6, where two data points at high levels are on the right-hand side of the black dashed line ($L_{in} = L_{out}$). Therefore, if the gain loss, as specified by Eq. (4.4), was assumed to be relative to the black dashed line, it would be overestimated. The effect of a pure conductive loss on categorical loudness scaling is assumed in the loudness model used in the present study to be the same as the effect of a pure IHC loss, as pointed out above. Therefore, HL_{OHC} would be underestimated in listener GF. The overestimation of the gain loss and the underestimation of HL_{OHC} lead to the bracketed data point in the upper left panel of Figure 4.5.

5 Prediction of consonant recognition in quiet for listener with normal and impaired hearing using an auditory model[1]

Abstract

Consonant recognition is assessed in normal-hearing (NH) and hearing-impaired (HI) listeners in quiet condition as a function of speech level using a nonsense logatome test. Average recognition rates, recognition rates of single consonants, and confusions are analyzed. A 'microscopic' model of speech recognition, which includes an auditory model and a speech recognizer, is used to model the results using different auditory models, different model configurations accounting for hearing impairment, and different model parameters reflecting the supra-threshold processing of the speech signals. Based on these model variations, the hypothesis is tested that the speech recognition performance predicted by the model is affected by changes of the compressive supra-threshold processing, which is often observed in HI listeners. The measurement results show poorer consonant recognition rates of HI listeners than found in NH listeners with only a few common consonant confusions among the two groups. The model accurately predicts the speech reception thresholds of the NH and two out of four HI listeners, but shows too steep psychometric functions of average consonant recognition. A modification of the supra-threshold processing in the auditory model, while keeping audibility constant, shows only little impact on predicted consonant recognition, whereas a more linear auditory processing produces slightly higher recognition rates.

[1] This chapter represents a journal article manuscript in preparation. Parts of the data were presented at the 157th meeting of the Acoustical Society of America, Portland, Oregon.

5.1 Introduction

The Speech Intelligibility Index (SII) (ANSI, 1997) is designed to predict the speech recognition performance in the presence of a background noise by calculating a weighted sum over the signal-to-noise-ratios (SNRs) of speech and noise in different frequency bands. As a 'macroscopic' approach to predict individual speech intelligibility, the SII is based on the audibility (expressed by the individual pure-tone thresholds) of the long-term spectrum of speech. In the absence of noise, fairly good correlations between predicted and observed speech recognition performance can be achieved using the SII or related models of human speech recognition for listeners with normal hearing and very different kinds of hearing losses (e.g., Dubno *et al.*, 1984; Pittman and Stelmachowicz, 2000; Stelmachowicz *et al.*, 2000; Sukowski *et al.*, 2010). Nevertheless, in all of these studies some individual listeners show also large discrepancies between predicted and observed speech recognition in quiet condition. Since the SII and related macroscopic models solely rely on the individual audibility, these discrepancies cannot be explained by the factor audibility. An altered supra-threshold processing of the speech signals, which is not assessed by pure-tone audiometry, is one of the factors that might contribute to these discrepancies. Among different supra-threshold factors that may influence the processing of speech (such as the ability to benefit from the temporal fine structure of the speech signal (Lorenzi *et al.*, 2006) or top-down interactive mechanisms in the auditory processing (Davis and Johnsrude, 2007), the individual amount of compression is one candidate to play a crucial role for speech perception of hearing-impaired (HI) listeners. It is very likely (1) that the compressive nonlinearity is affected if cochlear damage is present, as physiological evidence in animal studies suggests (Patuzzi *et al.*, 1989; Yates *et al.*, 1990; Ruggero and Rich, 1991) and (2) that this affects the processing of sound well above the audiometric threshold such as speech. Moreover, the individual amount of compression being one candidate to play a crucial role for speech perception is supported by a recent study of Rhebergen *et al.* (2010), who extended the SII using a compressive input/output (I/O)-function and who found an increase of the predictive power of this extended model compared to the standard SII.

The SII as a very simple model of human speech recognition focuses on the spectral parts of speech that listeners have access to if taking their respective absolute hearing threshold into account. However, this might not be sufficient for modeling the function and dysfunction of human speech recognition. One important advantage of a

‘microscopic’ model of speech recognition (Jürgens and Brand, 2009) compared to the ‘macroscopic’ SII is that the individual auditory *spectro-temporal* processing of speech is mimicked in more detail. This approach might be a more realistic model of human speech recognition, since it allows for implementing both audibility and supra-threshold factors such as different compressive properties directly using an auditory model that processes each single speech waveform separately. Hence, the influence of altered compressive properties on modeled speech recognition performance and especially on modeled phoneme recognition can be studied.

The way *how* to implement sensorineural hearing impairment and particularly supra-threshold factors into a model of speech recognition - more specifically into an auditory model - is still a matter of debate. For instance, in the SII, hearing impairment is implemented by assessing whether or not the speech signal in a frequency band with a positive SNR is above the audiometric threshold (ANSI, 1997). Similarly, many approaches to model the speech perception of hearing-impaired listeners consider the effect of the individual audibility only: Holube and Kollmeier (1996) and Jürgens and Brand (2009), for instance, used an external masking noise, spectrally shaped to the absolute hearing threshold, to limit the audibility of the speech signals in their microscopic models. Contrary to models that consider the audibility only, Plomp (1978) proposed to interpret any hearing loss as a combination of two parts, one that attenuates sounds (related to the audiometric thresholds) and one that distorts sounds. His model could appropriately describe major aspects of the effect of hearing impairment on speech intelligibility in noise. Kollmeier (1999) proposed four factors involved in sensorineural hearing loss, which could be adjusted separately in an auditory model: loss in audibility, loss of dynamic range, increase of an ‘internal noise’, and a factor that affects several binaural functions. Derleth *et al.* (2001) split the absolute hearing threshold into an attenuating and an expansive component that could be adjusted separately in their model. The model of Derleth *et al.* (2001) is capable of modeling modulation detection, modulation matching and forward masking of hearing-impaired listeners. Stadler (2009) used an adaptation-and-prediction procedure to individually adjust parameters of his auditory model to account best for speech recognition results of cochlear implant users. Jepsen (2010) used data from psychoacoustic masking experiments to adjust the compressive (I/O)-characteristic of the nonlinear filterbank in his auditory model and could show that his model can reproduce data from other psychoacoustic masking experiments quite well.

Moreover, macroscopic models such as the SII predict *average* speech intelligibility (usually expressed in terms of the speech reception threshold, SRT) only for an entire speech test and are not suitable for predicting single speech items, i.e., words, syllables, or phonemes (Kollmeier, 1990). In contrast, the microscopic model of Jürgens and Brand (2009) allows for the prediction of recognition rates of *single phonemes* and shows a good prediction of SRT and phoneme recognition for normal-hearing (NH) listeners in speech-shaped background noise. The analysis of recognition scores of single phonemes may contribute to illuminate the nature of the speech-recognition-problem that hearing-impaired listeners suffer from in an analytic way. This is supported by Phatak *et al.* (2009), for instance, who found that recognition scores of single speech items vary considerably between different HI listeners, who show similar SRTs. It is therefore more reasonable to assess additional speech recognition information such as phoneme recognition scores, than solely assessing the SRTs of a pool of words or sentences (as usually done in the clinical assessment of speech intelligibility). By using a nonsense syllable test, Bilger and Wang (1976) found that consonant recognition scores in quiet are highly reproducible over time and are significantly decreased in HI listeners. Patterns of consonant confusions were found to be weakly related to the audiogram. There is a tendency that people with similar audiometric thresholds group consonants using common patterns that could be assessed using similarity ratings (Walden and Montgomery, 1975) or phonetic features (Bilger and Wang, 1976). These patterns are characteristic for one out of three subject groups: mild-to-moderate hearing-impaired listeners, listeners with high-frequency loss, and severe hearing-impaired listeners. However, these two studies offer a far-away-from-perfect prediction of confusion patterns and other studies, for instance Lawrence and Byers (1969), find very subject-specific (idiosyncratic) phoneme confusions of HI listeners. The reason for finding such subject-specific phoneme confusion patterns can be that phoneme confusions might be strongly influenced by non-acoustical effects of the respective test design (i.e. better familiarity of listeners with one phoneme than with another one, similarity of the nonsense syllable with the next meaningful word). Even though these non-acoustical effects limit the usage of such test results to evaluate auditory models that rely primarily on acoustic features, it still appears worthwhile to test auditory models on this kind of data to separate the acoustic effects from the presumably non-acoustic effects.

The main goals of the present study can be summarized as follows. First, consonant recognition in quiet is assessed in NH listeners (Experiment I) and in individual HI listeners (Experiment II) as a function of speech level. Second, the measurement results are predicted using different versions of the microscopic model of speech recognition (Jürgens and Brand, 2009). Third, the possible range of predicted speech recognition scores of consonants is explored by changing the compressive properties of the microscopic model predicting average consonant recognition scores in individual HI listeners. Furthermore, the hypothesis is tested that using an individual compressive I/O-characteristic might improve the prediction of consonant recognition in quiet. This I/O-characteristic is adjusted using parameters derived from a supra-threshold measurement method, namely categorical loudness scaling (cf. Chapter 4).

The development of a model predicting consonant recognition rates and consonant confusions of an individual HI listener, which incorporates results from easy and fast audiological measurements such as categorical loudness scaling, is of high interest for rehabilitative purposes. Different signal processing strategies in hearing aids could be tested using such a model and an individually optimized hearing aid could be presented to the HI listener, which enhances particularly his or her poorly recognized consonants. Furthermore, such a prediction might be of importance to prescribe individualized auditory rehabilitation training on these poorly recognized consonants without very time-consuming measurements that assess the individual consonant recognition performance.

5.2 Experiment I: phoneme recognition in normal-hearing listeners

5.2.1 Method

5.2.1.1 Participants and apparatus

Ten listeners (eight male, two female) aged from 20 to 38 years participated in this experiment. Each listener's investigated ear showed pure-tone thresholds of not more than 15 dB HL at frequencies between 125 Hz and 8000 Hz using standard audiometry (IEC60645-1, 2002). Figure 5.1 shows the range of pure-tone thresholds as gray area. All listeners received a compensation for their participation in the experiments on an hourly basis. The apparatus for the assessment of the consonant recognition scores was the same as in Section 2.2.6.

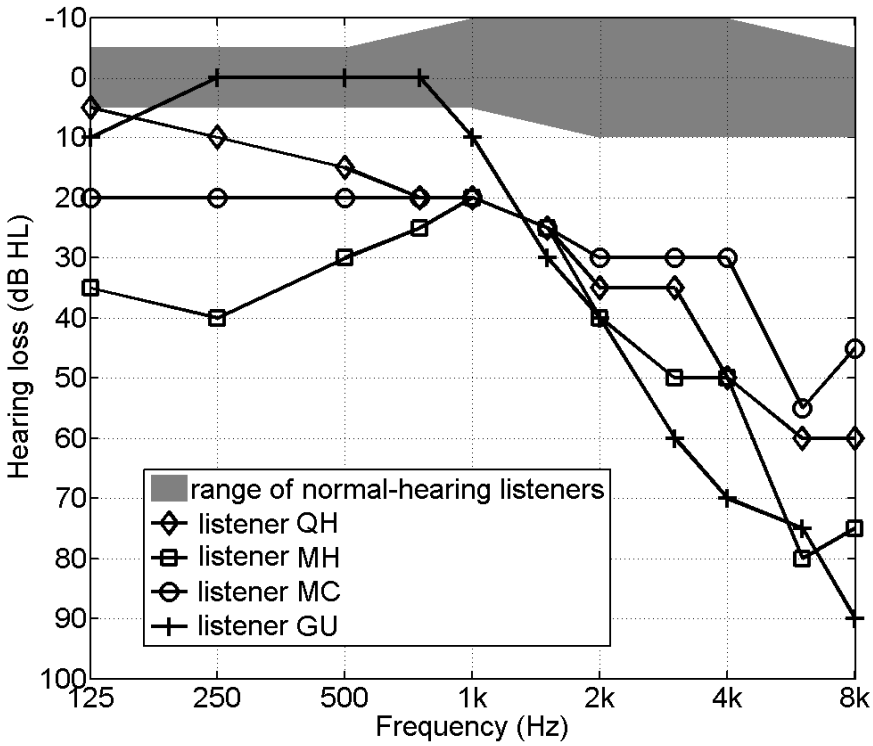

Figure 5.1: Range of pure-tone audiometric thresholds of ten normal-hearing listeners participating in Experiment I (gray area) and pure-tone audiometric thresholds of four hearing-impaired listeners participating in Experiment II (black solid lines).

5.2.1.2 Calibration

The calibration was performed using a Brüel&Kjaer (B&K) measuring amplifier (Type 2610), a B&K artificial ear (Type 4153), and a B&K microphone (Type 4192). All stimuli were free-field-equalized using an FIR-filter with 801 coefficients. The absolute speech level was defined as the root-mean-square level of the speech waveform without regarding optionally preceding or subsequent silence. The levels of different speech waveforms were adjusted in the digital domain after the free-field-equalization and not in the acoustic domain in the headphones. This means that different logatomes were presented to the listener at slightly different acoustical presentation levels (differences from -7 dB to +3 dB), whereas identical acoustic presentation levels might have been advantageous. However, the exact acoustic presentation level of each logatome used during the measurements was calculated and was also given to the model afterwards. Hence, highest comparability between measurement and modeling is ensured.

5.2.1.3 Speech tests

The recognition rates of 150 different logatomes (70 VCVs and 80 CVCs) from the Oldenburg Logatome (OLLO)[1] speech corpus (Wesker *et al.*, 2005), spoken by a male German speaker with 'normal' speech articulation (same waveforms as used in Chapter 2), were assessed monaurally via headphones at five different presentation levels (5, 10, 15, 20, 25 dB SPL). For each presentation level, the sequence of the 150 logatomes was randomly chosen. In order not to present all 150 logatomes subsequently within one list, the 150 recordings were split into two lists with 75 recordings and the order of presentation of the recordings within the two lists was shuffled. Then all ten resulting lists of all presentation levels were randomly interleaved for presentation. The response alternatives were displayed on a touch screen during and after the acoustical presentation of the test item; hence, the listener had to choose either from 10 CVC (vowel identification) or 14 VCV (consonant identification) response alternatives. The response alternatives for the presentation of a single logatome had the same preceding and subsequent phoneme (closed test). The middle phonemes of the logatomes were either vowels or consonants, which are listed below (represented with the International Phonetic Alphabet, IPA, 1999).

• Consonants:
/p/, /t/, /k/, /b/, /d/, /g/, /s/, /f/, /v/, /n/, /m/, /ʃ/, /ts/, /l/

• Vowels:
/a/, /a:/, /ɛ/, /e/, /ɪ/, /i/, /ɔ/, /o/, /ʊ/, /u/

Consonants are embedded in the vowels /a/, /ɛ/, /ɪ/, /ɔ/, and /ʊ/, respectively, and vowel phonemes are embedded in the consonants /b/, /d/, /f/, /g/, /k/, /p/, /s/, and /t/, respectively. The listeners were asked to choose the recognized logatome from the list of alternatives and were asked to guess if nothing was understood. The order of response alternatives shown to the subject was shuffled as well. To make the listener familiar with the measurement task, he or she finished two lists of 30 randomly chosen logatomes that were presented at 40 dB SPL (first training list) and 20 dB SPL (second training list) before the main measurement. Both vowel and consonant identification was measured in NH listeners. Since the topic

[1] The OLLO corpus is freely available at http://sirius.physik.uni-oldenburg.de.

of the present study is consonant identification, the vowel identification data is only presented briefly in Section 0. In a follow-up measurement session, the recognition rates of the 70 different VCVs were assessed at presentation levels of 35, 50, and 60 dB SPL using subjects with the same specifications as described in Section 5.2.1.1. The results of Experiment I are presented in combination with the model results in Section 5.6.1.

5.3 Experiment II: consonant recognition in hearing-impaired listeners

5.3.1 Method

5.3.1.1 Participants, apparatus, and calibration

Four hearing-impaired listeners (three female aged 52, 63, and 67 years, and one male aged 65 years) participated in this experiment. They showed mild-to-moderate symmetric hearing loss, i.e., threshold differences between the right and left ears did not exceed 20 dB for any tested frequency. Listener's audiometric thresholds, measured using standard audiometry (IEC60645-1, 2002), are shown in Figure 5.1 (black solid lines). The air-bone gap in all listener ears tested did not exceed 10 dB for all frequencies between 500 Hz and 4 kHz, thus indicating sensorineural hearing loss. Subjects GU and MH were hearing-aid users, whereas QH and MC were not. All listeners received a compensation for their participation in the experiments on an hourly basis. The apparatus used for the measurements was the same as in Section 2.2.6. The calibration of the speech signals was the same as in Section 5.2.1.2.

5.3.1.2 Speech tests

In a first measurement session, the recognition rates of 70 different VCVs spoken by a male German speaker with 'normal' speech articulation were assessed monaurally via headphones. The VCV recordings were the same as used in Experiment I for the assessment of consonant recognition of NH listeners. Presentation levels were chosen individually in a range between 25 to 65 dB SPL in 5 dB steps to optimally cover of the individual psychometric function of average consonant identification. Two test lists with 30 VCVs were measured in advance to familiarize the listeners with the measurement task. After finishing the first measurement session, an individual speech

level was chosen from one of the speech levels measured in the first session, which resulted in a recognition rate of about 54%, i.e. very close to the individual SRT of that HI listener. At this speech level, the recognition rates of all 70 different VCVs were measured for this HI listener nine times in a second measurement session to obtain a large amount of speech recognition data close to the individual SRT of average consonant recognition. This data serves in the following to infer a confusion matrix for each listener. The results of Experiment II are presented in combination with modeling results in Section 5.6.2.

5.4 Estimation of individual supra-threshold processing

In Chapter 4 a procedure is proposed to estimate the amount of outer hair cell loss (HL_{OHC}) from adaptive categorical loudness scaling data (ACALOS) (Brand and Hohmann, 2002) and pure-tone audiogram data using a loudness model. The method's output is a parameter k_{fit}, which is the proportion of HL_{OHC} from HL. Thus, HL_{OHC} can be obtained using

$$HL_{OHC} = HL \cdot k_{fit}. \qquad (5.1)$$

Since all listeners, who participated in the speech recognition experiments, also participated in the ACALOS measurements (the description of the measurement procedure can be found in Section 4.2.3.3), this procedure is used in the following to infer HL_{OHC} as a function of frequency. All listeners performed the ACALOS procedure three times as described in Section 4.2.3.3 with exception of listener GU, who performed ACALOS two times. Note that it is necessary to compute a transformation from sone (output of the loudness model) to categorical units (ACALOS output) for each frequency separately (Eq. (4.7)), because loudness curves slightly differ as a function of center frequency of the one-third octave band noises used in ACALOS (cf. Appell, 2002). The resulting values of k_{fit} and HL_{OHC} can be found in Table 5.1.

Table 5.1: Estimates of k_{fit} and outer hair cell loss HL_{OHC} from ACALOS data. The calculation was performed as described in Chapter 4.

Frequency	500 Hz		1000 Hz		2000 Hz		4000 Hz	
Listener	k_{fit}	HL_{OHC} (dB)	k_{fit}	HL_{OHC} (dB)	k_{fit}	HL_{OHC} (dB)	k_{fit}	HL_{OHC} (dB)
GU	0.50	0.0	1.00	10.0	1.00	40.0	0.78	54.4
MC	1.00	20.0	1.00	20.0	1.00	20.0	1.00	31.5
MH	0.56	16.7	0.42	8.4	1.00	30.0	0.76	32.9
QH	0.82	12.3	1.00	20.0	1.00	35.0	0.77	34.3

To obtain separate k_{fit} values as a function of frequency, a linear interpolation at the center frequencies of the filters was carried out for center frequencies ranging from 500 Hz to 4 kHz. If the center frequency was below 500 Hz k_{fit} was set to that value corresponding to 500 Hz and if it was above 4 kHz k_{fit} was set to that value corresponding to 4 kHz.

For a better overview regarding the absolute thresholds and the frequency-dependent amount of outer hair cell loss (HL_{OHC}) and inner hair cell loss (HL_{IHC}) in each individual listener, a graphical display is introduced, which is shown in Figure 5.2 as an example for a single HI listener. The lower black line shows the absolute audiometric threshold measured using standard audiometry (HL). The range between 0 dB HL and the absolute audiometric threshold is split in two parts, one corresponding to HL_{OHC} (dark gray) and one corresponding to HL_{IHC} (light gray), both estimated from ACALOS measurement. Note that HL_{IHC} + HL_{OHC} = HL holds for each frequency. HL_{IHC} and HL_{OHC} were used to optimally estimate the supra-threshold processing of sound in the model versions for the HI listeners (see Section 5.5.2).

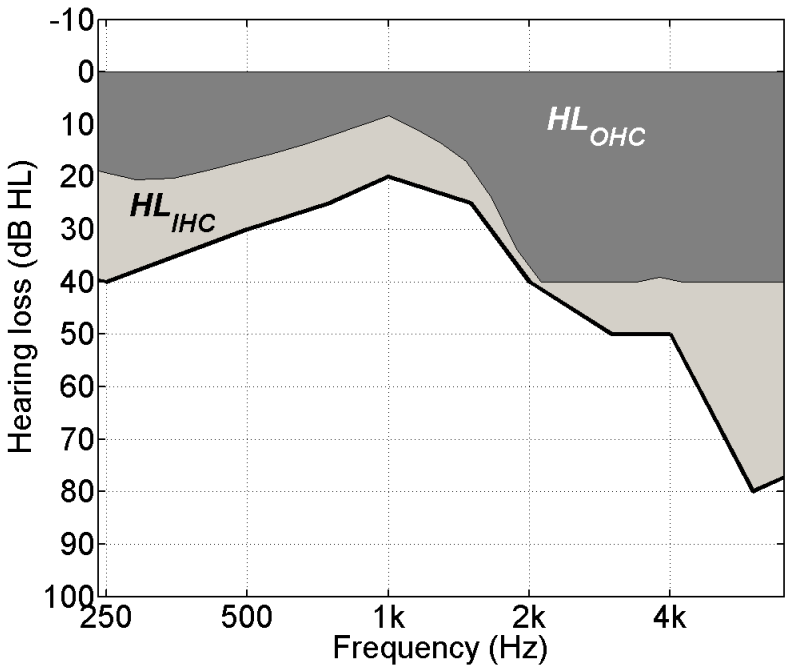

Figure 5.2: Overview on audiometric hearing loss and estimates of supra-threshold processing for the hearing-impaired listener MH. The lower black line shows the absolute audiometric threshold. The dark gray area denotes the amount of outer hair cell loss and the light gray area denotes the amount of inner hair cell loss as a function of frequency.

5.5 Modeling human speech recognition

5.5.1 Microscopic speech recognition model

The microscopic model (Jürgens and Brand, 2009) is used for the prediction of consonant recognition, i.e., the results of Experiments I and II. It is briefly described in this section, more details about the microscopic speech recognition model are presented in Chapter 2. Different auditory models are realized (specified in Section 5.5.2). Since the present study focuses on variations of the peripheral processing (i.e. the extraction of the internal representations from the speech signals), the recognizing stage was left unchanged in all model versions. This recognizing stage consists of a Dynamic-Time-Warp (DTW) speech recognizer that computes the Lorentzian distance measure (cf. Jürgens and Brand, 2009) of the internal representation to be recognized (test item) and the internal representations corresponding to the response alternatives (templates). The same speech waveforms, response alternatives, and speech levels as used in the measurements were also chosen for the model. Furthermore, a 'frozen-speech approach' was used for the recognition, i.e. the identical speech waveform used as test item is also contained within the templates (configuration B of the model of Jürgens and Brand, 2009). The same processing of the speech waveform to be recognized is also applied to the speech waveforms contained in the response alternatives. The recognition of each waveform was repeated nine times resulting in a total of ten recognition tasks per waveform, each time using a different temporal passage of the (external or internal) noise that was placed at different locations within the auditory model (see below). This kind of repetition of the recognition of each waveform was necessary to obtain enough speech recognition data for inferring confusion matrices and statistical tests.

