Analysis and optimization of psychophysical procedures in audiology

Thomas Brand

Bibliotheks- und Informationssystem der Universität Oldenburg
2000

Verlag/Druck/
Vertrieb:

Bibliotheks- und Informationssystem
der Carl von Ossietzky Universität Oldenburg
(BIS) - Verlag -
Postfach 25 41, 26015 Oldenburg
Tel.: 0441/798 2261, Telefax: 0441/798 4040
e-mail: verlag@bis.uni-oldenburg.de

ISBN 3-8142-0721-1

Analysis and optimization of psychophysical procedures in audiology

Thomas Brand

Bibliotheks- und Informationssystem der Universität Oldenburg
2000

Verlag/Druck/
Vertrieb:

Bibliotheks- und Informationssystem
der Carl von Ossietzky Universität Oldenburg
(BIS) - Verlag -
Postfach 25 41, 26015 Oldenburg
Tel.: 0441/798 2261, Telefax: 0441/798 4040
e-mail: verlag@bis.uni-oldenburg.de

ISBN 3-8142-0721-1

Preface

Physicists are notorious for spending most of their time in developing innovative measurement methods and – once they have successfully completed just a few measurements with their new methods – rapidly losing their interest. On a first glance, this might also hold for physicists being involved in such "unphysical" areas as psychophysics and audiology. The current dissertation by the physicist Thomas Brand hence might be considered as a further example of this tradition that already started with v. Helmholtz more than a hundred years ago. On a second glance, however, one may find that the opposite is true: The underlying problem (i.e. function and malfunction of the complex, nonlinear system "human auditory signal processing") is so complex and fascinating, but inaccessible to "old" approaches (such as standard audiometric techniques) that there is no other choice: Physics and quantitative measurement methods must continuously be introduced into audiology! Only by extensively exploiting these methods, a better quantitative understanding can be obtained on how our fascinating auditory system works and what can be done for the compensation of hearing disorders.

The thesis authored by Thomas Brand provides a substantial progress in this direction. He focuses on both efficient measurement methods for speech audiometry in noise (i.e. sentence tests) and the optimisation and quantitative modelling of loudness scaling and applies the methods extensively with human listeners. Both methods are fairly new and by far not generally accepted (although the current book will definitely contribute to it). Speech audiometry is the ultimate method to assess the "effective" hearing loss for speech communication and the potential gain of a hearing instrument. Since the most commonly employed speech tests are either outdated or consume too much time, the time–saving and statistically well–behaved adaptive methods introduced here by Thomas Brand will definitely go their way from physics laboratory to the audiological practice in the field – please read yourself!

Loudness scaling, on the other hand, has suffered in the past from several inadequate measurement procedures and from an insufficient understanding of the processes involved during the quantitative assessment of subjectively perceived loudness. Hence, the work performed here by Thomas Brand really sets the quantitative framework for

this procedure by, e.g., deriving the appropriate target functions, computing the statistical error and achievable estimation accuracy, and by introducing a new, time–efficient adaptive way of performing this important audiological measurement technique. Taken together, this work for the first time gives the theoretical framework for the innovative method of loudness scaling – an indispensable prerequisite, if a broad usage in practical audiology is targeted at – please read yourself!

Thomas Brand is the 17th Ph.D. candidate who finished his work in our working group and the last student from Göttingen who moved with the whole working group to Oldenburg in 1993. By laboriously serving as a subject in auditory experiments (amazing that his ears still perform so well even though the wearer is an excellent rock musician!) he "somehow managed to creep in into our group" (original citation T. Brand). Due to his broad range of interests and his excellent communication skills, though, within a short time he played a central role in our working group, especially with regard to all aspects of psychophysical testing methods, audiology, speech perception, signal processing – and the repair of bulky tube amplifiers! From this description, the reader can already assess how much fun it was to cooperate with Thomas Brand during his thesis work. Some of his enthusiasm for the subject described here will definitely cross over to the reader – please read yourself!

Oldenburg, March 2000 Birger Kollmeier

Contents

Chapter 1

Introduction

Hearing is one of the most important abilities of the human being. It enables us to orient ourselves, to hear and to play music and to communicate with other people. The auditory system has to fulfill difficult tasks to make these things possible:
It has to process very complex acoustical signals (like speech and music) and it has to separate them into different acoustical objects like talkers or musical instruments even if many of these complex signals are present at the same time. This ability to process and separate complex acoustical signals enables listeners with normal hearing to recognize speech even if the environmental noise has about 2 times (6 dB) more power than the speech of their interlocutor (e.g., Kollmeier and Wesselkamp, 1997; Wagener *et al.*, 1999a). If speech and noise come from different directions, the noise may have even 4 times (12 dB) more power (e.g., Bronkhorst and Plomp, 1989).
Moreover, the auditory system has to process a huge dynamic range of acoustical signals. The softest audible signals have a physical power of about $2 \cdot 10^{-5}$ N/m^2 and the loudest signals which can be processed by the auditory system have a physical power of about $20 \cdot 10$ N/m^2. That means that the auditory system is very sensitive and very robust at the same time by making very soft sounds audible and by protecting itself against very loud sounds.
However, the auditory system is very fragile and can easily be damaged (for instance, by too intense sounds or by some chronic diseases) – probably, because it is thus complex and sensitive. The ability to process complex acoustical signals decreases at very early stages of a hearing impairment. Accurate, efficient and valid measurement methods of normal and impaired hearing functions are required to detect these changes (diagnostic audiology) and to assist in the treatment or alleviation of hearing disorders (rehabilitative audiology). The current thesis is concerned with methodological factors central to both fields of audiology.
In many cases a decreased ability to recognize speech in noise is one of the first symptoms of an arising hearing impairment. Sentence intelligibility tests (e.g., Plomp and

Mimpen, 1979; Hagerman, 1982; Nilsson *et al.*, 1994; Kollmeier and Wesselkamp, 1997; Wagener *et al.*, 1999c) are able to assess speech recognition in noise directly. Since the recognition of speech is the most important function of our hearing system in daily life, the diagnosis of a reduced speech reception is the main criterion to quantify whether a hearing–impaired listener needs a hearing aid and whether a given hearing aid is fitted adequately. The current thesis proposes and evaluates efficient methods for measuring speech intelligibility in noise by concurrently estimating the speech reception threshold and the slope of the speech discrimination function. These two diagnostic parameters are essential in assessing a hearing loss.
A further problem which arises if the auditory system gets damaged is the reduced auditory range. Hearing–impaired people do not perceive soft sounds but loud sounds are perceived as loud as by normal–hearing listeners. This phenomenon is often called 'recruitment'. Recruitment makes it complicated to compensate for the hearing loss with hearing aids because the dynamic range of the input signals has to be compressed. Categorical loudness scaling procedures (e.g., Pascoe, 1978; Heller, 1985; Allen *et al.*, 1990; Hohmann and Kollmeier, 1995b) might be an adequate tool to diagnose recruitment and to fit automatic gain control systems in hearing instruments.
Although sentence intelligibility tests in noise and categorical loudness scaling procedures are necessary for an adequate audiological diagnosis, they do not belong to the standard audiological toolbox. This has several reasons: Both procedures have been introduced fairly recently, they incorporate a certain complexity in the measurement procedure (which usually requires a computer–controlled measurement setup), they require a considerable amount of measurement time, and their results are not yet considered as having a similar diagnostic value as, e.g., the audiogram. When sentence intelligibility tests are performed in noise, they have to be very accurate because the interindividual differences of SRT values are very small in competitive noise, i.e. SRT values differ by only about a few dBs between listeners. Categorical loudness scaling measurements take much longer than measuring a pure tone audiogram. Furthermore, they sometimes generate biased results. Since many sentence tests and loudness scaling procedures do not yield the desired accuracy in a tolerable amount of time and because of other reasons listed above, many audiologists do not use speech intelligibility tests in noise and loudness scaling procedures at all.

This study attempts to make both sentence intelligibility tests and categorical loudness scaling more accurate, more reliable and more efficient.
Although speech intelligibility tests and categorical loudness scaling aim at different diagnostic parameters, very similar statistical and empirical methods were employed in order to investigate and to improve these audiological procedures. Hence, both methods were treated in a similar way in this study: In both cases, adaptive procedures for stimulus level placements were applied. In both cases, these adaptive procedures were investigated and optimized using Monte–Carlo simulations. And in both cases, the opti-

mized procedures were evaluated using normal–hearing and hearing–impaired listeners. The aims of the adaptive procedures are completely different in sentence tests and in loudness scaling. In the sentence intelligibility tests, the aim is to present stimulus levels as near as possible at the optimal target levels in order to yield maximum efficiency. In categorical loudness scaling, the aim is to present stimulus levels in random order in the whole auditory range of the respective listener without presenting stimuli outside this range. In this way, biases due to context effects should be reduced and inefficient or annoying trials should be avoided.

Chapter 2 deals with the efficiency of sentence intelligibility tests. The accuracies of threshold and slope estimates were calculated as functions of the presentation levels on the basis of binomial theory. If only the threshold value should be estimated, the highest efficiency is yielded if the stimuli are presented close to the 50 % intelligibility threshold level (*sweetpoint*). This can be achieved using adaptive procedures which converge at 50 % intelligibility. If threshold and slope value should be estimated concurrently, an adaptive procedure yields the highest efficiency if it converges interleaved at two levels belonging to the intelligibilities 19 and 81 % (e.g., Wetherhill, 1963; Levitt, 1971). Some authors, however, have proposed adaptive procedures for concurrent threshold and slope estimates which converge only at one target (Hall, 1981; Leek *et al.*, 1992). In this study, adaptive procedures are proposed and evaluated which converge at the optimal targets (*pair of compromise*) in a randomly interleaved order. These adaptive procedures utilize the fact that in each sentence trial more the one statistically independent Bernoulli trial is performed. This allows for a faster convergence and for shorter track lengths compared to usual psychophysical tasks. The procedures were evaluated using Monte–Carlo simulations and measurements using the Göttingen sentence test (Kollmeier and Wesselkamp, 1997) and the Oldenburg sentence test (Wagener *et al.*, 1999c). These sentence tests differ in the mean number of Bernoulli trial which are performed per sentence trial.

Chapters 3 to 6 deal with categorical loudness scaling:
Since the accuracy of categorical loudness scaling can not be predicted by applying the binomial theory, as it was done in the case of sentence intelligibility tests in chapter 2, a different statistical model was necessary. Chapter 3 describes such a statistical model of categorical loudness scaling and its predictions concerning the reliability and accuracy of different loudness scaling procedures. The model was fitted to categorical loudness scaling data of normal–hearing and hearing–impaired listeners presented by Hohmann (Hohmann, 1993). Based on the model, Monte–Carlo simulations were performed which simulate the whole measurement. The impact of parameters like track length, number of response alternatives and fitting procedure on the reliability and accuracy of the measuring procedure was investigated. The number of outliers in the response statistics of the listener was varied in the simulations as well because some listeners have temporary lacks of attention which may bias the result of the measurement.

In chapter 4 the statistical model of categorical loudness scaling was used to design and too evaluate the Oldenburg–ACALOS (Adaptive CAtegorical LOudness Scaling) procedure which bases on the constant stimuli version of the Oldenburg loudness scaling procedure (Hohmann and Kollmeier, 1995b). The adaptive procedure attempts to be more efficient than other loudness scaling procedures (e.g., Pascoe, 1978; Heller, 1985; Allen *et al.*, 1990; Elberling and Nielsen, 1993; Hohmann and Kollmeier, 1995b; Ricketts and Bentler, 1996; Cox *et al.*, 1997; Rasmussen *et al.*, 1998; Keidser *et al.*, 1999) and to reduce biases due to context–effects. The rules and the step sizes of the adaptive procedures were optimized in order to yield an almost even distribution of presentation levels covering the whole auditory dynamic range of the listener without presenting stimuli outside this range. The evaluation measurements of the Oldenburg–ACALOS procedure with normal–hearing and hearing–impaired listeners showed that the use of a linear model loudness function produces loudness function estimates with a much smaller accuracy than predicted by the statistical model. The use of a non–linear model loudness function increased the accuracy of the procedure in a degree that it was consistent with the predictions.

This model loudness function was found in an empirical investigation into the shapes of loudness functions in categorical loudness scaling. This investigation is described in chapter 5. Several authors proposed different model loudness functions (e.g., Fechner, 1888; Stevens, 1956; Allen *et al.*, 1990; Nowak, 1990; Boretzki *et al.*, 1994; Hohmann and Kollmeier, 1995b; Heller *et al.*, 1997; Brand *et al.*, 1997c). The data which were collected during the evaluation of the Oldenburg–ACALOS procedure (cf. chapter 4) were used to derive reference loudness curves for the different normal–hearing and hearing–impaired listeners. The model function which yielded the smallest bias and the smallest intraindividual standard deviation in response level estimates consists of two straight lines which are connected at the 'medium'–level and smoothed in the transition area.

The use of this model function allows to detect differences in the shape of individual loudness functions. Chapter 6 investigates the influence of the stimulus type on the shape of the loudness function in normal-hearing and hearing-impaired listeners. The shape of the loudness function hardly depends on the center frequency but clearly on the bandwidth of stimuli. These findings are very consistent with other studies (e.g., Zwicker *et al.*, 1957) which applied loudness matching and found that loudness summation is maximal at medium levels and minimal at the limits of the auditory range. The influence of the stimulus type on the shape of the loudness function has important impacts on the input/output functions of hearing aids that aim at restoring the loudness in hearing–impaired listeners with recruitment.

Taken together, the innovative, computer–controlled methods proposed and evaluated in this thesis should supply the audiologist with sophisticated tools that advance the diagnosis and treatment of hearing impairment.

Chapter 2

Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests

ABSTRACT

The minimal standard deviations achievable for concurrent estimates of thresholds and psychometric function slopes as well as the optimal target values for adaptive procedures were calculated as functions of stimulus levels and track length on the basis of the binomial theory. Two approaches to adaptive procedures that converge at the optimal target values were introduced and tested with Monte–Carlo simulations. Both approaches utilize the fact that in some psychophysical tasks more than one statistically independent Bernoulli trial is performed per trial, as it is done, e.g., in sentence intelligibity tests that score each word separately. The first approach is a generalization of the adaptive procedure for sentence tests proposed by Hagerman and Kinnefors (1995). The second approach is a modification of the transformed up/down procedures proposed by Levitt (1971). The first approach yields higher efficiency. When the target discrimination value at which the adaptive procedure should converge is set to 50 %, a biasfree SRT estimate with a standard deviation only about 1.3 times higher than the theoretically optimally achievable value is reached using 20 trials. When the adaptive procedure converges at the target discrimination values 20 and 80 % in a randomized interleaved order, it achieves standard deviations in both SRT and slope estimates which are about 1.5 times higher than the optimal achievable values, with a small bias in slope estimates, using 30 trials. The adaptive procedures according to Hagerman and Kinnefors were evaluated

with normal–hearing and hearing–impaired listeners using two similar German sentence tests, namely the Göttingen sentence test (Kollmeier and Wesselkamp, 1997) and the Oldenburg sentence test (Wagener *et al.*, 1999c). These two tests differ in j factors which denote the number of statistically independently recognized elements per sentence trial (Boothroyd and Nittrouer, 1988). The measurements using the Göttingen sentences ($j \approx 2$) gave as accurate results as predicted by the simulations. The measurements using the Oldenburg sentences ($j \approx 4$) gave less accurate results than predicted which was probably due to a reduction of the j factor caused by the adaptive level placement.

INTRODUCTION

In many psychophysical questions, the experimenter is interested not only in the threshold but also in the slope of the psychometric function. Unfortunately, an accurate assessment of both threshold and slope requires much more measurement time than the assessment of the threshold alone. This study investigates how thresholds and psychometric function slopes can be assessed concurrently and as efficiently as possible. This study concentrates on sentence intelligibility tests, while some of the methods introduced here might be applied to other psychophysical tasks as well.

Sentence intelligibility tests for measurements of speech–reception thresholds (SRTs) in noise require a high precision of about 1 dB in order to differentiate between different acoustical situations (such as, e.g., spatial distributions of target speaker and interfering noise sources (Bronkhorst and Plomp, 1988; Peissig and Kollmeier, 1997)) and between different listeners (such as, e.g., normal–hearing listeners and hearing–impaired listeners with various impairments). On the other hand, high efficiency is necessary, because measurement procedures which take too much time before yielding reliable results are not practicable in clinical audiometry and hearing aid fitting.

Many sentence tests (e.g., Plomp and Mimpen, 1979; Hagerman, 1982; Nilsson *et al.*, 1994; Kollmeier and Wesselkamp, 1997; Wagener *et al.*, 1999a) show discrimination function slopes of between 15 and 25 %/dB which are considerably steeper than the values obtained with single–word tests. Since the standard deviation of SRT estimates is inversely proportional to the slope of the discrimination function, these sentence tests are better suited for efficient and reliable SRT measurements compared to single–word tests. An interfering effect, however, is that many hearing–impaired listeners show flatter discrimination functions than normal–hearing listeners. One reason might be that the speech materials are optimized in order to yield steep discrimination functions in normal–hearing listeners, which is not necessarily adequate to the individual's hearing impairment. A flatter discrimination function in a hearing–impaired listener indicates that the reception of the different speech elements such as phonemes is disturbed by noise in a different way than for normal listeners. A further audiological relevance of

the discrimination function slope is based on the fact that the SRT is related to an intelligibility value of 50 % which is a value too low for a satisfying conversation. Much more relevant to the listener's communication in noise is the range of SNRs, where intelligibility is beyond about 80 %. For a constant SRT level, this relevant level range increases with decreasing slope. Thus, a valid estimate of the SRT and the slope is required in order to assess speech recognition in noise.

While it is known that efficient SRT measurements are possible with sentence tests, this study investigates, whether an accurate concurrent estimate of SRT and slope is possible within a tolerable measuring time. The standard deviation of slope estimates is proportional to the slope of the underlying discrimination function. Typical slope values range from 0.2 dB^{-1} (20 % increase in intelligibility for 1 dB increase in SNR) for normal–hearing listeners to 0.05 dB^{-1} for severely hearing–impaired listeners. Aiming, e.g., at differentiating between high, intermediate and low slope values significantly, a relative intraindividual standard deviation of below 25 % of the actual slope value has to be reached.

The measurement time necessary to obtain sufficient accuracy in SRT and slope estimates is determined by the presentation levels, the speech material, and the number of sentences used in the measurement. The accuracy of SRT and slope estimates can be calculated as a function of the number of trials and of the placement of presentation levels based on the binomial theory. For this calculation, the predictability of the speech material is essential, which is characterized by the j factor that denotes the number of statistically independently percepted elements per sentence (Boothroyd and Nittrouer, 1988). Since the accuracy of estimates is inversely proportional to the square root of the number of statistically independent elements tested during the track, a higher j factor yields a higher accuracy within the same number of sentences, thus providing a more efficient estimate.

Adaptive procedures can be used to concentrate presentation levels in the range which yields the smallest standard deviations in SRT and slope estimates. A variety of adaptive procedures have been described in the literature, such as PEST (Parameter Estimation by Sequential Testing) (Taylor and Creelman, 1967), BUDTIF (Block–Up–and–Down, Two–Interval, Forced–choice) (Campbell, 1974) and UDTR (Up–Down Transformed Response) (Levitt, 1971). All of these procedures are designed for conditions with only two possible results per trial, e.g., *true* and *false* in forced choice tasks or *yes* or *no* in *yes–no* tasks. In sentence tests, however, each word can be scored independently, so that each trial has more than two possible outcomes. This additional information can be used to design adaptive procedures which converge more efficiently than the procedures mentioned above. In this study, two different approaches to adaptive procedures for sentence tests are investigated with respect to their efficiency of SRT and slope estimation. The first approach is a generalization of the procedure of Hagerman and Kinnefors (1995). In the second approach, modified versions of the transformed up–down procedures (Levitt,

1971) are used.

Kollmeier *et al.* (1988) investigated different adaptive staircase procedures with the help of Marcov chains. This approach permits to derive variances and biases of psychometric function parameter estimates given a fixed number of trials analytically without any random sampling. In this study, however, Monte–Carlo simulations are applied to investigate the accuracy and efficiency of the adaptive procedures.

To evaluate the results of the simulations, two different German sentence intelligibility tests, the Göttingen sentence test (Kollmeier and Wesselkamp, 1997) and the Oldenburg sentence test (Wagener *et al.*, 1999c; Wagener *et al.*, 1999b; Wagener *et al.*, 1999a) are applied to normal–hearing and hearing–impaired listeners. These two tests were recorded with the same talker and have nearly the same SRT and slope values in normal–hearing listeners, but differ in predictability.

In Sec. 2.1 the optimal presentation levels are calculated as well as the minimal standard deviations achievable at these levels in order to estimate either SRT values only or to estimate SRT and slope values concurrently. Five different adaptive procedures are presented which converge at these optimal presentation levels. In Sec. 2.4 these different adaptive procedures are compared with respect to their rate of convergence and accuracy using Monte–Carlo simulations. In Sec. 2.5 the two best adaptive procedures from the simulations are evaluated, using the Göttingen and the Oldenburg sentence test in normal–hearing and hearing–impaired listeners.

2.1 STATISTICAL METHOD

2.1.1 Discrimination function

The discrimination function describes the listener's speech intelligibility or discrimination, respectively, as a function of speech level. Speech intelligibility is defined as the probability p that the words of a sentence are repeated correctly by the listener. The speech level L may either refer to the sound pressure level of the speech signal in conditions without competitive noise or to the SNR, if the test is performed under noise. Similar definitions hold for psychometric functions in different psychophysical experiments.

The logistic discrimination function (2.1) is a close approximation to the normal distribution function which is based on the signal detection theory (Green and Swets, 1966).

$$p(L, L_{50}, s_{50}) = \frac{1}{1 + \exp(4 \cdot s_{50} \cdot (L_{50} - L))} \tag{2.1}$$

The parameter L_{50} denotes the speech reception threshold (SRT) which belongs to a 50 % probability of correct responses. It is equal to the mean μ of the normal distribution

function. The parameter s_{50} denotes the slope of the discrimination function at L_{50}. It is equal to $\frac{1}{\sqrt{2\pi}\,\sigma}$, with σ denoting the standard deviation of the normal distribution function. As the logistic function is much easier to compute than the normal distibution function, it is used in many studies and in this study as well to model the discrimination function in Monte–Carlo simulations and to fit the experimental data.

While the diagnostical utility of the L_{50} parameter is well established, the reduction of the slope s_{50} can be interpreted as an increased variability of the word–specific reception thresholds $L_{50,\mathrm{word}}$. This may be found in hearing–impaired listeners, because the sentence tests are usually optimized for normal–hearing listeners to produce a minimum variability in $L_{50,\mathrm{word}}$ across words in a test list. The relationship between the slope of the discrimination function and the distribution of the $L_{50,\mathrm{word}}$ values for different words can be calculated with a probabilistic model by Kollmeier (1990):
We assume that discrimination functions for single words differ in $L_{50,\mathrm{word}}$, but are approximately equal in s_{50}. If a sentence list consists of M different words, the discrimination function $p_{\mathrm{list}}(L)$ of this list is equal to the mean of the discrimination functions of the single words

$$p_{\mathrm{list}}(L) = \frac{1}{M}\sum_{i=1}^{M} p_{\mathrm{word}}(L - L_i), \tag{2.2}$$

with L_i denoting the L_{50} value of word i and $p_{\mathrm{word}}(L)$ denoting the logistic function with an L_{50} equal to 0 and an s_{50} equal to the common slope of the single words. For large M the sum in Eq. (2.2) can be substituted by a convolution between the model discrimination function of the single words p_{word} and the distribution density function of the threshold levels $f(L)$:

$$p_{\mathrm{list}}(L) = \int_{-\infty}^{\infty} p_{\mathrm{word}}(L - L') \cdot f(L')dL = p_{\mathrm{word}}(L) * f(L). \tag{2.3}$$

Furthermore, it is assumed that $p_{\mathrm{word}}(L)$ can be approximated by a normal distribution function (standard deviation σ_{word}) and that $f(L)$ is normal distributed (standard deviation σ_f). Hemce, it follows that:

$$\begin{aligned}\frac{dp_{\mathrm{list}}}{dL} &= \frac{\exp\left(\frac{-L^2}{2\sigma_{\mathrm{word}}^2}\right)}{\sigma_{\mathrm{word}} \cdot \sqrt{2\pi}} * \frac{\exp\left(\frac{-L^2}{2\sigma_f^2}\right)}{\sigma_f \cdot \sqrt{2\pi}} \\ &= \frac{\exp\left(\frac{-L^2}{2(\sigma_{\mathrm{word}}^2 + \sigma_f^2)}\right)}{\sqrt{2\pi(\sigma_{\mathrm{word}}^2 + \sigma_f^2)}}\end{aligned} \tag{2.4}$$

Thus, in good approximation $p_{\text{list}}(L)$ is a normal distribution function that can be approximated by Eq. (2.1) as well. Because of the convolution with $f(L)$, the slope of $p_{\text{list}}(L)$ decreases by

$$\frac{dp_{\text{list}}}{dL} = \frac{dp_{\text{word}}}{dL} \cdot \left(\sqrt{1 + \frac{\sigma_f^2}{\sigma_{\text{word}}^2}} \right)^{-1} . \tag{2.5}$$

In practice, the discrimination function of a whole test list is always estimated for a given listener. Therefore, p_{list} is denoted shortly with p in this study.

2.1.2 Variability of parameter estimates and optimal presentation levels

2.1.2.1 Binomial variance and predictability of speech

In a sentence test the response to each word can be assumed as a Bernoulli trial with the probability $p(L)$ of a correct outcome given by the discrimination function. Since the variance of one Bernoulli trial is $p(1-p)$, the standard deviation σ_p of an intelligibility estimate based on n trials is:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}} \tag{2.6}$$

Because of context effects prevalent in the perception of speech, the number n of statistically independent Bernoulli trials is smaller than the number of words. If one or more words of a sentence have been recognized, the probability to recognize the remainder increases. According to Boothroyd and Nittrouer (1988) the number of statistically independent elements in a sentence can be estimated by the j factor:[1]

$$j = \frac{\log(p_s)}{\log(p_w)}. \tag{2.7}$$

The value p_s denotes the probability to understand a sentence completely, i.e., to understand all words of a sentence correctly, and p_w denotes the probability to correctly understand each single word of the sentence separately. To take the context effect into account, the number of Bernoulli trials n in Eq. (2.6) has to be set to jN, with N denoting the number of sentences tested.

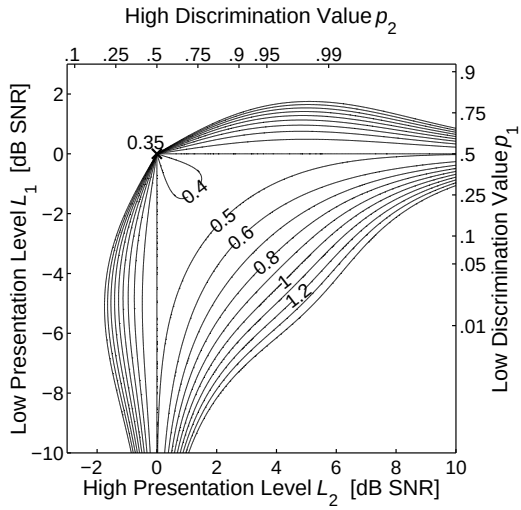

Figure 2.1: *Contour plot of the standard deviation $\sigma_{L_{50}}$ of the SRT estimation, when two different presentation levels are used. The contour lines show $\sigma_{L_{50}}$ in dB. The cross indicates the* sweetpoint*, i.e., the minimum of $\sigma_{L_{50}}$. The discrimination function is logistic with an L_{50} of 0 dB and an s_{50} of 0.2 dB^{-1}. $N_1 = N_2 = 10$ sentences were performed at each level. The number of statistically independent words per sentence is $j_1 = j_2 = 2.5$. The presentation levels are given both as levels in dB units and as discrimination values.*

2.1.2.2 Efficient SRT estimation – the *sweetpoint*

The main result of a sentence intelligibility test is the SRT value which is defined as the presentation level related to a certain arbitrarily chosen discrimination value. Since the standard deviation σ_{SRT} of the SRT estimate is minimal for the logistic discrimination function, if all trials are presented at $p = 0.5$ (Levitt, 1971; Laming and Marsh, 1988; Green, 1990), the L_{50} is often called *sweetpoint* and usually the SRT is defined as the level that is related to $p = 0.5$. To a first approximation, the value of σ_{SRT} is equal to the standard deviation of the intelligibility estimate σ_p divided by the slope of the

[1] The j factor is a comparatively rough model for the predictability of speech compared to other models (Boothroyd and Nittrouer, 1988; Bronkhorst *et al.*, 1993). However, the j factor has the advantage that it can be calculated from the sentence responses without any further measurements like intelligibility estimates of the isolated words of a test.

discrimination function at the SRT:

$$\sigma_{SRT} = \sqrt{\frac{p(1-p)}{Nj}} \cdot \left(\left. \frac{dp(L)}{dL} \right|_{SRT} \right)^{-1} \tag{2.8}$$

That means, the value of $\sigma_{L_{50}}$ at the *sweetpoint* is inversely proportional to s_{50}. If it is assumed that (1) j is independent of L and (2) that N sentences are presented at the *sweetpoint*, the asymptotic minimum value of $\sigma_{L_{50}}$ for large values of N is

$$\sigma_{L_{50}}(\text{sweetpoint}) = \frac{0.5}{s_{50} \cdot \sqrt{Nj}}. \tag{2.9}$$

The efficiency of an adaptive procedure in SRT estimates can be quantified by the normalized standard deviation $\hat{\sigma}_{L_{50}} = \frac{\sigma_{L_{50}}}{\sigma_{L_{50}}(\text{sweetpoint})}$, with $\sigma_{L_{50}}$ denoting the empirical standard deviation in SRT estimates provided by the procedure. The normalized standard deviation is equal to the reciprocal square root of the *efficiency* as defined by Taylor and Creelman (1967).

The considerations above imply that words are presented at a constant presentation level. However, at least two different presentation levels are required to estimate L_{50} and s_{50} concurrently. Since Eq. (2.1) has two free parameters, the discrimination function is completely determined by measurements at two levels. If two different presentation levels L_1 and L_2 are used in one set of measurements, two discrimination values p_1 and p_2 are estimated by N_1 and N_2 sentences, respectively. Consequently, L_{50} can be calculated by solving a system of two logistic functions and two variables L_{50} and s_{50}. According to the Gaussian error law the standard deviation $\sigma_{L_{50}}$ of L_{50} can be calculated as

$$\sigma_{L_{50}} = \sqrt{\frac{1}{j_1 N_1} \cdot \left(\frac{\partial L_{50}}{\partial p_1} \right)^2 \cdot p_1(1-p_1) + \frac{1}{j_2 N_2} \cdot \left(\frac{\partial L_{50}}{\partial p_2} \right)^2 \cdot p_2(1-p_2)}. \tag{2.10}$$

Fig. 2.1 shows $\sigma_{L_{50}}$ as a function of (L_1, L_2) and $(p_1(L_1), p_2(L_2))$, respectively. The numbers of statistically independent elements per sentence j_1 and j_2 are set to 2.5 and $N_1 = N_2 = 10$ sentences are performed per level. The L_{50} parameter is set to 0 dB and s_{50} is set to 0.2 dB^{-1}. These values refer to typical parameters of a test employing short, meaningful sentences. The minimal deviation $\sigma_{L_{50}}$ in this condition is found to be 0.35 dB. This value is accomplished, if both stimuli are presented exactly at the *sweetpoint* ($L_1 = L_2 = 0$ dB and $p_1 = p_2 = 0.5$, respectively). In this case, Eq. (2.10) is equivalent to Eq. (2.8). However, $\sigma_{L_{50}}(L_1, L_2)$ is not continuous in this point. If only one of the two stimuli is presented at the *sweetpoint*, σ_{L50} becomes 0.5 dB and the other stimuli has no influence on σ_{L50}.

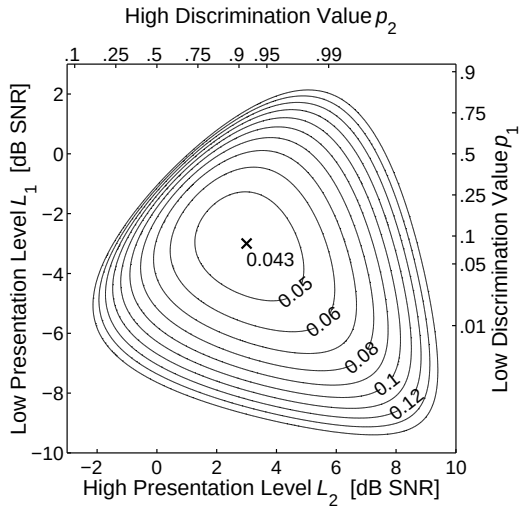

Figure 2.2: *Contour plot of the standard deviation $\sigma_{s_{50}}$ of the s_{50} estimation in the same condition and representation as given in Fig. (2.1). The contour lines show $\sigma_{s_{50}}$ in dB^{-1}. The cross indicates the* sweetpair, *i.e. the minimum of $\sigma_{s_{50}}$.*

2.1.2.3 Efficient slope estimation – the *sweetpair*

In analogy to Eq. (2.10) the standard deviation $\sigma_{s_{50}}$ of the s_{50} estimate can be derived as:

$$\sigma_{s_{50}} = \sqrt{\frac{1}{j_1 N_1} \cdot \left(\frac{\partial s_{50}}{\partial p_1}\right)^2 \cdot p_1(1-p_1) + \frac{1}{j_2 N_2} \cdot \left(\frac{\partial s_{50}}{\partial p_2}\right)^2 \cdot p_2(1-p_2)} \tag{2.11}$$

Numerical minimization of $\sigma_{s_{50}}$ gives the pair of discrimination values ($p_1 = 0.083, p_2 = 0.917$) with the minimal standard deviation of slope estimation. In analogy to the *sweetpoint* we denote this pair of stimulus levels as '*sweetpair*'. The value of $\sigma_{s_{50}}$ at the *sweetpair* is proportional to s_{50}. Under the simplifying assumption that $j_1 = j_2 = j$ is independent of the level and that $N_1 = N_2 = N$ sentences are performed at both levels of the *sweetpair*, the asymptotic minimum value of $\sigma_{s_{50}}$ for large values of N can be derived numerically as:

$$\sigma_{s_{50}}(\text{sweetpair}) = \frac{1.07 \cdot s_{50}}{\sqrt{Nj}}. \tag{2.12}$$

The efficiency of an adaptive procedure that estimates the slope of the discrimination function can be quantified by the normalized standard deviation $\hat{\sigma}_{s_{50}} = \frac{\sigma_{s_{50}}}{\sigma_{s_{50}}(\text{sweetpair})}$, with $\sigma_{s_{50}}$ denoting the standard deviation in slope estimates provided by the procedure.

As the *sweetpoint*, the *sweetpair* is independent of the actual L_{50} and s_{50} values and only depends on the mathematical form of the discrimination function. For the logistic discrimination function, the two discrimination values of the *sweetpair* are $p_1 = 0.083$ and $p_2 = 0.917$. These values are placed symmetrically around $p = 0.5$ at a relatively large distance. Thus, an adaptive procedure optimized to estimate s_{50} effectively has to present some trials at very high levels and some trials at very low levels. This may cause problems in practice, because some adaptive procedures have difficulties to converge at target discrimination values close to 0 and 1, respectively, and because some listeners' attention may be confused by these extreme discrimination values.

Fig. 2.2 shows $\sigma_{s_{50}}$ as a function of (L_1, L_2) and $(p_1(L_1), p_2(L_2))$, respectively. The conditions are the same as in Fig. 2.1. The minimal $\sigma_{s_{50}}$ is 0.043 dB^{-1} which is reached at the *sweetpair* (L_1 = -3.00 dB, L_2 = 3.00 dB and $p_1 = 0.083$, $p_2 = 0.917$, respectively). Wetherhill (1963) and O'Regan and Humbert (1989) calculated the same optimal values for p_1 and p_2 for the logistic function with the second derivatives of the likelihood function. Levitt (1971) calculated the optimal presentation levels for slope estimates for the cumulative normal psychometric function. He proposed to place observations at a distance of 1.57σ on either side of μ, that is, at $p = 0.058$ and $p = 0.942$, respectively. This shows that the form of the underlying discrimination function – whether logistic or normal – has only a small influence on the placement of the *sweetpair*.

2.1.2.4 Concurrent SRT and slope estimation – the *pair of compromise*

Since the discrimination values of the *sweetpoint* ($p = 0.5$) and of the *sweetpair* ($p_1 = 0.08$, $p_2 = 0.92$) differ considerably, an efficient concurrent SRT and slope estimate requires a compromise between the accuracies of L_{50} and s_{50} estimates. One possibility to quantify the accuracy of a common L_{50} and s_{50} estimate is to calculate the quadratic mean of the standard deviations $\sigma_{L_{50}}$ and $\sigma_{s_{50}}$ normalized by their respective minima, $\min(\sigma_{L_{50}})$ and $\min(\sigma_{s_{50}})$:

$$\sigma_{comp} = \frac{1}{\sqrt{2}} \cdot \sqrt{\left(\frac{\sigma_{L_{50}}}{\min(\sigma_{L_{50}})}\right)^2 + \left(\frac{\sigma_{s_{50}}}{\min(\sigma_{s_{50}})}\right)^2} \tag{2.13}$$

Fig. 2.3 shows σ_{comp} as a function of (L_1, L_2) and ($p_1(L_1)$, $p_2(L_2)$), respectively. The conditions were the same as in Figs. 2.1 and 2.2. The minimum of σ_{comp} is reached at $L_{1,comp}$ = -1.82 dB, $L_{2,comp}$ = 1.82 dB and $p_{1,comp} = 0.19$, $p_{2,comp} = 0.81$, respectively. Wetherhill (1963) and O'Regan and Humbert (1989) derived similar values ($p_{1,comp} = 0.176$ and $p_{2,comp} = 0.823$) from the second derivatives of the likelihood function. Levitt (1971) proposed similar values for a good compromise in estimating both σ and μ of the cumulative normal psychometric function. He proposed to place observations at distance σ on either side of μ, that is at $p = 0.159$ and $p = 0.841$.

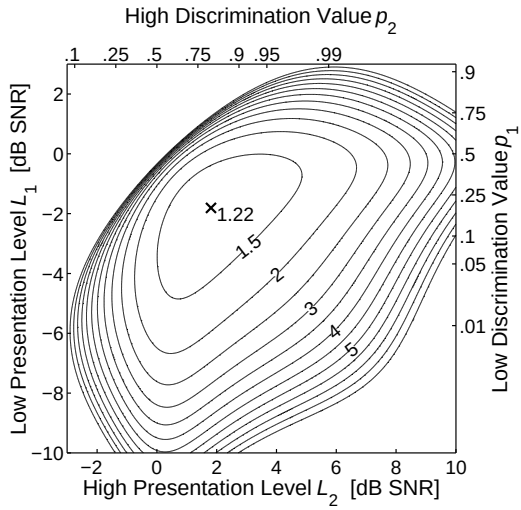

Figure 2.3: *Standard deviation σ_{comp} in the same condition and representation as given in Figs. (2.1) and (2.2). The contour lines show σ_{comp}. The cross indicates the* pair of compromise, *i.e. the minimum of σ_{comp}.*

2.1.3 Maximum likelihood estimator

In order to fit the discrimination function to the data, a maximum likelihood method was employed: If a test list with a total of m words is presented to the listener, the likelihood of a given discrimination function $p(L, L_{50}, s_{50})$ is

$$l(p(L, L_{50}, s_{50})) = \prod_{k=1}^{m} p(L_k, L_{50}, s_{50})^{c(k)}[1 - p(L_k, L_{50}, s_{50})]^{1-c(k)}, \tag{2.14}$$

with $c(k) = 1$, if the word k was repeated correctly and $c(k) = 0$, if the word k was not repeated correctly. The discrimination function with the maximum likelihood is determined by varying the parameters L_{50} and s_{50} until $\log(l(p(L, L_{50}, s_{50})))$ is maximal.

2.1.4 Monte–Carlo simulations

For each simulation condition of this study, 2,000 Monte–Carlo runs were performed for the logistic discrimination functions with s_{50} values of 0.05, 0.1, 0.15, 0.2, and 0.25 dB^{-1}. Without loss of generality, L_{50} was set to 0 dB in all simulations. The initial level of each simulation was chosen randomly from a uniform distribution with the limits ± 10 dB. To account for the j factor, a number of j Bernoulli trials with a probability of success $p(L)$

were performed per sentence. The discrimination value for each sentence was calculated by dividing the sum of the Bernoulli trials results by j. Thus, only integer values of j could be simulated, which is a simplification, because in real speech noninteger j factors occur as well.

2.2 EXPERIMENTAL METHOD

2.2.1 Apparatus

A computer–controlled audiometry workstation was used which was developed within a German joint research project on speech audiometry (Kollmeier *et al.*, 1992). A personal computer with a coprocessor board (Ariel DSP 32C) with 16–bit stereo AD–DA converters was used to control the complete experiment as well as stimulus presentation and recording of the listener's responses. The stimulus levels were adjusted by a computer–controlled custom–designed audiometer comprising attenuators, anti–aliasing filters, and headphone amplifiers. Signals were presented monaurally to the listeners with Sennheiser HDA 200 headphones with free–field equalization. The listeners were situated in a sound–insulated booth. Their task was to repeat each sentence presented over headphones as closely as possible. The instructor, also situated in the booth in front of the listener, marked each incorrectly repeated word. For this purpose, an Epson EHT 10S handheld computer was used with an LCD touchscreen on which the target sentence was displayed. The handheld computer was connected to the personal computer via serial interface. The exact test outcome (i.e., each correctly identified word for each listener) was stored by the computer for later statistical analysis.

2.2.2 Speech materials

Two similar sentence tests, the Göttingen and the Oldenburg sentence test, which differ mainly in predictability, were used in the measurements:
For both tests, speech and noise were added digitally at a predefined signal–to–noise ratio and converted to analog (16 bits, 25 kHz sampling frequency).

The Göttingen sentence test (Kollmeier and Wesselkamp, 1997) consists of 20 lists with 10 sentences each, recorded with a male unschooled speaker. The average number of words per sentence is 5. The predictability of the speech material is high, i.e., $j = 1.95$ at an SNR of -8 dB ($p = 0.21$) and $j = 2.38$ at an SNR of -4 dB ($p = 0.84$) (Kollmeier and Wesselkamp, 1997). The different lists have equivalent discrimination functions in a speech spectrum–shaped noise. Measurements with normal–hearing listeners with a noise level of 65 dB SPL gave an L_{50} of -6.23 dB SNR with a standard deviation of

0.27 dB between lists. The slope of the discrimination function at the SRT is 0.192 dB^{-1} with a standard deviation of 0.025 dB^{-1} between lists (Kollmeier and Wesselkamp, 1997).

The Oldenburg sentence test (Wagener *et al.*, 1999c; Wagener *et al.*, 1999a) consists of 120 lists with 30 sentences each recorded with the same speaker as the Göttingen sentence test. Each sentence consists of 5 words and has the same syntactical form: *Name verb number adjective object*. The predictability of the speech material is low, i.e., $j = 3.18$ at an SNR of -9 dB ($p = 0.22$) and $j = 4.29$ at an SNR of -5 dB ($p = 0.81$) (Wagener *et al.*, 1999a). The different lists have equivalent discrimination functions in a speech spectrum–shaped noise. Measurements with normal–hearing listeners with a noise level of 65 dB SPL gave an L_{50} of -7.11 dB SNR with a standard deviation of 0.16 dB between lists (Wagener *et al.*, 1999a). The slope of the discrimination function at the SRT is 0.171 dB^{-1} with a standard deviation of 0.0165 dB^{-1} between lists (Wagener *et al.*, 1999a).

2.3 ADAPTIVE PROCEDURES

2.3.1 Procedure A

A generalization of the procedure of Hagerman and Kinnefors (1995) is proposed which changes the presentation level of the subsequent sentence by:

$$\Delta L = -\frac{f(i) \cdot (prev - tar)}{slope} \tag{2.15}$$

The parameter *tar* denotes the target discrimination value at which the procedure should converge, *prev* denotes the discrimination value obtained in the previous sentence which is used as input to the adaptive level setting. The *slope* parameter is set to 0.15 dB^{-1} in this study which is a medium value for the sentence tests used in this study (cf. 2.2.2). The parameter $f(i)$ controls the rate of convergence. Its value depends on the number i of reversals of presentation level. The sequence $f(i)$ has to start at values above one to allow for large steps at the beginning of the adaptive procedure, where the current level might be far from the target. With increasing number of reversals $f(i)$ has to decrease to stabilize presentation levels near the target. The influence of $f(i)$ is investigated with Monte–Carlo simulations, cf. Sec. 2.4.1. The sequence $f(i) = 1.5 \cdot 1.41^{-i}$ yielded the optimal efficiency in the simulations. With $f(i) = 1$, $tar = 0.4$ and $slope = 0.2$ Eq. (2.15) is equal to the adaptive rule of Hagerman and Kinnefors (1995).
If only the SRT should be measured, *tar* is set to 0.5 (*sweetpoint*) to yield an optimal SRT estimate. In this condition, the adaptive procedure is called A1.
If L_{50} and s_{50} should be measured concurrently, two randomly interleaved tracks are

used which converge at $p_1 = 0.2$ and $p_2 = 0.8$ (*pair of compromise*).[2] This condition of the adaptive procedure using the targets 0.2 and 0.8 in a randomly interleaved order with random switch between two concurrent adaptive procedures is called A2.

2.3.2 Procedure B

In procedures A1 and A2 the calculation of ΔL is based on only one previous sentence, i.e., j Bernoulli trials. This may be insufficient for highly predictable speech materials such as the Göttingen sentence test which has a relatively small number of statistically independent elements per sentence ($j \approx 2$).[3] Transformed up/down procedures (Levitt, 1971) use more than one previous trial to calculate the next presentation level. Hence, they may be somewhat more reliable in their level placement. Table 2.1 shows the adaptive rules for five different transformed up–down procedures with the target intelligibilities $p = 0.16,\ 0.27,\ 0.5,\ 0.73,\ 0.84$ which are adapted to experiments with two Bernoulli trials per trial. Assuming that a sentence consists of two statistically independent elements, it follows that both Bernoulli trials have succeeded (denoted with '++' in Table 2.1), if a complete sentence is repeated correctly. If only some words of the sentence are repeated correctly, one Bernoulli trial has succeeded and the other has failed (denoted with '+−' and '−+'). If no word is repeated correctly, both Bernoulli trials have failed (denoted with '−−').
With these adaptive rules three procedures are implemented: Procedure B1 converges at $p = 0.5$ in order to yield efficient SRT estimates. Procedure B2 converges at $p = 0.84$ and 0.16 in randomly interleaved order in order to yield concurrent SRT and slope estimates, efficiently. These targets are very close to the discrimination values of the *pair of compromise*. However, the difference between targets is very large – it is even larger than in the *pair of compromise* ($p_{1,comp} = 0.19$, $p_{2,comp} = 0.81$). Hence, the somewhat closer targets probabilities 0.73 and 0.27 are used in procedure B3.

A disadvantage of procedures B2 and B3 is that their target discrimination values depend on j. The difference between the targets increases with increasing j. Therefore, they are adequate only for speech materials with a j of an almost constant value of 2.
In all procedures, the step size ΔL of the change in level is given by the sequence $f(i) = 4 \cdot 2^{-i}$, with i denoting the number of reversals of level. This sequence yields optimal results in the Monte–Carlo simulations in Sec. 2.4.1.

[2] The discrimination values of the *pair of compromise* ($p_{1,comp} = 0.19$ and $p_{2,comp} = 0.81$) were rounded corresponding to the integer values 1 and 4 of correctly repeated words in a five–word sentence.

[3] Note that procedures A1 and A2 are generalizations of the procedure of Hagerman and Kinnefors (1995) which was reported to be adequate for the test of Hagerman (1982) which has a j value of about 4.

Table 2.1: *Adaptive rules for modified transformed up/down procedures for experiments with two statistically independent elements per trial. A '+' indicates a correctly repeated element, a '−' indicates an element which was not repeated correctly.*

Target	increase	decrease	stay
$p = 0.5$	−−	++	+−
			−+

Targets	increase	decrease	Targets	increase	decrease
$p = 0.84$	−−	++ ++	$p = 0.73$	−−	++ ++
	−+			−+	++ +−
	+−			+−	++ −+
	++ −−			++ −−	
	++ −+				
	++ +−				
$p = 0.16$	−− −−	++	$p = 0.27$	−− −−	++
		+−		−− −+	+−
		−+		−− +−	−+
		−− ++			−− ++
		−− +−			
		−− −+			

2.4 SIMULATIONS

2.4.1 Rate of convergence

Since the discrimination function is fitted to the complete track, the efficiency of an adaptive procedure is optimal, if all stimuli during the whole track are placed as close to the target as possible.
The rate of convergence is controlled by $f(i)$ in Eq. (2.15) in procedures A1 and A2 and by the step size ΔL for procedures B1, B2 and B3, respectively. The exponential function $f(i) = a \cdot b^{-i}$ was chosen to parameterize the step size. It enables the occasionally employed rules of halving step size after each reversal ($b = 2$) and of halving step size after every second ($b = 1.41$), third ($b = 1.26$) and fourth ($b = 1.189$) reversal, respectively. Mean values and standard deviations of the presentation level L were derived as functions of trial number n by Monte–Carlo simulations.

Since procedures A1 and A2 are probably more adequate for low predictable speech materials, j was set to 4, which is a value typical of low predictable speech materials, such as the Hagerman sentence test (Hagerman, 1982; Hagerman, 1996) and the Oldenburg sentence test (Wagener *et al.*, 1999a). Simulations were carried out for all combinations of the parameters $a = 1, 1.5, 2, 2.5, 3, 4$ and $b = 1.189, 1.26, 1.41, 2$. The final value of

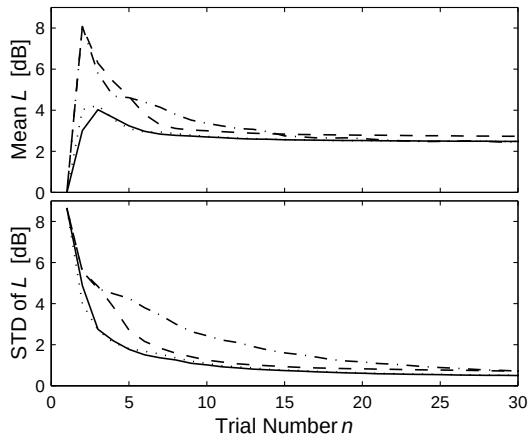

Figure 2.4: *Upper panel: Mean presentation level L of procedure A2 in Monte–Carlo simulations as a function of trial number n, for different parameter settings of the rate sequence* $f_i = a \cdot b^{-1}$. *The parameters are: Solid line:* $a = 1.5$, $b = 1.41$*; dotted line:* $a = 2$, $b = 1.41$*; dot–dashed line:* $a = 4$, $b = 1.189$*; dashed line:* $a = 4$, $b = 2$. *Lower panel: Related standard deviation of the presentation level L.*

$f(i)$ was limited to 0.1. This minimal value produced the best convergence for high trial numbers in a pilot study in which the final values 0.025, 0.05, 0.1 and 0.2 were tested. Because of the large number of tested conditions the results are only summarized briefly: Fig. 2.4 displays some examples. Mean values and standard deviations of the presentation level L are shown as functions of the trial number n for the adaptive procedure A2 and an underlying discrimination function with $s_{50} = 0.15$ dB^{-1}. The parameters $f_i = a \cdot b^{-i}$ shown in Fig. 2.4 are ($a = 1.5$, $b = 1.41$) (solid line), ($a = 2$, $b = 1.41$) (dotted line), ($a = 4$, $b = 1.189$) (dot–dashed) and ($a = 4$, $b = 2$) (dashed). The parameters ($a = 1.5$, $b = 1.41$) yielded the best convergence in this example. The same finding holds also for all other conditions that were tested for both procedure A1 and procedure A2 regarded over all of the tested slope values. However, the parameters ($a = 2$, $b = 1.41$) gave very similar results (the standard deviations were slightly smaller, but biases were slightly higher in the first 3 trials). Only parameter settings with $b = 2$ sometimes gave slightly better results for medium and high underlying slope values, but they caused high standard deviations and biases for an underlying slope value of $s_{50} = 0.05$. For all slope values, the most unfavourable condition was $a = 4$ and $b = 1.89$ (standard deviations were about 2 times higher than in the best condition). For these reasons, the procedures A1 and A2 are always used with the sequence $f(i) = 1.5 \cdot 1.41^{-i}$, with i denoting the number of reversals in the remainder of this study. The final value of $f(i)$

is set to 0.1.