5.5.2 Model versions to implement hearing impairment

The auditory model used in the present study is either the Perception Model (PeMo) for HI listeners (Derleth *et al.*, 2001) or the Computational Auditory Signal processing and Perception model (CASP, Jepsen *et al.*, 2008). Both, PeMo for HI listeners and CASP are extensions and improvements of the 'original' PeMo (Dau *et al.*, 1997). A model sketch of PeMo can be found in Chaper 2.2.1.1 and a direct comparison of the two models of the present study can be found in Figure 5.3. Gray blocks in Figure 5.3 denote blocks with free parameters that are adjusted in order to account for sensorineural hearing impairment. In comparison to PeMo for NH listeners, PeMo for

HI listeners consists of the same blocks, except for an attenuating block after the hair cell model, which accounts for hearing loss due to dysfunction of inner hair cells (IHCs), and an expansion block after this attenuation block, which accounts for hearing loss due to dysfunction of outer hair cells (OHCs). By adjusting the parameters in these two blocks the supra-threshold processing of the audio signal in HI listeners can be manipulated. For a listener with 0 dB HL at all audiometric frequencies, PeMo for HI listeners shows exactly the same processing as the original PeMo (Dau *et al.*, 1997).

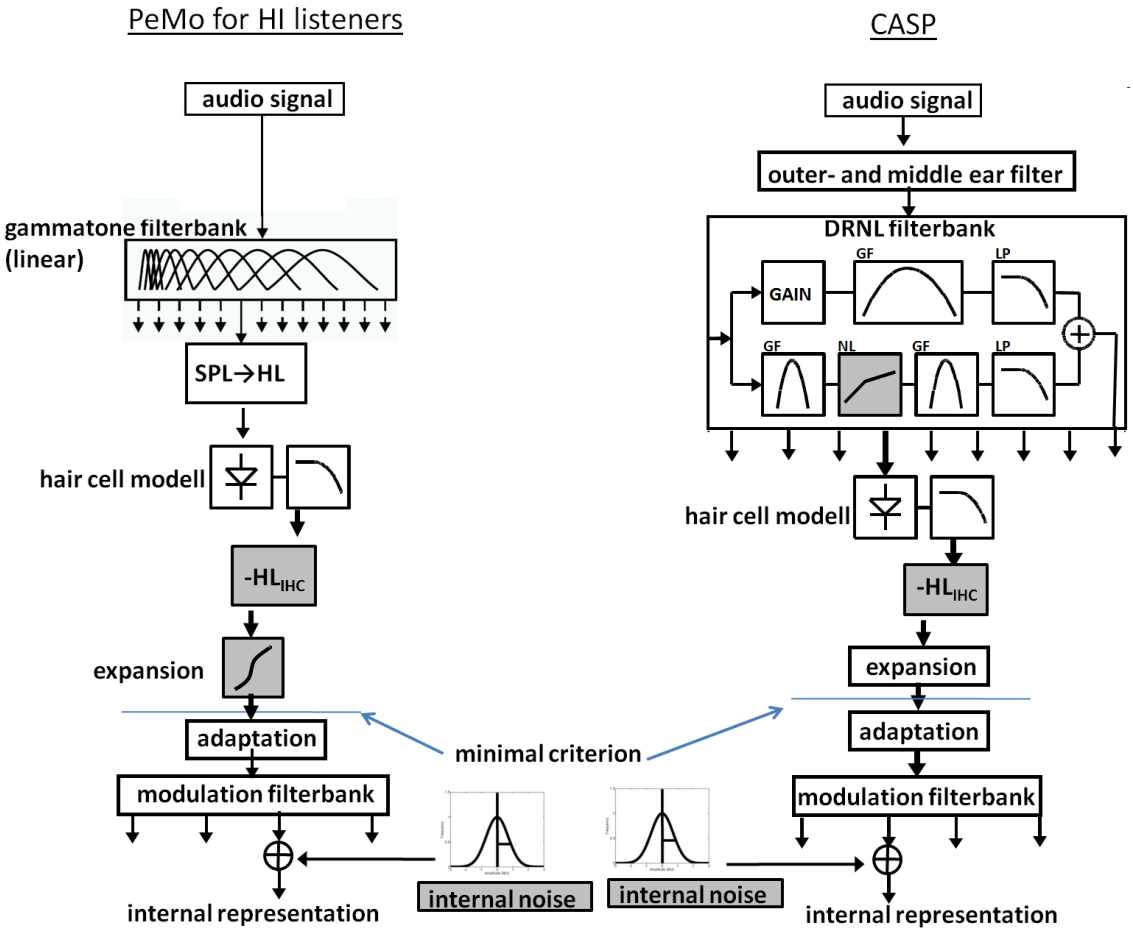

Figure 5.3: Comparison of the two extensions of the auditory model PeMo. Left: PeMo version for hearing-impaired (HI) listeners. Right: Computational Auditory Signal processing and Perception model (CASP). Gray highlighted blocks are changed in order to account for sensorineural hearing impairment (see text for details).

CASP differs more from the original PeMo. An outer- and middle ear FIR-filter attenuates the incoming audio signal in the same way as observed in the human outer and middle ear. A Dual-Resonance-NonLinear (DRNL) filterbank (Meddis *et al.*, 2001) models the complex I/O-characteristic of the basilar membrane (BM) with the motion of the stapes as input and the velocity of the BM as output. This DRNL filterbank consists

of two processing paths, one linear path with a broad frequency characteristic and one nonlinear path with a frequency characteristic tuned very sharply to the respective center frequency. For a pure-tone signal with a frequency equal or close to the respective center frequency, the linear path dominates the output at high input levels and the nonlinear path dominates the output at low input levels. Afterwards, the outputs of both paths are summed up. The gain that is given by the nonlinear processing path depends on the amplitude of the current sample, i.e. the DRNL filterbank works instantaneously. This gain is highest for low-amplitude samples, decreases as the amplitude rises, and is zero for high-amplitude samples. Hence, the DRNL filterbank reproduces the 'broken-stick' compressive nonlinearity of the BM's I/O function found in many physiological and behavioral (psychoacoustic) studies. Schematical I/O functions of a NH and a HI listener are shown in Figure 5.4.

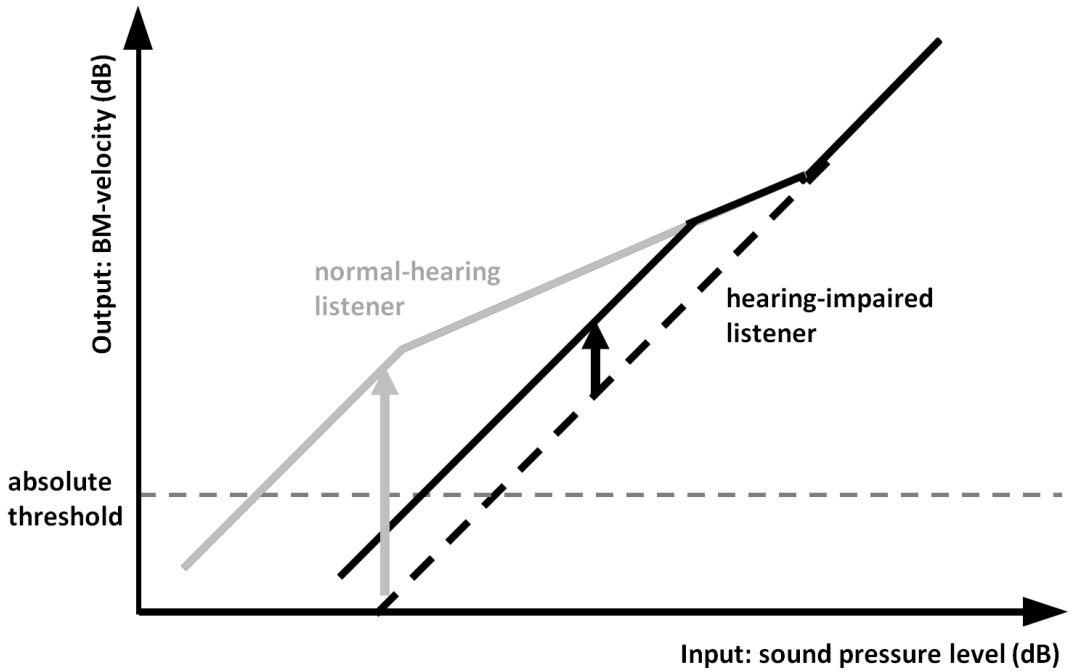

Figure 5.4: Schematic input/output function of a normal-hearing (gray) and a sensorineural hearing-impaired (black) listener. The gray and black arrows indicate the low-level gain, respectively. The black dashed black line shows a linear I/O function (input = output). The dashed grey line denotes the minimal BM-velocity necessary for the sensation of a sound.

The maximal gain at low input levels (referred to in the following as 'low-level gain') is denoted by the gray (NH) and black (HI) arrows and is defined as the difference between the low-level portion of the I/O function and the linear I/O function (input = output, black dashed line). This low-level gain, which is reduced if an outer hair cell

dysfunction is present (black arrow and cf. Chapter 4), can be easily adjusted using one parameter in the DRNL filterbank. The assumption that HL_{OHC} equals individual gain loss relatively to normal-hearing listeners is strongly supported by the results of Chapter 4. Using the same center frequencies and number of channels as in PeMo and PeMo for HI listeners, the audio signal gets split up in 27 frequency channels using this DRNL filterbank. A hair cell model extracts the temporal envelope in the frequency channels and an attenuation block accounts for the IHC part of the hearing loss. A subsequent expansion stage models the rate-intensity functions found in the auditory nerve fibers of animals for low-level stimulation and is realized by squaring the amplitude. In contrast to the approach used in PeMo for HI listeners, in CASP the expansion stage is left unchanged for NH and HI listeners. Adaptation loops and the modulation filterbank are the same as used in PeMo and PeMo for HI listeners.

Since it is not quite clear what is the best model version to account for the supra-threshold processing of normal-hearing and hearing-impaired listeners, five different model versions were realized. All model versions differ only in the 'peripheral' stage, i.e. the auditory model that computes the internal representation from the acoustic waveform, whereas the 'recognizing' stage remains unchanged. An overview about the model versions is given in Table 5.2. Furthermore, the model versions are explained in more detail below.

Table 5.2: Different model versions at a glance

	auditory model	implementation of hearing threshold	difference in supra-threshold processing from normal hearing
mC1	CASP	external hearing threshold simulating noise	none
mC2	CASP	external hearing threshold simulating noise	$HL_{OHC} = 0.8 \cdot HL$
mC3	CASP	internal hearing threshold simulating noise after DRNL filterbank	$HL_{OHC} = 0.8 \cdot HL$
mP1	PeMo for HI listeners	internal Gaussian noise after modulation filterbank	four different settings of HL_{OHC}
mC4	CASP	internal Gaussian noise after modulation filterbank	four different settings of HL_{OHC}

1. CASP version mC1 (Hearing-threshold simulating noise and NH processing)

   Model version mC1 uses a hearing-threshold simulating noise spectrally shaped to the individual audiogram (cf. Chapter 2 and Chapter 3) added to the acoustic waveform prior to entering CASP. This model version is closest to the model described in Jürgens and Brand (2009) and uses a supra-threshold processing as also used in the model for NH listeners. As no further information on supra-threshold processing is included in this model version, it only uses audibility specified by the pure-tone audiogram as a subject-specific parameter.

2. CASP version mC2 (Hearing-threshold simulating noise and HI processing, $HL_{OHC} = 0.8 \cdot HL$)

   Model version mC2 uses the same hearing-threshold simulating noise as mC1. In addition, the supra-threshold processing of the stimuli is adjusted to account for the processing observed in HI listeners. According to Moore and Glasberg (1997) it was assumed that 80% of the hearing loss, specified in dB by the pure-tone audiogram, accounts for outer hair cell loss and thus for a reduction of the gain in the low-level portion of the I/O-characteristic. Hence, the low-level portion of the I/O function of the DRNL filterbank is attenuated by $HL_{OHC} = 0.8 \cdot HL$. Since the output of the DRNL filterbank is the sum of the amplitudes of the linear and nonlinear processing path, an attenuation that is larger than the gain of the nonlinear part relative to the linear path at low levels results in an I/O function that is completely linear. If the attenuation was going to exceed the maximal possible attenuation it was limited to the maximum possible attenuation that was inferred from I/O functions of the CASP auditory model at different frequencies. The resulting values of the maximum possible attenuation are shown in Table 5.3.

Table 5.3: Maximum possible attenuation of the low-level portion of the I/O-characteristic in the CASP model

center frequency of DRNL filter (Hz)	236	488	761	1000	1470	2119	3799	7469
maximal attenuation (dB)	18.75	24.75	32.5	39	40	40	40	40

3. CASP version mC3 (internal noise added after DRNL and HI processing, HL_{OHC} = 0.8·HL)

Model version mC3 uses additive internal noise after the DRNL filterbank to simulate the absolute hearing threshold. This noise was generated by first, feeding external hearing-threshold simulating noise for the simulation of normal hearing (0 dB HL at audiometric frequencies between 125 Hz and 8 kHz) into the model and second, recording the noise in the frequency channels after the processing of the DRNL filterbank. In this model version, the recorded noise is added to the signal after the DRNL processing. For HI listeners with a pure OHC loss, the recorded noise is kept the same as for NH listeners. If IHC loss is present, the recorded noise is amplified by HL_{IHC}. The amount of HL_{OHC} and HL_{IHC} is calculated in the same way as in model version mC2.

4. PeMo version mP1 (Additive internal noise on the internal representation and different versions of supra-threshold processing)

Model version mP1 uses the PeMo model for HI listeners (Derleth *et al.*, 2001) as auditory model. For NH listeners, the suprathreshold processing in this model is the same as in PeMo (Dau *et al.*, 1997). The absolute hearing threshold is accounted for in this model as a minimum amplitude value of 10^{-5}, prior to the adaptation loops. If the amplitude of the processed signal is below this value it is set to this minimum value. Since this value corresponds to a threshold of 0 dB *SPL*, an additional stage was introduced directly after the gammatone filterbank, which attenuates the amplitude signal in such a way that the minimum amplitude value now corresponds to 0 dB *HL*. In this model version, hearing impairment is accounted for as follows. An OHC loss is implemented in the instantaneous expansion stage as proposed by Derleth *et al.* (2001). It expands the dynamics of the signal relatively to the dynamics in the supra-threshold processing of NH listeners. An IHC loss is implemented as an attenuation of the processed signal prior to the instantaneous expansion by HL_{IHC} in the same way as modeled in model version mC4.

For this model version it was necessary to limit the recognition performance by adding internal noise at the end of the auditory processing. The amount of this Gaussian noise at this position was adjusted in order to

reproduce the normal-hearing listeners' psychometric function of average consonant recognition as good as possible. The amount of Gaussian noise was kept constant for all HI listeners. Different combinations of HL_{IHC} and HL_{OHC} were realized in order to estimate the possible influence of such a supra-threshold parameter. Furthermore, the amount of OHC loss was estimated using ACALOS (cf. Chaper 4 and Table 5.1) and implemented as an additional model realization.

5. CASP version mC4 (Additive internal noise on the internal representation and different versions of supra-threshold processing)

Model version mC4 uses the CASP model and shows many similarities to version mP1. No external hearing-threshold simulating noise was used. A minimum value prior to the adaptation stage corresponds to 0 dB HL as a lower limit for the amplitude at this processing stage. An IHC loss is implemented in this model version by an attenuation and an OHC loss is implemented similar to version mC2 as an attenuation of the low-level portion of the I/O-characteristic of the DRNL filterbank, as proposed by Jepsen (2010). Similar to model version mP1, a Gaussian internal noise was necessary to limit the performance. The procedure for setting the amount of Gaussian noise at this stage was exactly the same as for model version mP1. Also in this model version, different combinations of HL_{IHC} and HL_{OHC} were realized.

5.6 Comparison of observed and predicted results

The most important data (obtained using different model versions and the variation of different model parameters) is graphically shown below in the following figures and described in some detail in the text. However, since the graphically presented data is only a small part of the completely gained model data, estimates are given for all model variations tested how the results of these models differ from the modeling results presented in the figures.

5.6.1 Modeling data of Experiment I

5.6.1.1 Average consonant recognition rates

Figure 5.5 shows average consonant recognition rates for ten normal-hearing listeners (gray error bars) and model mC1 (black symbols) as a function of speech level. The error bars denote the inter-individual standard deviation of the observed consonant recognition rates. The dashed gray line shows the random hitrate of 7.1%. A psychometric function according to Eq. (2.8) was fitted to the average data and plotted as a solid line by assuming chance at very low speech levels and perfect recognition at very high speech levels. This fit resulted in an SRT of 17.9 and 15.8 dB SPL and a slope of 4.2 %/dB and 8.4 %/dB respectively, the former for observed and the latter for predicted data, respectively. When observed and predicted data is compared, it can be stated that the model fits the average observed data of 5, 10, and 15 dB SPL quite well, whereas it slightly overestimates human performance at speech levels of 20 and 25 dB SPL. The SRT is predicted by the model within about 2 dB accuracy and the slope of the predicted psychometric function is slightly steeper than the slope of the observed psychometric function.

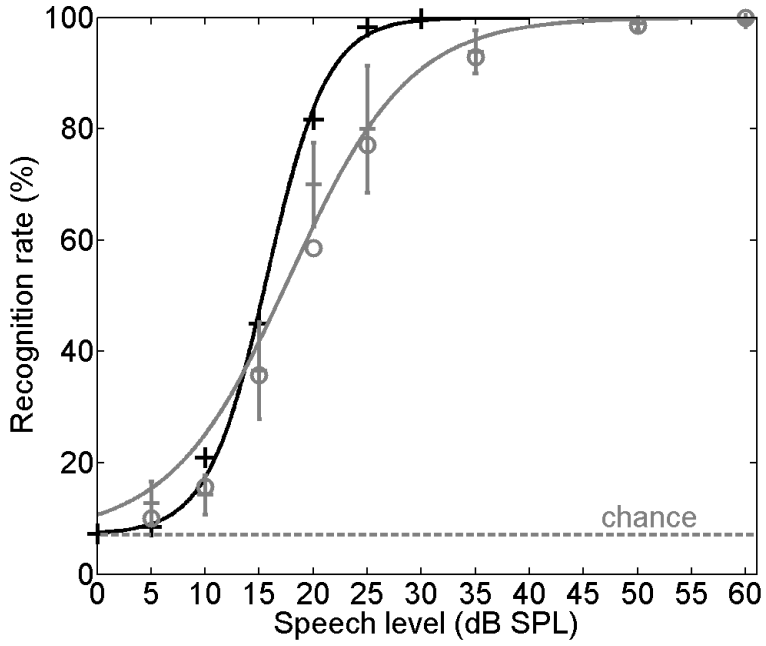

Figure 5.5: Psychometric functions of normal-hearing listeners (gray error bars and gray solid line) and model mC1 (black symbols and black solid line) for average consonant recognition in quiet. Error bars denote the inter-individual standard deviation of the observed consonant recognition rates. Additionally, the recognition rates of one (randomly chosen) normal-hearing listener are plotted for comparison (gray circles).

Additionally, the recognition rates of one (randomly chosen) normal-hearing listener are plotted in Figure 5.5 (gray circles) for comparison with the average data. The single-listener data shows an increase in recognition rate that is about as shallow as the consonant recognition data averaged over all listeners.

5.6.1.2 Confusion matrices

Figure 5.6 shows confusion matrices of consonant recognition at 15 dB SPL speech level of normal-hearing listeners (panel 1) and model mC1 (panel 2). The display is the same as in Figure 2.6. Comparing the same diagonal confusion matrix elements in panel 1 and 2, for some consonants like /t/, /g/, /b/,/p/,/ts/,/ʃ/,/l/ no significant differences between observed and predicted consonant recognition rates can be found by choosing the significance criterion described in Section 2.7. For most of the other consonants, the recognition rates predicted by the model are too high compared to the observed recognition rates, except for /k/.

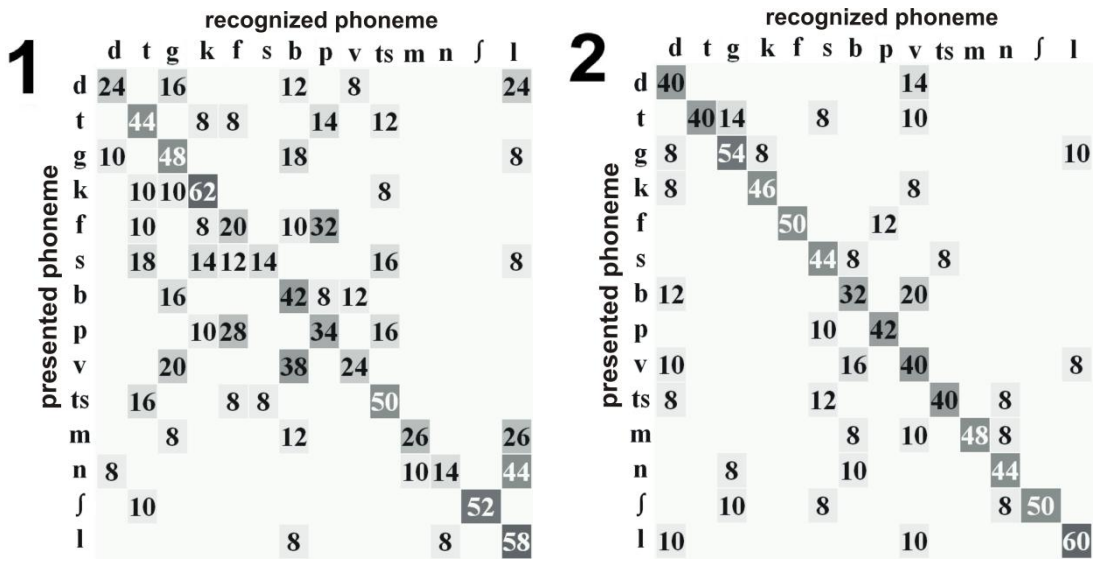

Figure 5.6: Consonant confusion matrices at 15 dB SPL, averaged over ten normal-hearing listeners (panel 1) and for model mC1 (panel 2). The display is the same as in Figure 2.6.