Since procedures B1, B2 and B3 were optimized for $j \approx 2$, simulations for these procedures were performed for $j = 2$. For these procedures all combinations of the parameters $a = 4$ dB, 6 dB, 8 dB and $b = 1.189$, 1.26, 1.41, 2, 3 were tested. The minimal value of $f(i)$ was limited to 0.25 dB. For all conditions examined, the combination of the parameters $a = 4$ dB and $b = 2$ yielded the best convergence under almost all conditions with the exception of certain conditions with higher values of a which caused slightly smaller standard deviations in the first 2 to 3 trials. After a maximum of 3 trials, however, an initial step size of $a = 4$ dB was always favorable. Higher initial step sizes, probably, allow a faster convergence at the target region at the beginning of the track. However, higher initial step sizes also run the risk of overshooting the target. The second exception were conditions with the smallest slope value ($s_{50} = 0.05$ dB^{-1}), where $b = 1.41$ gave a slightly better convergence for procedures A2 and A3. However, with increasing slope, $b = 2$ was superior.
For these reasons, in the remainder of this study the procedures B1, B2 and B3 are used with a step size sequence $f(i) = 4 \cdot 2^{-i}$ dB, with i denoting the number of reversals and a minimum value of $f(i) = 0.25$ dB.

2.4.2 Accuracy of fit

For each algorithm the intraindividual standard deviations and biases in L_{50} and s_{50} estimates were predicted using Monte–Carlo simulations. All combinations of $s_{50} = 0.05$, 0.10, 0.15, 0.20 and 0.25 dB^{-1} and $j_1 = j_2 = 1$, 2, 3, 4, 5 and $N = 10$, 20, 30, 40 were performed. In order to simplify the simulations, j was kept constant with the level. In the interleaved tracks $N_1 = N_2 = \frac{N}{2}$ sentences were performed per target.

2.4.2.1 L_{50} estimates

Fig. 2.5 shows the normalized standard deviations $\hat{\sigma}_{L_{50}}$ of the L_{50} estimates, i.e., $\sigma_{L_{50}}$ relative to its minimal value at the *sweetpoint* for the respective settings of N, j and s_{50} given by Eq. (2.8). Each panel shows $\hat{\sigma}_{L_{50}}$ as a function of N for all procedures for different settings of j and N, respectively.
Except for $s_{50} = 0.05\text{dB}^{-1}$, the $\hat{\sigma}_{L_{50}}$ values of the respective procedures for constant settings of $N \geq 20$ and j scarcely vary with the underlying s_{50} value, i.e. the panels of a given row in Fig. 2.5 are very similar. That means that the efficiency of the algorithms is scarcely influenced by the discrimination function slope. Under most conditions, $\hat{\sigma}_{L_{50}}$ decreases with increasing N, although $\hat{\sigma}_{L_{50}}$ is already normalized with $\frac{1}{\sqrt{N}}$. This is certainly due to the fact that the adaptive procedures converge at their specific targets

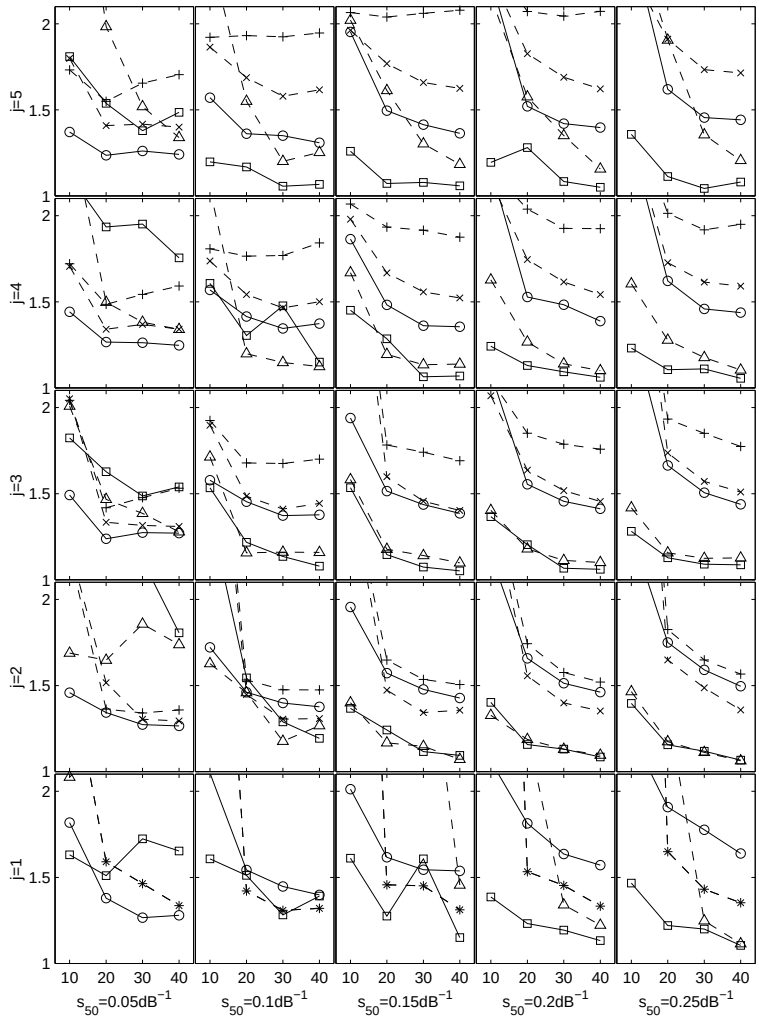

Figure 2.5: *Normalized standard deviations of L_{50} estimates as functions of N for different settings of s_{50} and j, respectively. The ordinate of each panel indicates $\hat{\sigma}_{L_{50}}$ (i.e. $\sigma_{L_{50}}$ divided by its minimum value at the sweetpoint according to Eq. (2.8)), the abscissa indicates N. Some conditions are not visible, since their ordinate value is outside the displayed range. Procedures: A1 (solid line with squares), A2 (solid line with circles), B1 (dashed line with triangles), B2 (dashed line with 'x'), B3 (dashed line with '+').*

during the track, which causes a larger portion of trials to be placed more efficiently.[4]
For all conditions with $s_{50} > 0.05$ dB^{-1}, procedures A1 and B1, which converge at the *sweetpoint*, yield better results than procedures A2, B2 and B3, which converge at the *pair of compromise*. The advantage of procedures A1 and B1 increases with increasing N. This is certainly due to the fact that the algorithms concentrate their presentation levels closer to their specific targets during the track and that the *sweetpoint* is better suited for threshold estimates.
Procedure A1 gives the best L_{50} estimates on average across all conditions. The $\hat{\sigma}_{L_{50}}$ values are always smaller than 1.6 for $N \geq 20$ and they decrease further with increasing j and with increasing N. Since $\hat{\sigma}_{L_{50}}$ is already normalized with $\frac{1}{\sqrt{j}}$, this means that the estimates can use the additional amount of statistically independent elements per trial even more than predicted by Eq. (2.9). Procedure B1 gives worse results than A1 for $j > 4$, which is probably due to the fact that B1 is optimized for $j = 2$, whereas A1 is optimized for large j factors. Nevertheless, A1 yields almost the same results as B1 for $j = 2$ and even better results for $j = 1$. Consequently, A1 is the optimal choice for efficient threshold estimates for all j factors.
For an underlying s_{50} value of 0.05 dB^{-1}, clearly different $\sigma_{L_{50}}$ values were derived for procedures A1, A2 and B1. Here, procedures A1 and B1 gave worse results than for other slope values and procedure A2 gave the smallest standard deviations of all procedures. These results contradict the theoretical expectation. They are probably due to the fact that the step size sequences of all procedures were optimized for a medium discrimination function slope of 0.15 dB^{-1} which is 3 times larger than 0.05 dB^{-1}. Probably, this mismatch between assumed and real slope caused the procedures to not work suitably.

2.4.2.2 Biases of L_{50} estimates

Fig. 2.6 shows the biases in L_{50} estimates as functions of N for the different settings of s_{50} and j, respectively. There is no tendency to systematic biases. This result has been expected, because all procedures generate symmetric level distributions around the L_{50}. The absolute value of the bias is maximal in procedures B2 and B3, for example 0.7 dB, for $j = 1$ and $N = 10$, and decreases rapidly with increasing N and increasing s_{50}. Since the biases in L_{50} estimates are always smaller than 10 % of the absolute standard deviation $\sigma_{L_{50}}$ in L_{50} estimates, they have no practical relevance. Probably, all L_{50} biases shown in Fig. 2.6 are caused by the statistical uncertainty of the Monte–Carlo simulations and can be reduced by a larger number of Monte–Carlo runs.

[4] The only conditions in which $\hat{\sigma}_{L_{50}}$ increases with increasing j are procedure B2 and B3 for large values of j. Since the target discrimination values of these procedures move to 0 and 1, respectively, with increasing j, the tracks remove from the *sweetpoint* causing a reduced efficiency.

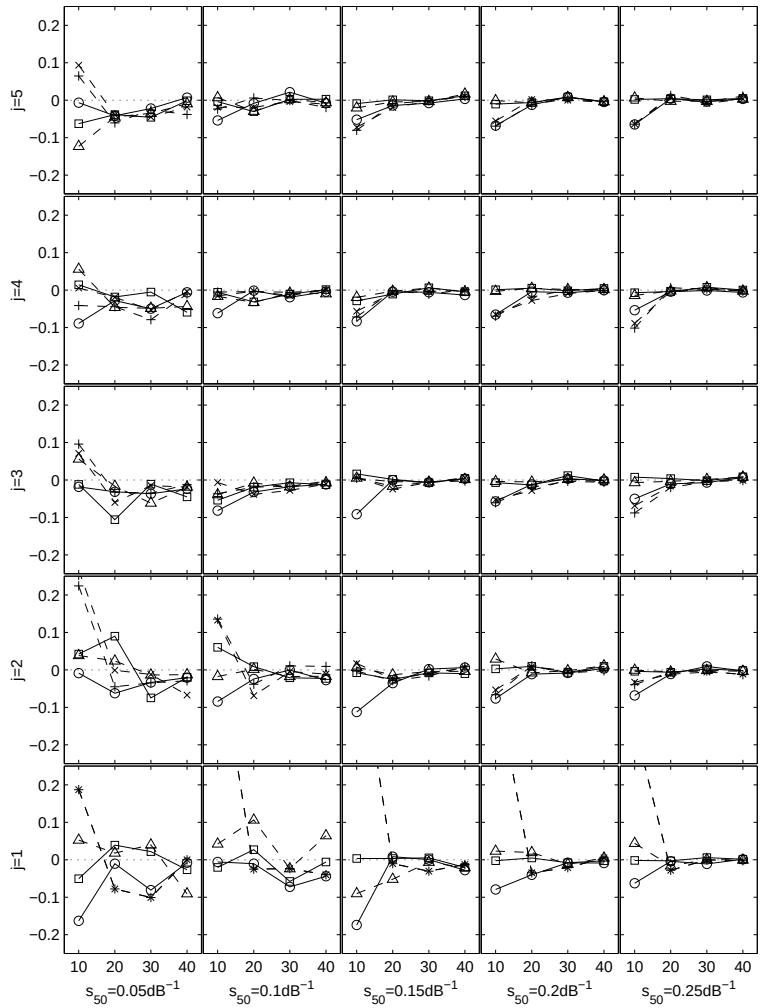

Figure 2.6: *Mean biases of L_{50} estimates as functions of N for different settings of s_{50} and j. The ordinate of each panel indicates the bias of L_{50} in dB SNR, the abscissa indicates N. Some conditions are not visible, since their ordinate value is outside the displayed range. Procedures: A1 (solid line with squares), A2 (solid line with circles), B1 (dashed line with triangles), B2 (dashed line with 'x'), B3 (dashed line with '+'). The horizontal dotted line denotes the zero bias line.*

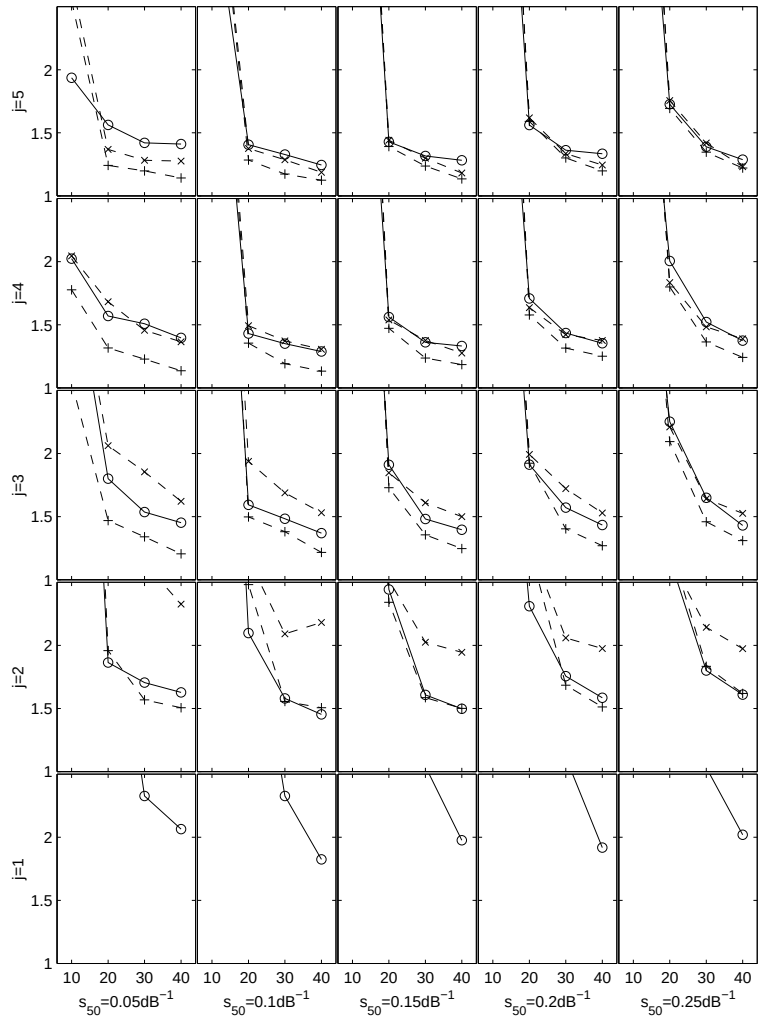

Figure 2.7: *Normalized standard deviations of s_{50} estimates as functions of N for different settings of s_{50} and j. The ordinate of each panel indicates $\hat{\sigma}_{s_{50}}$ (i.e. $\sigma_{s_{50}}$ divided by its minimum value at the* sweetpair *according to Eq. (2.12)), the abscissa indicates N. Some conditions are not visible, since their ordinate value is outside the displayed range. Procedures: A1 (solid line with squares), A2 (solid line with circles), B1 (dashed line with triangles), B2 (dashed line with 'x'), B3 (dashed line with '+').*

2.4.2.3 Standard deviations of slope estimates

Fig. 2.7 shows the normalized standard deviations of the s_{50} estimates as functions of N for the different settings of s_{50} and j.

As expected, the algorithms A1 and B1, which converge at the *sweetpoint* fail completely in slope estimates. For the respective equal values of j, N and s_{50} these procedures generate $\hat{\sigma}s_{50}$ values which are about 10 times larger than the results obtained with A2, B2 and B3, respectively, that converge at targets near the *pair of compromise*. For that reason, the results in slope estimates of A1 and B1 are out of the range of the ordinate in Fig. 2.7 and they are not further discussed in this section.
Although the $\hat{\sigma}_{s_{50}}$ values have been normalized by $1/\sqrt{N}$, all procedures yield decreasing $\hat{\sigma}_{s_{50}}$ values with increasing N. Since all procedures concentrate stimuli at the *pair of compromise* during the track, the standard deviation in slope estimates decreases faster than the $1/\sqrt{N}$ rule predicted by Eq. (2.12).
All procedures yield decreasing $\hat{\sigma}_{s_{50}}$ values with increasing j factor. That means, the standard deviation in slope estimates decreases faster than the $1/\sqrt{j}$ rule predicted by Eq. (2.12) as well. For procedure A2 this is probably due to the faster convergence at the target level which is caused by the more reliable estimation of ΔL according to Eq. (2.15) at each trial. For procedures B2 and B3 with increasing j the targets move from the *pair of compromise* towards the *sweetpair*, which probably leads to better slope estimates. The smallest standard deviations are generated by procedure B2 which has the largest distance between the targets.
Since procedure A2 always converges at the *pair of compromise* independently of j and produces also a sufficient accuracy in SRT estimates (cf. Fig. 2.5), this procedure is recommended for concurrent SRT and slope estimates. For $j = 4$ and $N \geq 20$, the value of $\hat{\sigma}_{s_{50}}$ decreases only slightly with increasing N and has a value of approximately 1.5. This value is slightly higher than 1.16 which is the value of $\hat{\sigma}_{s_{50}}$ predicted by Eq. (2.11) for a track in which all trials are placed optimally at the *pair of compromise*.

2.4.2.4 Relative biases in slope estimates

Fig. 2.8 shows the relative biases in s_{50} estimates. All estimates are positively biased. For all algorithms, the relative bias decreases with increasing j and increasing N. For $N = 10$ the relative bias is unacceptably high. For $j \geq 2$ and $N \geq 30$ the relative bias in slope estimation is smaller than 0.15 for all underlying slopes. Thus, it is recommended to use at least 30 sentences to get a reliable slope estimate.
Leek *et al.* (1992) proposed to apply correction factors to the measured slope values in order to reduce bias. The biases in slope estimates calculated in these simulations can also be applied as correction factors to slope values measured with sentence tests. Since the relative bias in slope estimates scarcely depends on the slope itself, the bias correction can be done very accurately.

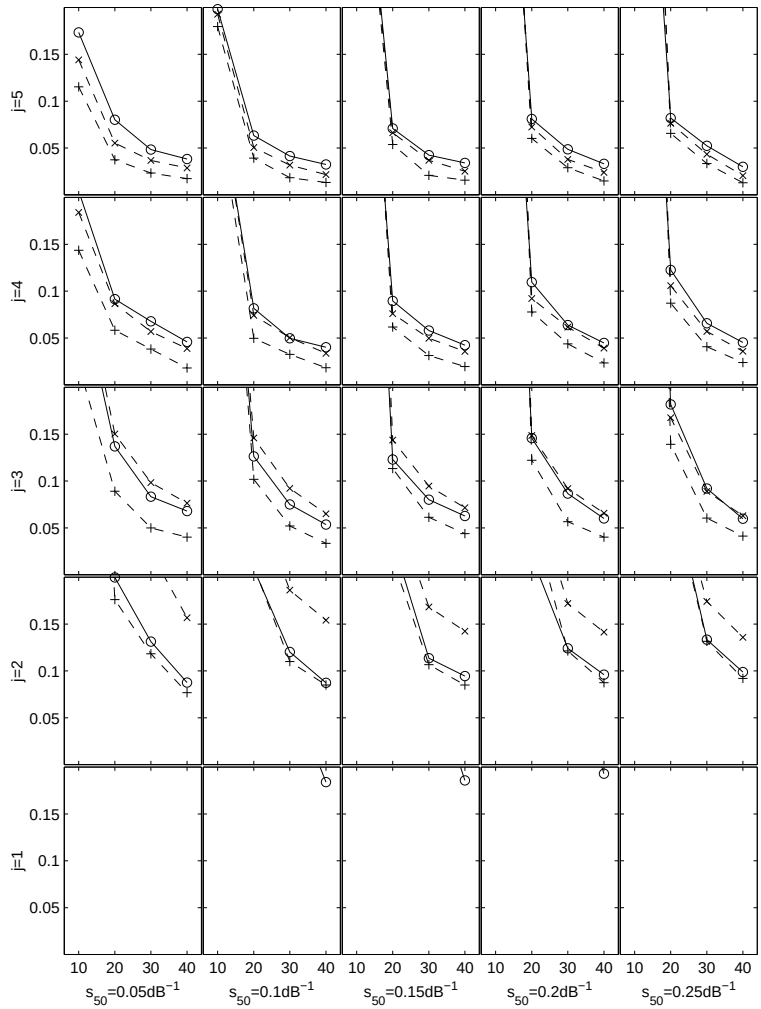

Figure 2.8: *Relative biases of s_{50} estimates as functions of N for different settings of s_{50} and j. The ordinate of each panel indicates the bias of s_{50} relative to the underlying s_{50} value, the abscissa indicates N. Some conditions are not visible, since their ordinate value is outside the displayed range. Procedures: A1 (solid line with squares), A2 (solid line with circles), B1 (dashed line with triangles), B2 (dashed line with 'x'), B3 (dashed line with '+').*

2.5 MEASUREMENTS

2.5.1 Subjects and measurement program

10 normal–hearing (4 male, 6 female; aged 16–32 years; median 24 years) and 10 hearing–impaired (5 male, 5 female; aged 37–75 years; median 66 years) listeners participated in the experiment. The normal–hearing listeners showed hearing thresholds better than 15 dB HL at the frequencies 0.125, 0.25, 0.5, 1, 1.5, 2, 3, 4, 6 and 8 kHz. The sentence tests were performed at their respective better ear. The hearing–impaired listeners showed different types and degrees of sensorineural hearing loss. Pure–tone hearing thresholds ranged from 5 dB HL up to more than 100 dB HL. (3 flat hearing losses at 50, 60 and 70 dB SPL, respectively; 2 pure high–frequency hearing losses; 5 sloping hearing losses.) No listener had prior experience in psychoacoustical experiments. All listeners were paid for their participation on an hourly basis.

All listeners performed 3 tracks each with both procedure A1 and procedure A2 and with both the Göttingen and the Oldenburg sentences. Each listener performed two sessions in which approximately half of the tracks were measured: At the beginning of the first session a pure–tone audiogram was measured. In the hearing–impaired listeners also a categorical loudness measurement using the interfering noise signal was performed. Thereafter, each listener performed 3 practice lists with 20 sentences which were not analysed (2 tracks using the Oldenburg sentences and procedures A1 and A2, respectively, and one track using the Göttingen sentences and procedure A2). Then the 4–6 tracks were performed using the different procedures and speech materials in pseudo–random order. At the beginning of the second session, which was performed on another day, a practice track with 30 trials using the Oldenburg sentences and procedure A2 was performed. Then the remainder of the tracks were performed in pseudo–random order.

In the normal–hearing listeners the interfering noise was presented at 65 dB SPL which is the typical 'medium' loudness level for this signal in normal–hearing listeners. In the hearing–impaired listeners the interfering noise was presented at the individual medium–loudness level which was determined by categorical loudness scaling (Hohmann and Kollmeier, 1995b) using the noise signal as stimulus. The signal level varied according to the respective signal–to–noise ratio.

2.5.2 Results

2.5.2.1 Individual SRT and slope estimates

Fig. 2.9 shows the individual L_{50} and s_{50} estimates for all listeners for both speech materials. The L_{50} estimates of the normal–hearing listeners show a relatively small

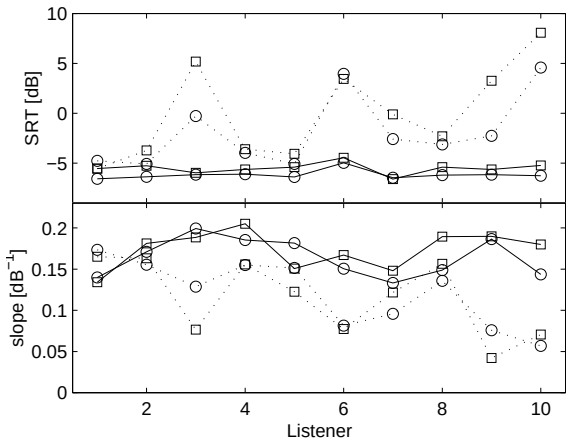

Figure 2.9: *Individual L_{50} (upper pannel) and individual s_{50} estimates for the normal–hearing listeners (solid lines) and hearing–impaired listeners (dotted lines). The measurements with the Göttingen sentence test are denoted by squares. The measurements with the Oldenburg sentence test are denoted with circles. The abscissa indicates the different listeners (normal–hearing and hearing–impaired, respectively).*

interindividual variability (± 1 dB) and the L_{50} estimates using the Oldenburg sentence test are about 1 dB smaller than those using the Göttingen sentence test. This is consistent with the reference values of -6.2 dB SNR (Kollmeier and Wesselkamp, 1997) and -7.1 dB SNR (Wagener *et al.*, 1999a), respectively. The interindividual variability of the L_{50} estimates is much larger in the hearing–impaired listeners. The estimates of the Göttingen and Oldenburg sentence tests are similar for each listener. There are, however, two listeners (3 and 9), who show significantly lower L_{50} estimates in the Oldenburg sentence test as cpmpared to the Göttingen sentence test. One reason for this effect might be that these listeners have difficulties to understand the Göttingen sentence test because of its somewhat higher speech rate.
The s_{50} estimates in the normal–heraing listeners are between 0.13 dB^{-1} and 0.21 dB^{-1} and there are no systematical differences between the Göttingen and Oldenburg sentence tests. In the hearing–impaired listeners, the range of s_{50} estimates is much larger (0.05 dB^{-1} to 0.18 dB^{-1}). In the hearing–impaired listeners, there is a clear relationship between L_{50} and s_{50}: Listeners with relatively low L_{50} value exhibit a relatively high s_{50} value and vice versa. That means, a pathological increase of the SRT is usually related to a decrease of discrimination function slope.

2.5.2.2 Accuracy of SRT and slope estimates

Table 2.2: *Intraindividual standard deviations and biases of L_{50} and s_{50} estimates for normal–hearing and hearing–impaired listeners using the Göttingen and the Oldenburg sentence tests. In order to make results comparable between listeners with different slope values, the intraindividual standard deviations and biases of the L_{50} estimates have been multiplied by the slope value of the respective listener before averaging across listeners. For the same reason, the standard deviations and biases of the s_{50} estimates have been divided by the slope value of the respective listener, i.e. standard deviations and biases are given relative to individual slope value. The corresponding predicted values from the Monte–Carlo simulations are shown as small numbers in parentheses below the observed values.*

	Göttingen sentence test				Oldenburg sentence test			
	normal–hear.		hear.–impaired		normal–hear.		hear.–impaired	
	A1	A2	A1	A2	A1	A2	A1	A2
s_{50}·STD(L_{50})	0.05 (0.08)	0.08 (0.10)	0.08 (0.08)	0.08 (0.10)	0.07 (0.05−0.09)	0.06 (0.06−0.07)	0.09 (0.05−0.09)	0.08 (0.06−0.07)
s_{50}·Bias(L_{50})	−0.01 (0)	0.01 (0)	−0.03 (0)	0.05 (0)	−0.01 (0)	0.00 (0)	−0.03 (0)	0.01 (0)
rel. STD(s_{50})	0.79 (−)	0.24 (0.22−0.26)	0.90 (−)	0.29 (0.22−0.26)	0.61 (−)	0.21 (0.15−0.16)	0.58 (−)	0.32 (0.15−0.16)
rel. Bias(s_{50})	0.66 (−)	0.08 (0.12−0.15)	0.97 (−)	0.10 (0.12−0.15)	0.33 (−)	0.09 (0.07)	0.40 (−)	0.17 (0.07)

Table 2.2 shows the mean intraindividual standard deviations and biases of the L_{50} and s_{50} estimates for the normal–hearing and the hearing–impaired listeners using the Göttingen and the Oldenburg sentence test, respectively. In order to make results comparable between listeners with different slope values, the intraindividual standard deviations and biases of the L_{50} estimates were multiplied by the slope value of the respective listener before averaging across listeners. For the same reason, the standard deviations and biases of the s_{50} estimates were divided by the slope value of the respective listener before averaging, i.e., the standard deviations and biases are given relative to the individual slope value. The individual s_{50} values for each listener were calculated by fitting the model function (2.1) to all trials which were performed with the specific listener and speech material. Because of the large number of trials that entered this estimate ($N = 180$ sentences), this s_{50} estimate can be assumed to be almost bias free. The corresponding predicted values from the Monte–Carlo simulations are shown as small numbers in parentheses below the observed values.[5] In cases where the results of

[5] The predicted values were calculated as follows:
Standard deviations of L_{50} estimates: Since 30 sentences were used per track, the minimal value of s_{50}·STD(L_{50}) at the *sweetpoint* is $\frac{0.5}{\sqrt{30 \cdot j}}$, with $j = 2$ for the Göttingen sentences and with $j = 4$ for the Oldenburg sentences (cf. Eq. 2.8). These values have been multiplied with the normalized standard

the Monte–Carlo simulations depended on the slope of the discrimination function, the minimal and the maximal predicted values are shown in Table 2.2.

Using the Göttingen sentences, the observed values of s_{50}·STD(L_{50}) shown in Table 2.2 are equal to or slightly smaller than the predicted values for both normal–hearing and hearing–impaired listeners. Using the Oldenburg sentences, the measured values of s_{50}·STD(L_{50}) are equal to or slightly larger than the predicted values. There are only small differences between speech materials.

As predicted by the Monte–Carlo simulations (cf. Sec. 2.4.2.2), the measurements show no tendency towards a significant bias of L_{50} estimates.

In the predictions, procedure A1 failed completely in slope estimates. For that reason, no predictions for slope estimates are shown in Table 2.2. The observed standard deviations and biases of s_{50} estimates using procedure A1 are unacceptably high as well. That means, procedure A1 is not suitable for slope estimates. For that reason, only the results of procedure A2 are discussed:
Using the Göttingen sentences the measured values of the relative standard deviation of s_{50} are consistent with the predicted values. The predicted improvement of s_{50} estimates using the Oldenburg sentences, however, does not occur, as the relative standard deviations of s_{50} have about the same size for both sentence tests. The relative standard deviations of s_{50} are smaller in the normal–hearing listeners than in the hearing–impaired listeners.
The relative bias of the s_{50} estimates is slightly smaller than predicted for both normal–hearing and hearing–impaired listeners, when the Göttingen sentences are used. This is not the case, when the Oldenburg sentences are used. Here, the measured values are slightly larger than predicted for the normal–hearing listeners. For the hearing–impaired listeners, the measured bias of s_{50} is clearly larger than predicted. The reasons for this unexpected discrepancy are discussed below.

---

deviations of the L_{50} estimates $\hat{\sigma}_{L_{50}}$ for the specific procedure, j factor and slope value s_{50} as displayed in Fig. 2.5 to obtain predictions of s_{50}·STD(L_{50}).
Bias of L_{50} estimates: The Monte–Carlo simulations predicted no significant bias of L_{50} estimates, cf. 2.4.2.2. Furthermore, no bias was expected, because the adaptive procedures produce level distributions symmetrical around the L_{50}. Therefore, all predicted biases of L_{50} were set to zero in Table 2.2.
Relative standard deviations of s_{50} estimates: Since 30 sentences were used per track, the minimal value of $\frac{\mathrm{STD}(s_{50})}{s_{50}}$ at the *sweetpair* is $\frac{1.07}{\sqrt{30 \cdot j}}$, with $j = 2$ for the Göttingen sentences and with $j = 4$ for the Oldenburg sentences (cf. Eq. 2.12). These values have been multiplied with the normalized standard deviations of the s_{50} estimates $\hat{\sigma}_{s_{50}}$ for the specific procedure, j factor and slope value s_{50} as displayed in Fig. 2.7 to obtain predictions of $\frac{\mathrm{STD}(s_{50})}{s_{50}}$.
Relative bias of s_{50} estimates: The predictions for the relative bias of s_{50} estimates were calculated in Sec. 2.4.2.4 (cf. Fig. 2.8).

2.6 DISCUSSION

Different authors have proposed different procedures for level placement in order to estimate threshold and psychometric function slope concurrently:
Hall (1981) used the PEST adaptive procedure with the midpoint of the psychometric function as target. Leek *et al.* (1992) used transformed up–down adaptive procedures either with the target $p = 0.71$ or with the target $p = 0.79$. Both approaches are based on the fact that the adaptive procedures do not converge immediately at their specific target, but wander up and down the level axis aimlessly. Indeed, the convergence has to be slow, because otherwise the presentation levels are placed too closely and the slope estimate gets inaccurate. Lam *et al.* (1996) proposed to present several trials at four different levels which are guessed or estimated by the investigator prior to measurement and which are newly estimated iteratively during the measurement until the estimates of threshold and slope stabilize.
In the current study the adaptive procedures converge at the *pair of compromise* which are the target levels enabling the most efficient concurrent threshold and psychometric function slope estimates. In this way, the theoretically optimal level placement is achieved, provided that the adaptive procedure converges optimally.
In those cases in which the efficiencies of the procedures are specified in the different studies, they can be compared to the procedures tested in this study: Hall's hybrid adaptive procedure (1981) yields an *efficiency* (as defined by Taylor and Creelman (1967)) of 68 % in threshold estimates after 50 trials. The efficiency of slope estimates is not specified in Hall's study. In the current study, procedure A2 yields an *efficiency* of 75 % in threshold estimates and of 39 % in slope estimates after 30 trials, if the Göttingen sentence test is used. Leek *et al.* (1992) specify the efficiency of the slope estimates of their procedure as the standard deviation of the logarithm of slope estimates. Therefore, their results cannot be directly compared to this study. Lam *et al.* (1996) specified the standard deviations of their procedure, but did not state how many iterations they used. Therefore, it is not possible to calculate the efficiency of their procedure.

The adaptive procedures presented in this study exploit the fact that a sentence trial has more than one statistically independent element. This allows for a more efficient convergence and data collection. At least in Monte–Carlo simulations, acceptable slope estimates can be obtained using at least 30 sentence trials. In usual forced–choice situations about 100 trials are necessary to obtain slope estimates with comparable accuracy (e.g., O'Regan and Humbert, 1989).[6]
The procedures A1 and A2 which are generalizations of the adaptive procedure of Hager-

[6] O'Regan and Humbert (1989) tested different constant stimuli distributions in respect of their accuracy in threshold and slope estimates. In Monte–Carlo simulations the best distributions produced relative standard deviation of slope estimates of about 0.14 using 100 trials. In this study, procedure A2 produced a relative standard deviations of slope estimates of about 0.16 in Monte–Carlo simulations for $j = 4$ using 30 trials.

man and Kinnefors (1995) are advantageous compared to the procedures B1, B2 and B3 which are modifications of transformed up/down procedures (Levitt, 1971). Procedures A1 and A2 converge efficiently at their target values independently of the predictability of speech material. Procedures B1, B2 and B3 converge only efficiently at their target values, if the number of statistically independent elements per sentence is approximately 2. For these reasons, it is recommended to use the adaptive procedure A1 for simple SRT measurements and procedure A2 for simultaneous SRT and slope measurements for all kinds of speech material.

The optimal final step sizes of the adaptive procedures found in this study are considerably smaller than those found by other authors (e.g., Green, 1989; Hagerman and Kinnefors, 1995).
The restriction of $f(i)$ to 0.1 in procedures A1 and A2 causes final step sizes which are only 13.3 % of those used by Hagerman and Kinnefors (1995). Note, however, that this final value of $f(i)$ is reached after the eighth reversal, when the target level can be estimated very accurately and there is no reason for large changes in presentation level. The final step size of 0.25 dB in procedures B1, B2 and B3, which is reached after the fourth reversal, is only an eighth of the step size of 2 dB, used by Nilsson (1994) and Plomp and Mimpen (1979). But if it is taken into account that the discrimination functions of sentence tests in noise are considerably steeper than psychometric functions in most other psychophysical tasks, this small final step size is consistent with other studies. Green (1989) found an optimal final step size of 2 dB in n–alternative forced choice tests with a psychometric function slope of s_{50} of 1.25 dB^{-1}. Scaling of this final step size to sentence tests with a typical slope of about 10 dB^{-1} gives the same final step size as derived from the simulations in this study.

Significant biases of slope estimates have been reported several times, especially, when the chance level (the probability to guess the right response) is non–zero and if small track lengths ($N < 200$) are used (O'Regan and Humbert, 1989; Leek *et al.*, 1992). Fortunately, in sentence test the chance level is nearly zero. Furthermore, for $j > 1$ more than one Bernoulli trial is performed per trial and thus, much smaller track lengths ($N = 30$) can be used. However, even in the Monte–Carlo simulations of this study the s_{50} estimate was positively biased in all conditions, but for $N \geq 20$ and $j \geq 3$ the relative bias in slope estimation was smaller than 10 %, which is acceptable for diagnostics. Further, these biases do not depend on the SRT and the discrimination function slope, respectively. Therefore, they can be compensated for by the correction factors which were calculated by the Monte–Carlo simulations.

Both simulations and measurements showed that slope estimates are much more sensitive to an inadequate level placement than SRT estimates are. This is very consistent with the standard deviations predicted on the basis of the binomial theory shown in Figs. 2.1 and 2.2. Fig. 2.1 shows that the L_{50} estimate is relatively robust, if stimuli are presented not very accurately at the sweetpoint. Fig. 2.2, however, shows that the s_{50} estimate

becomes very ineffective, if stimuli are presented near the sweetpoint and not near the sweetpair. In fact, each trial which is presented near the sweetpoint does not contribute to the slope estimate at all.

The measurements with the Göttingen sentence test showed intraindividual standard deviations and biases in L_{50} and s_{50} estimates which were very similar to the predicted values. However, most measurements were slightly more accurate than the predictions. This is striking, because measurements with 'real' listeners usually generate larger errors than simulations do, which is generally explained by 'human factors'. One reason for the unexpected accurate measurements might be that the intelligibility is calculated using weigthing factors for each word in the Göttingen sentence test (Kollmeier and Wesselkamp, 1997). These weigthing factors have been introduced to reduce the inhomogenieties in the test material. However, they might cause an increase in efficieny as well.

The measurements with the Oldenburg sentence test, on the other hand, showed less efficiency in L_{50} and s_{50} estimates than predicted by the Monte–Carlo simulatons. This might be due to a mismatch between the assumed and the real number of statistically independent elements per sentence (j factor):
In measurements using fixed signal–to–noise ratios, the mean j factor was 3.74 in the Oldenburg sentences (Wagener *et al.*, 1999a) and 2.17 in the Göttingen sentences (1997).[7] Since a higher j factor allows for a faster convergence of the adaptive procedure and since the number of statistically independent samples analyzed is higher, the simulations predicted a decrease of standard deviations and biases by 30 to 50 % between both tests for both L_{50} and s_{50} estimates. The fact that this decrease did not occur in the measurements is possibly due to a decrease of the j factor caused by the adaptive level placement. Table 2.3 shows the j factors derived from the adaptive tracks for the different procedures, speech materials and collectives of listeners. Under all conditions, the differences between normal–hearing and hearing–impaired listeners are insignificantly small. For both speech materials, procedure A1 generates j factors which are slightly smaller than the expected values which are $j = 2.17$ for the Göttingen sentences and $j = 3.74$ for the Oldenburg sentences. This indicates that an adaptive procedure converging at a discrimination value of 0.5 has only a small impact on the number of statistically independent elements per sentence in comparison with fixed level tracks. Procedure A2, however, generates clearly smaller j factors. This procedure generates much larger steps of presentation level in order to converge at the discrimination values 0.2 and 0.8 in a randomly interleaved order. It is likely that these large steps influence the attention of the subject. In a fixed or almost fixed level track, the subject knows roughly what will happen in the next trial. Thus, he/she likely spends equal attention to each word of each

[7] Both Wagener *et al.* (1999a) and Kollmeier and Wesselkamp (1997) measured j at discrimination values of approximately 0.2 and 0.8, respectively. The values given here are the mean values of the observed j factors.

sentence. In an interleaved track with large level steps, however, subjects might easily give up to even discriminate some words at very low levels. This decreases the probability p_w to recognize single words in sentences which cannot be recognized completely. According to Eq. 2.7, this decreases the j factor as well.

Table 2.3: *j factors for the Göttingen and Oldenburg sentences for the different procedures.*

	Göttingen sentences		Oldenburg sentences	
	A1	A2	A1	A2
normal–hearing listeners	1.97	1.62	3.10	2.04
hearing–impaired listeners	1.89	1.57	2.99	2.03

A further possible reason for the unexpectedly poor results produced by the Oldenburg sentences might be a training effect. Since all lists of the Oldenburg sentence test consist of the same vocabulary, the L_{50} estimates were observed to decrease by 2 dB within the first 6 tracks due to training (Wagener *et al.*, 1999a). The main portion of this training effect occurs within the first two lists. In the Göttingen sentences, no comparable training effect was observed. In order to reduce this training effect in the current study, each listener absolved a training consisting of two tracks employing the Oldenburg sentences before the actual measurements. This and the fact that Oldenburg and Göttingen sentence lists were measured in an interleaved order aimed at reducing the training effect. Nevertheless, a certain training effect might have occurred in this study anyway, i.e., the L_{50} estimates of a specific listener might have decreased with increasing number of tracks.
Fig. 2.10 shows the mean difference between the L_{50} estimates of the different tracks and the first track averaged across all listeners. A small training effect of about 0.7 dB within the whole 6 tracks occurred only in the normal–hearing listeners with the Oldenburg sentences there. In the hearing–impaired listeners, no training effect was observed nor in the tracks using the Göttingen sentences. This means that the precautions to reduce training effects, i.e. prior training and mixing of speech materials, did work well and the standard deviations of L_{50} and s_{50} estimates were not deteriorated by a training effect.

This study uses word scoring, i.e. each word of a sentence is scored separately. Other sentence tests score whether or not the complete sentence is repeated correctly (e.g., Plomp and Mimpen, 1979; Nilsson *et al.*, 1994). The latter implies one statistically independent element per sentence (j=1). The simulations of this study, however, show that this scoring method is the worst condition for efficient L_{50} and s_{50} estimates, because the adaptive procedures cannot work properly and the fitting is based on too

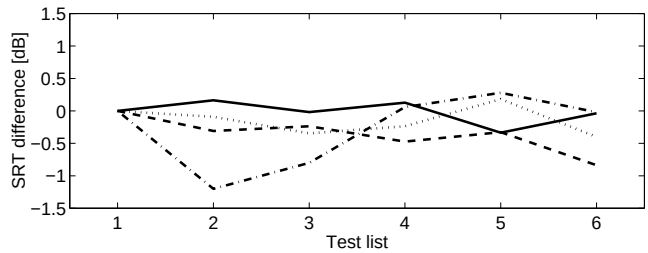

Figure 2.10: *'Training effect': Difference between L_{50} estimates of repeated tracks and the first track as a function of the number of tracks performed. Normal–hearing listeners with Göttingen sentences (solid), normal–hearing listeners with Oldenburg sentences (dashed), hearing–impaired listeners with Göttingen sentences (dash–dotted) and hearing–impaired listeners with Oldenburg sentences (dotted).*

few samples. Although no measurements with sentence scoring have been performed in this study, we assume that sentence scoring is ineffective in measurements as well. Word scoring only takes little additional experimental effort than sentence scoring, but this effort seems to be justified by the gain in efficiency and accuracy.
Plomp and Mimpen (1979) proposed to use 13 sentences to obtain a standard deviation in SRT estimates of approximately 1 dB with normal–hearing listeners in noise. Nilsson *et al.* (1994) proposed to use 10 sentences for SRT measurements and obtained a standard deviation of 1.52 dB in noise with normal–hearing listeners. However, if decreased slope values for hearing–impaired listeners are accounted for as well as the fact that differences of only a few dB in SRTs in noise can be very important for diagnostics and hearing aid fitting, it appears necessary to achieve smaller standard deviations in SRT measurements.
For single SRT measurements, we therefore propose to use procedure A1 with at least 20 sentences to get reliable results with a standard deviation of less than 1 dB for listeners with $s_{50} \geq 0.10$ dB^{-1}.
If also the slope should be measured, the use of procedure A2 with at least 30 sentences is recommended. In this case, a standard deviation in SRT estimation of less than 1 dB and a relative standard deviation in slope estimation of 20 to 30 % are achieved.

2.7 SUMMARY AND CONCLUSIONS

Adaptive procedures for efficient threshold and discrimination function slope estimates were introduced and evaluated. They utilize the fact that more than one statistically independent Bernoulli trial is sampled in each sentence trial, if word scoring is applied.

This fact allows for much more efficient SRT and slope estimates compared to usual forced choice tasks. To obtain a reliable bias–free SRT estimate with a standard deviation of less than 1 dB, it is recommended to use at least 20 sentence trials and an adaptive procedure that converges at a discrimination value of $p = 0.5$ with decreasing step size. However, a reliable estimate of discrimination function slopes using sentence tests is problematic, because clinical diagnostics requires measuring times much shorter than those typical of psychoacoustics. At least 30 sentences are necessary (more are recommended) as well as an adaptive procedure that converges at the discrimination values of $p_1 = 0.2$ and $p_2 = 0.8$ in randomly interleaved order to obtain concurrent estimates of SRTs and discrimination function slopes. Thus, a relative standard deviation of slope estimates of 20 to 30 % and a relative bias of about 10 % within 30 sentence trials can be obtained.
Theoretically, sentence tests with lowly predictable speech materials (such as the Oldenburg sentence test (Wagener *et al.*, 1999c)) should be more efficient in estimating SRTs and discrimination function slopes than sentence tests with highly predictable speech materials (such as the Göttingen sentence test (Kollmeier and Wesselkamp, 1997)). In measurements with normal–hearing and hearing–impaired listeners, however, the expected advantage of the Oldenburg sentence test did not occur. This was, probably, due to a reduced j factor caused by the adaptive level placement.

ACKNOWLEDGMENTS

We would like to thank Anita Gorges, Kirsten Wagener and Birgitta Gabriel for performing the measurements. Thanks are also due to Karin Zimmer and Angelika Sievers for helpful comments on the manuscript.

This study was supported by DFG KO 942/13–1 and by the CEC supporting action NATASHA.

Chapter 3

Statistical Model of the Accuracy and Validity of Categorical Loudness Scaling Procedures

ABSTRACT

Categorical loudness scaling procedures in audiology should provide maximum information within a minimum amount of time. Therefore, they have been modified in many features by different researchers. In this study, a simple statistical model of the decision process in categorical loudness scaling tasks is introduced and tested which enables the simulation of the whole measuring process in order to predict the accuracy and validity of categorical loudness scaling procedures. The model is based on the reproducibility of single categorical loudness ratings which was derived from measurements with normal–hearing and hearing–impaired listeners presented by Hohmann (1993). From the data an 'optimistic' and a 'pessimistic' estimate of the response characteristic are derived that should account for the differences in reproducibilities between listeners. The optimistic estimate has a standard deviation of 4 categorical units (cu) on a 50 unit scale in repeated ratings of identical stimuli. This is a typical value for normal–hearing and many hearing–impaired listeners. In several hearing–impaired listeners, however, the reproducibility is less. The pessimistic estimate of the response characteristic assumes a standard deviation of 7 cu and a 5 % probability of outliers. Based on these two estimates of the underlying response characteristics, Monte–Carlo simulations of categorical loudness scaling measurements were performed. The influence of the fitting procedure, the track length, and the number of response alternatives on the estimates of the 'medium' level L_{25} and the slope m of the loudness function was investigated. If a probability of at least 7 % outliers can be expected, a robust fitting procedure, e.g.,

a maximum likelihood fit with Lorentzian merit function is recommended rather than a least-squares fit. Minimum numbers of about 10 response alternatives and about 15 trials are needed to yield suitable loudness function estimates for practical purposes.

3.1 INTRODUCTION

For the diagnosis of loudness recruitment in hearing-impaired listeners and for an adequate fitting of hearing aids with automatic gain control, it is desirable to determine the individual auditory dynamic range or, preferably, the individual shape of the loudness function. Such a measurement which only takes a few minutes may determine the fitting of a hearing aid which is used for years. Therefore, the reliability and the accuracy of the fitting procedure has a strong influence on the acceptance of the hearing aid. Since categorical loudness scaling might support the hearing–aid fitting, the factors determining its accuracy and reproducibility are important for practical application and are investigated in this study.

The classical, number–based procedures to measure loudness functions like the 'ratio scale' procedure (e.g., Stevens, 1957) or the different kinds of magnitude estimation and magnitude production (e.g., Hellman and Zwislocki, 1963; Hellmann and Meiselman, 1993) require some training and are somewhat difficult to handle in audiology. Furthermore, in audiology, one is often interested in the question, how loud a stimulus is perceived in terms of 'soft' and 'loud' rather than in the ratios of the loudnesses of different stimuli. Therefore Pascoe (1978) and Heller (1985) proposed to measure loudness functions using a category scale. These early approaches have been modified in many ways by different researchers (e.g., Hellbrück and Moser, 1985; Allen *et al.*, 1990; Elberling and Nielsen, 1993; Hohmann and Kollmeier, 1995b; Ricketts and Bentler, 1996; Cox *et al.*, 1997; Brand *et al.*, 1997a; Rasmussen *et al.*, 1998; Keidser *et al.*, 1999). These modifications concerned parameters such as order and range of stimulus levels, track length, response scales (e.g., number of response alternatives), fitting procedure, and model loudness function. Some modifications (such as the reduction of the number of response alternatives or the reduction of track length) were made in order to simplify the handling of the procedure. Other modifications (namely different stimulus level placements) were made to obtain more reliable and/or more efficient results. In this chapter, the effect of track length, number of response alternatives and fitting procedure is investigated. The effect of the stimulus level placement on accuracy is investigated in Chap. 4. The effect of the model function is investigated in Chap. 5.

The whole measuring process is modeled in this study by Monte-Carlo simulations based on empirical response statistics. This approach allows for a detailed investigation of the influence of different parameters of the measuring procedure and may possibly be applied for categorical scaling procedures of different psychophysical magnitudes. Data

of Hohmann (1993) are used to derive the typical response statistics of normal–hearing and hearing–impaired listeners empirically.