For these consonants, namely /d/, /f/, /s/,/v/,/m/, and /n/, at least one confusion exists in the same row of the observed data, whose recognition rate is significantly above chance level, i.e. higher than 17%. The most prominent confusions out of these are /n/→/l/ (44%) and /v/→/b/ (38%). These confusions cannot be observed in the confusion matrix produced by the model. Only the confusion

/b/→/v/ in panel 2 is significantly above chance level. A correlation of the two diagonals of panel 1 and 2 results in $r^2 = 0.08$, $p = 0.34$ using Pearson's correlation coefficient r. When comparing different diagonal elements within the same matrix it can be stated that there are many combinations of phonemes, whose recognition rates differ significantly within the confusion matrix of panel 1, but only very few combinations of phonemes, whose recognition rates differ significantly within the confusion matrix of panel 2. This means that panel 2 shows more or less the same recognition rates around 45% for all consonants. Moreover, if Figure 5.6 panel 2 is compared to a random confusion matrix that consists of the same average recognition rate of 45% (mean of the diagonal elements) no significant differences (concerning the significance criterion described in Section 2.7) can be observed. However, if Figure 5.6 panel 1 is compared to a random confusion matrix that consists of the same average recognition rate of 37%, significant differences can be observed both on the diagonal and in some non-diagonal elements (namely /n/→/l/, /v/→/b/, /m/→/l/, /p/→/f/, /f/→/p/, and /d/→/l/).

5.6.1.3 Special confusion patterns

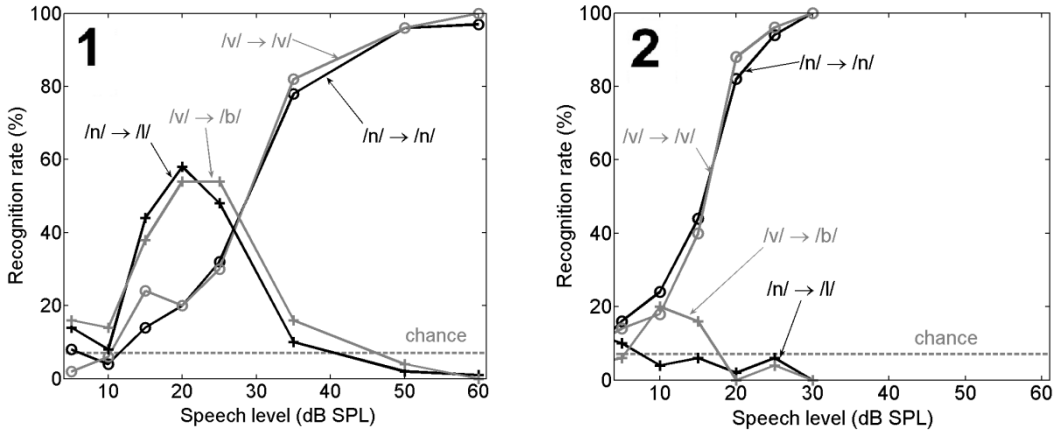

Figure 5.7: Recognition rates of consonants /n/ (black circles) and /v/ (gray circles) and of confusions /n/→/l/ (black crosses) and /v/→/b/ (gray crosses) averaged for ten normal-hearing listeners (panel 1) and for model mC1 (panel 2).

To illuminate more closely the most prominent confusions /n/→/l/ and /v/→/b/ that were found in Figure 5.6 panel 1, Figure 5.7 shows the percentages of these two matrix elements (gray and black crosses). Furthermore, the percentages of the correct recognized phoneme in this row of the confusion matrix, i.e. the diagonal element (gray and black circles) /n/→/n/ and /v/→/v/, as a function of speech level are

displayed. Panel 1 presents the recognition rates of the ten normal-hearing listeners and panel 2 the recognition rates of the model mC1. Starting in the upper right corner of panel 1 and decreasing the speech level, the recognition rates of /n/ and /v/ decrease and reach chance level at about 10 dB SPL. The recognition rates of the confusions /n/→/l/ and /v/→/b/ are close to 0% at 50 and 60 dB SPL. They increase and cross the curves of the recognition rates of /n/ and /v/ as the speech level is decreased. A maximum can be observed for 20 to 25 dB SPL and the recognition rates drop to chance level again at 10 dB SPL. Unfortunately, two gray curves are not statistically independent of each other, because the data points are part of the same row of the confusion matrix, i.e. if the recognition rate of the confusion is increased the recognition rate of the correct response is automatically decreased (cf. Chapter 2.7). Therefore, no statements can be given if the recognition rate of the respective confusion is significantly different from that of the correct response. In general, it can be stated that average NH listeners show a sensory 'morphing' of the consonant /v/ to /b/ and a morphing of /n/ to /l/ at speech levels close to and closely above the SRT as the speech level is decreased. Sensory 'morphing' was defined by Phatak *et al.* (2008) as a confusion significantly exceeding the recognition of the presented sound. Note that this morphing is performed only by attenuating the speech level without any distortion of the speech waveform. Model mC1 (Figure 5.7, panel 2) shows higher recognition rates than observed for the correct response at all speech levels. The recognition rates of the confusions /n/→/l/ and /v/→/b/ never cross the corresponding curves of the correct responses. This means that the model is not capable of modeling these very prominent confusions, especially not capable of modeling the sensory morphing, and hence the recognition rate of the correct response of these phonemes is also predicted inappropriately.

5.6.2 Modeling data of Experiment II

5.6.2.1 Average consonant recognition rates

Each of the four panels of Figure 5.8 shows average consonant recognition rates of one HI listener (gray crosses) as a function of speech level. A psychometric function according to Eq. (2.8) is fitted to the data, respectively, in the same way as in Figure 5.5 with the following exception. The maximal achievable average recognition score of the fitting function employed here is implemented as a third free parameter in the fitting

routine, rather than assuming perfect recognition at high speech levels. The reason for this is that some HI listeners do not show near-to-perfect recognition rates at speech levels as high as 60 or 65 dB SPL, in contrast to NH listeners (cf. Figure 5.5). SRTs of HI listeners are in the range from 35.5 to 44.8 dB SPL and slopes are in the range from 2.6 to 4.6 %/dB. Thus, much higher SRTs and much shallower slopes are observed when comparing these values to those of NH listeners. Four predicted psychometric functions using model mP1 are plotted in each panel. Black dashed lines show fitted psychometric functions that denote the range of possible average recognition rates, which can be achieved when HL_{OHC} is varied and HL is kept constant. This range extends from assuming $HL_{OHC} = HL$, i.e. 100% outer hair cell loss (upward triangles) which means an as-linear-as-possible processing, via 80% outer hair cell loss (downward triangles) to 0% outer hair cell loss, i.e. $HL_{IHC} = HL$ (squares) which means the compressive properties are preserved as in normal hearing.

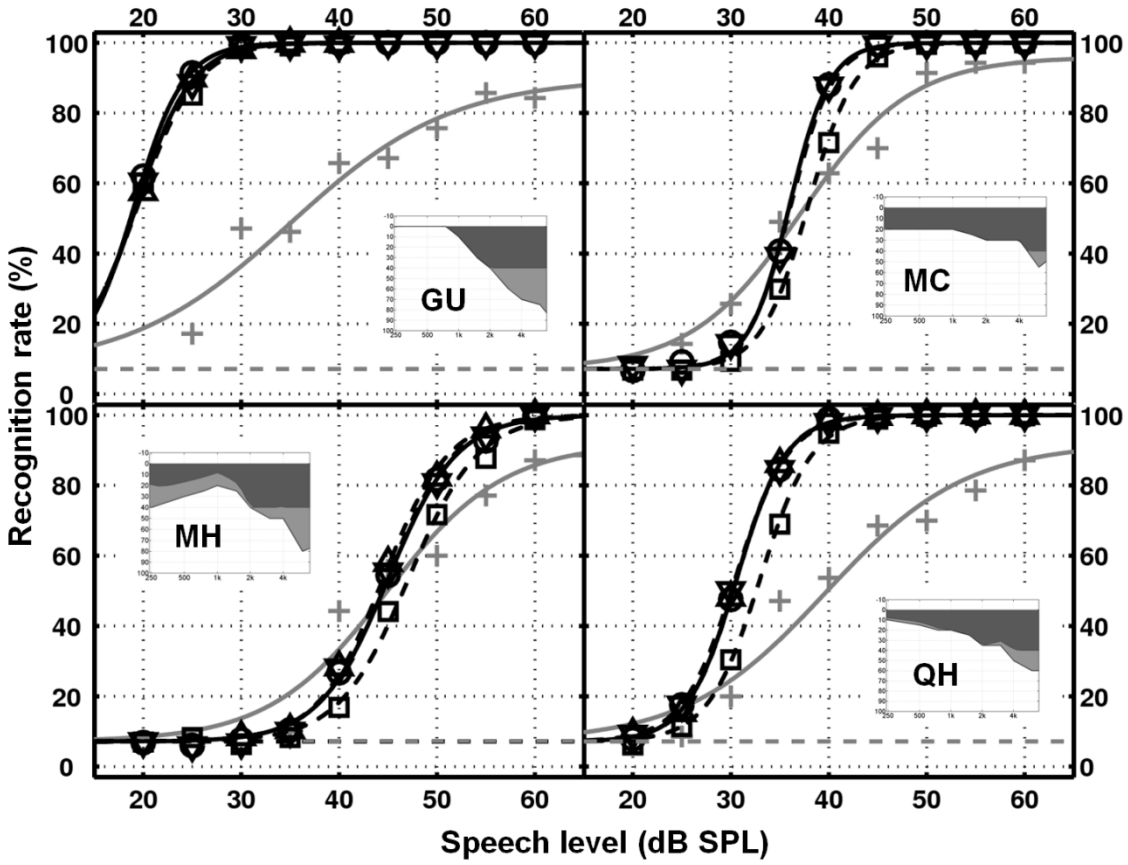

Figure 5.8: Each panel displays average consonant recognition rates of one hearing-impaired listener (gray symbols and lines), whose audiometric data and estimation of outer hair cell loss is shown in the inlet of the panels. Black dashed lines depict the range of possible average recognition rates that can be obtained by varying the parameter HL_{OHC} within model mP1. Downward triangles: $HL_{OHC} = HL$, upward triangles: $HL_{OHC} = 0.8 \cdot HL$, squares: $HL_{OHC} = 0$. Black circles and black solid lines show model predictions using model mP1 with HL_{OHC} estimated from the ACALOS (i.e. using the inlet of the panels).

Furthermore, the black circles and black solid lines represent the predicted psychometric function that was obtained by individually and frequency-specifically adjusting the compressive properties of the model as described in Section 5.4 by using the data from ACALOS. For subjects MC and MH a good agreement between the observed and predicted SRTs is obtained. For subject GU and QH the predicted average consonant recognition rates are much higher than observed. This results in a difference between observed and predicted SRT of 16 dB for GU and 8 dB for QH. Values of SRT and slope for different subjects and model variations can be found in Table 5.4. Additionally, squared correlation coefficients r^2 are given to estimate the goodness of the prediction of each model variation using Pearson's correlation coefficient r. These r^2 values were calculated by first, pooling the recognition rates of the four listeners and then, correlating the observed average recognition rates to the corresponding predicted average recognition rates paired according to speech level and listener.

Table 5.4: SRTs and slopes of the psychometric functions of average consonant recognition predicted using different model variations (rows of the table, see text for details) and for different subjects (columns of the table). The model version is denoted in the first column. Additionally, the results observed in the measurements (first row) and the squared correlation coefficients r^2 between predicted and observed average consonant recognition rates (last column) are shown.

Model version		GU		MC		MH		QH		
		SRT (dB SPL)	slope (%/dB)	SRT (dB SPL)	Slope (%/dB)	SRT (dB SPL)	slope (%/dB)	SRT (dB SPL)	slope (%/dB)	r^2
observed		35.5	2.6	36.4	4.6	44.8	3.9	39.7	3.2	1†
mC1		23.0	7.8	41.3	10.6	51.9	6.5	36.6	8.2	0.62
mC2		23.5	7.7	40.3	11.4	44.2	9.2	35.5	9.2	0.67
mC3		23.4	7.7	37.3	12.3	49.2	7.2	36.6	9.1	0.73
mP1	0%	19.4	7.7	37.9	9.7	47.1	6.5	33.0	8.7	0.63
	80%	19.3	8.2	36.1	10.8	44.9	6.7	30.5	8.6	0.62
	100%	19.3	8.3	36.0	11.3	44.5	6.9	30.6	9.2	0.62
	ACALOS	18.7	8.6	36.0	11.3	45.0	6.6	30.6	9.2	0.61
mC4	0%	21.9	6.4	40.7	6.7	52.7	4.6	36.1	6.1	0.60
	80%	21.5	6.4	36.3	9.0	42.3	7.0	34.1	7.9	0.73
	100%	21.7	6.4	34.4	9.3	40.7	7.0	33.0	8.2	0.75
	ACALOS	21.7	6.0	34.4	9.3	43.6	7.4	33.3	7.5	0.75

†: by definition

Thus, 31 data points are used for the calculation of the correlation coefficients. Note that the precondition for a correlational analysis, the statistical independence of the single data points is not valid in this case, since recognition rates were pooled over listeners. Therefore, no statements can be given regarding the statistical significance of these r^2

values. Instead, the correlation coefficient serves in the following as a tool to evaluate which model version qualitatively shows the best predictive power. r^2 values are in the range of 0.61 to 0.75. The highest r^2 values are obtained using model mC4 and HI listener's supra-threshold processing with HL_{OHC} = HL or HL_{OHC} estimated from ACALOS measurements. However, even with no change in the supra-threshold processing compared to normal hearing (model mC1) an r^2 value of 0.62 is obtained. When comparing the model versions incorporating HL_{OHC} estimated by the audiogram and model versions incorporating HL_{OHC} estimated from ACALOS (e.g. mC4 with HL_{OHC} = HL compared to mC4 with HL_{OHC} estimated from ACALOS measurements) no improvement of the match between observed and predicted SRTs is obtained. The slopes of the predicted psychometric functions are steeper than observed, whereas they tend to get shallower for a greater amount of hearing loss, which is indicated by the prediction of average recognition rates for subject MH, who shows the shallowest psychometric function.

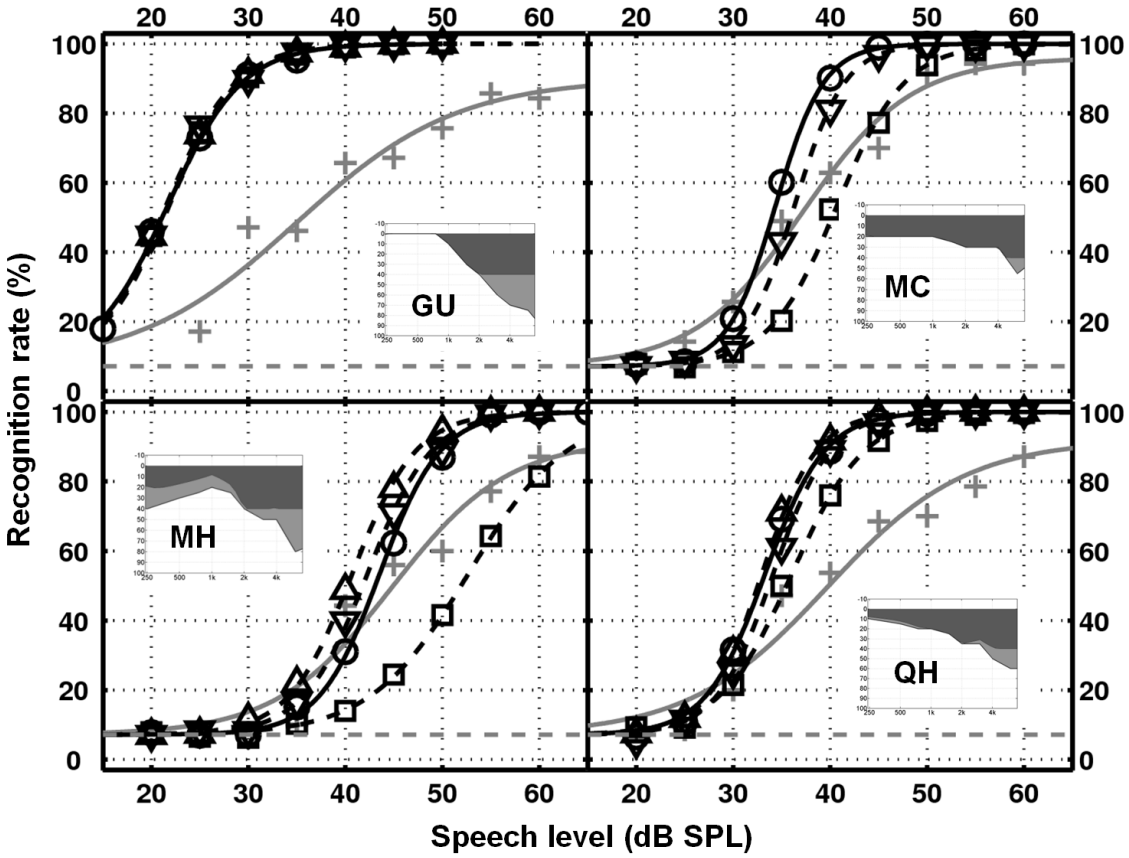

Figure 5.9: Same as Figure 5.8 but for model mC4.

The range of possible average recognition rates when adjusting the parameter HL_{OHC} in the model is quite small, below 1 dB for GU and 2 to 3 dB for MC, MH and QH. The slope of the psychometric function is not affected by the variation of HL_{OHC}. A more linear processing in the model, i.e. 100% OHC loss, results in higher recognition rates and a lower SRT for each one of the listeners.

Figure 5.9 shows predicted average consonant recognition rates using model mC4 (black symbols and black lines) in comparison to observed consonant recognition rates (gray crosses and gray solid lines). In general, the same results are obtained compared to model mP1 with the following exception. The range of possible SRTs of average consonant recognition, which can be obtained by varying the parameter HL_{OHC} within the model mC4, varies considerably between the subjects (indicated by the black dashed lines). Whereas for GU the range is below 1 dB, it amounts to 3 dB for QH, 6 dB for MC and 9 dB for MH. Furthermore, also in this model a more linear processing in the model (i.e. 100% OHC loss in model versions mP1 and mC4) results in higher recognition rates and a lower SRT for each listener.

5.6.2.2 Confusion matrices

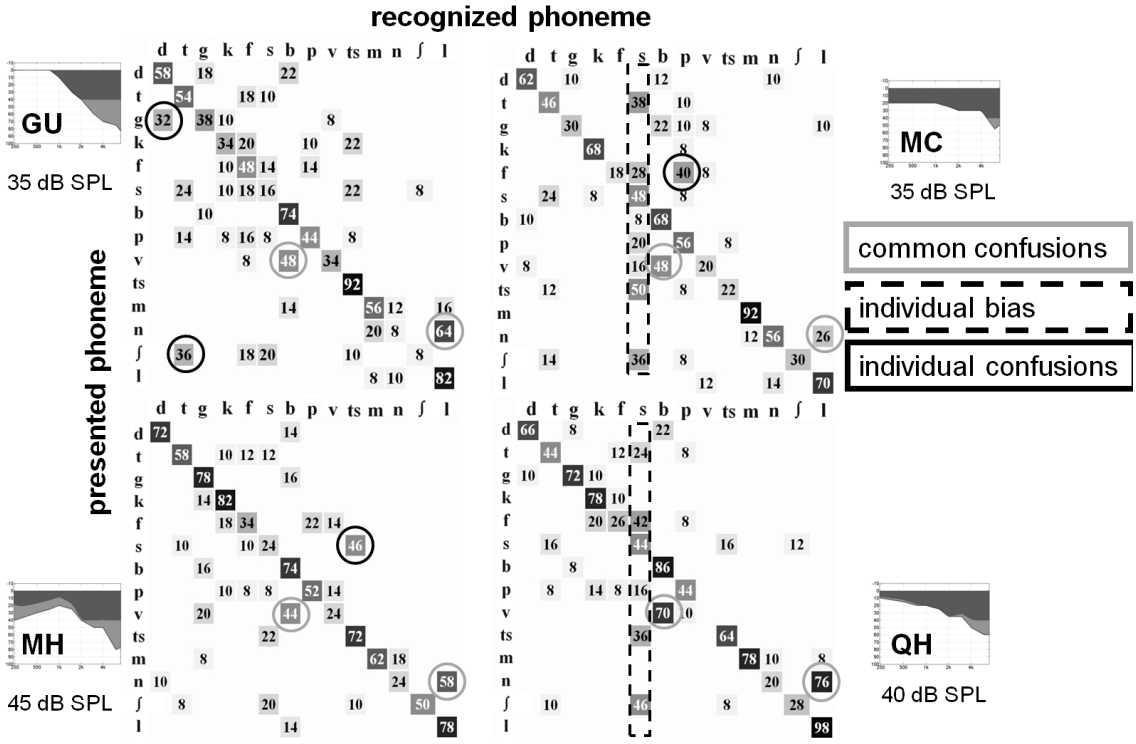

Figure 5.10: Observed confusion matrices of four HI listeners at the individual SRT of average consonant recognition in quiet. Highlighted elements are mentioned in the text. The display is the same as in Figure 5.6. Close to the confusion matrices are plots of the audiometric thresholds of the HI listeners.

Each panel of Figure 5.10 presents a confusion matrix for one out of the four HI listeners, which was assessed at a constant speech level. This speech level was chosen individually to be that level closest to the individual SRT of the average consonant recognition rates (stated next to each confusion matrix in Figure 5.10). Furthermore, the audiometric thresholds and supra-threshold processing information is given using the display introduced in Section 5.4. Some elements that will be discussed in the following are highlighted for a better overview. A glance at the diagonals reveals that there are some similarities between HI listeners, as well as between HI and NH listeners (compare to Figure 5.6 panel 1). High recognition rates are observed for /l/ in all listeners and for /k/ in all listeners except for GU. Low recognition rates are observed for /v/ in all listeners and for /n/ in all listeners except for MC. /b/ is fairly well recognized, whereas /p/ and /t/ show medium recognition rates in all HI listeners. The other consonants do not show common patterns concerning their recognition rates. In general, consonants are recognized across a very large range of possible recognition rates in each one of the listeners, ranging from nearly 0% to close-to-perfect recognition. Common confusions that are significantly above chance level are /n/→/l/ and /v/→/b/, indicated by gray circles and also observed in NH listeners (cf. Figure 5.6 panel 1). Additionally, there are individual confusions that are prominent in one listener and less prominent or absent in other listeners (black circles). For instance/g/→/d/ and /ʃ/→/t/ are confusions observed in GU, but not in MH, whereas /f/→/p/ is a confusion observed in MC and /s/→/ts/ is a confusion observed in MH. Furthermore, an individual bias can be observed in MC and QH (black dashed rectangle); no matter what high-frequency consonant (like /t/, /f/, /s/, /ts/, or /ʃ/) is presented, these listeners show a tendency to recognize /s/, revealed by high recognition rates in this column of their confusion matrix. If these observed consonant confusion matrices are compared to random confusion matrices that consist of the same average recognition rate (mean of the diagonal elements) respectively, significant differences can be found both for the diagonal and the non-diagonal elements: The diagonal elements shows a much bigger range of recognition rates than would be expected in a random confusion matrix. Non-diagonal elements that are significantly (>17%) above chance level are more frequent than expected in a random confusion matrix.

Figure 5.11 displays predicted confusion matrices for the four HI listeners using model mC4, whereas the supra-threshold processing was estimated using

ACALOS. Note that, due to the deviance of observed and predicted SRTs for subjects GU and QH, for these subjects predicted confusion matrices are shown at speech levels that best match the SRT of the *predicted* rather than the *observed* psychometric function, i.e. 20 dB SPL for GU and 35 dB SPL for QH, rather than 35 dB SPL and 40 dB SPL as in Figure 5.10. This was done because confusion matrices at the same speech levels as chosen for the observed data revealed close-to-perfect recognition of the model, i.e. only the diagonal is present with recognition rates between 90% and 100%. A glance at the diagonal of each panel of Figure 5.11 reveals that consonants are predicted to be recognized across only a quite small range of possible recognition rates, much smaller than the range of the observed recognition rates. Exactly the same elements that are highlighted in Figure 5.10 are also highlighted in Figure 5.11 for a direct comparison of single elements between observed and predicted confusion matrices.