Since the track length determines both measuring time and accuracy, it is of fundamental importance for any clinical application. A small number of response alternatives (such as 7 (Allen *et al.*, 1990)) is easier to realize in the experimental setup than a large number (such as 51 (Hellbrück and Moser, 1985)). On the other hand, an insufficient number of response alternatives inevitably causes losses in accuracy. The quantification of the measuring accuracy using Monte-Carlo simulations enables to find the minimum track length and the minimum number of response alternatives which are required to yield the desired measuring accuracy.
Possible temporary lacks of the listener's attention can generate responses away from the actual loudness function which may strongly influence the measuring procedure. A maximum likelihood estimator which assumes a Gaussian data distribution (least-squares fit or Gaussian fit, respectively) applies a large weight to such outliers and may lead to biased results. Therefore, a more robust maximum likelihood estimator which assumes a Lorentzian data distribution (Lorentzian fit) is investigated in this study as well. The minimum probability of outliers which requires the Lorentzian fit rather than the Gaussian fit is derived by Monte-Carlo simulations.

3.2 STATISTICAL METHOD

3.2.1 Statistical modeling

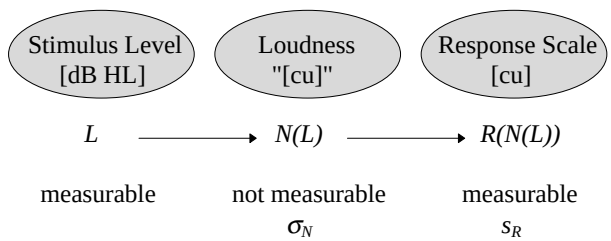

Figure 3.1: *Statistical model of categorical loudness scaling.*

The aim of categorical loudness scaling is to determine the relationship between the stimulus level and the loudness. However, the loudness is not measurable directly because it is a subjective impression. Therefore, in categorical loudness scaling, the loudness is quantified by ratings on a category scale. Since two identical presentations are not necessarily rated equally, the listener's decision process in categorical loudness scaling

tasks includes stochastic processes which are modeled as follows (cf. Fig. 3.1): In each trial, the loudness perception can be described as a stochastic projection $L \rightarrow N$ of a known stimulus level L on the subjective loudness $N(L)$. The rating process can be described as a further projection $N \rightarrow R$ from the internal loudness impression on the categorical loudness $R(N(L))$. The responses R are quantified in terms of categorical units (cu) from 0 cu ('inaudible') to 50 cu ('too loud'). Since the loudness impression $N(L)$ is not measurable directly, its unit can be chosen arbitrarily. For simplification, we quantify N in terms of categorical units as well. In contrast to R, N is not limited to values between 0 and 50 cu because the subjective impression is not limited like the response scale.

It cannot be determined how the respective parts of the observed variability are distributed between the projection $L \rightarrow N$ and the projection $N \rightarrow R$. For simplification, we assume that all variability of the decision process is introduced in the first projection. Furthermore, we assume that the second projection is linear apart from the fact that it limits and rounds the continuous loudness impression to a restricted and discrete response scale.

The loudness $N(L)$ is given by the loudness function $F(L)$ and a random variable x which is added to $F(L)$. We assume x to be normally distributed with the standard deviation $\sigma_N(L)$ and the mean 0. Furthermore, we assume that the listener might have lacks of attention which produce outliers in the response statistics, i.e. some responses are totally random and have no relation to the presentation level. We assume that an outlier is given by a random variable y that is uniformly distributed in the interval (0 cu, 50 cu) and that an outlier occurs with the probability p_{out}. That means, the variability of the measuring process is modeled by a switching process between the two random variables x and y. This switching process itself is described by a third random variable z which is binomially distributed and has the probability p_{out}. Thus, the assumed subjective loudness is given by:

$$N(L) = \begin{cases} F(L) + x & \text{for} \quad z = 1 \\ y & \text{for} \quad z = 0 \end{cases} \tag{3.1}$$

The resulting density function is equal to $1 - p_{\text{out}}$ times the normal density function (with $\mu = F(L)$ and $\sigma = \sigma_N$) plus a rectangular pulse with the hight $\frac{p_{\text{out}}}{50}$ from 0 to 50 cu. Note, that for $p_{\text{out}} > 0$, the mean of $N(L)$ is not equal to $F(L)$ but shifted towards 25 cu.

The second projection $N \rightarrow R$ is modeled as a unity projection which limits values of N which are outside the range from 0 to 50 to the corresponding limits of the scale and which quantizes N according to the number of response alternatives n_{alt}:

$$R(N) = \begin{cases} 0 & \text{for} \quad N \leq 0 \\ \frac{50}{n_{\text{alt}}-1} \cdot \text{floor}(0.5 + (n_{alt} - 1) \cdot \frac{N}{50}) & \text{for} \quad 0 < N < 50 \\ 50 & \text{for} \quad N \geq 50 \end{cases} \tag{3.2}$$

(The function floor(x) rounds x to the nearest integer towards minus infinity.)
Since $N(L)$ is a subjective impression, the parameters $\sigma_N(L)$ and p_{out} cannot be derived directly from empirical data. In Sec. 3.4 they are estimated from the variability in the loudness ratings $R(L)$ which can be described by the standard deviation $s_R(L)$ of repeated loudness ratings of identical stimuli with the level L. Because of the limiting effect of Eq. (3.2), $s_R(L)$ is smaller than $\sigma_N(L)$, especially at the limits of the scale.

3.2.2 Monte-Carlo simulations

According to the statistical model, the simulation of the whole measuring process using the Monte-Carlo method requires the knowledge of the loudness function $F(L)$ and of the parameters σ_N and p_{out}. Although the loudness function can be parameterized by different model functions more adequately (cf. Chap. 5), a simple straight line[1] with only two parameters L_{25} and m is used in this study in order to simplify the simulations:[2]

$$F(L) = 25 + m(L - L_{25}) \tag{3.3}$$

Without loss of generality, the underlying loudness function parameters are set to $L_{25} = 65$ dB HL and $m = 0.7\ \frac{\text{cu}}{\text{dB}}$ in all simulations. The setting of L_{25} has no effect on the actual results of the simulations because Eq. 3.3 is invariant to scaling on the level axis. If m is varied, the standard deviation $\sigma_{L_{25}}$ and the bias $b_{L_{25}}$ of the L_{25} estimates multiplied by m remain constant. Therefore, $\sigma_{L_{25}}$ and $b_{L_{25}}$ are quantified in terms of normalized standard deviation $m\sigma_{L_{25}}$ and normalized bias $mb_{L_{25}}$. The standard deviation σ_m and the bias b_m of m estimates remain constant as well, if they are normalized by m. Therefore, σ_m and b_m are expressed in terms of $\frac{\sigma_m}{m}$ and $\frac{b_m}{m}$.

The stimuli are placed optimally, i.e., covering the full dynamic range of the subject, at the levels which correspond to the categorical loudness values 4, 11, 18, 25, 32, 39 and 46 cu.[3] If not stated differently, each stimulus level is presented twice per track (That means 14 trials are presented per track). 2,000 tracks (Monte-Carlo runs) are simulated per condition.
The parameters σ_N and p_{out} are varied in the simulations according to the requirements of the actual investigation. The mean value of σ_N is derived from empirical data of normal–hearing and hearing–impaired listeners in Sec. 3.4.

[1] Cf. model function 5.2 in Chap. 5.

[2] The effect of different model functions on the accuracy of the loudness function estimate is investigated in Chap. 5. Non–linear model functions are able to fit individual loudness functions much better than the linear model function used in this study. However, the effects of the basic measurement parameters that are investigated in this study (i.e. fitting procedure, track length and number of response alternatives) can be predicted using a linear model function without loss of generality. Furthermore, in the evaluation of the predictions a broadband stimulus (a complete sentence) is used which generates loudness functions that can be fitted very well with a linear loudness function, cf. Chap. 6.

[3] The effect of level placement on the accuracy of loudness function estimates is not investigated here, but in Chap. 4.

3.3 EXPERIMENTAL METHOD

3.3.1 Procedure

The constant stimuli version of the Oldenburg loudness scaling procedure (Hohmann and Kollmeier, 1995b) was used. It includes two parts. The auditory dynamic range of the individual listener is estimated by presenting an ascending level sequence in the first part of the measurement. The loudness function is assessed by presenting stimuli covering the so determined full auditory dynamic range in the second part.
The first part uses an ascending stimulus level sequence with an initial level of 0 dB HL and a step size of 5 dB. The subject's task is to press a certain response button as soon as the stimulus is audible. After the subject has pressed the response button, the level is further increased in 15 dB steps up to 85 dB and in 5 dB steps beyond 85 dB. Now, the listener is asked to press another response button 'too loud' immediately when the stimulus is perceived as too loud. In case that the listener does not press the response button, the sequence stops at 120 dB HL.
In the second part of the procedure, the loudness function is estimated. Two stimuli are presented at each of 7 different levels which are equidistantly distributed on a dB–scale between the limits of the dynamic range estimated in the first part of the procedure. The subject rates the loudness either using the 1–step–scale procedure or using the 2–step–scale procedure (cf. Sec. 3.3.2). The stimuli are presented in a pseudo–random order, in a way that the maximum difference of subsequent presentation levels is smaller than half of the dynamic range of the sequence in order to avoid context effects which are due to the tendency of many listeners to rate the current stimulus relatively to the previous stimulus. Moreover, only sequences with an initial level smaller than the final level are allowed. This causes the stimulus sequences to have an increasing tendency. After completion of the track a model function is fitted to the data by a modified least-squares fit (cf. Sec. 3.3.3).

3.3.2 Response scales

In the 1–step–scale procedure, the subjects have to indicate their ratings on the response scale which is shown in Fig. 5.1, i.e. eleven response alternatives including five named loudness categories, four not named intermediate response alternatives and two named limiting categories. The named response categories are 'sehr leise' ('very soft'), 'leise' ('soft'), 'mittel' ('medium'), 'laut' ('loud') and 'sehr laut' ('very loud') and correspond to 5, 15, 25, 35 and 45 cu (categorical units) as shown on the left side of Fig. 3.2. The 4 not named response alternatives are used to increase the total number of response alternatives. The not named response alternatives are used to increase the total number of response alternatives. They are indicated with horizontal bars with increasing

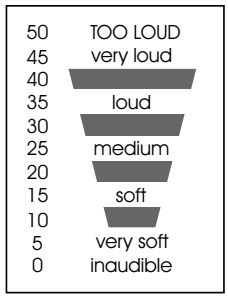

Figure 3.2: *Category scale with 11 response alternatives used by the subjects to rate the loudness. The numbers on the left side indicate the categorical units (cu) which are used for data storage and analysis. They were not visible to the listener.*

length for increasing loudness and are placed between the named loudness categories. They correspond to the categorical units 10, 20, 30 and 40 cu, respectively. The two limiting categories are named 'unhörbar' ('inaudible') and 'ZU LAUT' ('TOO LOUD') and correspond to 0 and 50 cu. The 1–step–scale procedure was proposed by Hohmann and Kollmeier (1995b).

In the 2–step procedure, the stimulus is presented twice to the subject. After the first presentation, the subject has to rate the loudness using the five named response categories of the 1–step procedure. The categories are related in the same way to the categorical units as indicated by Fig. 3.2. Thereafter, the stimulus is repeated and the subject has to rate it again using a 10–step subdivision of the main category chosen in the initial rating. This 2–step scaling procedure is very similar to the procedure proposed by Heller (1985).

3.3.3 Fitting

Two different modes of a maximum likelihood estimator are used to fit the model function to the data:
In the first mode it is assumed that the loudness ratings $R_i(L)$ for the level L are Gaussian distributed with a constant standard deviation $\sigma_R(L)$. The resulting Gaussian merit function (i.e. the negative logarithm of the likelihood function) is:

$$d_{\text{Gauss}} = \sum_i \Delta_i^2, \tag{3.4}$$

with $\Delta_i = R(L_i) - F(L_i)$. The maximum likelihood estimator is calculated by varying the parameters L_{25} and m of the model function $F(L)$ until the minimum of Eq. (3.4) is reached. Since the quadratic deviations between data points and model function are

minimized, this maximum likelihood estimator is often called least-squares fit. According to Eq. (3.4) the deviations Δ_i between data points and model function are weighted linearly with Δ_i. Thus, possible outliers which deviate strongly from $F(L)$ are largely weighted and may bias the fit.

In the second mode of the maximum likelihood estimator, this effect is minimized by assuming that the data points $R_i(L)$ are Lorentzian distributed. The resulting merit function is:

$$d_{\mathrm{Lorentz}} = \sum_i \log\left(1 + \frac{1}{2}\Delta_i^2\right). \tag{3.5}$$

The different merit functions d_{Gauss} and d_{Lorentz} are shown in Fig. 3.3. Since the Lorentzian merit function shows less increase with increasing deviation between data and model function, possible outliers which deviate strongly from the remaining data points are weighted less than by the Gaussian merit function. Therefore, the Lorentzian merit function is more robust dealing with outliers.

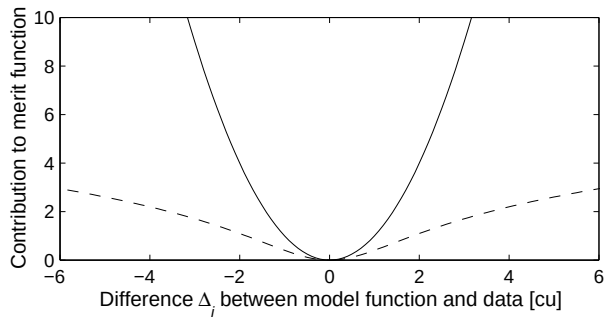

Figure 3.3: *Contribution of one data point to the Gaussian merit function* $d_{\mathrm{Gauss}}(\Delta_i)$ *(solid) and contribution of one data point to the Lorentzian merit function* $d_{\mathrm{Lorentz}}(\Delta_i)$ *(dashed), with* Δ_i *denoting the distance between model function and data point.*

The limitation of the response scale from 0 to 50 cu has to be considered in the definition of Δ_i at the limits of the scale:
If a stimulus beyond the uncomfortable level or below the hearing threshold level is presented, the listener is restricted to the responses 'too loud' (50 cu) or 'inaudible' (0 cu), respectively, according to Eq. (3.2). To consider this, Δ_i is set to zero if $F(L_i)$ is outside the category range from 0 to 50 cu and the listeners rates L_i with the corresponding limit of the scale because in this case the listener uses the response alternative which is closest to the subjective loudness impression (see e.g. data point at the right end of Fig. 3.4). On the other hand, the limited response scale has no impact on the response

if the listener responds with a category $R(L_i)$ which is within the limits of the scale (i.e. $5 \leq R(L_i) \leq 45$). In this case, a model function with $F(L_i) < 0$ cu or $F(L_i) < 50$ cu deviates strongly from the response value. Consequently, in such a case, the full distance between model function and data point is used (see e.g. data point at the left end of Fig. 3.4). That means, the definition of Δ_i used in this study is:[4]

$$\Delta_i = \begin{cases} 0 & \text{for } F(L_i) < 0 \quad \wedge \quad y_i = 0 \\ 0 & \text{for } F(L_i) > 50 \quad \wedge \quad y_i = 50 \\ y_i - F(L_i) & \text{else} \end{cases} \tag{3.6}$$

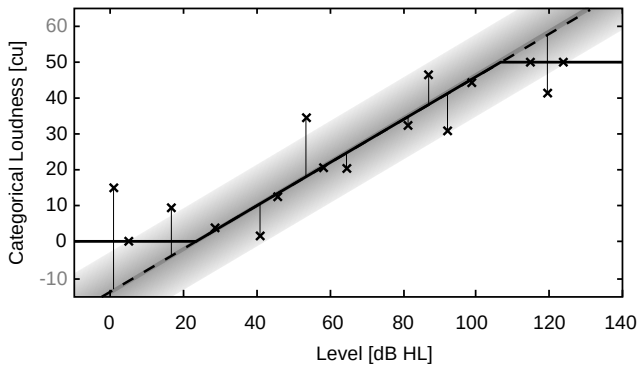

Figure 3.4: Definition of the distance Δ_i between model function and data for a limited response scale (indicated as vertical lines). The gray bar indicates an assumed underlying standard deviation σ_N in the perceived loudness N which is constant with level.

3.3.4 Apparatus

A computer–controlled audiometry workstation was used which was developed within a German joint research project on speech audiometry (Kollmeier *et al.*, 1992). A personal computer with a coprocessor board (Ariel DSP 32C) with 16–bit stereo AD–DA converters was used to control the complete experiment as well as stimulus presentation and recording of the subject's responses. The stimulus levels were adjusted by a computer–controlled custom–designed audiometer comprising attenuators, anti–aliasing

[4] A different approach to consider the limited response scale is to limit also $F(L)$ to values between 0 and 50 cu. In this way, smaller Δ_i values are generated if the subject's response is between 0 and 50 cu while $F(L)$ implies that L_i is outside the dynamic range of the subject. However, this different approach causes biases in loudness function estimates, especially at high levels if non–linear model functions are used as it is done in Chap. 5. Furthermore, limiting the model loudness function contradicts the idea that the internal loudness impression is not limited to values below 50 cu and that the responses of the listeners are restricted by the scale only.

filters, and headphone amplifiers. The signal was presented monaurally to the subjects with Beyer Dynamics DT 48 headphones in the fitting study (Sec. 3.4). In the evaluation study the signal was presented monaurally with Sennheiser HDA 200 headphones (Sec. 3.6). The headphones were calibrated to free–field sensitivity level of sinusoidals (ISO 389 (1991)) according to Richter (1992). The subjects were seated in a sound–insulated booth. Their task was to rate the loudness of each stimulus using an Epson EHT 10S handheld computer with an LCD touchscreen showing the response scale. The handheld computer was connected to the personal computer via serial interface. The subjects loudness ratings for each stimulus were stored for later statistical analysis.

3.3.5 Stimuli

Two stimuli were used in this study:
1) A third–octave band of noise with a center frequency of 1 kHz. It was generated from a random noise with Gaussian amplitude statistic, 5 s duration and 44.1 kHz sampling rate. The signal was transformed to the frequency domain by an FFT. All FFT–coefficients outside the desired band were set to zero and the resulting signal was transformed back to the time domain by an inverse FFT. A segment of 2 s duration was selected randomly and windowed with 100 ms $\cos^2$ ramps. During each trial, the noise was presented twice with a silent interstimulus interval of 1 s duration.
2) The German sentence 'Der Bahnhof liegt sieben Minuten entfernt.' ('The railway station is seven minutes away.') which is a sample of the Göttingen sentence test (Kollmeier and Wesselkamp, 1997). It has a duration of 2.3 s and was presented once per trial. It was calibrated in sound-pressure level (SPL) (sampling rate 25 kHz). That means that in the specifications of the measurement procedure (Sec. 3.3.1) 'HL' has to be substituted by 'SPL'. The sentence gives a hearing threshold of 12.2 dB SPL for normal–hearing subjects.

3.4 FITTING OF STATISTICAL MODEL

3.4.1 Experimental data

The response characteristics of normal–hearing and hearing–impaired listeners have to be derived to fit the statistical model in order to perform Monte–Carlo simulations. Data presented by Hohmann (1993) are used to derive the standard deviations $\sigma_N(L)$ in subjective loudness ratings.
9 normal–hearing and 10 hearing–impaired listeners participated in the experiment. The stimulus was a third–octave band of noise with a center frequency of 1 kHz (cf. Sec. 3.3.5). The measurements were performed using both the single–step procedure

with 11 response alternatives and a two–step procedure with 51 response alternatives. In both cases 7 different presentation levels were applied twice during one track in random order. In the normal–hearing subjects the stimulus levels were evenly distributed over a range of 40 dB HL to 90 dB HL, i.e., the pre–measurement was omitted. Each subject performed 10 tracks. Since each level was rated twice per track, 20 loudness ratings were obtained for each stimulus level and each subject. Further details are described in Hohmann (1993).

3.4.2 Normal–hearing subjects

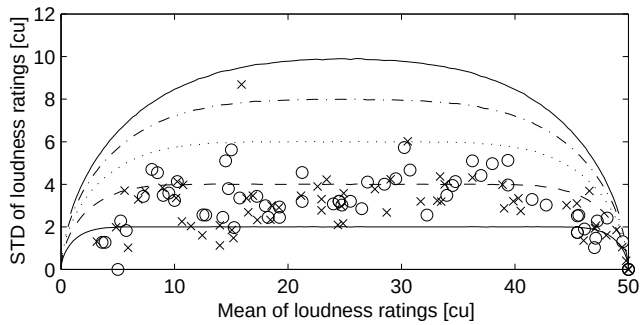

Figure 3.5: *Measured and simulated intraindividual standard deviations s_R of repeated loudness ratings as functions of their respective mean values. The crosses and circles show the values calculated from the 20 repeated ratings per level and ear in the 9 normal–hearing subjects (data taken from Hohmann (1993)). The data from the one–step procedure with 11 response alternatives are indicated by circles, the data from the two–step procedure with 51 response alternatives are indicated by crosses. The lines show simulated s_R values for different constant values of σ_N. Each data point is based on 20,000 simulations with the same stimulus level. The underlying σ_N values are: 10 cu (upper solid line), 8 cu (dash-dotted line), 6 cu (dotted line), 4 cu (dashed line) and 2 cu (lower solid line).*

Fig. 3.5 shows the intraindividual standard deviations s_R of the repeated loudness ratings of the same stimulus level L for the normal–hearing subjects from the measurements by Hohmann (1993). The abscissa value of each data point indicates the mean $\bar{R}$ of the 20 loudness ratings for a given subject, ear and level. The ordinate indicates the corresponding standard deviation s_R. Circles indicate the data from the one–step procedure and crosses the data from the two–step procedure.

The s_R values are approximately equal to 4 cu in the medium categorical loudness

range and decrease near the limits of the category scale. No significant differences in the distributions of the s_R values can be seen between the one–step procedure with 11 response alternatives and the two–step procedure with 51 response alternatives. This indicates that the response statistics derived from this data can be used to predict the accuracies of categorical loudness scaling procedures with up to about 50 response alternatives.

Near the limits of the category range, the s_R values decrease towards 0 cu, because a mean value equal to the limit can only occur if all values are equal to the mean. On the other hand, the reproducibility of the loudness ratings might be higher at the limits of the dynamic range than at medium loudness values, because the limits of the dynamic range might be defined best in terms of subjective loudness. To decide whether the decrease of s_R is only due to the limited response scale or whether it is supported by a possibly smaller standard deviation σ_N of the subjective loudness near the limits of the dynamic range, Monte-Carlo simulations with constant settings of $\sigma_N = 2, 4, 6, 8, 10$ cu were performed. In these simulations, the stimulus level was varied in order to obtain a mean subjective loudness impression N in the range from -10 to 60 cu. The probability of outliers p_{out} was set to zero. For a large number of 20,000 runs the loudness ratings $R(L)$ were calculated. The lines in Fig. 3.5 show the standard deviations $s_R(L)$ of the simulated loudness ratings as a function of their mean value $\bar{R}(L)$. The graph for an underlying σ_N equal to 4 cu approximates closest the data of the normal–hearing subjects. Near the limits of the scale, in the ranges 0 cu $\leq \bar{R} \leq$ 5 cu and 45 cu $\leq \bar{R} \leq$ 50 cu, respectively, the simulated s_R values are only slightly higher than the observed values in Fig. 3.5. This indicates that normal–hearing subjects can reproduce loudness ratings near the limits of the response scale slightly better than in the intermediate loudness range. Taken together, it may be concluded that σ_N in the normal–hearing subjects is about 4 cu and that it decreases only slightly near the limits of the category range.

3.4.2.1 Differences in reproducibility

The measured values of s_R in Fig. 3.5 show a large variability for a given $\bar{R}$. This might be due to both a real interindividual variability of s_R and to the relatively small number of 20 repeated measurements per subject which only allows for a relatively imprecise estimate of s_R. Monte–Carlo simulations with a constant σ_N of 4 cu and no outliers were performed in order to investigate the influence of the small number of repeated measurements. The only difference to the simulations in Sec. 3.4.2 is that the number of runs per simulated level was set to 20, which is the same value as in the measurements. The resulting s_R values as a function of $\bar{R}$ are shown in Fig. 3.6. The variability in the simulated s_R values in Fig. 3.6 is only slightly smaller than the variability of the s_R values in the measurements (cf. Fig. 3.5) that results both from

the interindividual differences in s_R and from the small number of samples. Thus, the interindividual standard deviation in the reproducibility of loudness ratings in the normal–hearing subjects cannot be derived from this data basis.

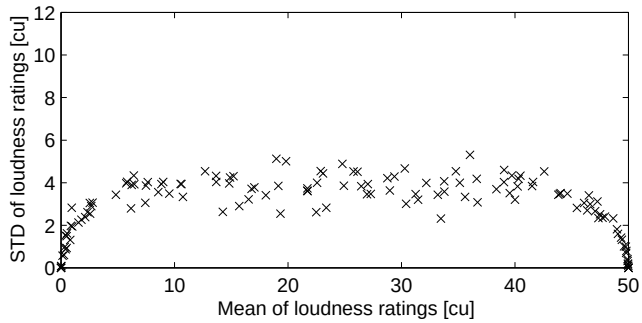

Figure 3.6: *Simulated intraindividual standard deviations s_R of repeated loudness ratings as functions of their respective mean values. A constant underlying standard deviation of subjective loudness $\sigma_N = 4$ cu and 20 Monte-Carlo runs per data point were used.*

3.4.3 Hearing–impaired subjects

Fig. 3.7 shows the observed intraindividual standard deviations s_R in the repeated ratings of the same stimulus level L for the hearing–impaired subjects plotted in the same way as in Fig. 3.5. The s_R values in the hearing–impaired subjects show a considerably higher variability than in the normal–hearing subjects. Many s_R values are considerably higher than the mean value of 4 cu found in the normal–hearing subjects. As in the normal–hearing subjects, no difference in the reproducibility of the loudness ratings can be seen between the one–step procedure with 11 response alternatives and the 2–step procedure with 51 response alternatives. The scatter plot shows no similarity to the distribution in Fig. 3.6 which was generated using a constant underlying σ_N value. This indicates that there are large interindividual differences in the reproducibility of loudness ratings in hearing-impaired subjects. To a certain degree, they may be due to outliers.

3.4.4 Conclusions for statistical modeling

The individual σ_N values for normal–hearing and hearing–impaired subjects derived in Secs. 3.4.2 and 3.4.3 show a large variability. Two different estimates of the response characteristics are used in the following simulations, to account for this effect. An

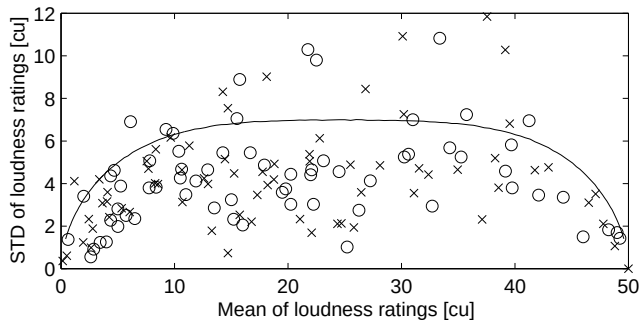

Figure 3.7: Intraindividual standard deviations s_R of repeated loudness ratings as functions of their respective mean values in 10 hearing-impaired subjects.

'optimistic' response statistic is estimated with $\sigma_N = 4$ cu and a probability of outliers $p_{\text{out}} = 0$. This response statistic is typical for the normal–hearing subjects and many of the hearing–impaired subjects. A 'pessimistic' response statistic is estimated with $\sigma_N = 7$ cu and $p_{\text{out}} = 0.05$. This response characteristic is considerably more variable than the mean response characteristic in the normal–hearing subjects derived above (Sec. 3.4.2) and also more variable than the response characteristic of most hearing–impaired subjects (cf. Sec. 3.4.3). In this way, a conservative estimate of the measuring accuracy is achieved and it can be expected that in most real measurements the accuracy is better than the results of simulations based on the pessimistic response characteristic.

3.5 PREDICTIONS

3.5.1 Robustness against outliers

In Sec. 3.3.3 it has been stated that the Lorentzian fit is more robust against outliers in the data than the Gaussian fit. In this section the differences between the accuracies of the Lorentzian and the Gaussian fitting procedure for different probabilities of outliers p_{out} are calculated using Monte–Carlo simulations. In this way, the limit of p_{out} which requires the Lorentzian fit rather than the Gaussian fit is calculated.

The underlying standard deviation of the subjective loudness $\sigma_N(L)$ is set to 4 cu as in the optimistic estimate of the response characteristic. The probability of outliers p_{out} is varied from 0 to 0.25. Fig. 3.8 shows the normalized standard deviations $m\sigma_{L_{25}}$ and $\frac{\sigma_m}{m}$ and normalized biases $mb_{L_{25}}$ and $\frac{b_m}{m}$ of the L_{25} and m estimates as functions of p_{out}. The Gaussian fit yields smaller standard deviations and biases in both L_{25} and m

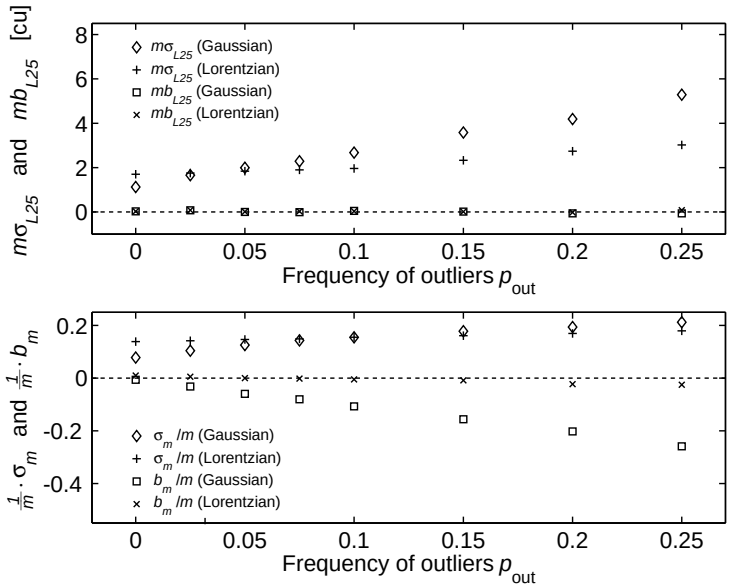

Figure 3.8: *Upper panel: Normalized standard deviation $m\sigma_{L_{25}}$ and normalized bias $mb_{L_{25}}$ of L_{25} estimates as a function of the probability of outliers p_{out} with Gaussian fit and with Lorentzian fit. Lower panel: Relative standard deviation $\frac{\sigma_m}{m}$ and relative bias $\frac{b_m}{m}$ of m estimates as a function of p_{out} with Gaussian fit and Lorentzian fit.*

estimates if p_{out} is equal to 0. Especially, the standard deviation of the m estimate is approximately 50 % smaller than in the Lorentzian fit. Thus, the Gaussian fit is clearly favorable in subjects with a small incidence of outliers. With increasing p_{out}, however, the accuracy of L_{25} and m estimates decreases more using the Gaussian fit than using the Lorentzian fit. Especially, the standard deviation of L_{25} estimates and the bias of m estimates increase rapidly using the Gaussian fit. Therefore, the Lorentzian fit is clearly favorable for p_{out} values above 0.07.

3.5.2 Track length

Monte–Carlo simulations were performed in which the track length n was systematically varied in order to quantify the influence of n on the accuracy of the measurements. The optimistic response characteristic ($\sigma_N = 4$ cu, $p_{\text{out}} = 0$) and the pessimistic response characteristic ($\sigma_N = 7$ cu, $p_{\text{out}} = 0.05$) were used. In the first case, the Gaussian fit and in the latter case the Lorentzian fit was used, to derive the parameters L_{25} and m from each simulated track. Fig. 3.9 shows the normalized standard deviations and

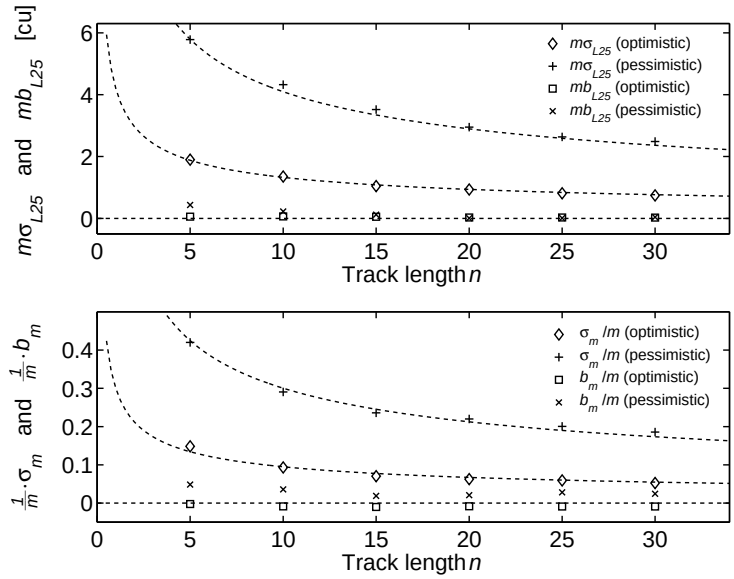

Figure 3.9: *Upper panel: Normalized standard deviation* $m\sigma_{L_{25}}$ *and normalized bias* $mb_{L_{25}}$ *of* L_{25} *estimates as a function of the track length* n*. Lower panel: Relative standard deviation* $\frac{\sigma_m}{m}$ *and relative bias* $\frac{\sigma_m}{m}$ *of* m *estimates as a function of* n*. In both panels the results are shown of simulations based on both the optimistic and the pessimistic estimate of the response characteristic. The dashed lines are proportional to* $\frac{1}{\sqrt{n}}$.

biases of the L_{25} and m estimates as a function of n. The standard deviations of both L_{25} and m decrease very closely to $\frac{1}{\sqrt{n}}$ which is indicated by dashed lines. In both L_{25} and m estimates the biases are much lower than the standard deviations. Therefore, the estimates can be regarded as almost bias free.

If more than 15 trials are used, $m\sigma_{L_{25}}$ is below 2 cu in all subjects with a response characteristic similar to the optimistic estimate, cf. Fig. 3.9. That means, that the intraindividual standard deviation in L_{25} estimates is below 4 dB in these subjects because loudness function slope values are typically about 0.5 $\frac{\text{cu}}{\text{dB}}$ and higher. The intraindividual standard deviation in m estimates is about 10 % of the actual m value in these subjects using about 15 trials.

In case of a response characteristic similar to the pessimistic estimate, 15 trials yield an intraindividual standard deviation in L_{25} estimates below 8 dB. The intraindividual standard deviation in m estimates is about 25 % of the actual m value under this condition.

That means, a track length of about 15 trials, as it is used in the constant stimuli version of the Oldenburg loudness scaling procedure (14 trials), is sufficient for diagnostics and hearing aid fitting in listeners who have a normal response characteristic. In listeners who have difficulties in rating loudness accurately, the accuracy of the procedure might be insufficient. This, especially, holds for the estimation of the loudness function slope.

3.5.3 Number of response alternatives

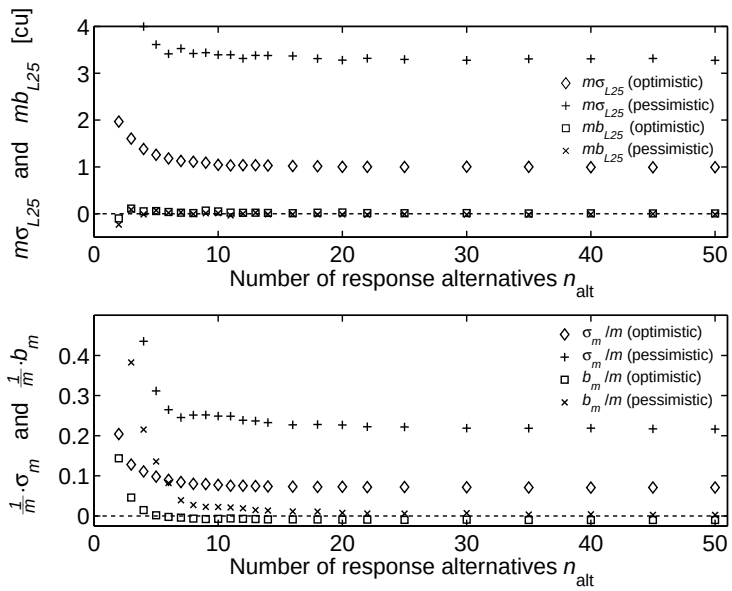

Figure 3.10: *Upper panel: Normalized standard deviation* $m\sigma_{L_{25}}$ *and normalized bias* $mb_{L_{25}}$ *of* L_{25} *estimates as a function of the number of response alternatives* n_{alt}. *Lower panel: Relative standard deviation* $\frac{\sigma_m}{m}$ *and relative bias* $\frac{\sigma_m}{m}$ *of* m *estimates as a function of* n_{alt}.

Monte–Carlo simulations with both the optimistic response characteristic ($\sigma_N = 4$ cu, $p_{\mathrm{out}} = 0$) and the pessimistic response characteristic ($\sigma_N = 7$ cu, $p_{\mathrm{out}} = 0.05$) were performed in order to quantify the influence of the number of response alternatives n_{alt} on the accuracy of the measurement. In the first case, the Gaussian fit and in the latter case the Lorentzian fit was used. To reduce stochastic effects, the number of Monte–Carlo runs was increased to 5,000 in these simulations.

Fig. 3.10 shows the normalized standard deviations and biases of the L_{25} and m estimates as a function of n_{alt}. The standard deviations of both L_{25} and m estimates decrease

with increasing number of response alternatives until $n_{\mathrm{alt}} \approx 10$ is reached. A further increasement of n_{alt} gives no further increase in accuracy. For n_{alt} values below 10, considerably large biases in m estimates occur. This bias is largest if the pessimistic estimate is used.

3.6 EVALUATION OF PREDICTIONS

Table 3.1: Predicted and measured intraindividual standard deviations of L_{25} and m estimates (normalized with m).

	predicted		measured	
	optimistic	pessimistic	norm.–hear.	hear.-imp.
$m \cdot \sigma_{L_{25}}$	1.1 cu	3.6 cu	1.3 cu	2.0 cu
$\frac{\sigma_m}{m}$	0.07	0.24	0.09	0.12

Measurements with 10 normal–hearing listeners and 10 listeners with sensorineural hearing impairment were performed in order to evaluate the predictions of the statistical model. A German sentence was used as stimulus, cf. Sec. 3.3.5.[5] Each listener performed 10 repeated loudness scaling tracks (5 per ear). None of the listeners did prior participate in the measurements which were used to fit the statistical model (Sec. 3.4). The modified least–squares fit (Gaussian merit function as described in Sec. 3.3.3) was used to fit the model function to the data. The track length was 14 (2 trials at 7 levels), according to Sec. 3.3.1. The number of response alternatives n_{alt} was 11, according to Fig. 3.2.

Table 3.1 shows the measured intraindividual standard deviations of the L_{25} and m estimates (normalized with m), as well as predicted values. (The standard deviation was calculated for each ear and than averaged across all subjects.) In the predictions the optimistic and the pessimistic estimate of the underlying response statistic was used.

In the normal–hearing listeners, the normalized intraindividual standard deviations of both L_{25} and m estimates are slightly higher than the corresponding optimistic predictions. This difference might be due to the fact that an optimal even stimulus level distribution covering the full auditory dynamic range was assumed in the predictions, whereas the pre–measurement often produces a suboptimal estimate of the auditory dynamic range in the measurements.

[5] This German sentence has an almost linear loudness function in both normal–hearing and hearing–impaired listeners, cf. Chap. 6. Therefore, the linear model function which is used in this study in order to simplify the predictions can be fitted adequately to the measurement data, as well. In narrowband stimuli, a non–linear model function is recommended, cf. Chapters 5 and 6.

In the hearing–impaired listeners, the normalized standard deviations are higher than in the normal–hearing subjects. This was already expected from the data by Hohmann (1993) which were used in Sec. 3.4, where the hearing–impaired group showed a higher standard deviation than the optimistic estimate as well. However, the standard deviations are clearly smaller than the pessimistic predictions. This indicates that the hearing–impaired listeners in this study showed a more consistent response behavior than expected on the basis of the pessimistic estimated response statistic that assumes 5 % 'outliers'.

3.7 DISCUSSION

In both normal–hearing and hearing–impaired listeners no significant difference in the reproducibility of loudness ratings was observed between the one–step scale with 11 and the two–step scale with 51 response alternatives. A similar result was derived by Hohmann (1993) who fitted a linear model function with a variable offset for small levels (cf. Eq. 5.3) to the same data and observed equal standard deviations of the model parameters L_{25} and m within subjects for both scales.

A simple statistical model which was fitted to the response statistics of normal–hearing and hearing–impaired listeners was used to predict the influence of different parameters in loudness scaling procedures on the accuracy of the measurement:

The Gaussian fit generated smaller standard deviations and biases than the Lorentzian fit in both L_{25} and m estimates if the probability of outliers in the response characteristic was small, i.e., below about 7 %. Thus, in subjects with rare lacks in attention the Gaussian fit is preferable. If the incidence of outliers increases to values above 7 %, the Lorentzian fit gives a better accuracy. Especially, the standard deviation of the L_{25} estimates and the bias in m estimates is considerably smaller than using the Gaussian fit. The Lorentzian fit is preferable in subjects who may have frequent lacks in attention, which may be the case in many clinical situations. Most subjects who participated in this study showed constant attention, so that the Gaussian fit was adequate. However, it has to be taken into account that all subjects participated voluntarily and had a high level of motivation. Further, they became very familiar with the procedure during the study.
A different method to generate a robust fit is to manually remove data points that deviate strongly from the remainder before the model function is fitted. This method may combine the advantages of both, the Gaussian fit and the Lorentzian fit, i.e., small standard deviations in estimates for low p_{out} and robustness against outliers. However, it requires empirical rules for identifying outliers and may bias the results towards the expectations of the experimenter.

The most important factor influencing the accuracy of the measurement is the track length n. Theoretically, the standard deviation of estimates of parameters of the loudness function, such as L_{25} and m is proportional to $\frac{1}{\sqrt{n}}$ for large track lengths. The results of the simulations were in accordance to this rule. The constant stimuli version of the Oldenburg loudness scaling procedure (Hohmann and Kollmeier, 1995b) uses 14 trials while the adaptive version (cf. Chap. 4) uses about 20 trials.[6] In subjects with a normal response statistic ($\sigma_N \approx 4$ cu, $p_{\text{out}} \approx 0$), these track lengths are sufficient since the L_{25} estimates have an intraindividual standard deviation below about 2 cu divided by m (cf. Fig. 3.9) and the m estimates have an intraindividual standard deviation of about 0.1 relative to m. In subjects with a more variable response statistic, however, the accuracy in the estimates decreases. In subjects with a response statistic which is as variable as the pessimistic estimate ($\sigma_N \approx 7$ cu, $p_{\text{out}} \approx 0.05$), the L_{25} estimates have an intraindividual standard deviation of almost 3 cu divided by m and the m estimates have a relative intraindividual standard deviation of about 0.3. Especially, the accuracy of the m estimate is very poor and probably not sufficient for many applications.

Another factor influencing the accuracy of the measurement is the number of response alternatives n_{alt}. The simulations revealed that $n_{\text{alt}} = 10$ is sufficient to achieve the maximum accuracy of parameter estimates. A smaller number of response alternatives would force 'accurate' subjects to give inaccurate responses. This would lead to higher intraindividual standard deviations of L_{25} and m estimates and to a large positive bias in m estimates, cf. Fig. 3.10. These findings are in accordance with Allen *et al.* (1990) who used a scale with 7 response alternatives and reported that many of their subjects suggested that the number of response alternatives was insufficient. However, in the Oldenburg loudness scaling procedure with 11 response alternatives, many subjects complained about an insufficient number of response alternatives as well. Especially, normal–hearing subjects who scaled narrow–band stimuli frequently reported that the number of response alternatives was insufficient for soft stimuli. That means that in contrast to the results of the simulations, a minimum number of 10 response alternatives may be insufficient to get the optimal measuring accuracy. The reason for this mismatch between the simulations and the complaints of the subjects is probably that the loudness function was assumed as a straight line in the simulations. However, the measurements reported in Chap. 5 and 6 show that most normal–hearing people have loudness function for narrow band signals which is upwardly concave. For the 1 kHz stimulus used in this study, the usual slope of the loudness function is about 0.3 $\frac{\text{cu}}{\text{dB}}$ below the L_{25} level and about 0.9 $\frac{\text{cu}}{\text{dB}}$ above the L_{25}. That means, the dynamic range which is related to the lower half of the loudness range is about 3 times wider than the dynamic range which is related to the upper half of the loudness range. Since σ_N is

[6] The measuring time, however, is about the same in both procedures since the pre–measurement which is necessary in the constant stimuli procedure is omitted in the adaptive procedure. Consequently, the adaptive procedure gives more accurate results using the same measuring time.

approximately constant with level, the minimal number of response alternatives in the lower loudness range has to be increased by a factor of about $\frac{3}{2}$. Since the response alternatives should be placed equidistantly in the loudness range (Parducci and Perret, 1971; Poulton, 1989), the number of response alternatives has to be increased by about $\frac{3}{2}$ in the upper half of the loudness range as well. Thus, the minimum number of response alternatives which yields the optimal measuring accuracy in normal–hearing subjects is expected to be about $10 \cdot \frac{3}{2} = 15$. If this is taken into account, the number of 11 response alternatives which is used in the Oldenburg loudness scaling procedure has to be regarded as an compromise between measurement accuracy and minimum complexity of the procedure which causes losses in accuracy, at least in some normal–hearing subjects.

In addition to the statistical aspects that were investigated in this study (i.e., the number of response alternatives, the track length and the fitting procedure) the accuracy of the measurement is also influenced by psychological aspects such as context effects. Although the subject is asked to rate the absolute loudness, each trial is rated in the context of the preceding trials (Garner, 1954; Parducci, 1965; Poulton, 1968; Poulton, 1989; Gabriel, 1996) which can cause different types of biases (for a review see Poulton (1989)). In Chap. 4 an adaptive procedure for categorical loudness scaling is presented, which was designed in order to reduce these bias effects. The Monte–Carlo method for categorical loudness scaling procedures which was introduced here is used in Chap. 4 to optimize this adaptive procedure.

Another factor which may decrease the measuring accuracy is the distribution of stimulus levels. In the simulations of this study this distribution was set in an optimum way, i.e., equidistantly on a dB–scale covering the full dynamic range of the subject. If the measuring procedure generates a sub–optimal level distribution, e.g., because the pre–measurement gave an inaccurate estimate of the dynamic range of the subject, the measuring accuracy is worse than predicted by the simulations of this chapter. The effect of the level distribution on the accuracy is investigated empirically in Chap. 4.

The evaluation measurements with 10 normal–hearing listeners using a broadband stimulus (German sentence) generated standard deviations of L_{25} and m estimates which were only slightly larger than the predicted values. The hearing–impaired listeners showed somewhat larger standard deviations which was expected because hearing–impaired listeners sometimes have difficulties to reproduce loudness estimates accurately.
As mentioned above, the statistical model introduced here is also used to predict the accuracy of an adaptive loudness scaling procedure, cf. Chap. 4. This procedure uses a different stimulus placement, a different model loudness function and a different number of stimuli. The evaluation measurements of this procedure with 10 normal–hearing and 10 hearing–impaired listeners also showed accuracies of loudness function parameter estimates which were very similar to the predicted values. This consistence between predictions and measurements in Chap. 4 indicates that the statistical model might be appropriate to predict the accuracy of arbitrary categorical loudness scaling procedures.

3.8 SUMMARY AND CONCLUSIONS

A simple statistical model is developed which enables to predict the accuracy of arbitrary categorical loudness scaling procedures using Monte-Carlo simulations. The model requires the knowledge of the subjects loudness function and of his/her response statistic, i.e. the reproducibility of single loudness rating trials. The underlying response statistic is derived from repeated measurements by Hohmann *et al.* (1993). In normal–hearing subjects the response statistic can be characterized by a standard deviation of categorical loudness ratings which is equal to 4 cu. In the hearing–impaired subjects this standard deviation ranges between 3 cu and 10 cu. The simulations give the following results:

- If the incidence of outliers in the subject's responses exceeds 7 % the Lorentzian fit is more robust than the Gaussian fit (least–squares fit).
- The standard deviations of the estimates of the loudness function parameters L_{25} and m decrease proportional to $\frac{1}{\sqrt{n}}$, with n denoting the track length. The actual standard deviations of the estimates depend on the respective reproducibility of the loudness ratings of the subject. It is recommended to use at least 14 trials to yield a reliable loudness function estimate. In normal–hearing and many–hearing impaired listeners this track length yields intraindividual standard deviations of about 2 cu divided by m in L_{25} estimates and a relative intraindividual standard deviation of about 0.1 in m estimates. However, in subjects with a response statistic which is more variable than normal the standard deviation of L_{25}, estimates increases up to 5 cu divided by m and the standard deviation of m estimates increases up to 0.3 relative to m.
- To avoid losses in accuracy at least 10 response alternatives are necessary. Otherwise standard deviations increase and in subjects with outliers in their responses the m estimates are strongly biased. In many subjects the accuracy may be further improved if the number of response alternatives is increased up to 15. An even further increasement of the number of response alternatives generates no further improvement of accuracy.

ACKNOWLEDGEMENT

We would like to thank Stefan Uppenkamp for helpful comments on the manuscript.

This study was supported by BMFT, PT AUG and the DFG.

Chapter 4

Design and Evaluation of an Adaptive Procedure for Categorical Loudness Scaling

ABSTRACT

In categorical loudness scaling procedures the presentation levels have to be adjusted to the individual auditory dynamic range of the subject. Different categorical loudness scaling procedures realize this either by employing pre–measurements which determine the stimulus range of the following measurement phase or by employing strictly ascending level sequences which start at the hearing threshold and stop when the subject rates the stimulus as being 'too loud'. Pre–measurements are inefficient and are found to produce inaccurate results. Strictly ascending level sequences are known to bias the loudness function estimate. In this study, an adaptive procedure for categorical loudness scaling is introduced and evaluated which yields high efficiency because a pre–measurement is omitted and which uses randomized presentation levels in order to reduce biases. The procedure aims at generating evenly distributed stimulus levels in pseudo–random order covering the full auditory dynamic range of the subject with a minimal number of presentation levels outside this range. The novel procedure is based on the constant stimuli version of the "Oldenburg loudness scaling procedure" (Hohmann and Kollmeier, 1995b). It has been named "Oldenburg – ACALOS" (Oldenburg – Adaptive CAtegorical LOudness Scaling). The adaptive procedure is evaluated both with Monte–Carlo simulations and with repeated measurements on 10 normal–hearing and 10 sensorineural hearing–impaired subjects obtained with both the constant stimuli and the adaptive version of the Oldenburg loudness scaling procedure. The desired even distribution of stimulus levels is better approximated by the adaptive version than by the constant

stimuli version in both groups of subjects. The adaptive procedure yields considerably smaller intraindividual standard deviations in loudness estimates than other procedures. The intraindividual standard deviation of the response level estimates (i.e., the levels corresponding to certain loudness categories) is less than 5 dB for all categories using a total number of about 20 trials.