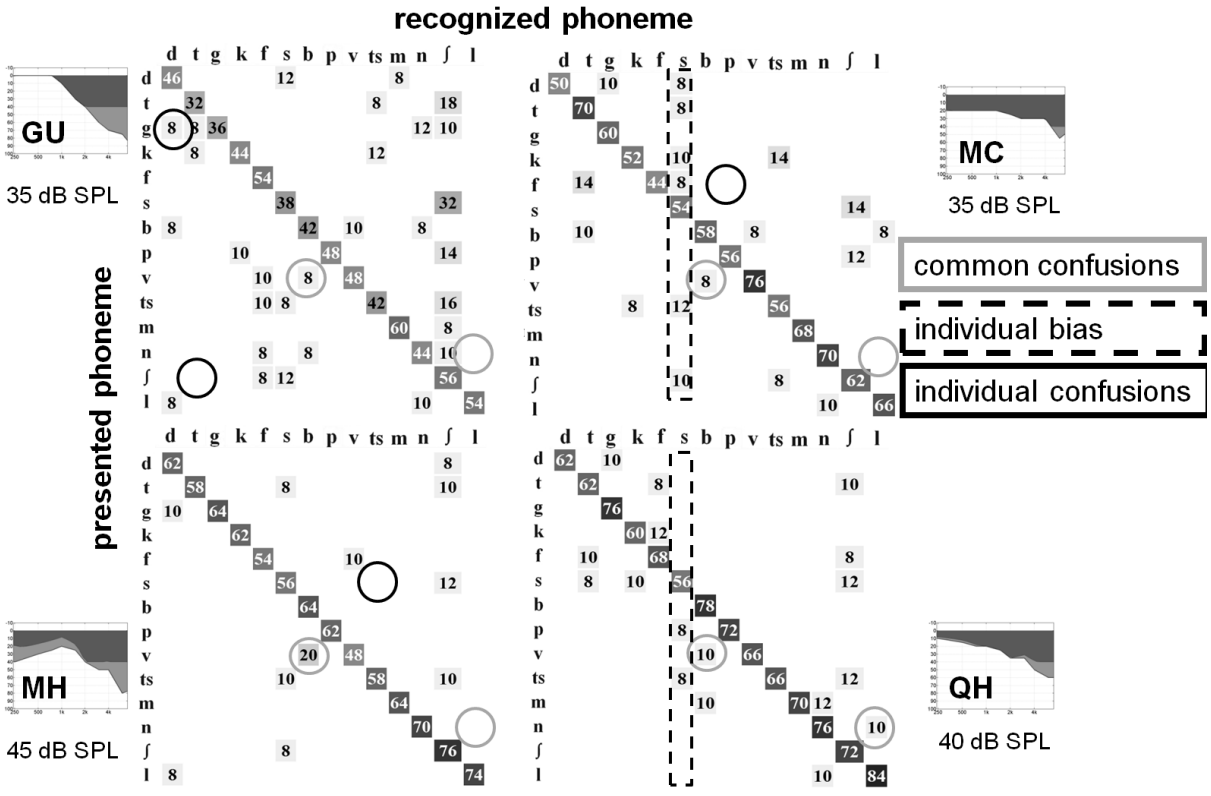

Figure 5.11: Predicted confusion matrices of the four hearing-impaired listeners using model version mC4 (supra-threshold processing was estimated using ACALOS measurements). The display is the same as in Figure 5.10. The highlighted elements are at the same positions as in Figure 5.10 for comparison.

Almost no common confusions /n/→/l/ and /v/→/b/ can be observed (gray circles) except for subject MH. In the predicted confusion matrix of MH, the recognition rate of /v/→/b/ is just above chance level. The individual confusions (black circles) that were observed in the measurements are completely absent and no significant bias towards /s/ is present in the predicted confusion matrices (black dashed rectangle). However, there is a bias towards /ʃ/ in the predicted confusion matrices of subjects GU and MC.

The other model variations specified in Table 5.4 qualitatively show the same results. If these predicted consonant confusion matrices are compared to random confusion matrices that consist of the same average recognition rate (mean of the diagonal elements) respectively, no significant differences can be found both for the diagonal and the non-diagonal elements: The range of recognition rates on the diagonal matches the range expected in a random confusion matrix and the occurrence of non-diagonal elements with recognition rates significantly (> 17%) above chance is as frequent as expected in a random confusion matrix.

5.7 General discussion

The main findings of the two experiments can be summarized as follows. First, as expected from the difference in absolute threshold, NH listeners show lower SRTs of average consonant recognition in quiet than HI listeners. Second, close to the individual SRTs of average consonant recognition, the confusion matrices of NH and HI listeners exhibit only a few common characteristics, mainly in the form of the common confusions /n/→/l/ and /v/→/b/. In addition to these few common characteristics, HI listeners show subject-specific (in the literature termed ‘idiosyncratic’, cf. Lawrence and Byers, 1969) patterns of recognition and confusions. Third, a microscopic model of speech recognition appropriately predicts the SRT of average consonant recognition of NH listeners in quiet and the SRT of two out of four HI listeners, but the predicted psychometric functions are steeper than the observed psychometric functions. Fourth, a variation of the I/O function of the model within physiologically reasonable ranges shows only little impact on the predictions, whereas a mostly linear processing results in slightly higher recognition rates than a compressive processing. Implementing the I/O function that was estimated using categorical loudness scaling does not show an

improvement of the predictions compared to estimating the I/O function using the pure-tone audiogram. In general, different model versions show only little differences in the speech recognition predictions. In the following, these findings and their implications will be discussed in more detail.

5.7.1 Audibility

Based on the results of the present study, higher SRTs of average consonant recognition in quiet for HI listeners compared to NH listeners are expected to result mainly from the individual frequency-dependent audibility of the speech. This is supported by the fairly good prediction of the SRTs of model versions mC1 and mC2, because the only subject-specific parameter in these models is audibility. These two model versions use an external noise, spectrally shaped to the individual audiogram, and they show good predictions of the SRT for NH listeners and two out of four HI listeners. For the other two HI listeners, the predicted SRTs are higher than those of NH listeners, but still lower than the SRTs that were actually observed. The identification of audibility as the main reason for higher SRTs is in line with a couple of other studies. For instance Dubno and Schaefer (1992) found no consistent differences in average consonant recognition of HI listeners and noise-masked normal-hearing (NMNH) listeners in quiet, who listened to a noise spectrally shaped to the individual hearing threshold during speech presentation. Zurek and Delhorne (1987) found an identical trend in the results for consonant recognition in noise and concluded that the primary source of difficulty in listening in noise for HI listeners, aside from the noise itself, is audibility. These studies suggest that masking consonants using an audibility-limiting noise in combination with normal hearing might be a good model for consonant recognition of hearing-impaired listeners (referred to in the following as the NMNH model). The NMNH model is very close to the implementation of hearing impairment used within the SII (ANSI, 1997). The SII sums up the frequency spectra of the speech that are both, above absolute auditory threshold and above the level of an optional background noise using a weighted sum to determine a measure of speech intelligibility. In the literature the NMNH model is often used to predict the performance of HI listeners in various psychoacoustic tasks, which results in a good reproduction of HI listener's performance for gap-detection within tonal signals, gap-duration discrimination, and detection of brief tones in modulated noise, but not for temporal integration (for an overview see Reed *et al.* (2009)). However, some drawbacks can be identified with the NMNH model: First, it does not seem to be very realistic to assume an external noise for the

limitation of audibility, because HI listeners do not report hearing a noise in quiet condition. Second, several studies presented data that cannot be explained purely in terms of audibility. For instance, Plomp (1986) highlighted the importance of supra-threshold processing deficits for the modeling of SRTs of HI listeners. Similarly, the microscopic model of the present study cannot reproduce the SRTs of the HI listeners GU and QH using model versions mC1 and mC2. Using the same argument, Humes *et al.* (1987) found differences between HI listeners and their NMNH pendants in two out of four cases when comparing their ability to recognize nonsense syllables in quiet. Particularly, the types of *errors* made by HI and NMNH listeners differed. Fabry and Van Tasell (1986) found both, noise-masking as well as attenuation, respectively, to be a good model for consonant recognition of the HI ear of only three out of six unilaterally HI listeners. Ching *et al.* (1998) predicted speech intelligibility for HI listeners with very different audiometric thresholds and NH subjects listening to low-pass filtered speech from a nonsense syllable test and found that audibility cannot explain the speech recognition performance of most the HI listeners. A recent study of Li (2009) predicted the impact of audibility on consonant perception of five HI listeners using an 'extended speech banana model', which means that the temporal and spectral audibility of certain consonants was investigated in terms of the pure-tone audiogram. This study found equivocal results. In two out of the five HI listeners a good match between predicted and observed consonant recognition rates was found, in two HI listeners the matching was found to be very poor. Li (2009) hypothesized that in HI listeners with poor model predictions high amounts of distortions in the processing of speech might be present due to dead regions along the BM, i.e. a region on the BM with no remaining outer and inner hair cells. This hypothesis might also hold for subject GU in the present study, because a very steep slope in the audiogram at high frequencies might be an indication for a high-frequency dead region. However, at least as a first approximation, the NMNH model seems to be an adequate model to predict important aspects of the consonant recognition of HI listeners.

5.7.2 Compression

5.7.2.1 Comparison of model versions

The model version mC1 does not discriminate between a NMNH listener and an HI listener, since external noise is used to individually limit the audibility of the model followed by a NH supra-threshold processing. The other model versions (mP1, mC2,

mC3, and mC4) do discriminate between NMNH and HI listeners in terms of a different supra-threshold processing of the signals. It might be assumed that model version mC2 is a double-implementation of hearing-impairment since two factors are included, (1) audibility using an external hearing threshold simulating noise, and (2) an increase of the absolute threshold caused by a reduction of the low-level gain of the I/O function. A closer look reveals that this is not the case, since the second factor does only affect the hearing threshold if a constant minimal amplitude is assumed in the model, i.e. a threshold value, which limits the dynamic range. This constant minimal amplitude is used, e.g., in model version mC4 to adjust the dynamic range according to the individual audibility assessed by the pure-tone audiogram. The model version mC2 does not use such an absolute threshold value; thus, audibility is only determined by the hearing threshold simulating noise. In the same way as the waveform to be processed, the hearing threshold simulating noise is less amplified in the auditory model if the low-level gain is reduced. Model version mC3 seems to be more realistic than mC1 and mC2 because no external noise is needed to limit audibility. The additive internal noise after the DRNL filterbank limits audibility and can be interpreted as a physiological noise, e.g. spontaneous firing rates of neurons or the low-level noise produced by biological processes or the thermal motion of the basilar membrane. Also, mC4 and mP1 seem to be more realistic than mC1 and mC2, since they use an internal noise that was also proposed by Kollmeier (1999). A possible interpretation of such an internal noise at the end of the peripheral processing is that processing errors (represented by the internal noise) limit the performance in the auditory system.

In all model versions of the present study, the assumption is made that it is valid to vary both, audibility in terms of an external noise and compressive properties using the I/O function independently of each other. This assumption is supported by a study of Gregan *et al.* (2010), who showed that external noise does not affect the compressive properties of NH listeners when estimating the I/O function using psychoacoustics.

5.7.2.2 Direct relations between speech recognition and compression

Indirect relations between speech recognition and compression can be found in studies incorporating hearing-aids. However, due to the non-conformity of compressive algorithms to the compressive nonlinearity in the auditory system, relations found in these studies must be interpreted carefully if a comparison to the results of the present study is done. Therefore, only a qualitative discussion between the relations of results of

the present study and results of studies incorporating hearing-aids is undertaken in the appendix of this chapter (Section 5.10.2).

However, also a couple of recent studies investigate a *direct* relation between BM compression and speech recognition. For example Horwitz *et al.* (2007) measured growth-of-masking for nonsense speech and found their results consistent with individual tonal growth-of-masking functions. Their interpretation was that nonlinearities such as variations of compression play a role in the understanding of speech in noise for NH listeners. Rhebergen *et al.* (2010) showed that a compressive I/O function can be implemented within the SII and that the predictive power of this 'extended' speech intelligibility model for HI listeners in quiet is improved. Since Rhebergen *et al.* (2010) adjusted the compressive I/O function individually using audiometric data only and not by incorporating results from other supra-threshold measurements, the amount of the impact on the prediction of using additional supra-threshold information was not investigated. Brown *et al.* (2010) implemented 'efferent' processing in their speech recognition model that consists of an auditory model (that uses the same DRNL filterbank as used in the model versions mC1, mC2, mC3, and mC4 of the present study) and an automatic speech recognition system. Efferent processing was implemented by attenuating the signal in the nonlinear processing path of the DRNL filterbank, which resulted in a more linear I/O function of the entire DRNL filter. Contrary to the manipulation of the DRNL filterbank in the model versions of the present study, their manipulation also affected samples within the compressive high-level portion of the I/O function, because the attenuation was kept constant regardless of the sample amplitude. Using an automatic speech recognition (ASR) system, Brown *et al.* (2010) found that efferent attenuation (i.e. a more linear processing) improves the speech recognition of this ASR system in pink noise. They interpreted the results as a beneficial effect of the efferent system for the recognition of speech in noise. The goal of the study of Brown *et al.* (2010) was not a comparison between the model performance and the performance of human listeners as in the present study. However, their results are in agreement with the results obtained with model versions mC1 and mC2 of the present study, which use an external noise, since slightly higher recognition rates (i.e. lower SRTs) are found for a more linear processing in all HI listeners. A beneficial effect of a more linear supra-threshold processing is also found in model versions mP1 and mC4 when the parameter HL_{OHC} is adjusted. Furthermore, Jepsen (2010) implemented a compressive I/O function into his auditory model and adjusted the I/O function by using parameters extracted from psychoacoustic

masking experiments. He predicted phonetic features of consonant recognition of HI listeners using a rhyme test. Although the predicted results match fairly well the observed data, unfortunately, it is not quite clear how much of the effect can be attributed to audibility only and how much can be attributed to individual supra-threshold processing, since HI listeners showed different audiometric thresholds and the impact of parameters on the predictions was not analyzed systematically.

The beneficial effect of a more linear processing on predicted speech recognition found in the present study might be explained in terms of the SNR in different frequency bands that is accessible to the DTW speech recognizer. At least for the model versions using external noise (mC1 and mC2) it is very likely that the compression in the auditory model changes the SNR in single frequency bands (cf. Hagerman and Olofsson, 2004) in a way that the nonlinearity (1) provides less gain to high speech amplitudes emerging from the noise and (2) provides more gain to the lower amplitudes of the noise. This means that a compression effectively decreases the SNR in the frequency bands and thus decreases the possibility for the speech recognizer to match speech patterns. A more linear processing (as in mC2) does not change the SNR and therefore the possibility for the speech recognizer to match speech patterns is also left unchanged.

5.7.3 Phoneme recognition rates and confusions

5.7.3.1 Results in the framework of the literature

When comparing NH and HI listeners' confusion matrices, some common characteristics and some differences can be found. Common characteristics could be attributed to the same amount of audibility of the cues important for specific consonants at the individual SRT. This hypothesis is supported for instance by a study of Sher and Owens (1974), who showed that subjects with a high-frequency hearing loss listening to normal speech and NH subjects listening to high-frequency-attenuated speech exhibit comparable recognition and confusions of specific consonants. Since it was not the goal of the present study to ensure equal audibility of HI and NH listeners at the SRT, but to investigate undistorted consonant recognition as a function of speech level, it is reasonable to also expect differences between the resulting confusion matrices, because the audibility of the speech cues available to HI and NH listeners at the individual SRT in quiet surely differs. The identification and isolation of cues important for specific consonants is not easy and shows some progress over the last about 30 years for NH

listeners (Dubno and Levitt, 1981; Li *et al.*, 2010) but much less progress for HI listeners (Dubno *et al.*, 1982; Li, 2009). The here-found subject-specificity (idiosyncrasy) of confusion patterns of HI listeners is in line with several studies in quiet (e.g., Lawrence and Byers, 1969) and in noise (Phatak *et al.*, 2009). Furthermore, some parallels can be drawn concerning common patterns of consonant recognition and consonant confusions in the literature. These parallels will be discussed in the following.

A morphing of the reception of one consonant to another consonant was also found by Phatak *et al.* (2008) when assessing consonant recognition of NH listeners in white noise as a function of SNR. For instance, Phatak *et al.* (2008) found a morphing from /f/→/b/ and from /f/→/m/. The type of confusion also depended on the individual utterance used for the assessment of the recognition rates in their study. Unfortunately, no plausible explanation for their finding was given. Since the results presented in the present study were assessed in quiet, the morphing /v/→/b/ in NH and HI listeners appears plausible when the absolute hearing threshold is taken into account: As the speech level is decreased, more speech energy of the voiced consonant /v/ slides below the absolute hearing threshold. As the plosive /b/ is mainly characterized by a stop in the waveform of the VCV, i.e. a short period of silence between the preceding and subsequent vowel, it is very likely that an attenuated /v/ is perceived as /b/ if parts of the consonant speech energy fall below threshold. The confusion /v/→/b/ is also found to be very prominent for NH listeners in quiet by Gelfand *et al.* (1992). In general, confusions between /v/ and /b/ are reported in some studies on consonant recognition both in noise for NH listeners (Woods *et al.*, 2010) and in quiet for HI listeners (Phillips *et al.*, 2009). The confusion /n/→/l/ is very rarely reported in the literature, since /l/ as a lateral approximant consonant is often omitted in nonsense syllable tests as a response alternative. Thus, the question arises whether or not this confusion is only characteristic for the specific speech utterances used (e.g. if this confusion depends on the speaker etc.). Comparable measurements with NH listeners performed with a larger data set were presented in Meyer (2009). The speech data included utterances from four different speakers (two male and two female) without dialect and six different speech articulation styles (labeled ‘normal’, ‘fast’, ‘slow’, ‘soft’, ‘loud’, and ‘question’). This data set was compiled based on the OLLO speech corpus, and also included the speech material used in the present study. As in the present study, Meyer (2009) found the confusion

/n/→/l/ to be the most prominent confusion in his analysis. The confusion /v/→/b/ was also found, but not as prominent as observed in the present study. This might be explained by the fact that the assessment in Meyer (2009) was carried out in noisy conditions with speech levels substantially above absolute hearing threshold. The fact that Meyer (2009) identified the same confusions with a larger amount of utterances indicates that the confusion /n/→/l/ is presumably a typical confusion for German consonants, and not specific to the data set with a single talker employed in this study. Furthermore, the confusion /n/→/l/ was also found for NH listeners in noise by Woods *et al.* (2010). The relatively low recognition rate for the confusion /v/→/b/ in the assessment of Meyer (2009) compared to the recognition rates of /v/→/b/ in the present study might be due to the background noise that affects the low-level phoneme /v/ already at low SNRs. This background noise causes /v/ to be perceived as either /v/ or /b/ without a bias towards /b/, as for NH and HI listeners in silence in the present study.

5.7.3.2 Possible explanations for mismatches between observation and prediction

The microscopic model of phoneme recognition is found to be insufficient for predicting consonant *confusions* for NH listeners in quiet. Confusions could also not be predicted appropriately using the same model for NH listeners in noise (Jürgens and Brand, 2009) and therefore it is plausible that the model moreover is not sufficient for predicting the recognition rates of consonants that show a strong 'competing consonant' or even a morphing to another consonant (as e.g. for /d/, /f/, /s/,/v/,/m/, and /n/) in general. One may hypothesize that the frequency of occurrence of consonants in German language might play a role for this bias towards these phonemes, since the German listeners employed in the present study might be slightly more familiar with consonants they hear more frequently. However, the distribution of phonemes in spoken German language shown in Meier (1967) shows that /n/ is more frequent than /l/, and /v/ is more frequent than /b/. This order is completely reversed to what would be expected in support of this hypothesis. Thus, there is no evidence found for a benefit in recognition abilities in quiet condition if more frequent consonants are presented. However, it may be that some phoneme *combinations* of the present study are more familiar to the participants than other phoneme combinations, which might result in a

certain bias towards the more familiar phoneme combinations. To the author's knowledge, no systematic reference exists to support or reject such a hypothesis.

One may furthermore hypothesize that the sequence of learning consonants in early childhood might result in more familiarity of early-learned consonants than of later-learned consonants. An indicator for a characteristic sequence of learning phonemes can be the sequence of consonants *produced* by small children, since no study of consonant reception tests with German children exists, to the author's knowledge. Hacker (2002) stated that nasals and plosives are earlier produced by children than fricatives, and labial consonants; alveolar consonants are earlier produced than velar and palatal consonants. Since all consonants that are matter of interest here (/v/, /n/, /b/, and /l/) belong to the group of earlier-learned consonants, no evidence is found in support of the hypothesis of a benefit in recognition of early-learned consonants.

However, one reason for the inability of the model to predict confusions might be the implementation of modulation frequency channels. In the present model version, the modulation channels have the same weighting as the frequency channels and are thus warped in time by the Dynamic-Time-Warp speech recognizer as if they were independent of the temporal development of speech. However, it might be that this assumption does not hold for the speech recognition process in human listeners, because modulations might be affected by the effort of the brain to match the incoming internal representations of speech to previously stored speech representation. As the modulation spectrum was found to be a strong predictor for phoneme confusions of NH listeners (Gallun and Souza, 2008), this might be an important approach to improve the model. Another reason for the inability of the model to predict confusions is the deterministic design of the recognizing stage of the model. A frozen-speech approach was used in the present study and in the study of Jürgens and Brand (2009) with the aim of finding the best match of the high performance of phoneme recognition of NH listeners in noise that is by far not reached by common automatic speech recognition systems. However, Ewert and Dau (2004) showed that the auditory model PeMo (the predecessor to the models used in the present study) cannot predict amplitude modulation detection of deterministic signals when using an internal Gaussian noise. This presumably also holds for both, PeMo for HI listeners and CASP. Another approach to model the detector stage of the model more realistically is needed to account for this problem. Aspects that should be incorporated in such a model are speech production, speech articulation and a

verbal working memory, since each one of these factors plays an important role in human speech recognition (Hickok and Poeppel, 2007).

Nevertheless, the auditory models used in the present study are developed to predict psychoacoustic experimental results of NH listeners when a noise is present (or preceding or succeeding), so it is straightforward to expect best predictions also for phoneme recognition of NH listeners in noise. The absence of a range of speech levels in the predicted results, in which some of the consonants are morphed to other consonants, gives rise to the hypothesis that if the model 'hears' something, i.e. gets acoustic cues above the absolute hearing threshold, it instantly recognizes it correctly, probably due to the frozen speech approach. On the other hand, if human listeners perceive the same cues they may attribute the cue to an incorrect response alternative. To support this hypothesis the following additional analysis is done. In the observed data of NH listeners in quiet, consonant confusions that show recognition rates significantly above chance level (>17%) are regarded as 'correctly recognized' additionally to the recognition rate of the diagonal element of the confusion matrix. This leads to new 'diagonal' elements of the confusion matrices (i.e. to the percentages counted as 'correctly recognized') shown by the light gray bars in Figure 5.12.