4.1 INTRODUCTION

The categorical loudness scaling method (e.g., Pascoe, 1978; Heller, 1985; Allen *et al.*, 1990; Hohmann and Kollmeier, 1995b) enables the assessment of loudness as a function of presentation level for individual subjects and different stimuli. Loudness functions are used for the diagnosis of loudness recruitment in clinical audiology (e.g., Allen *et al.*, 1990; Hellmann and Meiselman, 1993; Kießling *et al.*, 1993; Kießling, 1995; Hohmann and Kollmeier, 1995b; Launer, 1995) and to determine the input/output characteristics of hearing aids with automatic gain control (e.g., Pascoe, 1978; Geller and Margiolis, 1984; Hellbrück and Moser, 1985; Margiolis, 1985; Moore *et al.*, 1992; Kießling *et al.*, 1995). Furthermore, categorical loudness scaling provides the basis for the development of loudness models predicting the loudness of various stimuli in normal–hearing and hearing–impaired subjects (Launer, 1995; Blum, 1998).
An advantage of categorical loudness scaling in audiology is its ability to estimate not only the hearing threshold level (HTL) and the uncomfortable level (UCL) but also the shape of the whole loudness function using stimulus levels in the complete auditory dynamic range. Pure–tone audiomety determines only the extremes of the auditory dynamic range, i.e., HTL and sometimes also UCL, using extreme stimulus levels. These extremes do not occur frequently in daily life situations. At least they should not occur frequently in a well fitted hearing aid. Consequently, it seems inadequate to base diagnostics and hearing aid fitting on the measurement of these extreme values. Furthermore, the measurement of HTL and UCL alone gives no information about the shape of the loudness function in between. Different subjects can differ in their loudness functions even if they have the same HTL and UCL values.

However, clinical diagnostic is still mostly based on the pure–tone audiogram as well as the prescriptive rules used in hearing–aid fitting (McCandless and Lyregaard, 1983; Byrne and Dillon, 1986; Berger *et al.*, 1988; Cornelisse *et al.*, 1995). This is probably due to the fact that measurement time necessary to assess the loudness function with categorical scaling is long when compared to pure–tone audiometry. This study is an approach to overcome this problem using an optimized adaptive procedure for categorical loudness scaling which is more efficient than conventional loudness scaling methods.

Each loudness scaling procedure has to ensure that the number of inaudible stimuli and of too–loud stimuli is low, because the first are ineffective and the latter annoy the listener

and can be dangerous if they are much too loud. Further, the stimuli should be evenly distributed within the limits that span the whole individual auditory dynamic range to reduce bias effects (Parducci and Perret, 1971; Montgogomery, 1975). The different categorical loudness scaling procedures available today (e.g., Pascoe, 1978; Heller, 1985; Allen *et al.*, 1990; Elberling and Nielsen, 1993; Hohmann and Kollmeier, 1995b; Ricketts and Bentler, 1996; Cox *et al.*, 1997; Rasmussen *et al.*, 1998; Keidser *et al.*, 1999) solve this problem either using a pre–measurement, which roughly determines the individual auditory dynamic range before the actual data collecting phase, or by repeated ascending level sequences which stop when the uncomfortable level is reached.
The use of a pre–measurement (e.g., Heller, 1985; Allen *et al.*, 1990; Elberling and Nielsen, 1993; Hohmann and Kollmeier, 1995b; Ricketts and Bentler, 1996; Rasmussen *et al.*, 1998) has the advantage that the stimulus levels in the data collecting phase can be randomized which reduces biases. On the other hand it has the disadvantage that the auditory dynamic range is often estimated inproperly by the pre–measurement. In such cases, the stimulus levels are not set optimal in the data collecting phase which might cause significant biases, because loudness estimates depend on the range of presentation levels (e.g., Teightsoonian, 1973; Marks and Warner, 1991; Gabriel, 1996). Many subjects systematically shift their response criterion to cover the dynamic range of stimuli using the entire scale (Poulton, 1989). Further, the pre–measurement needs time but produces no data for later analysis, in other words, it is ineffective.
Ascending level sequences (e.g., Pascoe, 1978; Cox *et al.*, 1997; Keidser *et al.*, 1999) have the advantage that they estimate the upper limit of the dynamic range in every ascending run, that means they are flexible if the subject changes his/her criterion for too–loud stimuli during the track. On the other hand, ascending level sequences cause significant biases in loudness function estimates and the scaling results deviate significantly from those of descending or randomly selected level sequences (e.g., Gabriel, 1996; Jenstad *et al.*, 1997; Kollmeier, 1997).

The Oldenburg–ACALOS (Adaptive CAtegorical LOudness Scaling) procedure presented in this study is an approach to realize high efficiency and small biases using an adaptive stimulus selection strategy which needs no pre–measurement and randomizes presentation levels. It is based on the constant stimuli version of the Oldenburg loudness scaling procedure, which uses a pre–measurement consisting of an ascending level sequence to roughly estimate the auditory dynamic range of the subject (Hohmann and Kollmeier, 1995b). As stated above, this initial ascending level sequence was found to produce a bias, i.e. the estimates of the uncomfortable levels were too low. This bias generates an inadequate choice of stimulus levels in the subsequent data collection.
In the adaptive procedure, the pre–measurement is omitted in order to reduce measurement time. Instead, the auditory dynamic range of the subject is estimated coarsely during the first few trials by estimating the upper and the lower limit of the individual dynamic range concurrently, i.e. without an ascending or decreasing level sequence. In

contrast to the constant stimuli version, the actual dynamic range is recalculated several times during the track. By this means, the adaptive procedure attempts to adapt presentation levels to the individual auditory dynamic range using the previous responses. Hence, this procedure aims at preventing the subject from adapting responses to a fixed predefined range of stimulus levels.

In Chap. 3 a method was developed to simulate different categorical loudness procedures with Monte–Carlo simulations. This method is used here to optimize the adaptive procedure with respect to the desired distribution of stimulus levels and to estimate the accuracy of the procedure. Subsequently, the optimized procedure was evaluated with 10 normal–hearing and 10 hearing–impaired listeners.

4.2 CONSTANT STIMULI AND ADAPTIVE PROCEDURE

4.2.1 Response scale

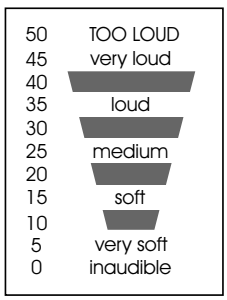

Figure 4.1: *Category scale with 11 response alternatives used by the listeners to rate the loudness. The numbers on the left side indicate the categorical units (cu) which are used for data storage and analysis. They were not visible to the subject.*

Both constant stimuli and adaptive procedure use the same response scale. A schematical view of the scale is given in Fig. 4.1. The scale consists of eleven response alternatives including five named loudness categories, four not named intermediate response alternatives and two named limiting categories. The named response categories are ‘sehr leise’ (‘very soft’), ‘leise’ (‘soft’), ‘mittel’ (‘medium’), ‘laut’ (‘loud’) and ‘sehr laut’ (‘very loud’) and correspond to 5, 15, 25, 35 and 45 cu (categorical units) as shown on the left side of Fig. 4.1. The not named response alternatives are used to increase the total number of response alternatives. They are indicated with horizontal bars with increasing

length for increasing loudness and are placed between the named loudness categories. They correspond to the categorical units 10, 20, 30 and 40 cu, respectively. The two limiting categories are named 'unhörbar' ('inaudible') and 'ZU LAUT' ('TOO LOUD') and correspond to 0 and 50 cu.

Prior to testing, each listener was instructed verbally by the experimenter (cf. A.2). During the instruction, the response box was practically demonstrated and any questions were clarified.

4.2.2 Constant stimuli procedure

The constant stimuli version of the Oldenburg loudness scaling procedure (Hohmann and Kollmeier, 1995b) includes two parts. The auditory dynamic range of the individual listener is estimated by presenting an ascending level sequence in the first part of the measurement. The loudness function is assessed by presenting stimuli covering the so determined full auditory dynamic range in the second part.
The first part uses an ascending stimulus level sequence with an initial level of 0 dB HL and a step size of 5 dB. The listener's task is to press the response button as soon as the stimulus is audible. Then the level is further increased in 15 dB steps up to 85 dB and increased in 5 dB steps beyond 85 dB. Now, the listener is asked to press another response button 'too loud' immediately when the stimulus is perceived as too loud. In case that the listener does not press the response button, the sequence stops at 115 dB HL.
In the second part of the procedure the loudness function is estimated. Two stimuli are presented at each of 7 diffrent levels which are distributed equidistantly on a dB–scale between the limits of the dynamic range estimated in the first part of the procedure. The listener rates the loudness using the scale described above. The stimuli are presented in pseudo–random order in a way that the maximum difference of subsequent presentation levels is smaller than half of the dynamic range of the sequence in order to avoid context effects which are due to the tendency of many listeners to rate the current stimulus relatively to the previous stimulus. After completion of the track a model function is fitted to the data by a modified least–squares fit, cf. Sec. 4.3.1.[1]

4.2.3 Adaptive procedure

Most adaptive procedures in psychoacoustics converge at certain given target values in the response domain to achieve maximum accuracy in threshold estimates (e.g., Taylor and Creelman, 1967; Levitt, 1971; Campbell, 1974). Such procedures are often used in

[1] In the original constant stimuli version of the Oldenburg loudness scaling procedure a different model function was used, i.e. a linear function with variable offset at low levels (cf. Eq. 5.3).

loudness comparison methods as well, usually in randomly interleaved sequences of trials (e.g., Jesteadt, 1980; Buus *et al.*, 1999; Verhey, 1989) However, converging at specific targets would be not adequate for categorical loudness scaling experiments, because this would cause significant response biases due to range– and context–effects (e.g., Poulton, 1989). An adaptive categorical loudness scaling procedure should fulfill the following demands in order to avoid biases:
1) The stimulus spacing should be subjectively equal (Parducci and Perret, 1971; Poulton, 1989).
2) The whole individual auditory dynamic range of stimuli should be employed in order to reduce range equalizing biases. When this goal is achieved, the range of presentation levels is subjectively equal to each subject.
3) Presentation levels outside of the auditory dynamic range have to be avoided, because inaudible stimuli reduce efficiency and too–loud stimuli may harm the subject.
4) The order of stimulus levels should be randomized.

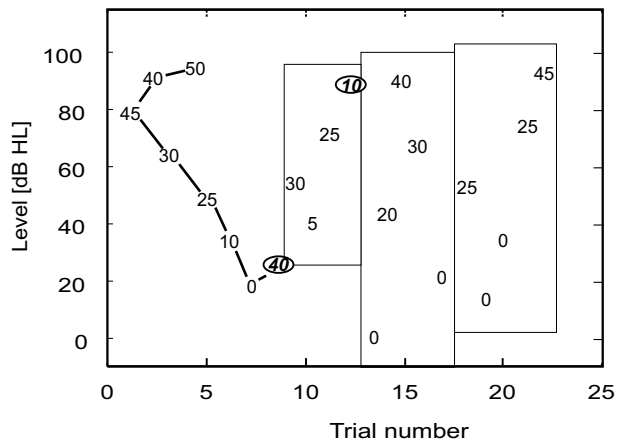

Figure 4.2: Example of a track produced by the adaptive procedure in a Monte–Carlo simulation, cf. Sec. 4.3.2. The responses are indicated with numbers between 0 ('inaudible') and 50 ('too loud'). The numbers that are marked with ellipses indicate outliers in the response statistics. The abscissa value of each data point indicates the trial number. The ordinate value of each data point indicates the presentation level. Those presentation levels which belong to the same block of the adaptive procedure in the second part of the procedure are combined by rectangles. The upper and lower limits of the rectangles correspond to the limits of the estimated auditory range per block.

The adaptive version of the Oldenburg loudness scaling procedure (Brand *et al.*, 1997a; Brand *et al.*, 1997b) was designed to fulfill the demands mentioned above.
The procedure consists of two different phases. However, the listener is not aware that

there are two phases, since his/her task is the same – to scale loudness – in both phases. The dynamic range of the listener is roughly estimated in the first phase. More data are collected in the second phase and the estimate of the dynamic range in which the stimuli are presented is updated two times.

The first phase starts with a stimulus at 80 dB HL. When the initial stimulus is inaudible or too loud, it is increased or decreased, respectively, in 15 dB steps until a response between inaudible and too loud is achieved. Thereafter, two interleaved sequences of stimuli begin. The first sequence increases the stimulus level in 5 dB steps until the response 'too loud' is given or the maximum level of 120 dB is reached. The final stimulus level of this sequence is the first estimate of the upper limit of the dynamic range. The second sequence decreases the stimulus level in 15 dB steps until it is inaudible or the limit of 0 dB HL is reached. In the first case the level is increased again with 5 dB steps until it is audible. The final stimulus level of this sequence is used as the first estimate of the lower limit of the dynamic range.

The second phase consists of two blocks according to the following procedure: Five stimulus levels, i.e. L_5 ('very soft'), L_{15} ('soft'), L_{25} ('medium'), L_{35} ('loud') and L_{45} ('very loud'), are estimated by linear interpolation between the estimated limits of the dynamic range. For the first block the limits estimated within the first phase are used. Before the second block starts, the limits are recalculated by fitting a linear function to the results of all previous trials in the track including the first phase with a modified least–squares fit (cf. Sec. 4.3.1). The five levels in each block are rounded to integer values and presented in pseudo–random order, in a way that subsequent levels do not differ more than half of the dynamic range of the sequence. After the second block, a model function is fitted to all data with a modified least–squares fit. To increase the measurement accuracy, more than 2 blocks can be performed in the second phase, i.e., the procedure of the second block can be repeated several times. Fig. 4.2 gives a sketch of three subsequent blocks.

There are two exceptions to these rules: Firstly, the estimated value of L_5 is not presented in the first block since there have been several trials near L_5 during the first phase. Secondly, the listener is protected against stimuli which are too loud. If the adaptive rule requires a stimulus level which was rated as 'too loud' before, the stimulus level is limited to maximal 5 dB above the level which was rated as 'too loud'.[2]

[2] It may seem inadequate to present a stimulus with a larger level than a stimulus which was prior rated as 'too loud'. However, experimental findings from the constant stimuli procedure indicate that many subjects shift their criterion for 'too loud' stimuli upwards during the track. Note, that the adaptive procedure demands a stimulus which is louder than a stimulus prior rated as 'too loud' only in those cases in which the preceding 'too loud' rating was inconsistent with the majority of the remaining responses.

4.3 EXPERIMENTAL METHOD

4.3.1 Model function and fitting

During the evaluation of the adaptive procedure using Monte–Carlo simulations, a linear model function with the parameters L_{25} (level related to 25 cu ('medium')) and m (slope) was used.

$$F(L) = 25 + m(L - L_{25}) \tag{4.1}$$

However, the evaluation of the procedure with human subjects showed that the linear model function is not adequate, because the obtained loudness functions showed an increasing slope. Hence, a different model function was used to fit the measurements:

$$F(L) = \begin{cases} 25 + m_{\mathrm{lo}}(L - L_{\mathrm{cut}}) & \text{for} \quad L \leq L_{15} \\ \mathrm{bez}(L, L_{\mathrm{cut}}, L_{15}, L_{35}) & \text{for} \quad L_{15} < L < L_{35} \\ 25 + m_{\mathrm{hi}}(L - L_{\mathrm{cut}}) & \text{for} \quad L \geq L_{35} \end{cases} \tag{4.2}$$

It consists of two linear parts with independent slope values m_{lo} and m_{hi}. The two parts are connected at the L_{cut}. The transition area between the loudness categories L_{15} ('soft') and L_{35} ('loud') is smoothed with a Bezier fit denoted with $\mathrm{bez}(L, L_{\mathrm{cut}}, L_{15}, L_{35})$. The exact form of the Bezier smoothing is given in Appendix A.1.[3] This model function provided the best fit to experimental data (cf. Chap. 5).

The model function $F(L)$ was fitted to the data $y_i(L_i)$ using a modified least–squares fit, i.e., by minimizing $\sum_i \Delta_i^2 = \sum_i (y_i(L_i) - F(L_i))^2$. To account for the limited range of the response scale, the difference between model function and data was defined as:

$$\Delta_i = \begin{cases} 0 & \text{for} \;\; F(L_i) < 0 \quad \wedge \quad y_i = 0 \\ 0 & \text{for} \;\; F(L_i) > 50 \quad \wedge \quad y_i = 50 \\ y_i - F(L_i) & \text{else} \end{cases} \tag{4.3}$$

4.3.2 Monte–Carlo simulations

The Monte–Carlo simulation method described in Chap. 3 was used to evaluate the adaptive procedure. The response characteristic of the simulated subject was specified by the two parameters σ_N and p_{out}. The parameter σ_N denotes the standard deviation within loudness ratings of different presentations of an identical stimulus. The parameter p_{out} denotes the frequency of outliers, i.e., the frequency of random responses which have

[3] Because of the smoothing in the medium range, the L_{cut} parameter in Eq. (4.2) does not represent the medium loudness level L_{25} but the level where the two linear parts would meet if they were not smoothed. L_{25} is always specified in this study because it can be calculated from L_{cut}, m_{lo} and m_{hi} and because L_{25} can better be interpreted than L_{cut}.

no relation to the stimulus level. Two different estimates of the underlying response statistics of the subjects were assumed in the simulations: 1) an 'optimistic' estimate with $\sigma_N = 4$ cu and $p_{\text{out}} = 0$ % which is typical of normal–hearing and most hearing–impaired listeners, cf. Chap. 3. 2) a 'pessimistic' estimate with $\sigma_N = 7$ cu and $p_{\text{out}} = 5$ %. This response statistic simulates subjects who have difficulties to rate loudness accurately and that show an inconsistent response behaviour.

4.3.3 Stimulus

A third–octave band of noise with a center frequency of 1 kHz was used as stimulus. The signal was generated from a random noise with Gaussian amplitude statistic, 5 s duration and 44.1 kHz sampling rate. The signal was transformed to the frequency domain by an FFT. All FFT–coefficients outside the desired band were set to zero and the resulting signal was transformed back to the time domain by an inverse FFT. A segment of 2 s duration was selected randomly and windowed with 100 ms $\cos^2$ ramps. During each trial, the noise was presented twice with a silent interstimulus interval of 1 s duration.

4.3.4 Apparatus

A computer–controlled audiometry workstation was used which was developed within a German joint research project on speech audiometry (Kollmeier *et al.*, 1992). A personal computer with a coprocessor board (Ariel DSP 32C) with 16–bit stereo AD–DA converters was used to control the complete experiment as well as stimulus presentation and recording of the subject's responses. The stimulus levels were adjusted by a computer–controlled custom–designed audiometer comprising attenuators, anti–aliasing filters, and headphone amplifiers. Signals were presented monaurally to the listeners with Sennheiser HDA 200 headphones. The headphones were calibrated to free–field sensitivity level of sinusoidals (ISO 389 (1991)) according to Richter (1992). The subjects were seated in a sound–insulated booth. Their task was to rate the loudness of each stimulus presented using an Epson EHT 10S handheld computer with an LCD touchscreen showing the response scale. The handheld computer was connected to the personal computer via serial interface. The subjects loudness ratings for each stimulus were stored for later statistical analysis.

4.3.5 Subjects and measurement program

10 normal–hearing listeners (5 male, 5 female; aged 24–57 years; median 28 years) and 10 hearing–impaired listeners (6 male, 4 female; aged 22–76 years; median 58 years)

participated in the experiment. The hearing threshold of the normal–hearing listeners was lower than 15 dB HL at the standard audiometric frequencies from 125 Hz to 8 kHz. Three of the normal–hearing listeners were members of the research group. The other listeners had no prior experience in psychoacoustical experiments and were paid on an hourly basis. The hearing–impaired subjects showed sensorineural hearing losses of drifferent degrees. Their audiograms ranged between 15 and 85 dB at 500 Hz and between 15 and 95 dB at 4 kHz They had never performed loudness scaling prior to the experiment. All subjects performed 10 loudness scaling tracks on each ear with both the constant stimuli and the adaptive procedure. The measurements were performed in blocks of 5, i.e., 5 tracks with the same procedure and ear were performed successively followed by a block with a different ear and/or procedure. The order of these blocks was randomized between subjects.

4.4 SIMULATIONS

Monte–Carlo simulations were performed to adjust the rules of the adaptive procedure in order to generate the desired uniform distribution of presentation levels covering the full auditory dynamic range of the subject without presenting stimuli outside this range. The resulting procedure has already been described above (Sec. 4.2.3.) Further, the accuracy which can be expected by the procedure was calculated by the simulations. A linear loudness function was assumed which was parameterized by L_{25} (level related to 25 cu) and m (slope of the loudness function). Each simulated adaptive track consisted of an initial phase of approx. 6 trials and two blocks with four and five trials, cf. Sec. 4.2.3.

Fig. 4.3 shows the mean frequency of stimulus levels per track generated by the adaptive procedure in a Monte–Carlo simulation with 2,000 runs. The underlying loudness function was assumed to be linear with a hearing threshold level (HTL) of 30 dB HL and an uncomfortable level (UCL) of 100 dB HL. The 'pessimistic' estimate of the subject's response characteristic ($\sigma_N = 7$ cu, $p_{\text{out}} = 5$ %) was used. The distribution of stimulus levels shows several peaks which are due to the first phase of the procedure which always starts at 80 dB and uses only certain step sizes. The maximum distance between the peaks is 15 dB in each track due to the rules of the procedure The peaks are approximately uniformly distributed in the auditory dynamic range. Between the peaks there is a continuous distribution of stimulus levels which is approximately uniformly distributed. This distribution was generated by the second phase of the procedure which produces various integer level values depending on the respective estimate of the auditory dynamic range of the subject. Note, that only a very small number of stimuli was presented above the uncomfortable level. This indicates that the rules of the adaptive procedure to minimize the number of uncomfortable loud stimuli have the desired effect and subjects are well protected against too high stimulus levels. Since no similar rule

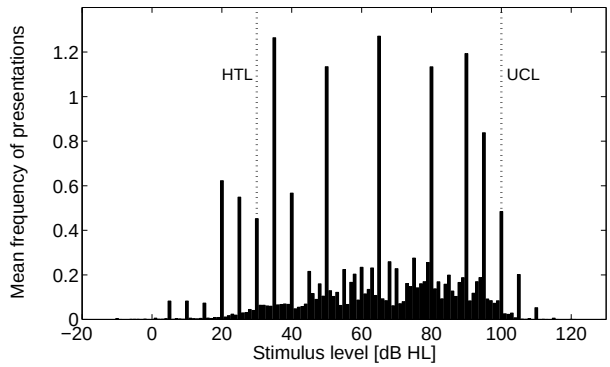

Figure 4.3: *Mean frequency of stimulus levels per track generated by the adaptive procedure in a Monte–Carlo simulation with 2,000 runs. The hearing threshold level (HTL) of the simulated subject was set to 30 dB HL, the uncomfortable level (UCL) was 100 dB HL (both are indicated by dotted lines).*

was introduced in the adaptive procedure for subthreshold levels, the number of stimuli below the hearing threshold level is higher than the number of stimuli above the discomfort level.
Altogether the distribution of stimulus levels is fairly equal and it can be assumed that all response alternatives are utilized by the subject with the same frequency.

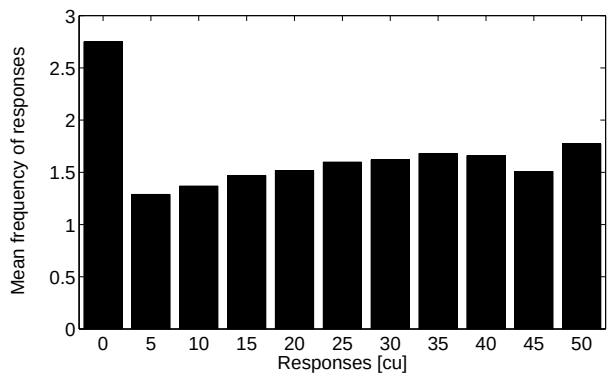

Figure 4.4: *Mean frequency of responses per track generated by the simulated subject in the Monte–Carlo simulation shown in Fig. 4.3.*

Fig. 4.4 shows the mean frequency of the different responses per track generated by the simulated subject in the same simulations as shown in Fig. 4.3. Only the responses

0, 5, 10, ... 50 cu are possible due to the response scale of the Oldenburg loudness scaling procedure. All responses show an approximately uniform distribution with a frequency of about 1.5 responses per track, except for the response 'inaudible' (0 cu) which has a mean frequency of 2.5. Peaks like in the distribution of presentation levels (Fig. 4.3) do not appear in the distribution of responses.
Since the response 'inaudible' has to appear at least once per track due to the rules of the adaptive procedure, the frequency of 2.75 for 0 cu indicates that on average 1.75 trials are presented 'ineffectively' below threshold. The response 'too loud' has to appear at least once per track as well. The rules of the adaptive procedure concerning the uncomfortable level, however, cause that on average only 0.75 additional stimuli are presented above the uncomfortable level.