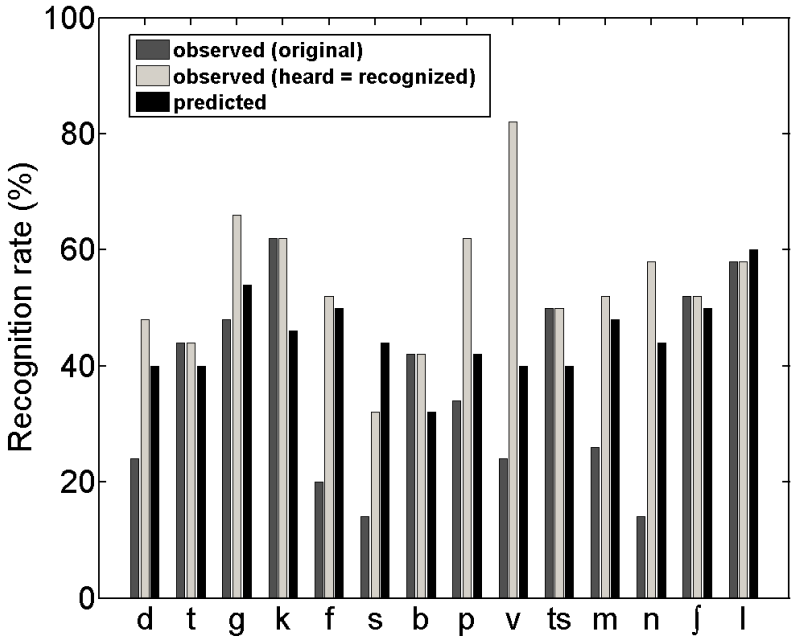

Figure 5.12: Recognition rates of single consonants of normal-hearing listeners at 15 dB SPL speech level, dark gray: observed, i.e. diagonal elements of Figure 5.6 panel 1, light gray: observed if each significant confusion is counted as 'correctly recognized, black: predicted, i.e. diagonal elements of Figure 5.6 panel 2.

For a comparison, also the diagonals of the two panels of Figure 5.6 are plotted as dark gray (observed) and black (predicted) bar plots. Using this 'heard = recognized' assumption, only the 'observed' recognition rates of /k/, /p/, and /v/ differ significantly from their predicted pendants, rather than the recognition rates of seven out of the 14 consonants when comparing the original observed recognition rates to their predicted pendants. Thus, a much better match is obtained. However, the correlation coefficient between observed ('heard = recognized') and predicted recognition rates (r^2 = 0.06, p = 0.42) is not substantially increased over the correlation coefficient between observed (original) and predicted recognition rates. This may be simply due to the very small range of predicted recognition rates (it ranges only from 40% to 60%) of the consonants at the SRT. Computing a new psychometric function using this hypothesis, an SRT of 16.9 dB SPL and a slope of 3.3 %/dB is obtained. Even if the correlation between model and observation is not substantially improved and also the psychometric function is not perfectly predicted, the prediction is qualitatively better concerning the SRT and the recognition rates of single phonemes than quantified in Section 5.6.1.2. This indicates that parts of the non-perfect-matching of the model might be explained using the assumption that the model equals 'hearing' with 'recognizing', whereas NH listeners do not.

5.7.4 Sensorineural hearing impairment

Some important characteristics of sensorineural hearing impairment such as elevated absolute hearing thresholds and a changed compressive supra-threshold processing are included in the versions of the microscopic model used in the present study. Especially model versions using CASP might account for some important characteristics in consonant recognition of HI listeners, since a realistic I/O function of the BM is included. This is one aspect of phoneme perception deemed to be important by some authors, e.g., Allen *et al.* (2009), who hypothesized that the onsets of, e.g., stop consonants are enhanced by normal OHC function and are degraded in HI listeners, who show an impaired OHC function. However, there are a bundle of factors that are not regarded in the microscopic model, but which may play a role in consonant recognition of NH and especially of HI listeners. Noise and even a different (compressive) suprathreshold processing of the stimuli might only be rough approximations to what is supposed to play a role in the multitude of forms of (sensorineural) hearing impairment. Factors of these diverse forms will be listed in the following.

First, presbyacusis (that is very likely to be present in most of the HI listeners of the present study) is assumed to be attributed to the degeneration of cells in the stria vascularis, which affects the voltage important for the triggering of action potentials in the inner hair cells (Gates and Mills, 2005). Unfortunately, it is not clear if the impacts of this physiological finding on the effective signal processing in the auditory models used in the present study are regarded appropriately. Second, neither the NH nor the HI auditory model used in the present study regards phase-locking and temporal fine-structure (TFS) of the speech signals. However, Lorenzi *et al.* (2006) and Hopkins and Moore (2010) showed that NH listeners can use TFS, whereas HI listeners often cannot benefit from TFS in speech recognition tasks. Furthermore, TFS was particularly found useful for consonant identification (Sheft *et al.*, 2008). Third, it is not clear how the degeneration of nervous connections in the auditory pathway, which are no longer used (e.g., non-fed spiral ganglion cells due to destroyed inner hair cells), accounts for a change in the processing of speech. The knowledge about the impact of this factor might be of importance to correctly predict speech recognition performance after rehabilitation using hearing-aids or a cochlear implant, because it was shown that the time of deafness preceding the usage of a cochlear implant is inversely related to the performance of the listener to understand speech in quiet after implantation (Gomaa *et al.*, 2003). Fourth, the auditory models used in the present study are bottom-up models, i.e., the processing of the signal of succeeding blocks does not influence the processing of the preceding blocks. However, Davis and Johnsrude (2007) showed that top-down interactive mechanisms within auditory networks play an important role in explaining the perception of spoken language. Sixth, there are other cognitive and auditory factors that contribute for the recognition of speech, like attention, cognitive skills, and factors related to age (for a review cf. Houtgast and Festen, 2008). Further research is needed to investigate the role of each one of these factors to obtain knowledge about how to incorporate them in computational models with the aim to get a better understanding of human speech recognition, especially speech recognition of hearing-impaired listeners.

A comparison of the representation of phonemes in the cortex of animals is presented by Mesgarani *et al.* (2008), who show that invasive techniques in animals might be useful to better understand human speech recognition. However, this approach may suffer from many shortcomings, because simply animals do not understand speech, i.e. they have not learned speech their whole life. Hence, it is very likely that their neuronal connections are not formed to extract speech cues optimally, like humans presumably can. Furthermore, the dysfunction of parts of the animal auditory system

might show different impacts on their ‘internal representations’, which do not correspond to the impacts present in human listeners. Therefore, it may be more appropriate to further improve existing models of the human auditory periphery and of human speech recognition and to include recent findings about the functioning and the dysfunction of physiological parts of the auditory pathway. The present study is one step towards the long-term aim of a more sophisticated model of the speech processing of NH and HI listeners.

5.8 Conclusions

Consonant recognition in quiet was assessed in ten normal-hearing (NH) and four hearing-impaired (HI) listeners as a function of speech level. A microscopic model of speech recognition was used to predict the consonant recognition results using different model versions for auditory processing and different versions of supra-threshold processing of the speech signals. The results of this study can be summarized as follows.

(1) Poorer consonant recognition than in NH listeners is observed in HI listeners, quantified by a higher speech reception threshold (SRT). Close to the individual SRT in quiet, a simple attenuation of speech level does in some cases not only reduce the recognition of a consonant, but it can also perceptually morph this consonant into another consonant. This morphing might be a problem for HI listeners at levels close to conversational speech levels.

(2) Consonant confusion matrices, inferred close to the individual SRT, reveal some common characteristics between NH and HI listeners in the form of common confusions. HI listener’s confusion matrices show a large amount of subject-specificity.

(3) The microscopic model of speech recognition can appropriately predict the SRT of the NH and of two out of the four HI listeners. For the other two HI listeners lower SRTs (i.e. better speech recognition performance) are predicted than are observed. Predicted psychometric functions are always steeper than observed. This effect is more pronounced in the HI listeners. The model can account for the recognition rates of those consonants of NH listeners in quiet, which do not show a strong ‘concurring’ consonant in the observations, i.e. which are not confused by a specific other consonant close to the SRT.

(4) The model cannot account for confusions or the morphing of consonants to other consonants. Parts of this non-matching are plausible if it is assumed that the model equals ‘detecting a speech cue’ with ‘recognizing the correct speech item’. This assumption is not valid for the recognition of some consonants by human listeners.

(5) An altered supra-threshold processing in the models considered here shows only little impact on predicted consonant recognition if the same audibility of the signal is assumed. A more linear supra-threshold processing shows higher recognition rates than a more compressive supra-threshold processing. Implementing the I/O function estimated by the supra-threshold measurement technique adaptive categorical loudness scaling, does not result in a significant improvement of the prediction of individual consonant recognition. Furthermore, only small differences are observed when comparing the performance of the different model versions.

At large, the model (besides matching general SRTs for NH and two out of four HI listeners) fails to predict the specificities of confusions observed in the empirical data. This may partially be due to non-acoustic factors not yet included in the model. Such factors might be a familiarity with certain response alternatives (combinations of vowels and consonants) and the proximity of the presented speech item to meaningful speech fragments as well as attention or cognitive skills. It will be a challenge to incorporate such factors in new numerical models of human speech recognition.

5.9 Acknowledgements

This work was supported by SFB TRR 31 “The active auditory system”. The author would like to thank Eugen Rasumow, Michel Hahn, and Sven Kissner for the execution of the measurements. Many thanks to Morten Jepsen for making available the CASP model and special thanks to Birger Kollmeier, Thomas Brand, Stefan Fredelake, and Bernd Meyer for comments on an earlier version of the manuscript.

5.10 Appendix

5.10.1 Vowel recognition of normal-hearing listeners

In Experiment I, vowel recognition was assessed in NH listeners at speech levels of 5, 10, 15, 20, and 25 dB SPL, additionally to consonant recognition. Since the present study focuses on consonant recognition, the vowel recognition results are presented only briefly in this section for the sake of completeness.

5.10.1.1 Average vowel recognition rates

Figure 5.13 displays observed vowel recognition rates averaged over ten NH listeners (gray crosses) and predicted average vowel recognition rates using model mC1 (cf. Section 5.5.2). Fitting a psychometric function according to Eq. (2.8) to observed vowel recognition rates results in an SRT of 8.9 dB SPL and a slope of 6.5 %/dB. Fitting a psychometric function to predicted vowel recognition rates results in an SRT of 14.8 dB SPL and 9.1 %/dB. Comparing observed and predicted vowel recognition rates, the model mC1 underestimates the average SRT by about 6 dB and shows a slightly steeper psychometric function.

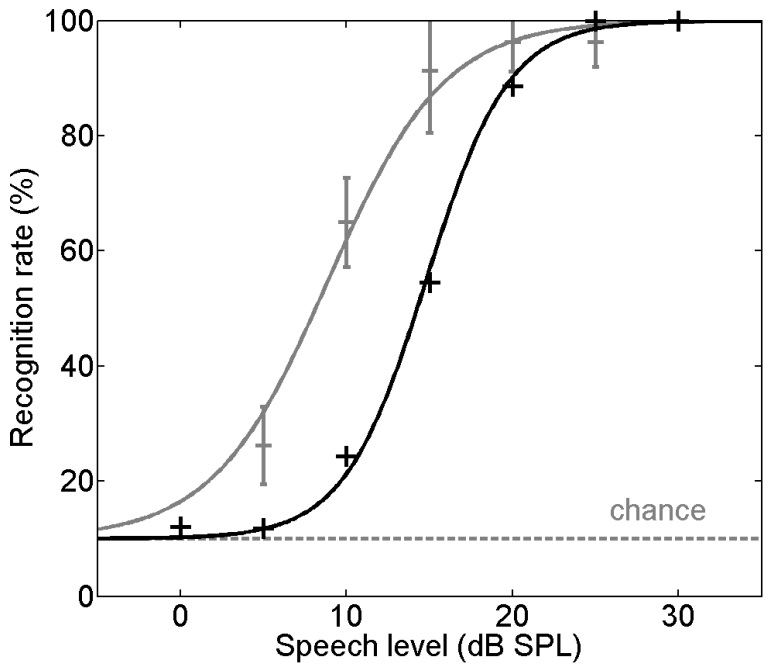

Figure 5.13: Psychometric functions of normal-hearing listeners (gray symbols) and model mC1 (black symbols) for average vowel recognition in quiet. Error bars denote the inter-individual standard deviation of the observed consonant recognition rates.

Compared to average consonant recognition (cf. Figure 5.5), vowels show higher recognition rates at the same speech level in both, NH listeners and model mC1. This effect is much more pronounced for the observed data (difference in SRT between average consonant and vowel recognition is 9 dB) than for the predicted data (difference is 1 dB).

5.10.1.2 Confusion matrices

Figure 5.14 shows confusion matrices of vowel recognition close to the SRT of NH listeners (10 dB SPL, panel 1) and model mC1 (15 dB SPL, panel 2). The display is the same as in Figure 5.6. Confusion matrices at different speech levels for both, observation and prediction are displayed, because predicted and observed SRTs differ by 6 dB. These confusion matrices show some common characteristics. For example, an element-wise comparison of the same confusion matrix elements between the predicted and observed vowels /a/, /ʊ/, /a:/, /e/, and /o/ shows that the recognition rates of these vowels (i.e. the diagonal elements) do not differ significantly from each other. A few common confusions can be observed, e.g., /a/→/a:/ and /a:/→/a/.

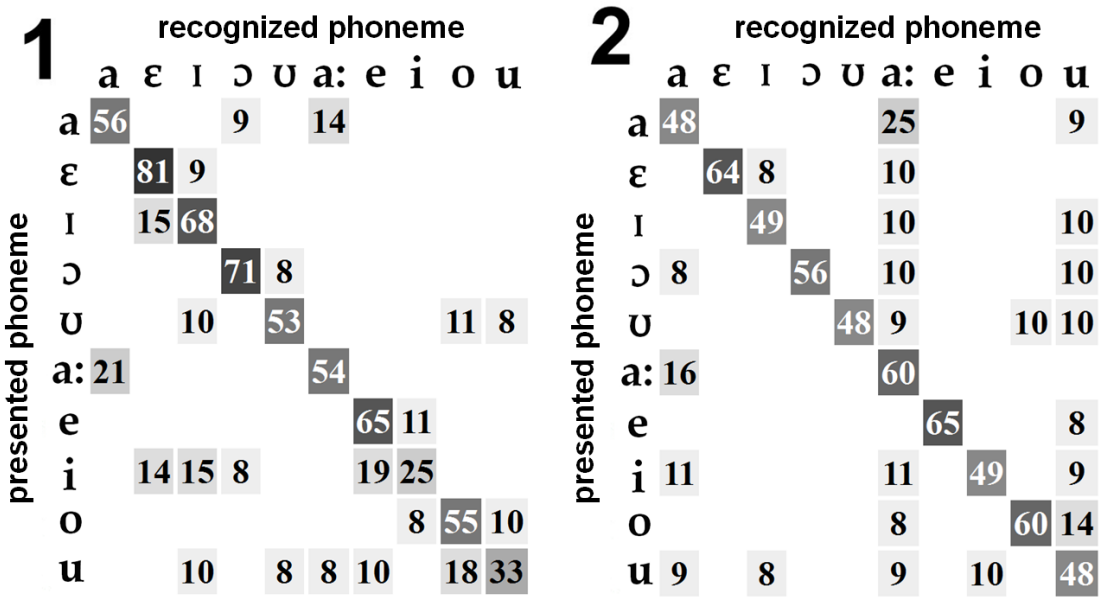

Figure 5.14: Vowel confusion matrices averaged for ten normal-hearing listeners at 10 dB SPL, (panel 1) and at 15 dB SPL for model mC1 (panel 2). The display is the same as in Figure 5.6.

However, there are also differences. For instance, the 'competitor' /e/ is clearly prominent in panel 1 if /i/ is presented, but this competitor is completely absent in

panel 2. Of course, this also affects the recognition rate of the correct response /i/, which is also not correctly predicted. To shortly summarize the results presented in this section, the statements made here fit to the statements made when investigating consonant recognition. The model is able to appropriately predict recognition rates of some consonants. If a competitor phoneme is present in the observed data the model fails to appropriately predict the recognition rate of the correct response. The only exceptions are the vowels /a/ and /a:/; for these vowels both, the confusion and the correct response could be modeled appropriately if confusion matrices close to the SRT are compared.

5.10.2 Relations between speech recognition and compression in hearing-aid studies

To the author's knowledge, the present study is the first study that systematically investigates the impact of adjusting the I/O function in a microscopic model on the prediction of speech recognition in quiet. Therefore, it is hard to discuss the present results concerning the impact of this adjustment on the prediction of consonant recognition in the context of the literature. However, many studies have investigated the influence of compressive, linear, or expansive algorithms in hearing-aid processing on speech intelligibility of HI listeners. Because of multiple parameters such as attack and release time constants of the compressor and technical restrictions such as a frequency-dependent maximal possible gain in hearing-aids, it is hard to make detailed comparisons between the following studies and the present study. However, some general tendencies can be observed that will be listed below.

The result that a linear processing results in higher recognition rates than a compressive processing is, e.g., in line with recent results that mutual information (a measure for speech intelligibility) is smaller when using a fast compression than when using a linear algorithm in a hearing aid (Leijon and Stadler, 2008). Furthermore, Meyer *et al.* (2009) found that HI listeners benefit in their speech recognition performance in noise more by a linear than by a compressive algorithm in a hearing aid.

The hypothesis of a decrease of SNR and thus decreased speech intelligibility as a consequence of compression (stated in Section 5.7.2.2) is also in line with a couple of other hearing-aid studies. For instance, Walker *et al.* (1984) and van Buuren *et al.* (1999) showed that both compression and expansion of a noisy speech signal fail to show better speech intelligibility than linear amplification. Rosengard *et al.* (2005b) found that mild-to-moderate simulated HI listeners do not benefit in speech

intelligibility from slow-acting compression neither in noise nor in quiet. It is important to also regard the influence of compressive time constants on the speech intelligibility results. In a fast compressive algorithm (i.e. using small attack and release time constants), speech intelligibility is more detrimentally affected, especially at low signal-to-noise ratios (Hohmann and Kollmeier, 1995; Hansen, 2002). The results of Goedegebure (2005) somehow contradict the results of the present study, because he found improvements for speech intelligibility in quiet using a relatively fast-acting compressive algorithm compared to a linear algorithm. However, also this study showed a detrimental effect of the compressive algorithm on speech intelligibility in noisy conditions. Even if in all of these hearing-aid studies different results were obtained concerning the role of compression for speech intelligibility in quiet, they all show either detrimental or at best no effects for speech intelligibility in noise. Since in all model versions of the present study, noise was used to limit either audibility (mC1, mC2, mC3) or the speech recognition performance of the model (mP1 and mC4), which was necessary because a frozen-speech approach was chosen for the model, the result of a detrimental effect of compression on the speech recognition rates is in line with these studies.

6 Summary and concluding remarks

In this dissertation, a microscopic model has been proposed to predict phoneme and sentence recognition of normal-hearing and hearing-impaired listeners in both noisy and in quiet conditions. Contrary to standard models of human speech recognition that use the long-term spectrum of speech and noise, as for instance the Speech Intelligibility Index (SII), this approach mimics the temporal processing of speech waveforms in auditory filter bands by using an auditory model. The most important results of the research presented in this dissertation are:

- The microscopic model accurately predicts average and single phoneme recognition of normal-hearing listeners in noise, using a 'frozen-speech' approach, i.e. identical speech waveforms both for the token to be recognized and the token included within the speech memory as a reference pattern (Chapter 2). Furthermore, the microscopic model accurately predicts average consonant recognition of normal-hearing listeners, and of two out of four hearing-impaired listeners in quiet conditions (Chapter 5).

- The performance gap between human and automatic speech recognition was quantified to amount to 12-13 dB using different model configurations of the microscopic speech recognition model. Different perceptual distance measures affect the predicted speech recognition rates using the frozen-speech approach (Chapter 2 and 7).

- The model is not capable of correctly predicting consonant *confusions* both in noise and in quiet conditions (Chapter 2 and 5). This mismatch between human and modeled speech recognition might partly be explained by the observation that the model predictions of confusions depend less on the speech level or SNR than the recognition scores of human listeners: If the model detects a cue characteristic for one specific phoneme, it performs an "optimum pattern match" to recognize the correct consonant. For human listeners, however, a cue that is characteristic for one phoneme at low speech levels or SNRs might be interpreted as belonging to a *different* consonant, which leads to a confusion of one consonant with another one. Therefore, the model is also not capable of modeling the observed

‘morphing’ of consonants: In the speech recognition experiments with human listeners some consonants are found to be ‘morphed’ to other consonants as a function of the absolute speech level (Chapter 5).

- The microscopic model predicts speech reception thresholds (SRTs) in noisy condition with about the same accuracy as the SII, concerning the variability introduced by different audiometric thresholds using external, hearing threshold simulating noise, (Chapter 3).

- Parameters characterizing supra-threshold processing could be consistently inferred from psychoacoustic measurements in a mixed group of normal-hearing and hearing-impaired listeners. Two measurement methods were used: temporal masking curves (TMCs) and adaptive categorical loudness scaling (ACALOS). ACALOS was found to be much faster than TMC and almost as reliable. It should further be used as a standard tool in clinical applications to assess the individual auditory processing beyond the audiogram (Chapter 4).

- Audibility appears to be the dominant factor in consonant recognition of hearing-impaired listeners for modeling speech recognition in quiet. Improvements of the predictions by implementing supra-threshold processing deficits to model hearing impairment are limited. No improvement of the predictions was found by implementing supra-threshold parameters estimated individually using ACALOS (Chapter 5).

The microscopic model has been shown to account for some important details observed in human speech recognition. In noisy conditions, normal-hearing listeners show only very few confusions in the form of “competing phonemes”, i.e. phonemes that are very often confused with the presented phoneme in a systematic, not simple random way. Since the model is not capable of modeling the effect of such competing phonemes, the best match between model and measurements is found in conditions with very few confusions, i.e. for normal-hearing listeners in noise, where very few competing phonemes are empirically found. If competing phonemes occur in the measurements, as shown for instance in normal-hearing listeners or hearing-impaired listeners in quiet conditions, some mismatches between model and measurement can be observed. It will be a challenge to accurately model the recognition of competing phonemes using more sophisticated models of human speech recognition. A very recent step towards the long-term aim of developing such a sophisticated model of human speech recognition was

done by Meyer and Kollmeier (2010) by training an automatic speech recognition (ASR) system using human confusion data. Meyer and Kollmeier (2010) could show that the match between ASR data and human data is improved if the ASR training is based on the confusions made by human listeners. The corresponding stage to be manipulated in the microscopic model used here, in order to perform a similar match to human performance, would be the speech memory, i.e. the recognizing stage of the model.

Such a revised microscopic model of speech recognition should also account for the observed 'morphing' of one consonant to another, i.e. the dependence of the phoneme perception on the absolute speech level. This morphing might be one key problem in understanding the mechanisms of the speech recognition problem that hearing-impaired listeners suffer from. In normal-hearing or mild hearing-impaired listeners this morphing may be restricted to low speech levels, but in moderate to severe hearing-impaired listeners this morphing might occur even at conversational speech levels and might not vanish if speech is amplified in level (by, e.g., using hearing aids). In turn, such a morphing presumably results in confusions of words with other words, which rely on morphed key phonemes. The confusions /n/→/l/ and /v/→/b/ (which were found to be very prominent in Chapter 5) might induce the confusions such as 'mind'→'mild', 'not'→'lot', 'when'→'well', and 'Kevin'→'cabin' or the German 'Wald'→'bald'. If these confusions are adequately modeled by a more sophisticated model extended to the model described here, a very useful application will become feasible: A hearing aid processing strategy could be incorporated prior to the processing of the model, and hearing aid parameters could be adjusted in such a way that the recognition rates of poorly recognized phonemes are particularly increased. Such a procedure would offer the possibility of testing and optimizing various hearing-aid strategies and parameters in order to get the best speech recognition performance, without expensive and time-consuming hearing tests with human listeners.

Therefore, the microscopic model presented in this dissertation should be seen as one first step towards more comprehensive speech recognition models and their applications in the future. Refined models of the auditory periphery or a more sophisticated speech recognition stage might reveal a better match with measurement results. The frozen-speech approach used for the current model version might also be revised for each new model version to be considered. An advantage of this frozen-speech approach is that only a little amount of speech material is required for the

training of the speech recognition stage. Hence, this modeling concept can easily be used for speech tests with only a small amount of recorded training materials.

Some interesting questions remain that could not be answered in the present dissertation, such as: How large are the respective contributions of the peripheral preprocessing stages to the modeled speech recognition results? How large is the contribution of the recognizer back-end stage? Which one of these two is the key stage for a more appropriate modeling of HSR? To date, some comments can be given in the following regarding these questions and what would be a way to answer them.

Auditory models that are closer to the physiology, e.g., Heinz *et al.* (2001) or Meddis and O'Mard (2006), might reveal more direct connections between the dysfunction assumed at a specific stage of the model and speech intelligibility. However, it is unclear if such models that aim at describing the observed *physiological* findings are equally well suited to describe the *functional* deficits in sensory processing as the functional models used here. Moreover, since speech recognition is a very complex process in humans that is influenced by a large number of parameters, it seems advisable to limit the modeling to a small number of successive functional units of "effective" signal processing and a few number of altered processing parameters. This view would argue for a model of human speech recognition on the peripheral side that consists of a cascade of (more or less dysfunctional) processing stages with a limited number of processing parameters with only a limited explanatory value for physiological observations. With respect to the subsequent recognition back-end stage, one could argue that a more realistic modeling of the recognition stage should incorporate some of the strategies also used in automatic speech recognition, for example Hidden Markov Models (HMMs). Such an argument might be correct in the sense that these strategies form more generalized models of the human speech memory. However, such models are the result of an optimizing process that aims at producing best automatic speech recognition performance without necessarily reflecting the details of human speech recognition (Meyer, 2009). Also, the training and data representation of modern ASR techniques deviate substantially from the assumed functioning of the human auditory processing indicating that the potential of modeling HSR with ASR techniques is limited.

Another interesting (but to date rather notional) approach for a more realistic modeling of the recognition stage can be found in very recent results from olfactory sensory research: Wiechert *et al.* (2010) used recurrent neural networks to account for

pattern recognition of different odors. They hypothesize that the mechanisms inherent in these recurrent neural networks are probably relevant for pattern processing in various brain areas and thus presumably also for the pattern recognition of speech. Recurrent neural networks are deemed to be physiologically plausible, are very likely to be present in many vertebrates and the study of Wiechert *et al.* (2010) showed that an efficient decorrelation of sensory input using sparse recurrent neural networks is possible. Interestingly, similar recurrent neural network architectures have been used in (robust) automatic speech recognition systems in the past (Tchorz *et al.*, 1997) – but with only a limited success.

At the very end of this dissertation, some possible applications and links to perform further research on the microscopic model presented in this thesis are listed in the following.

The microscopic model of sentence recognition presented in Chapter 3 might be used to model the Speech Reception Thresholds (SRTs) of sentence tests in different languages. Thus, differences in the observed SRTs from one language to another language (using native or non-native listeners) might be explained by the model using the same auditory front-end but different back-ends that are trained in a different way for describing native or non-native listeners, respectively. The same model could be extended using models for context effects in speech recognition (cf. Bronkhorst *et al.*, 2002), and it may be possible to model the benefit from the context of speech material in speech intelligibility.

It is reasonable to reduce the computational load of the microscopic model, at least for modeling sentence recognition. In the present version of the model presented in Chapter 3, a psychometric function of speech recognition is computed by using the complete speech material (all sentences presented to the listener during the measurement session) at each SNR used in the measurements. This was done in order to calculate the SRT and the slope of the psychometric function as precisely as possible. However, it is more reasonable to let the model act like a human listener during the measurements, i.e. to use the same adaptive testing algorithm (as in the measurements) for the assessment of the SRT. This would also lead to a drastic reduction in computational load.

To date, the microscopic model has inspired a study of modeling speech recognition of cochlear implant users (Fredelake *et al.*, 2010). In the study of Fredelake

et al. (2010) explicit advantage is taken of the fact that the model uses the *temporal processing* of the speech waveforms, because approaches using the spectra of speech and noise such as the SII cannot be applied if a cochlear implant is incorporated. Furthermore, this model might be used to model human speech recognition in fluctuating background noises, since it is very likely that the 'resolution' of single spectro-temporal speech patterns (that are not masked energetically by the fluctuating background noise) plays a major role in such an acoustic environment. A recent study using the SII (Meyer *et al.*, 2009) has shown that a consideration of the temporal fluctuations in both speech and noise results in an improvement of the prediction of individual speech intelligibility.

By and large, this dissertation has presented a microscopic model of human speech recognition, (1) whose particular stages resemble the stages assumed to be relevant in human speech recognition, and (2) which is capable of modeling some important aspects of phoneme and sentence recognition in normal-hearing and hearing-impaired listeners. Moreover, the assessment of supra-threshold processing deficits and the investigation of how these deficits affect speech recognition using this microscopic model have shown new insights into the auditory processing performed by normal-hearing and hearing-impaired listeners and can be taken as starting point for further research.

7 Appendix: Modeling the human-machine gap in speech reception: microscopic speech intelligibility prediction for normal-hearing subjects with an auditory model[1]

Abstract

In this study speech intelligibility in noise for normal-hearing subjects is predicted by a model that consists of an auditory preprocessing and a speech recognizer. Using a highly systematic speech corpus of phoneme combinations (logatomes) allows for the analysis of response rates and confusions of single phonemes. The predicted data is validated by listening tests using the same nonsense speech material. If testing utterances that are not identical to those in training material are used, the psychometric function in noise is predicted with an offset of 13 dB to higher signal-to-noise ratios (SNR). This is consistent with the man-machine performance gap between human speech recognition (HSR) and automatic speech recognition (ASR) (Meyer *et al.*, 2007a). However, this offset reduces to 4 dB in a second model design with identical recordings for training and testing. Furthermore, predicted confusion matrices are compared to those of normal-hearing subjects with the second model design.

[1] This paper was published as Jürgens et al. (2007) and was presented at the 8th annual conferences of the International Speech Communication Association (Interspeech, Antwerp, Belgium). This paper is a predecessor to Chapter 2 of this dissertation.

7.1 Introduction

Typical models that predict speech intelligibility in noise for normal-hearing subjects, as, e.g., the Speech Intelligibility Index (SII) (ANSI, 1997), analyze the long-term spectra of speech and noise separately in different frequency channels. The outcome of these models can be transformed to the speech reception threshold (SRT), which gives the SNR of 50% speech intelligibility and the slope of the psychometric function. Recognition rates and confusions of phonemes cannot be studied using these models.

The model proposed here is based on an idea of Holube and Kollmeier (1996) and consists of a psychoacoustically motivated preprocessing of the time-signal and a standard dynamic-time-warp (DTW) speech recognizer (Sakoe and Chiba, 1978). By determining the distances between a test utterance and training utterances "on a perceptual scale" the utterance with the least distance is taken as the recognized one.

For prediction and validation we used the context-free speech database Oldenburg Logatome Corpus (OLLO) (Wesker *et al.*, 2005). It contains 70 different vowel-consonant-vowel (VCV) and 80 CVC logatomes composed of German phonemes. Each logatome was recorded 18 times by each speaker. 6 different speech articulation styles are included: "slow", "normal", "fast", "loud", "quiet" and "questioning". The use of this corpus allows systematical investigations of phoneme recognition rates and confusions. At the same time it avoids that human listeners can use any semantic knowledge for intelligibility.

7.2 Measurements

7.2.1 Method

10 clinically normal-hearing subjects (7 male, 3 female) aged between 19 and 37 years were employed. The intelligibility of 150 logatomes was measured in a sound isolated booth at different signal-to-noise-ratios. All recordings were taken from the OLLO database and were spoken by a single German speaker with speech variability "normal". The 150 recordings were randomly split into two lists of the same length for each of the 5 SNRs and the resulting 10 lists were randomly interleaved for presentation. The speech was presented at a level of 60 dB SPL via Sennheiser HDA 200 headphones that were free-field equalized using a FIR-filter with 801 coefficients. A non-modulated running noise with speech-like frequency spectrum was used (ICRA-1 noise, Dreschler *et al.*, 2001). All audio signals were presented diotically. Response alternatives for a

single logatome had the same preceding and subsequent phoneme (closed test); hence, the subject had to choose from 10 or 14 alternatives when a CVC or a VCV was presented, which one was recognized.

7.2.2 Results

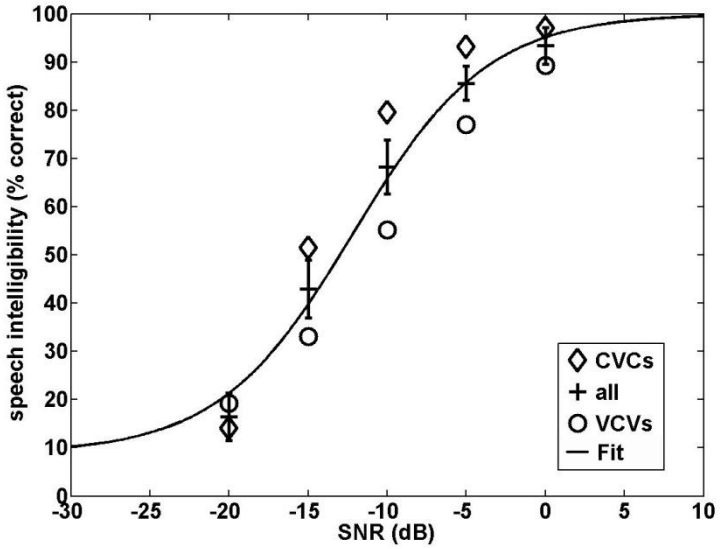

Figure 7.1: Psychometric function for normal-hearing subjects measured with logatomes in ICRA-1 noise at 5 fixed SNR respectively. Error bars show the interindividual standard deviation for 10 subjects. The fitted function is shown for comparison.

Figure 7.1 shows the results of the speech intelligibility test plotted versus the SNR. Every symbol represents the mean intelligibility of CVCs, VCVs or all logatomes for 10 subjects. The error bars show the inter-individual standard deviations. The model function given in Eq. (7.1) was fitted to the data by varying the free parameters SRT (SNR at 55% intelligibility) and s (slope of the psychometric function at the SRT).

$$\Psi(x) = \frac{1-g}{1+\exp(4 \cdot s \cdot (SRT - x))} + g \qquad (7.1)$$

Here: x: SNR, g: guessing probability (g = 8.9 %) and Ψ: intelligibility. The fit was performed by maximizing the likelihood under the assumption that the recognition of each logatome is a Bernoulli trial (cf. Brand and Kollmeier, 2002). This yielded a slope of (5.4 ± 0.6) %/dB and a SRT of (-12.2 ± 1.1) dB. Note that CVCs have always a higher intelligibility than VCVs except for -20 dB SNR.

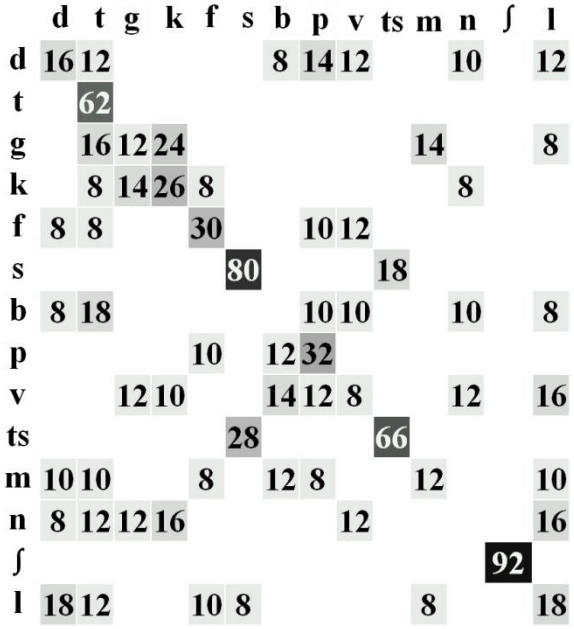

Figure 7.2: Confusion matrix (response rates in %) for normal-hearing subjects at -15 dB SNR, measured with consonants embedded in logatomes. Row: presented phoneme, column: recognized phoneme. Gray scales denote different grades of response rates. Response rates below 8% are not shown.

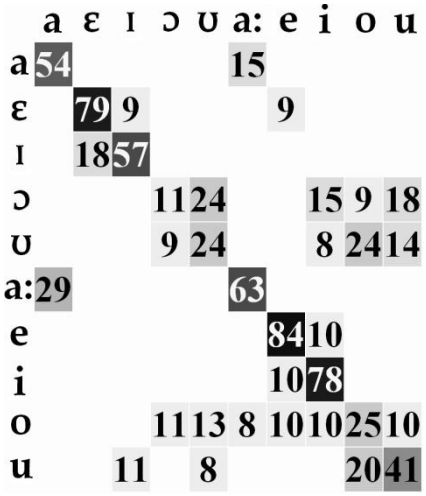

Figure 7.3: Confusion Matrix for normal-hearing subjects at -15 dB SNR, measured with vowels embedded in logatomes. The display is the same as in Figure 7.2.

Figure 7.2 and Figure 7.3 show the confusion matrices of consonants and vowels for all 10 subjects. Due to the design of OLLO each middle consonant was presented 5 times and every vowel 8 times at a given SNR to each subject. Hence, the overall number of presentations of each phoneme for these matrices are 50 and 80 respectively. The SNR was chosen to -15 dB, which corresponds to an intelligibility of 33% (VCV) and 52% (CVC). Each row symbolizes the presented phoneme and each

column the recognized one. Correct recognized phonemes are shown as diagonal elements of the matrices. Due to clarity all entries below 8% were left blank.

Corresponding to Figure 7.2, fricative consonants like /ʃ/, /s/, and /ts/ are recognized best, whereas voiced consonants like /n/, /v/, and /b/ are recognized worst or not at all. Note the big variance between the diagonal elements of /n/ and /ʃ/. Unvoiced plosive consonants like /p/, /t/, and /k/ are recognized at significantly higher recognition rates than voiced ones (/b/, /d/, /g/). There are almost no confusions between consonants with very high frequency content as /ʃ/, /ts/ and those with low one. However, there does not seem to be a systematic pattern of confusions.

There is some kind of clustering in the vowel confusion matrix (Figure 7.3): /ɔ/, /ʊ/, /o/ and /u/ are recognized worst and there are many confusions between them. The next cluster is /a/, /a:/ with no significant confusions with other vowels. The vowels best recognized are /ɛ/, /ɪ/, /e/ and /i/.

7.3 The perception model

7.3.1 Specification

The perception model applied in this study was initially developed by Dau *et al.* (1996a) and it was further on used to model many different psychoacoustical experiments with different masking conditions as well as modulation detection tasks in an extended version (Dau *et al.*, 1997). In this study this extended version is combined with a standard DTW speech recognizer to mimic the decision process in a closed speech intelligibility test.

Figure 7.4 shows the model structure. The level of the template speech waveform is set to 60 dB SPL and both the background ICRA-noise and a hearing threshold simulating noise for normal-hearing listeners is added. The resulting waveform is filtered using a gammatone filterbank (Hohmann, 2002) with 27 frequency channels between 236 Hz and 8 kHz equally spaced on an ERB-scale. The filter-outputs are half-wave rectified and low pass filtered at 1 kHz in a hair cell model. After processing with five consecutive adaptation loops with time constants chosen as in Holube and Kollmeier (1996) the signal is again filtered by a modulation filterbank that

consists of 4 modulation filters: one low pass at 2.5 Hz and three band passes with center frequencies of 5, 7.5 and 10 Hz and bandwidths of 5 Hz, respectively.

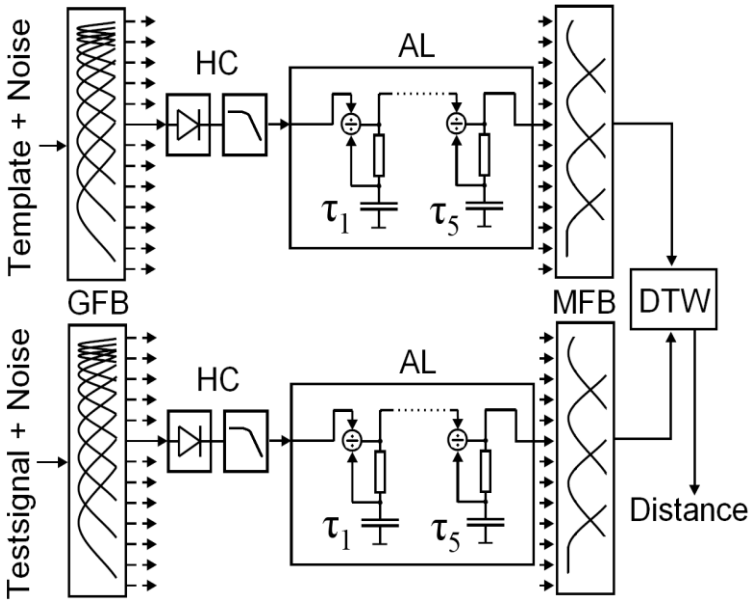

Figure 7.4: Scheme of the speech intelligibility model. The model calculates the distance between both, the template waveform and the testsignal waveform after preprocessing in the same way. GFB: gammatone filterbank, HC: hair cell modell, AL: adaptation loops, MFB: modulation-filterbank, DTW: Dynamic-Time-Warp speech recognizer.

The outcome is an "internal representation" (IR) of the time signal. The testsignal + noise waveform is preprocessed in the same way by the perception model. Note that "noise" in this scheme means running ICRA background noise added to a running hearing threshold simulating noise for normal-hearing subjects. All samples of the training vocabulary were equalized to the same length before processing by attaching silence. This was done to rule out a possible discrimination cue due to the individual length of the speech recordings.

The IR of the template and the IR of the testsignal are the inputs of the speech recognizer that calculates the Euclidian distance between the two versions. To allow for a mismatch in the temporal structure between sample and template a DTW algorithm (Sakoe and Chiba, 1978) performs local stretching and compression of the time axes of both IRs in order to achieve a minimal distance. The logatome with the least distance is chosen as the recognized one. The response alternatives given to the model were the same as for HSR.

Two model configurations were realized in this study:

- In configuration A there were 5 IRs per logatome as templates. None of the 5 original recordings was identical to the tested time signal. The logatome that yielded the minimum mean distance of all 5 IRs was chosen as the recognized one. This mimics a realistic task for common speech recognizers because the exact acoustic utterance is unknown.
- Model configuration B contained a single IR per logatome as a template, whereas the original speech material was identical to that of the test signal. Thus the resulting IRs differ only in the initially added background noises. In contrast to configuration A, this configuration disregards the natural variability of speech, thus it assumes perfect knowledge of the "template" to be matched with the DTW algorithm.

There are many combinations possible to select speech material from OLLO for performing these model calculations. For these two model configurations the speech recognizing task was calculated 10 times using each time a new combination of speech recordings spoken by the same speaker.

7.3.2 Model predictions and comparison with listening tests

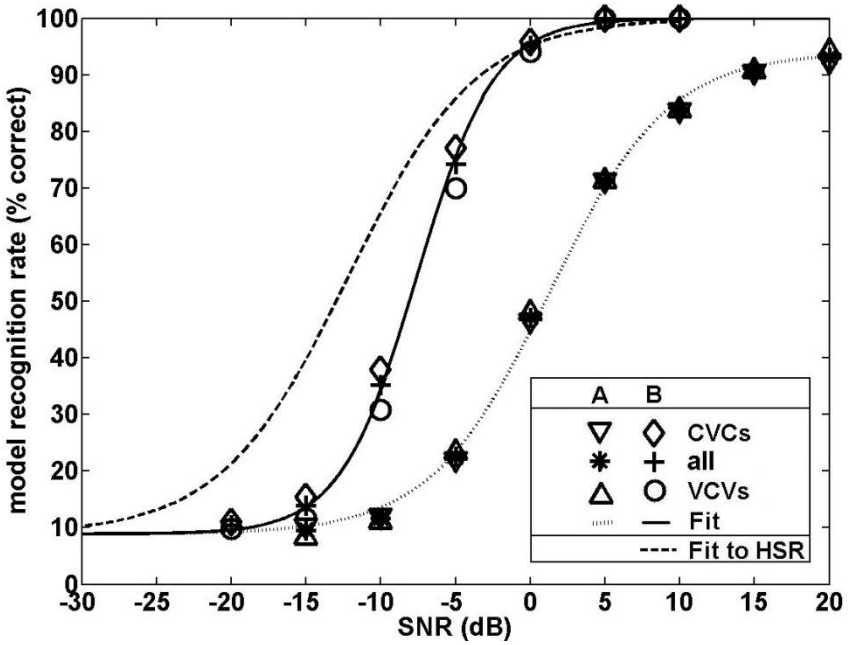

Figure 7.5: Predicted psychometric functions for model configurations A and B derived with utterances of logatomes in ICRA-noise at fixed SNR respectively. For comparision inserted: fit to measured normal-hearing psychometric function (HSR) from Figure 7.1.

The resulting psychometric functions of the ASR experiments are shown in Figure 7.5. Additionally, the fitted psychometric function for normal-hearing subjects from Figure 7.1 is plotted as a reference. Configuration A shows the same recognition rates for CVCs and for VCVs. The resulting SRT calculated by a fitted psychometric function is 1.3 dB and thus is more than 13 dB higher than that in HSR. It was assumed that in this model configuration, which closely resembles ASR tasks, 100% model recognition rate can never be achieved even without background noise. This is due to the inherent speech variability that is still a problem in ASR tasks (Lippmann, 1997). To include this fact a third parameter (the difference between 100% and the saturation recognition rate of the model) was introduced into the fitting routine. With a slope of 5.8 %/dB the reference slope is reproduced quite well.

A much better prediction of the normal-hearing psychometric function is achieved with model configuration B. The order of CVC and VCV as well as the upper part of the reference curve is modelled correctly. 100 % recognition rate is reached at 10 dB SNR. The slope (8.9 %/dB) deviates slightly from the reference, the SRT (-7.6 dB) is much closer to human listeners SRT, but still there is a gap of 4.6 dB between them.

	d	t	g	k	f	s	b	p	v	ts	m	n	ʃ	l
d	22						8		16		16	20		
t		26					10				22	12		8
g			26				10		8		14	18		8
k			10	20			10				18	16		8
f					12		10		10		18	18		14
s						54					12			10
b							30				16	14		18
p							12	18	12		10	22		10
v	8								24		22	18		8
ts							8		8	38	16	14		
m							10				32	14		18
n							10	8	16		14	26		10
ʃ												8	72	
l							10		8		18	18		32

Figure 7.6: Consonant confusion matrix for model configuration B at -10 dB SNR, The display is the same as in Figure 7.2.

In the following only confusion matrices for model configuration B are evaluated and compared to HSR confusion matrices. The SNR was chosen to -10 dB to

ensure about the same intelligibility as for human listeners. Figure 7.6 and Figure 7.7 show these confusion matrices. Comparing Figure 7.2 to Figure 7.6 the same consonants /ʃ/, /ts/, and /s/ are recognized best by the model, but that high human recognition rates like 92% for /ʃ/ are not reached. However, some consonants like /n/, /v/, and /b/ are recognized even better by the model than by human listeners. There is no significant difference between the model recognition rates for unvoiced and voiced plosives. Overall the "contrast" of the model matrix between the diagonal elements is worse than in HSR. This is also the case for the model confusion matrix for vowels: The clustering found in Figure 7.3 could not be reproduced. At -10 dB SNR the overall recognition rate of CVC utterances is significantly worse than for normal-hearing subjects at -15 dB SNR (38% compared to 52%). However, the phonemes /ɔ/ and /ʊ/ are recognized slightly better by the model than in HSR. The characteristic nearly uniform columns at /o/ and /u/ provide an indication that these phonemes are the most probable vowels to recognize by presenting any vowel at such low SNRs.

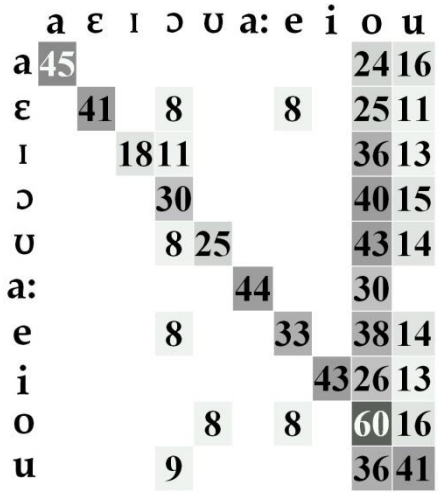

Figure 7.7: Vowel confusion matrix for model configuration B at -10 dB SNR. The display is the same as in Figure 7.2.

7.4 Discussion

Two model configurations were employed, one taking the natural variability of speech into account, the other one disregarding it. Our results show that there is only a chance of predicting the psychometric function for normal-hearing listeners by ignoring the variability of speech itself, i.e. taking identical speech test and training utterances as

inputs for the model. Conversely this gives an indication that speech variability is not crucial to speech intelligibility of normal-hearing subjects at high SNRs. Human speech recognition is as perfect and in some phonemes better than the prediction as if the listener knew the audio signal before the recognition process. However, speech variability is crucial to a model that does not hold the exact speech recording in its training vocabulary.

Although confusion matrices of HSR and ASR are quite similar (especially the consonant phoneme ones), those of the model show a smaller contrast between highly and poorly recognized phonemes. This can be an indication that human listeners use more information from high frequencies to discriminate nonsense speech material than it is done by this model. In each ASR confusion matrix there is a bias favouring some phonemes, like /o/ and /u/ in Figure 7.7, independent of the type of the presented phoneme. This bias could be corrected by changing the selection criteria, which would probably also be done by human listeners during the measurement procedure.

7.5 Conclusions

This speech intelligibility model is based on the time signal of speech and consists of a psychoacoustically motivated preprocessing and a simple speech recognizer. It is capable of predicting essential aspects of speech intelligibility of normal-hearing subjects. By considering the intrinsic variability of speech the modeled SRT is 13 dB higher than human listeners show. This is consistent with findings of other studies exploring the human-machine gap (e.g., Meyer *et al.*, 2007a). Introducing a perfect knowledge about the speech signal to recognize allows for predicting the psychometric function with a much smaller offset. This refers to the “optimal detector” concept required to model human perception assuming that the “world knowledge” yields an optimal template in each HSR experiment. In addition, it was possible to detect characteristic differences between phoneme confusion matrices of HSR and ASR.

Future studies should investigate speech intelligibility of hearing-impaired subjects and also should analyse the influence of the loss of dynamic range at the hearing-impaired on speech intelligibility in a microscopic way.

7.6 Acknowledgements

We would like to thank the EU HearCom Project, the ‘Förderung wissenschaftlichen Nachwuchses des Landes Niedersachsen’ (FwN) and SFB/TR 31 ‘Das aktive Gehör’ (URL: http://www.uni-oldenburg.de/sfbtr31) for funding the research reported in this paper.

8 Bibliography

Al-Salim, S. C., Kopun, J. G., Neely, S. T., Jesteadt, W., Stiegemann, B., and Gorga, M. P. (**2010**). "Reliability of categorical loudness scaling and its relation to threshold," Ear Hear. **31**, pp. 567-578.

Allen, J. B., Hall, J. L., and Jeng, P. S. (**1990**). "Loudness growth in 1/2-octave bands (LGOB) - A procedure for the assessment of loudness," J. Acoust. Soc. Am. **88**, pp. 745-753.

Allen, J. B., Regnier, M., Phatak, S., and Li, F. P. (**2009**). "Nonlinear cochlear signal processing and phoneme perception," Concepts and Challenges in the Biophysics of Hearing, pp. 93-105.

ANSI (**1969**). "ANSI S3.5-1969 American national standard methods for the calculation of the articulation index," standard of the American National Standard Institute, Washington, D.C.

ANSI (**1997**). "ANSI S3.5-1997 Methods for calculation of the Speech Intelligibility Index," standard of the American National Standard Institute, Washington, D.C.

Anweiler, A. K., and Verhey, J. L. (**2006**). "Spectral loudness summation for short and long signals as a function of level," J. Acoust. Soc. Am. **119**, pp. 2919-2928.

Appell, J. (**2002**). "Loudness models for rehabilitative audiology," PhD thesis, Medizinische Physik, Carl-von-Ossietzky-Universität, Oldenburg.

Barker, J., and Cooke, M. (**2007**). "Modelling speaker intelligibility in noise," Speech Communication **49**, pp. 402-417.

Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**, pp. 331-342.

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

Brand, T., and Hohmann, V. (**2001**). "Effect of hearing loss, centre frequency, and bandwidth on the shape of loudness functions in categorical loudness scaling," Audiology **40**, pp. 92-103.

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

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

Brand, T., Wittkop, T., Wagener, K., and Kollmeier, B. (**2004**). "Vergleich von Oldenburger Satztest und Freiburger Wörtertest als geschlossene Versionen," (Comparison of the Oldenburg sentence test and the Freiburg word test as closed versions), Proceedings of the 7th annual meeting of the Deutsche Gesellschaft für Audiologie (DGA), Leipzig.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001**). "Binaural processing model based on contralateral inhibition. I. Model structure," J. Acoust. Soc. Am. **110**, pp. 1074-1088.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2002**). "A time-domain binaural signal detection model and its predictions for temporal resolution data," Acta Acustica United with Acustica **88**, pp. 110-112.

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

Brown, G. J., Ferry, R. T., and Meddis, R. (**2010**). "A computer model of auditory efferent suppression: implications for the recognition of speech in noise," J. Acoust. Soc. Am. **127**, pp. 943-954.

Chalupper, J., and Fastl, H. (**2002**). "Dynamic loudness model (DLM) for normal and hearing-impaired listeners," Acta Acustica United with Acustica **88**, pp. 378-386.

Chi, T. S., Gao, Y. J., Guyton, M. C., Ru, P. W., and Shamma, S. (**1999**). "Spectro-temporal modulation transfer functions and speech intelligibility," J. Acoust. Soc. Am. **106**, pp. 2719-2732.

Ching, T. Y., Dillon, H., and Byrne, D. (**1998**). "Speech recognition of hearing-impaired listeners: predictions from audibility and the limited role of high-frequency amplification," J. Acoust. Soc. Am. **103**, pp. 1128-1140.

Christiansen, T. U., Dau, T., and Greenberg, S. (**2006**). "Spectro-temporal processing of speech - An information-theoretic framework," *International Symposium on Hearing, Cloppenburg*, edited by B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. M. Verhey (Springer, New York).

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

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

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

Davis, M. H., and Johnsrude, I. S. (**2007**). "Hearing speech sounds: top-down influences on the interface between audition and speech perception," Hear. Res. **229**, pp. 132-147.

Derleth, R.-P. (**1999**). "Temporal and compressive properties of the normal and impaired auditory system," PhD-thesis, Medizinische Physik, Carl von Ossietzky Universität, Oldenburg.

Derleth, R. P., Dau, T., and Kollmeier, B. (**2001**). "Modeling temporal and compressive properties of the normal and impaired auditory system," Hear. Res. **159**, pp. 132-149.

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

Dubno, J. R., and Levitt, H. (**1981**). "Predicting consonant confusions from acoustic analysis," J. Acoust. Soc. Am. **69**, pp. 249-261.

Dubno, J. R., Dirks, D. D., and Langhofer, L. R. (**1982**). "Evaluation of hearing-impaired listeners using a nonsense-syllable test. 2. Syllable recognition and consonant confusion patterns," J. Speech Hear. Res. **25**, pp. 141-148.

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

Dubno, J. R., and Schaefer, A. B. (**1992**). "Comparison of frequency selectivity and consonant recognition among hearing-impaired and masked normal-hearing listeners," J. Acoust. Soc. Am. **91**, pp. 2110-2121.

Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (**2007**). "Estimates of basilar-membrane nonlinearity effects on masking of tones and speech," Ear Hear. **28**, pp. 2-17.

Elberling, C. (**1999**). "Loudness scaling revisited," J. Am. Acad. Audiol. **10**, 248-260.

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

Emiroğlu, S. (**2007**). "Timbre perception and object separation with normal and hearing impaired," PhD thesis, Medizinische Physik, Carl-von-Ossietzky-University, Oldenburg.

Emiroğlu, S., and Kollmeier, B. (**2008**). "Timbre discrimination in normal-hearing and hearing-impaired listeners under different noise conditions," Brain Res. **1220**, pp. 199-207.

Ewert, S. D., and Dau, T. (**2004**). "External and internal limitations in amplitude-modulation processing," J. Acoust. Soc. Am. **116**, pp. 478-490.

Fabry, D. A., and Van Tasell, D. J. (**1986**). "Masked and filtered simulation of hearing loss: effects on consonant recognition," J. Speech Hear. Res. **29**, pp. 170-178.

Fletcher, H., and Galt, R. H. (**1950**). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, pp. 89-151.

Florentine, M., and Buus, S. (**1981**). "An excitation-pattern model for intensity discrimination," J. Acoust. Soc. Am. **70**, pp. 1646-1654.

Fowler, E. P. (**1950**). "The recruitment of loudness phenomenon," Laryngoscope **60**, pp. 680-695.

Fredelake, S., Hohmann, V., Haumann, S., Büchner, A., Lenarz, T., and Kollmeier, B. (**2010**). "Modellierung der Sprachverständlichkeit bei Versorgung mit einem Cochleaimplantat," (Modeling of speech intelligibility for cochlear implant users), Proceedings of the 13th annual meeting of the Deutsche Gesellschaft für Audiologie (DGA), Frankfurt am Main.

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

Gallun, F., and Souza, P. (**2008**). "Exploring the role of the modulation spectrum in phoneme recognition," Ear Hear. **29**, pp. 800-813.

Gates, G. A., and Mills, J. H. (**2005**). "Presbycusis," Lancet **366**, pp. 1111-1120.

Gelfand, S. A., Schwander, T., Levitt, H., Weiss, M., and Silman, S. (**1992**). "Speech recognition performance on a modified nonsense syllable test," Journal of Rehabilitation Research and Development **29**, pp. 53-60.

Ghitza, O., and Sondhi, M. M. (**1997**). "On the perceptual distance between speech segments," J. Acoust. Soc. Am. **101**, pp. 522-529.

Glasberg, B. R., and Moore, B. C. (**1992**). "Effects of envelope fluctuations on gap detection," Hear. Res. **64**, pp. 81-92.

Goedegebure, A. (**2005**). "Phoneme compression - processing of the speech signal and effects on speech intelligibility in hearing-impaired listeners," PhD thesis, Erasmus Universiteit, Rotterdam.

Gomaa, N. A., Rubinstein, J. T., Lowder, M. W., Tyler, R. S., and Gantz, B. J. (**2003**). "Residual speech perception and cochlear implant performance in postlingually deafened adults," Ear Hear. **24**, pp. 539-544.

Gregan, M. J., Nelson, P. B., and Oxenham, A. J. (**2010**). "Effects of background noise level on behavioral estimates of basilar-membrane compression," J. Acoust. Soc. Am. **127**, pp. 3018-3025.

Hacker, D. (**2002**). "Phonologie," in *Sprachtherapie mit Kindern (Speech therapy for children)*, edited by S. Baumgartner, and I. Füssenich (Ernst Reinhardt GmbH & Co. KG Verlag, München), pp. 13-62.

Hagerman, B., and Olofsson, A. (**2004**). "A method to measure the effect of noise reduction algorithms using simultaneous speech and noise," Acta Acustica United with Acustica **90**, pp. 356-361.

Hansen, M. (**2002**). "Effects of multi-channel compression time constants on subjectively perceived sound quality and speech intelligibility," Ear Hear. **23**, pp. 369-380.

Hant, J. J., and Alwan, A. (**2003**). "A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise," Speech Communication **40**, pp. 291-313.

Heinz, M. G., Zhang, X., Bruce, I. C., and Carney, L. H. (**2001**). "Auditory nerve model for predicting performance limits of normal and impaired listeners," Acoustics Research Letters Online **2**, pp. 91-96.

Hellman, R. P., and Zwislocki, J. (**1961**). "Some factors affecting estimation of loudness," J. Acoust. Soc. Am. **33**, pp. 687-694.

Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. **87**, pp. 1738-1752.

Hickok, G., and Poeppel, D. (**2007**). "The cortical organization of speech processing," Nat. Rev. Neurosci. **8**, pp. 393-402.

Hohmann, V., and Kollmeier, B. (**1995**). "The effect of multichannel dynamic compression on speech intelligibility," J. Acoust. Soc. Am. **97**, pp. 1191-1195.

Hohmann, V. (**2002**). "Frequency analysis and synthesis using a Gammatone filterbank," Acta Acustica United with Acustica **88**, pp. 433-442.

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

Hopkins, K., and Moore, B. C. (**2010**). "The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects," J. Acoust. Soc. Am. **127**, pp. 1595-1608.

Horwitz, A. R., Ahlstrom, J. B., and Dubno, J. R. (**2007**). "Speech recognition in noise: estimating effects of compressive nonlinearities in the basilar-membrane response," Ear Hear. **28**, pp. 682-693.

Houtgast, T., and Steeneken, H. J. M. (**1984**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, pp. 1069-1077.

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

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

Humes, L. E., Dirks, D. D., Bell, T. S., and Kincaid, G. E. (**1987**). "Recognition of nonsense syllables by hearing-impaired listeners and by noise-masked normal hearers," J. Acoust. Soc. Am. **81**, pp. 765-773.

IEC60645-1 (**2002**). "Electroacoustics - Audiometric equipment - Part 1: Equipment for pure-tone audiometry," standard of the International Electrotechnical Commission, Geneva, Switzerland

IPA (**1999**). *The Handbook of the International Phonetic Association* (Cambridge University Press).

ISO 16832 (**2006**). "Acoustics — Loudness scaling by means of categories," standard of the International Organization for Standardization, Geneva, Switzerland.

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

Jepsen, M. L. (**2010**). "Modeling auditory processing and speech perception in hearing-impaired listeners," PhD thesis, Department of Electrical Engineering, Technical University of Denmark, Copenhagen.

Jürgens, T., Brand, T., and Kollmeier, B. (**2007**). "Modelling the human-machine gap in speech reception: microscopic speech intelligibility prediction for normal-hearing subjects with an auditory model," Proceedings of the Interspeech, Antwerp, Belgium, pp. 410-413.

Jürgens, T., and Brand, T. (**2009**). "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," J. Acoust. Soc. Am. **126**, pp. 2635-2648.

Kiessling, J., Steffens, T., and Wagner, I. (**1993**). "On the clinical applicability of loudness scaling," Audiol. Acoust. **32**, pp. 100-115.

Kollmeier, B. (**1990**). "Meßmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache," (Method, modeling and improvement of speech intelligibility), Habilitation thesis, Fachbereich Physik, Georg-August-Universität, Göttingen.

Kollmeier, B., Müller, C., Wesselkamp, M., and Kliem, K. (**1992**). "Weiterentwicklung des Reimtests nach Sotschek" (Further development of the rhyme test according to Sotschek) in *Moderne Verfahren der Sprach-audiometrie*, *Audiologische Akustik,* edited by B. Kollmeier (Median-Verlag, Heidelberg), pp. 216-237.

Kollmeier, B. (**1997**). *Hörflächenskalierung - Grundlagen und Anwendung der kategorialen Lautheitsskalierung für Hördiagnostik und Hörgeräteversorgung (Auditory sensation area scaling – Basics and applications of categorical loudness scaling for diagnostics and hearing aid fitting)* (Median-Verlag, Heidelberg).

Kollmeier, B. (**1999**). "On the four factors involved in sensorineural hearing loss," in *Psychophysics, Physiology and Models of Hearing*, edited by T. Dau, V. Hohmann, and B. Kollmeier (World Scientific, Singapore), pp. 211-218.

Kryter, K. D. (**1962**). "Methods for calculation and use of the Articulation Index," J. Acoust. Soc. Am. **34**, pp. 1689-1697.

Lancaster, H. O. (**1958**). "The structure of bivariate distributions," Annals of Mathematical Statistics **29**, pp. 719-736.

Launer, S. (**1995**). "Loudness perception in listeners with sensorineural hearing impairment," PhD thesis, Medizinische Physik, Carl-von-Ossietzky Universität, Oldenburg.

Launer, S., Hohmann, V., and Kollmeier, B. (**1997**). "Modeling loudness growth and loudness summation in hearing-impaired listeners," in *Modeling sensorineural hearing loss*, edited by W. Jesteadt (Lawrence Erlbaum Associates, Mahwah, New Jersey), pp. 175-185.

Lawrence, D. L., and Byers, V. W. (**1969**). "Identification of voiceless fricatives by high frequency hearing impaired listeners," J. Speech Hear. Res. **12**, pp. 426-434.

Leijon, A., and Stadler, S. (**2008**). "Fast amplitude compression in hearing aids improves audibility but degrades speech information transmission," Research report, Sound and Image Processing Lab., School of Electrical Engineering, Stockholm, Sweden.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, Suppl 2:467+.

Li, F. (**2009**). "Perceptual cues of consonant sounds and impact of sensorineural hearing loss on speech perception," PhD thesis, Electrical and Computer Engineering, University of Illinois, Urbana-Champaign.

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

Lippmann, R. P. (**1997**). "Speech recognition by machines and humans," Speech Communication **22**, pp. 1-15.

Lopez-Poveda, E. A., Plack, C. J., Meddis, R., and Blanco, J. L. (**2005**). "Cochlear compression in listeners with moderate sensorineural hearing loss," Hear. Res. **205**, pp. 172-183.

Lopez-Poveda, E. A., and Alves-Pinto, A. (**2008**). "A variant temporal-masking-curve method for inferring peripheral auditory compression," J. Acoust. Soc. Am. **123**, pp. 1544-1554.

Lopez-Poveda, E. A., Johannesen, P. T., and Merchan, M. A. (**2009**). "Estimation of the degree of inner and outer hair cell dysfunction from distortion product otoacoustic emission input/output functions," Audiological Medicine **7**, pp. 22-28.

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

Magnusson, L. (**1996**). "Speech intelligibility index transfer functions and speech spectra for two Swedish speech recognition tests," Scand. Audiol. **25**, pp. 59-67.

Marquardt, D. W. (**1963**). "An algorithm for least-squares estimation of nonlinear parameters," J. Soc. Appl. Math. **11**, pp. 431-441.

Mauermann, M., Long, G. R., and Kollmeier, B. (**2004**). "Fine structure of hearing threshold and loudness perception," J. Acoust. Soc. Am. **116**, pp. 1066-1080.

Meddis, R., O'Mard, L. P., and Lopez-Poveda, E. A. (**2001**). "A computational algorithm for computing nonlinear auditory frequency selectivity," J. Acoust. Soc. Am. **109**, pp. 2852-2861.

Meddis, R., and O'Mard, L. P. (**2006**). "Virtual pitch in a computational physiological model," J. Acoust. Soc. Am. **120**, pp. 3861-3869.

Meddis, R., Lecluyse, W., Tan, C. M., and Panda, M. R. (**2010**). "Beyond the audiogram: identifying and modelling patterns of hearing deficits," in *The Neurophysiological Bases of Auditory Perception - 15th International Symposium on Hearing, Salamanca,* edited by E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis (Springer, New York), pp. 631-640.

Meier, H. (**1967**). *Deutsche Sprachstatistik (Statistics of German language)* (Georg Olms Verlagsbuchhandlung, Hildesheim).

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (**2008**). "Phoneme representation and classification in primary auditory cortex," J. Acoust. Soc. Am. **123**, pp. 899-909.

Messing, D., Delhorne, L., Bruckert, E., Braida, L., and Ghitza, O. (**2008**). "Consonant discrimination of degraded speech using an efferent-inspired closed-loop cochlear model," Proceedings of the Interspeech, Brisbane, Australia, pp. 1052-1055.

Meyer, B., and Wesker, T. (**2006**). "A human-machine comparison in speech recognition based on a logatome corpus," Proceedings of the Workshop on Speech Recognition and Intrinsic Variation, Toulouse, France.

Meyer, B., Brand, T., and Kollmeier, B. (**2007a**). "Phoneme confusions in human and automatic speech recognition," Proceedings of the Interspeech, Antwerp, Belgium, pp. 1485-1488.

Meyer, B. (**2009**). "Human and automatic speech recognition in the presence of speech-intrinsic variabilities," PhD thesis, Medizinische Physik, Carl-von-Ossietzky Universität, Oldenburg.

Meyer, B., and Kollmeier, B. (**2010**). "Learning from human errors: Prediction of phoneme confusions based on modified ASR training," Proceedings of the Interspeech conference, Makuhari, Japan.

Meyer, R., Kollmeier, B., and Brand, T. (**2007b**). "Predicting speech intelligibility in fluctuating noise," Proceedings of the 8th EFAS Congress joint meeting with the 10th annual meeting of the Deutsche Gesellschaft für Audiologie (DGA), Heidelberg.

Meyer, R. M., Brand, T., and Kollmeier, B. (**2009**). "Prediction of speech intelligibility in fluctuating noise for listeners with normal and impaired hearing," Proceedings of the NAG/DAGA international conference on acoustics, Rotterdam, The Netherlands.

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

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

Moore, B. C. J. (**1998**). *Cochlear hearing loss* (Whurr Publishers Ltd, London).

Moore, B. C., Vickers, D. A., Plack, C. J., and Oxenham, A. J. (**1999**). "Inter-relationship between different psychoacoustic measures assumed to be related to the cochlear active mechanism," J. Acoust. Soc. Am. **106**, pp. 2761-2778.

Moore, B. C. J. (**2003**). "Speech processing for the hearing-impaired: successes, failures and implications for speech mechanisms," Speech Communication **41**, pp. 81-91.

Moore, B. C. J., and Glasberg, B. R. (**2004**). "A revised model of loudness perception applied to cochlear hearing loss," Hear. Res. **188**, pp. 70-88.

Nelson, D. A., Schroder, A. C., and Wojtczak, M. (**2001**). "A new procedure for measuring peripheral compression in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **110**, pp. 2045-2064.

Oxenham, A. J., and Plack, C. J. (**1997**). "A behavioral measure of basilar-membrane nonlinearity in listeners with normal and impaired hearing," J. Acoust. Soc. Am. **101**, pp. 3666-3675.

Patuzzi, R. B., Yates, G. K., and Johnstone, B. M. (**1989**). "Outer hair cell receptor current and sensorineural hearing loss," Hear. Res. **42**, pp. 47-72.

Pavlovic, C. V. (**1987**). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. **82**, pp. 413-422.

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

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

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

Phillips, S. L., Richter, S. J., and McPherson, D. (**2009**). "Voiced initial consonant perception deficits in older listeners with hearing loss and good and poor word recognition," J. Speech Lang. Hear. Res. **52**, pp. 118-129.

Pittman, A. L., and Stelmachowicz, P. G. (**2000**). "Perception of voiceless fricatives by normal-hearing and hearing-impaired children and adults," J. Speech Lang. Hear. Res. **43**, pp. 1389-1401.

Plack, C. J., Drga, V., and Lopez-Poveda, E. A. (**2004**). "Inferred basilar-membrane response functions for listeners with mild to moderate sensorineural hearing loss," J. Acoust. Soc. Am. **115**, pp. 1684-1695.

Plack, C. J., and Howgate, S. (**2010**). "A behavioral measure of the cochlear changes underlying temporary threshold shifts," Association for Research in Otolaryngology Midwinter Meeting (Anaheim, CA).

Plomp, R. (**1976**). *Aspects of tone sensation* (Academic Press, London).

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

Plomp, R. (**1986**). "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," J. Speech Hear. Res. **29**, pp. 146-154.

Press, W., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (**1992**). *Numerical recipes in C* (Cambridge University Press, Cambridge).

Reed, C. M., Braida, L. D., and Zurek, P. M. (**2009**). "Review article: review of the literature on temporal resolution in listeners with cochlear hearing impairment: a critical assessment of the role of suprathreshold deficits," Trends Amplif. **13**, pp. 4-43.

Rhebergen, K. S., and Versfeld, N. J. (**2005**). "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**, pp. 2181-2192.

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

Rhebergen, K. S., Lyzenga, J., Dreschler, W. A., and Festen, J. M. (**2010**). "Modeling speech intelligibility in quiet and noise in listeners with normal and impaired hearing," J. Acoust. Soc. Am. **127**, pp. 1570-1583.

Rohdenburg, T., Goetze, S., Hohmann, V., Kammeyer, K., and Kollmeier, B. (**2008**). "Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, Nevada, pp. 2449-2452.

Rosengard, P. S., Oxenham, A. J., and Braida, L. D. (**2005a**). "Comparing different estimates of cochlear compression in listeners with normal and impaired hearing," J. Acoust. Soc. Am. **117**, pp. 3028-3041.

Rosengard, P. S., Payton, K. L., and Braida, L. D. (**2005b**). "Effect of slow-acting wide dynamic range compression on measures of intelligibility and ratings of speech quality in simulated-loss listeners," J. Speech Lang. Hear. Res. **48**, pp. 702-714.

Ruggero, M. A., and Rich, N. C. (**1991**). "Furosemide alters organ of corti mechanics: evidence for feedback of outer hair cells upon the basilar membrane," J. Neurosci. **11**, pp. 1057-1067.

Sachs, L. (**1999**). *Angewandte Statistik (Applied statistics)* (Springer, Berlin, Heidelberg).

Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing **ASSP-26**, pp. 43-49.

Scharenborg, O. (**2007**). "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," Speech Communication **49**, pp. 336-347.

Sheft, S., Ardoint, M., and Lorenzi, C. (**2008**). "Speech identification based on temporal fine structure cues," J. Acoust. Soc. Am. **124**, pp. 562-575.

Sher, A. E., and Owens, E. (**1974**). "Consonant confusions associated with hearing loss above 2000 Hz," J. Speech Hear. Res. **17**, pp. 669-681.

Sohn, W. (**2001**). "Schwerhörigkeit in Deutschland, Repräsentative Hörscreening-Untersuchung bei 2000 Probanden in 11 Allgemeinpraxen," (Hearing impairment in Germany, representative screening investigation with 2000 volunteers in 11 doctor's offices), Z. Allg. Med. **77**, pp. 143-147.

Sroka, J. J., and Braida, L. D. (**2005**). "Human and machine consonant recognition," Speech Communication **45**, pp. 401-423.

Stadler, S., Leijon, A., and Hagerman, B. (**2007**). "An information theoretic approach to predict speech intelligibility for listeners with normal and impaired hearing," Proceedings of the Interspeech, Antwerp, Belgium, pp. 398-401.

Stadler, S. (**2009**). "Probabilistic modelling of hearing: speech recognition and optimal audiometry," Licentiate thesis, Sound and Image Processing Laboratory, School of Electrical Engineering, KTH, Stockholm, Sweden.

Steeneken, H. J. M., and Houtgast, T. (**1980**). "Physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, pp. 318-326.

Steinberg, J. C., and Gardner, M. B. (**1937**). "The dependence of hearing impairment on sound intensity," J. Acoust. Soc. Am. **9**, pp. 11-23.

Stelmachowicz, P. G., Hoover, B. M., Lewis, D. E., Kortekaas, R. W., and Pittman, A. L. (**2000**). "The relation between stimulus context, speech audibility, and perception for normal-hearing and hearing-impaired children," J. Speech Lang. Hear. Res. **43**, pp. 902-914.

Stern, R., Acero, A., Liu, F. H., and Ohshima, Y. (**1996**). "Signal processing for robust speech recognition," in *Automatic Speech and Speaker Recognition*, edited by C.-H. Lee, F. K. Soong, and K. K. Paliwal (Springer, Berlin).

Stevens, S. S. (**1957**). "On the psychophysical law," Psychol. Rev. **64**, pp. 153-181.

Sukowski, H., Meyer, R. M., Thiele, C., Brand, T., Wagener, K. C., Lesinski-Schiedat, A., and Kollmeier, B. (**2010**). "Sprachverständlichkeitsvorhersagen auf der Grundlage des Speech Intelligibility Index im Rahmen der HNO-ärztlichen Begutachtung bei angezeigter beruflicher Lärmschwerhörigkeit," (Prediction of speech intelligibility using the Speech Intelligibility Index in the framework of the assessment of job-related hearing impairment) in Proceedings of the 13th annual meeting of the Deutsche Gesellschaft für Audiologie, Frankfurt am Main, Germany.

Tchorz, J., Kasper, K., Reininger, H., and Kollmeier, B. (**1997**). "On the interplay between auditory-based features and locally recurrent neural networks," Proceedings of the Eurospeech, Rhodes, Greece, pp. 2075-2078.

Tchorz, J., and Kollmeier, B. (**1999**). "A model of auditory perception as front end for automatic speech recognition," J. Acoust. Soc. Am. **106**, pp. 2040-2050.

Turner, C. W., Souza, P. E., and Forget, L. N. (**1995**). "Use of temporal envelope cues in speech recognition by normal and hearing-impaired listeners," J. Acoust. Soc. Am. **97**, pp. 2568-2576.

van Buuren, R. A., Festen, J. M., and Houtgast, T. (**1999**). "Compression and expansion of the temporal envelope: evaluation of speech intelligibility and sound quality," J. Acoust. Soc. Am. **105**, pp. 2903-2913.

Wagener, K., Brand, T., and Kollmeier, B. (**1999a**). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests," (Development and evaluation of a sentence test for the German language I: design of the Oldenburg sentence test), Zeitschrift für Audiologie/Audiological Acoustics **38**, pp. 4-15.

Wagener, K., Brand, T., and Kollmeier, B. (**1999b**). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests," (Development and evaluation of a sentence test for the German language III: evalutation of the Oldenburg sentence test), Zeitschrift für Audiologie/Audiological Acoustics **38**, pp. 86-95.

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

Walker, G., Byrne, D., and Dillon, H. (**1984**). "The effects of multichannel compression/expansion amplification on the intelligibility of nonsense syllables in noise," J. Acoust. Soc. Am. **76**, pp. 746-757.

Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., and Kollmeier, B. (**2005**). "Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines," Proceedings of the Interspeech, Lisboa, Portugal, pp. 1273-1276.

Wiechert, M. T., Judkewitz, B., Riecke, H., and Friedrich, R. W. (**2010**). "Mechanisms of pattern decorrelation by recurrent neuronal circuits," Nat. Neurosci **13**, pp. 1003-1010.

Wojtczak, M., and Oxenham, A. J. (**2009**). "Pitfalls in behavioral estimates of basilar-membrane compression in humans," J. Acoust. Soc. Am. **125**, pp. 270-281.

Woods, D. L., Yund, E. W., Herron, T. J., and Cruadhlaoich, M. A. (**2010**). "Consonant identification in consonant-vowel-consonant syllables in speech-spectrum noise," J. Acoust. Soc. Am. **127**, pp. 1609-1623.

Yasin, I., and Plack, C. J. (**2003**). "The effects of a high-frequency suppressor on tuning curves and derived basilar-membrane response functions," J. Acoust. Soc. Am. **114**, pp. 322-332.

Yates, G. K., Winter, I. M., and Robertson, D. (**1990**). "Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range," Hear. Res. **45**, pp. 203-219.

Zurek, P. M., and Delhorne, L. A. (**1987**). "Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment," J. Acoust. Soc. Am. **82**, pp. 1548-1559.

Zwicker, E. (**1977**). "Procedure for calculating loudness of temporally variable sounds," J. Acoust. Soc. Am. **62**, pp. 675-682.

9 Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die mich bei der Erstellung dieser Arbeit unterstützt haben und die damit auch zu meiner persönlichen und professionellen Weiterentwicklung während meiner Promotionszeit beigetragen haben.

Ganz besonders bedanken möchte ich mich bei Dr. Thomas Brand, Prof. Dr. Dr. Birger Kollmeier und Dr. Stephan Ewert, die die wissenschaftliche Betreuung dieser Doktorarbeit übernommen haben.
Dr. Thomas Brand danke ich herzlich für die vielen Ideen, Tipps, wissenschaftlichen Diskussionen und besonders für die kritischen Nachfragen. Von der kritischen Auseinandersetzung mit – vielleicht im ersten Moment unerwarteten – Forschungsergebnissen hat diese Arbeit enorm profitiert. Um einen Einblick in die Tätigkeit als Lehrender in der Medizinischen Physik zu bekommen, war es außerdem sehr hilfreich, die Chance zu bekommen schon als Doktorand Vorlesungs- und Übungsvertretungen zu übernehmen.
Prof. Dr. Dr. Birger Kollmeier danke ich für die vielen hilfreichen Diskussionen nach Vorträgen und die extrem guten Vorschläge zum Aufbau und der Verbesserung von wissenschaftlichen Papieren. Trotz der Vielzahl an Themen und zu betreuenden Forschungsprojekte, die er täglich bearbeitet, sind Diskussionen über meine Arbeit immer genau im richtigen Maße in die inhaltliche Tiefe gegangen. Seine stete Arbeit des Aufbaus, des Erhalts und der Weiterentwicklung des Hörforschungsstandorts Oldenburg und die Möglichkeit des eigenständigen wissenschaftlichen Arbeitens haben die Grundlagen geschaffen diese Arbeit zu schreiben. Weiterhin konnte ich sehr von den vielen nationalen und internationalen Kontakten und Vernetzungen profitieren, die er im Laufe der Zeit aufgebaut hat und pflegt.
Dr. Stephan Ewert möchte ich danken für die Unterstützung bei Durchführung und Interpretation der psychoakustischen Experimente in dieser Arbeit und für die vielen Hinweise zum Schreiben wissenschaftlicher Papiere.
Herzlichen Dank auch an PD Dr. Volker Hohmann für die Übernahme der Aufgabe des Zweitgutachters dieser Arbeit.

Darüber hinaus danke ich der gesamten Arbeitsgruppe Medizinische Physik der Universität Oldenburg für die tolle, abwechslungsreiche Zeit. Immer wenn es an

irgendeiner Stelle Fragen oder Unklarheiten gab, habe ich einen Experten in der Arbeitsgruppe gefunden, der Support geben konnte.
Herausheben möchte ich einige Arbeitsgruppenmitglieder, die in besonderer Weise zum Gelingen dieser Arbeit beigetragen haben: Ralf Meyer, der als sehr angenehmer Bürokollege sowohl für Fragen und Diskussionen immer ein offenes Ohr hat als auch eine ruhige Atmosphäre zum konzentrierten Arbeiten ermöglicht hat. Stefan Fredelake danke ich für den ingenieurwissenschaftlichen Blickwinkel auf Fragestellungen dieser Arbeit, starken Einsatz bei der Weiterentwicklung von Modellen und natürlich eine besondere Freundschaft. Bernd Meyer möchte ich für sehr gewissenhafte Hinweise, Nachfragen und Ratschläge bezüglich der menschlichen und maschinellen Phonemerkennung im OLLO-Sprachkorpus danken. Jörg-Hendrik Bach und Hendrik Kayser danke ich für technische Unterstützung bei der Nutzung des Rechenclusters „Schroeder“ und die Geduld, mir auch bei der zehnten Nachfrage zum selben Problem nochmal eine freundliche Antwort per Email zurückzuschreiben. Weiterhin danke ich allen Mitgliedern des Teilbereichs Sprache und Audiologie (SprAud) der Arbeitsgruppe für vielfältige, themenreiche Bläschentreffen, bei denen auch der Spaß nicht zu kurz kam, und Unterstützung bei Paperentwürfen und Probevorträgen im Vorfeld zu Tagungen.
Danke an die „Eck-Mensarunde um 12:20 Uhr“, in wechselnder Zusammensetzung bestehend aus Bernd, Stefan, Jörg-Hendrik, Hendrik, Miriam, Anke, Melanie, Mathias und anderen für jeweils den neuesten intern-Arbeitsgruppen und extern-Arbeitsgruppen Flurfunk und dafür, dass ich in der Mittagspause ein wenig Ablenkung von meiner Arbeit bekommen habe.
Frank Grunau, Anita Gorges, Susanne Garre, Ingrid Wusowski, Katja Warnken und Annegret Bullermann-Wessels danke ich für die immer offen stehende Tür bei technischen Fragen und Verwaltungsangelegenheiten.
Den studentischen und wissenschaftlichen Hilfskräften Eugen Rasumow, Angela Josupeit und Sven Kissner möchte ich danken für die gewissenhafte Durchführung der von mir vorbereiteten Hörexperimente mit Probanden und das geduldige Warten vor den Hörkabinen, bis die doch teilweise recht langen Experimente von den Probanden durchgeführt waren.

Außerhalb der Arbeitsgruppe geht mein Dank an Andreas Gerbrand, der mich vor und während der gesamten Promotionszeit begleitet hat. Seine inhaltlichen Nachfragen, vor allem aber seine immer wieder hervorragenden konzeptionellen Impulse bezüglich der

Darstellung und Ordnung der wissenschaftlichen Erkenntnisse dieser Arbeit haben mir sehr stark geholfen, mich auf das Wesentliche zu konzentrieren.
Meinen Eltern Brigitte und Huldreich-Meno Jürgens möchte ich dafür danken, dass sie mir das Physikstudium ermöglicht haben und dass sie mir zusammen mit meinen Brüdern David und Philipp Jürgens stets einen starken familiären Rückhalt geboten haben.
Herzlichen Dank außerdem an die HörTech gGmbH und die Hörzentrum Oldenburg GmbH, dort besonders an Dr. Kirsten Wagener, Dr. Michael Schulte, Kerstin Sommer, Müge Kaya, Ania Mohadjer-Soleimani, Jessica Beeken, Niklas Grunewald und Matthias Vormann für das Praktikum „Audiologe vom Dienst“. Im Rahmen dieses Praktikums wurde ich an die Arbeit mit normal- und schwerhörenden Probanden und die praktische Durchführung von wissenschaftlichen Studien in der Hörforschung herangeführt.
Für wissenschaftliches Interesse an meiner Arbeit, Tipps, Hinweise und Impulse danke ich Torsten Dau, Morten Jepsen, Ray Meddis, Bernhard Seeber, Christopher Plack, Barbara Shinn-Cunningham, Koen Rhebergen, Svante Stadler und Arne Leijon.

Vielen Dank auch an Amy Beeston und Mani Swaminathan für Hinweise zur englischen Sprache in wissenschaftlichen Papieren und das Korrekturlesen von Teilen dieser Arbeit. Last, but not least danke ich allen Probanden, die mit ihrer Teilnahme diese Arbeit erst ermöglicht haben.

Diese Dissertation profitierte von Geldern des SFB TRR 31 „Das aktive Gehör“, der „Förderung des wissenschaflichen Nachwuchses des Landes Niedersachsen“, des BMBF Projektes „Modellbasierte Hörgeräte“, des HearCom EU-Projektes und des internationalen Graduiertenkollegs „Neurosensorik“.

10 Lebenslauf

Tim Jürgens

geboren am 25.5.1979 in Wilhelmshaven

Staatsangehörigkeit: deutsch

06/1998:	Abitur an der Liebfrauenschule Cloppenburg
09/1998 – 06/1999:	Wehrdienst
09/1999 – 03/2001:	Physikstudium (Dipl.) an der Georg-August-Universität Göttingen
04/2001 – 09/2005:	Physikstudium (Dipl.) an der Carl-von-Ossietzky Universität Oldenburg
09/2005:	Diplomarbeit: „Photoströme und Photolumineszenz von Cu(In,Ga)Se_2-Solarzellen mit lateraler sub-µm-Auflösung“
09/2005 – 12/2005:	Wissenschaftlicher Mitarbeiter des Sonderforschungsbereiches SFB TRR 31 „Das aktive Gehör“ der Universität Oldenburg
10/2005 – 01/2006:	Praktikum als „Audiologe vom Dienst“ bei der Hörzentrum Oldenburg GmbH und beim evangelischen Krankenhaus Oldenburg
01/2006 – 12/2008:	Wissenschaftlicher Mitarbeiter am Institut für Physik der Universität Oldenburg auf der Stelle zur „Förderung des wissenschaftlichen Nachwuchses“; 6 Semester Betreuer im Anfängerpraktikum Physik für Bachelorstudenten
seit 02/2006:	Promotionsstudium im Bereich Medizinische Physik an der Carl-von-Ossietzky Universität Oldenburg
04/2008 – 07/2008:	Vertretungsweise Übernahme von Teilen der Lehrveranstaltung „Einführung in die biomedizinische Physik und Neurophysik“ an der Universität Oldenburg
01/2009 – 09/2010:	Wissenschaftlicher Mitarbeiter des Sonderforschungsbereiches SFB TRR 31 „Das aktive Gehör“

11 Erklärung

Hiermit versichere ich, dass ich die vorliegende Dissertation selbständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Dissertation hat weder in Teilen noch in ihrer Gesamtheit einer anderen wissenschaftlichen Hochschule zur Begutachtung in einem Promotionsverfahren vorgelegen. Teile der Dissertation wurden bereits veröffentlicht bzw. sind zur Veröffentlichung eingereicht, wie an den entsprechenden Stellen angegeben.

Oldenburg, den 22. September 2010

12 List of abbreviations

3I-3AFC: three Interval, three Alternative Forced Choice

ACALOS: Adaptive CAtegorical Loudness Scaling

AI: Articulation Index

AL: Adaptation Loops

ANSI: American National Standards Institute

ASR: Automatic Speech Recognition

B&K: Brüel&Kjaer

BM: Basilar Membrane

BMBF: BundesMinisterium für Bildung und Forschung

BP: BreakPoint between linear and compressive portions in the I/O function

CASP: Computational Auditory Signal processing and Perception model

CR: Compression Ratio

CU: Categorical Untis

CVC: Consonant-Vowel-Consonant

$D(i, j)$: Distance matrix

dB: deciBel

Δd: Difference between internal representations

ΔHL : standard error of the measurement of HL

DLM: Dynamic Loudness Model

DRNL: Dual Resonance NonLinear

DTW: Dynamic-Time-Warp

ERB: Equivalent Rectangular Bandwidth

EU: European Union

f: frequency

f_{mod}: modulation frequency

f_p: frequency of the probe

f_m: frequency of the masker

FwN: Förderung des wissenschaftlichen Nachwuchses

g: guessing probability

G: low-level Gain

GFB: Gammatone Filter Bank

GL: Gain Loss

$\langle G_{NH} \rangle$: average Gain in Normal-Hearing listeners

GOM: Growth Of Masking

HC: Hair Cell

HearCom: Hearing in the Community

HI: Hearing-Impaired

HI-H: Hearing-Impaired listeners with loss at Higher frequencies

HI-LH: Hearing-Impaired listeners with loss at Low and High frequencies

HL: Hearing Level

HL_a: attenuating component of Hearing Loss

HL_{exp}: expanding component of Hearing loss

HL_{IHC}: Hearing Loss due to loss or dysfunction of Inner Hair Cells

HL_{OHC}: Hearing Loss due to loss or dysfunction of Outer Hair Cells

HL_{tot}: total Hearing Loss

HMM: Hidden Markov Model

HSR: Human Speech Recognition

ICRA: International Collegium for Rehabilitative Audiology

IEC: International Electrotechnical Commission

I/O: Input/Output

IPA: International Phonetic Alphabet

IR: Internal Representation

IR_{templ}: template‘s Internal Representation

IR_{test}: test signal’s Internal Representation

ISO: International Organization for Standardization

k: ratio of HL_{OHC} to HL

k_{fit}: fitted k

kHz: kiloHertz

L_{25}: Level that corresponds to medium loudness (25 CU)

L: Level (or SNR)

L_{in} : input Levels

L_{out} : output Levels

MFB: Modulation Filter Bank

MFCC: Mel Frequency Cepstral Coefficients

mC1: CASP model version including hearing-threshold simulating noise and NH processing

mC2: CASP model version including hearing-threshold simulating noise and HI processing

mC3: CASP model version including internal noise added after the DRNL stage and HI processing

mC4: CASP model version including internal noise after the internal representation and different versions of supra-threshold processing

mP1: PeMo version for HI listeners including internal noise after the modulation filterbank

m_{high}: slope of the upper portion of the loudness function

m_{low}: slope of the lower portion of the loudness function

ms: milliseconds

MU: Model Units

n: number of data points below the lower knee point of the I/O function

N: loudness

N_{CU}: categorical loudness

NH: Normal-Hearing

NMNH: Noise-Masked Normal-Hearing

OMA: Oldenburg Measurement Applications

OLLO: OLdenburg LOgatome speech corpus

p: significance level (e.g., of r)

PC: Personal Computer

PDF: Probability Density Function

PeMo: Perception Model

φ: Pearson's phi-index

π: probability

Ψ: psychometric function

r: Pearson's correlation coefficient

RMS: Root-Mean-Square

σ_{inter}: inter-individual standard deviation

σ_{intra}: intra-individual standard deviation

$\sigma_{lin,fit}$: variability of the vertical position of the fit's linear region

SFB/TR: SonderForschungsBereich TransRegio

SII: Speech Intelligibility Index

SL: Sensation Level

SNR: Signal-to-Noise Ratio

SPIN: SPeech In Noise

SPL: Sound Pressure Level

SprAud: Sprache und Audiologie

SRT: Speech Reception Threshold

STI: Speech Transmission Index

TFS: Temporal Fine Structure

TMC: Temporal Masking Curve

VCV: Vowel-Consonant-Vowel

V_{offset}: Vertical offset

y: vertical position of data points of the I/O function

y_{fit}: vertical position of the fit