Table 4.1: Standard deviations of L_{25} and m estimates derived from Monte–Carlo simulations for the constant stimuli and the adaptive procedure. The 'optimistic' estimate ($\sigma_N = 4$ cu; $p_{\text{out}} = 0$ %) and the 'pessimistic' estimate ($\sigma_N = 7$ cu; $p_{\text{out}} = 5$ %) of the underlying response statistic were used.

	adaptive		constant	
	'optimistic'	'pessimistic'	'optimistic'	'pessimistic'
STD(L_{25})	1.5 dB	3.6 dB	1.4 dB	3.4 dB
STD(m)	0.06 $\frac{\text{cu}}{\text{dB}}$	0.12 $\frac{\text{cu}}{\text{dB}}$	0.07 $\frac{\text{cu}}{\text{dB}}$	0.12 $\frac{\text{cu}}{\text{dB}}$

Table 4.1 shows the standard deviations of the L_{25} and m estimates generated. The table also shows standard deviations calculated by simulations for the constant stimuli procedure. The 'optimistic' estimate ($\sigma_N = 4$ cu; $p_{\text{out}} = 0$ %) and the 'pessimistic' estimate ($\sigma_N = 7$ cu; $p_{\text{out}} = 5$ %) of the underlying response statistic were used. The standard deviations of L_{25} and m estimates depend on the slope m of the underlying loudness function. The values shown in Table 4.1 are calculated for $m = 0.7\frac{\text{cu}}{\text{dB}}$. To estimate the standard deviation of L_{50} estimates for different m values, the STD(L_{50}) values shown in Table 4.1 have to be divided by $(m/0.7)\ \frac{\text{dB}}{\text{cu}}$. The STD($m$) values have to be multiplied by $(0.7/m)\ \frac{\text{cu}}{\text{dB}}$.
For the constant stimuli procedure, the stimulus levels were set optimally, i.e., 7 even distributed levels in the auditory dynamic range of the subject, in order to obtain a reference of the maximum accuracy which can be achieved. (In a real measurement, this is only the case if the pre–measurement would has been performed perfectly.) Each level was presented twice, i.e., 14 trials per track were assumed in the constant stimuli procedure. The adaptive procedure used on average 15.1 trials per track. For both optimistic and pessimistic estimate of the response statistics, the standard deviations of L_{25} and m estimates generated by the adaptive procedure were approximately equal to those of the constant stimuli procedure. That means, that the adaptive procedure yielded nearly the same accuracy as the constant stimuli procedure with nearly the

same track length. This is remarkable because the adaptive procedure had to find out the individual dynamic range during the track whereas the constant stimuli procedure always used the predefined optimal stimululs level distribution. In real measurements, the accuracy and efficiency of the constant stimuli procedure is therefore expected to be considerably smaller.

4.5 MEASUREMENTS

4.5.1 Distribution of stimulus levels and responses

As described above, 10 loudness scaling tracks were performed for each ear of 10 normal–hearing and 10 hearing–impaired using both the constant stimuli and the adaptive procedure. Both procedures generated stimulus level distributions which included levels near threshold in all tracks. In only 3.5 % of the measurements with the constant stimuli procedure none of the responses 'inaudible' or 'very soft' occurred (2.5 % in the normal–hearing subjects and 4.5 % in the hearing–impaired subjects). In the adaptive procedure the responses 'inaudible' or 'very soft' occurred at least once in each track.
In 29 % of the tracks with the constant stimuli method the uncomfortable level was not reached, i.e., none of the responses 'very loud' or 'too loud' occurred (20 % in the normal–hearing subjects and 37 % in the hearing–impaired subjects). In many of these cases the pre–measurement resulted in an uncomfortable level estimate of less than 90 dB HL, although the second part of the measurement indicated an uncomfortable level of more than 100 dB HL.
In the adaptive procedure the uncomfortable level was not reached only in those cases where it exceeded 115 dB HL, since the maximum stimulus level was restricted to this value. This occurred in 10 % of the cases (9 % in the normal–hearing subjects and 10 % in the hearing–impaired subjects).
Taken together, this means that the adaptive procedure better covers the full auditory dynamic range than the constant stimuli procedure.

Fig. 4.5 shows the mean frequency distribution of the response alternatives per track for the 10 normal–hearing subjects and both for the constant stimuli and for the adaptive procedure. The adaptive procedure generated on average 20.9 trials per track whereas the constant stimuli procedure always generated 14 trials per track.[4] For both procedures, the response 'inaudible' (0 cu) had a mean frequency per track which is much lower than in the simulation, cf. Sec. 4.4. Note, however, that in the simulations, the underlying hearing threshold was set to 30 dB HL whereas it is about 0 dB HL in the normal–hearing listeners employed in the experiment. For both the constant stimuli and

[4] Since the pre–measurement is omitted in the adaptive procedure, both procedures take about the same measurement time.

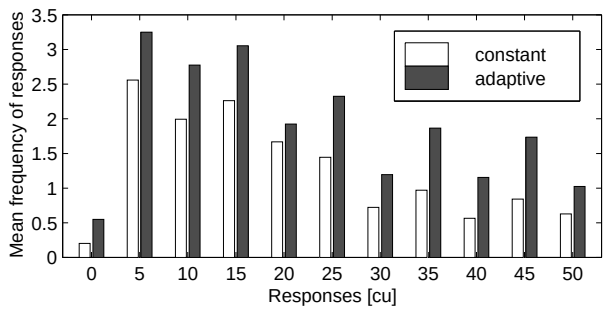

Figure 4.5: *Mean frequency of the different response alternatives per track in the 10 normal–hearing subjects for the constant stimuli and for the adaptive procedure.*

the adaptive procedure, the minimal stimulus level was limited to 0 dB HL which can be expected to be audible by about the half of the normal–hearing subjects. Consequently, the experimental tracks had only a probability of 0.5 to evoke the response alternative 'inaudible' at the lowest possible presentation level. In the adaptive procedure, the measured frequency 0.55 of the response 'inaudible' is in good agreement with the expected value, because the adaptive procedure usually presented a stimulus with 0 dB HL once per track in normal–hearing subjects.[5] The lowest stimulus level during the data collection part of the constant stimuli procedure is set to the lowest level which was rated as 'audible' in the preceding ascending level sequence which itself has an initial value of 0 dB HL. Thus, in many cases the lowest stimulus level in the data collection was beyond 0 dB HL. Consequently, the frequency of the response 'inaudible' was much lower than 0.5 in the constant stimuli procedure.

The mean frequencies of responses shown in Fig. 4.5 are not evenly distributed but decrease with increasing loudness in both procedures. This is probably due to the fact that in normal–hearing listeners the loudness function of an 1 kHz narrowband signal usually has an increasing slope (cf. Sec. 4.5.2). Hence, the level range related to loudness values below 25 cu is wider than the level range of loudness values above 25 cu. Since both procedures generate equal distributions of stimulus levels on a dB–scale, more stimuli are generated in the lower than in the upper loudness range. The frequency of responses shows a larger decrease with level in the constant stimuli than in the adaptive procedure. This is probably due to the fact that in the constant stimuli procedure more cases occur in which no stimuli are presented near the uncomfortable level, see above.

[5] In the adaptive procedure, a stimulus with 0 dB HL occurs in the first phase once when all trials of the descending branch have been rated as being audible. In the second phase 0 dB HL can occur once at most (namely in the second block), but this was unlikely since the lowest target loudness value is 5 cu.

The frequencies of the named response alternatives (5, 15, 25, 35 and 45 cu) tended to be larger than the frequencies of the respective neighbored unnamed response alternatives (10, 20, 30 and 40 cu) in both procedures. This indicates that the named response alternatives were chosen more frequently by the normal–hearing subjects. A similar effect was found by Keidser *et al.* (1999) who used a scale with 5 labeled and 18 not labeled response alternatives.

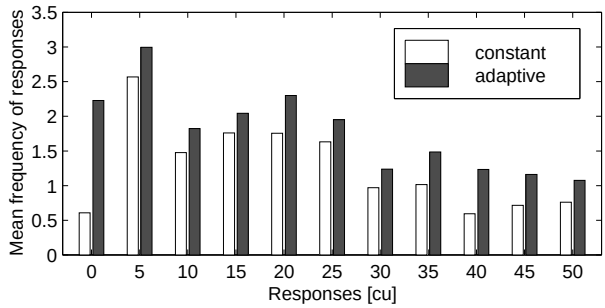

Figure 4.6: *Mean frequency of response alternatives per track in the 10 hearing–impaired subjects for the constant stimuli and for the adaptive procedure.*

Fig. 4.6 shows the mean frequency of the response alternatives per track in the 10 hearing–impaired subjects for the constant stimuli and for the adaptive procedure. Here, the adaptive procedure generated on average 19.5 trials per track. The frequency of responses decreased with increasing loudness at a lower rate than found in the normal–hearing subjects (Fig, 4.5). This effect is certainly due to the more linear loudness functions found in hearing–impaired (see below). The decrease was more pronounced in the constant stimuli procedure than in the adaptive procedure, which is probably due to the same reason as in the normal–hearing subjects.
The tendency to choose the named response alternatives more frequently is not as obvious as in the normal–hearing subjects.

4.5.2 Loudness function estimates

Since the loudness functions of both normal–hearing and hearing–impaired subjects showed an increasing slope in the measurements, the linear model function produced estimates that deviated much more from the data than expected from the simulations. For that reason, a new model function was introduced which has independent slope values for low and high stimulus values (cf. Sec. 4.3.1).[6]

[6] The advantages of this model function are discussed in more detail in Chap. 5.

Table 4.2: Median values of the estimated loudness function parameters, for both normal–hearing subjects (NH) and hearing–impaired subjects (HI) and for both adaptive and constant stimuli procedure.

	median(L_{25}) [dB]	median(m_{lo}) [$\frac{\mathrm{cu}}{\mathrm{dB}}$]	median(m_{hi}) [$\frac{\mathrm{cu}}{\mathrm{dB}}$]
NH (adaptive)	80.3	0.27	1.17
NH (constant)	78.1	0.26	1.06
HI (adaptive)	88.6	0.44	1.70
HI (constant)	82.6	0.56	1.42

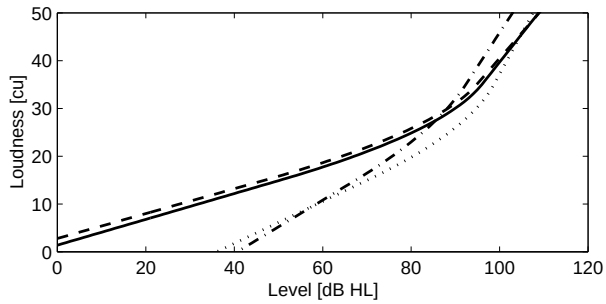

Figure 4.7: Loudness functions with the median parameters displayed in Table 4.2. Normal–hearing subjects with adaptive procedure (solid), normal–hearing subjects with constant stimuli procedure (dashed), hearing–impaired subjects with adaptive procedure (dotted), hearing–impaired subjects with constant stimuli procedure (dash–dotted).

Table 4.2 shows the median values of the estimated loudness function parameters, for both normal–hearing and hearing–impaired subjects and for both the adaptive and the constant stimuli procedure. For each ear, the median values of the parameter estimates were calculated across all 10 tracks with a given procedure (adaptive or constant stimuli, respectively). The median values from these values across all ears (normal–hearing or hearing–impaired, respectively) are displayed in Table 4.2.
In both normal–hearing and hearing–impaired subjects the adaptive procedure generated higher L_{25} estimates. In the hearing–impaired subjects this difference was relatively large (about 7 dB). In both normal–hearing and hearing–impaired subjects the loudness functions estimated by the adaptive procedure were slightly more upwardly concave (i.e, the ratio of m_{hi} to m_{lo} is larger) than those estimated by the constant stimuli procedure. Fig. 4.7 shows loudness functions with the parameters displayed in Table 4.2.

Figures 4.8 and 4.9 show the histograms of the L_{25}, m_{lo} and m_{hi} estimates for both normal–hearing and hearing–impaired subjects and for both constant stimuli and adap-

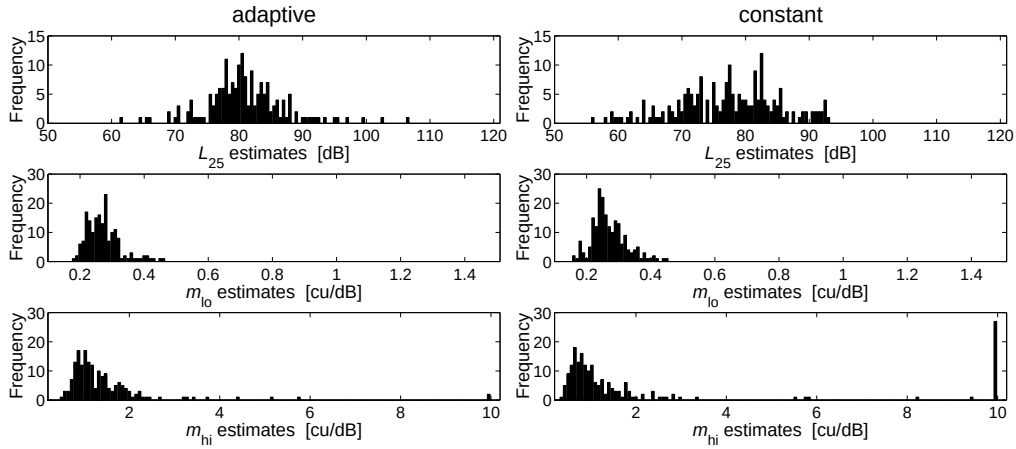

Figure 4.8: *Histograms of parameter estimates in normal–hearing subjects. The left three panels show the results of the adaptive procedure. The right three panels show the results of the constant stimuli procedure. Estimates beyond the displayed ordinate range are displayed at the respective upper limit of the ordinate.*

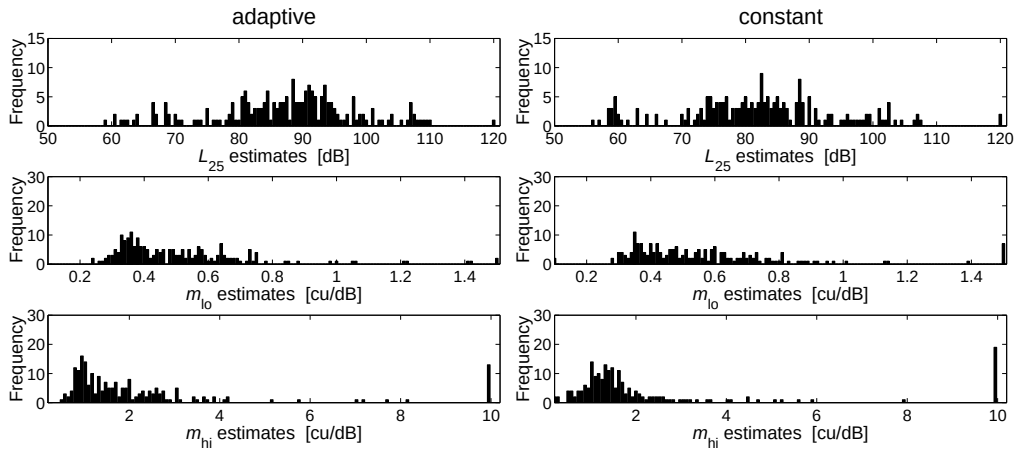

Figure 4.9: *Histograms of parameter estimates in the hearing–impaired subjects. The left three panels show the results of the adaptive procedure. The right three panels show the results of the constant stimuli procedure. Estimates beyond the displayed ordinate range are displayed at the respective upper limit of the ordinate.*

tive procedure. Both groups of subjects measured 200 tracks with both adaptive and constant stimuli procedure. Some of these 200 tracks produced extremely high m_{hi} es-

timates up to extrem values above 1,000 $\frac{\text{cu}}{\text{dB}}$. These estimates have to be regarded as outliers and are displayed at the respective upper limit of the ordinate. These outliers were probably due to an inaccurate fit of the upper part of the loudness function which was based on insufficient data because not enough stimuli were presented at high levels. In the constant stimuli procedure, these outliers occured more frequently than in the adaptive procedure, certainly, because the pre–measurement of the constant stimuli procedure often generated too–small estimates of the uncomfortable level.

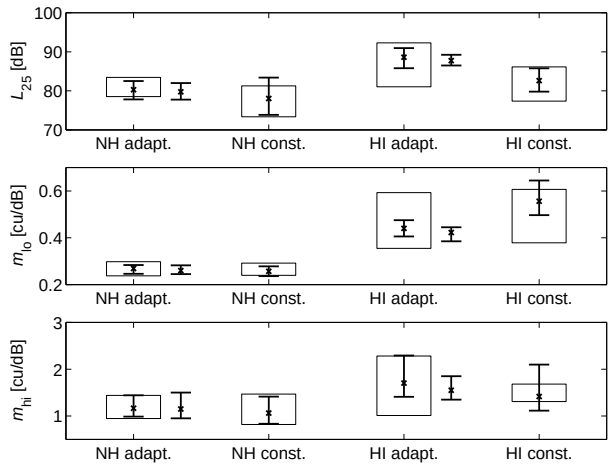

Figure 4.10: *Inter– and intraindividual quartile ranges of loudness function parameter estimates for normal–hearing (NH) and hearing–impaired (HI) listeners with the adaptive and the constant stimuli procedure. The crosses indicate the interindividual median values. The boxes indicate the interindividual 25 % and 75 % percentiles. The errorbars in the boxes indicate the intraindividual 25 % and 75 % percentiles. The crosses and errorbars on the right of the errorbars of the adaptive procedure show predicted intraindividual median values and quartile ranges, calculated with Monte–Carlo simulations using the 'optimistic' estimate of the response statistic.*

Fig. 4.10 shows the inter– and intraindividual quartile ranges of loudness function parameter estimates for normal–hearing and hearing–impaired listeners with the adaptive and the constant stimuli procedure. The crosses indicate the interindividual median values (median values across subjects/ears of individual median values across repeated measurements with the same subject/ear). The boxes indicate the interindividual quartile ranges (median values across subjects/ears of individual 25 % and 75 % percentiles across repeated measurements with the same subject/ear). The errorbars in the boxes indicate the intraindividual quartile ranges (median values across all subjects/ears of

individual 25 % and 75 % percentiles; before calculating the median values, the individual quartile ranges were shifted by the difference between individual median value and interindividual median value.).
In the normal–hearing subjects the intraindividual and interindividual quartile ranges of all parameters have the same magnitude. That means, none of the procedures is able to detect significant differences in these parameters between normal–hearing subjects.
In the hearing–impaired subjects, however, both procedures generated intraindividual quartile ranges which are considerably smaller than the interindividual quartile ranges for the estimates of L_{25} and m_{lo}. For the adaptive procedure, the intraindividual quartile range of m_{hi} has the same magnitude as the interindividual quartile range. For the constant stimuli procedure the intraindividual quartile range of m_{hi} is even somewhat larger than the interindividual quartile range. The large intraindividual quartile ranges in the m_{hi} estimates are certainly due to the smaller number of stimuli which are placed at higher loudness values. Since the loudness functions were upwardly concave and the stimulus level distribution was even on the level scale, the density of responses between 25 and 50 cu is smaller than between 0 and 25 cu (cf. Figs. 4.5 and 4.6). Consequently, the fit of the upper part of the loudness function is based on a smaller number of data points than the lower part. In the constant stimuli version, this effect is stronger because the pre–measurement often generates too low UCL estimates. This causes that no stimuli near the real UCL are presented during the data collection.
As a consequence, both procedures can detect differences between the subjects of this hearing–impaired group significantly for L_{25} and m_{lo} but not for m_{hi}. For all three parameters the adaptive procedure gives more significant results since it generates higher ratios between interindividual and intraindividual quartile ranges than the constant stimuli procedure.
The crosses and errorbars on the right of the errorbars of the adaptive procedure show predicted intraindividual median values and 25 % and 75 % percentiles, calculated with Monte–Carlo simulations using the 'optimistic' estimate of the response statistic. The predicted and the measured quartile ranges have about the same size in the normal–hearing subjects. This was expected, because the optimistic estimate of the response statistic is typical of normal–hearing subjects, cf. Chap. 3. The predicted quartile ranges are slightly smaller than the measured quartile ranges in the hearing–impaired subjects. This was expected, because hearing–impaired subjects sometimes show response statistics less accurate than the optimistic estimate, cf. Chap. 3.

Fig. 4.11 shows the intraindividual standard deviations of the L_x estimates, i.e., of the estimates of the presentation levels which are related to a given loudness category x. For both normal–hearing and hearing–impaired listeners the adaptive procedure yielded 20% to 50% smaller standard deviations. The intraindividual standard deviations of the L_x estimates are more accurate in the hearing–impaired listeners than in the normal–hearing listeners. This is probably due to the fact that hearing–impaired listeners usually show

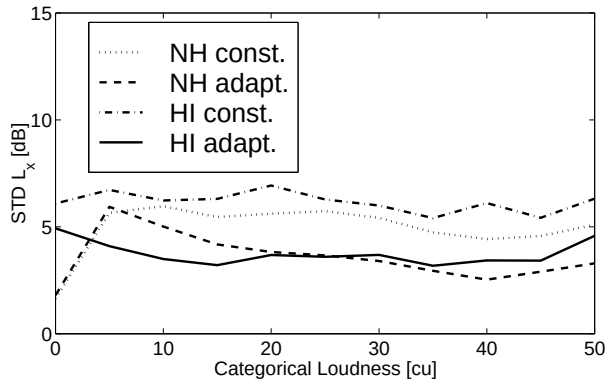

Figure 4.11: *Intraindividual standard deviations of L_x estimates for normal–hearing (NH) and hearing–impaired (HI) subjects with the adaptive and the constant stimuli procedure.*

increased slopes of loudness function and that the standard deviation of L_x estimates is inversely proportional to the slope. Obviously, this effect overrules the less consistent response behaviour of the hearing–impaired listeners which should lead to a less precise loudness function estimate.

4.6 DISCUSSION

The Oldenburg–ACALOS procedure differs in two points from the earlier constant stimuli version: 1) it uses an adaptive stimulus placement and 2) it uses a model loudness function which has independent slope values for the lower and the upper loudness range, whereas the original constant stimuli version uses a linear loudness function with a variable offset for low levels. Both modifications lead to an improvement of the measurement accuracy.

Using the adaptive procedure, the standard deviations of the estimates of the response level L_x, i.e. the level corresponding to the categorical loudness x cu amount to 4 dB for both normal–hearing and hearing impaired subjects (Figs. 4.10 and 4.11). These standard deviations are much larger than the standard deviations of the L_{25} estimates calculated with Monte–Carlo simulations shown in Table 4.1. In the simulations, a linear function (Eq. (4.1)) was assumed as underlying loudness function and used as model function for fitting. Since this linear model function only has 2 free paramters, it is much more stable to fit than model function (4.2) with 3 free parameters which was used to fit the measurements. However, since the experimental data showed an increasing slope with increasing level, model function (4.1) was not adequate and model function (4.2)

gave a much closer fit to the experimental data, cf. Chap. 5.

The standard deviations of the adaptive Oldenburg loudness scaling procedure are equal or slightly smaller than those generated by the NAL–ACLS procedure (Keidser *et al.*, 1999). While the NAL–ACLS needs 60 trials, the adaptive Oldenburg loudness scaling procedure needs about 20 trials making the latter much more efficient. The LGOB procedure generated within–subject standard deviations of 2.9 dB in normal–hearing subjects using 45 trials plus pre–measurement (Allen *et al.*, 1990). Under the assumption that the standard deviations decrease with $1/\sqrt{n}$ with increasing number of trials n this would correspond to 4.3 dB at 20 trials. Hence, the efficiency of the LGOB procedure is comparable to that of the adaptive Oldenburg procedure. However, the LGOB procedure requires a pre–measurement which is omitted in the adaptive Oldenburg procedure which makes the latter slightly more efficient. Rasmussen *et al.* 1998 used on average 123 trials per frequency and generated intrasubject intersession standard deviations of 4–8 dB with a maximum at medium levels. Robinson and Gatehouse (1996) used 64 trials and generated intraindividual intersession standard deviations which decreased from 7 to 2.8 dB with increasing level. However, the intersession standard deviations of Rasmussen *et al.* and Robinson and Gatehouse cannot be compared directly with the intrasession standard deviations measured in this study.

Constant stimuli and adaptive procedure differ slightly in loudness function estimates, although they were used with the same model loudness function. In general, the loudness functions estimated by the adaptive procedure tend to have larger L_{25} values (cf. Fig. 4.7). The estimate of m_{hi} is increased in the adaptive procedure compared to the constant stimuli procedure in the hearing–impaired subjects. There are four possible reasons for these differences in the procedures:
1) The adaptive procedure generates a slightly different stimulus spacing than the constant stimuli procedure. The stimulus spacing effects loudness function estimates, especially, if scales with few response alternatives are used like in this study (Parducci and Wedell, 1989; Poulton, 1989): The constant stimuli procedure always presents 7 different stimulus levels (twice each) which are evenly distributed on a dB–scale. The adaptive procedure usually presents more different stimulus levels, i.e. it produces a smaller spacing. (Most stimulus levels are presented only once per track, some stimulus levels might be presented more than once during the track, cf. Fig. 4.3).[7] The rules of the adaptive procedure, however, guarantee that the level distribution can only slightly be skewed on a dB–scale. Further, the probability that the level distribution does not cover the whole auditory dynamic range of the subject is much lower than in the constant stimuli

[7] Note, however, that the adaptive procedure never presents similar stimulus levels successively but always with some completely different stimulus levels in between. This makes it difficult to memorize the exact loudness impressions of the previous trials of a similar level. Therefore, we believe that there are only small biases produced by the adaptive procedure. In the constant stimuli procedure, however, the stimulus level of the previous trial is frequently repeated which might causes a bias.

procedure. For these reasons, we believe that the adaptive procedure generates smaller biases than the constant stimuli procedure.
2) There is a tendency in many subjects to rate the first stimulus near 'medium' (Heller, 1990). The adaptive procedure always begins with a presentation level of 80 dB HL. The constant stimuli procedure, however, usually begins with soft levels, because the level sequence is restricted to have an increasing tendency.
3) The adaptive procedure generates randomized stimulus levels whereas the constant stimuli procedure generates randomized levels with an increasing tendency which may cause a context–bias as well.
4) The pre–measurement of the constant stimuli procedure often leads to a stimulus level distribution which does not cover the whole auditory dynamic range. This causes a bias, because listeners tend to use the whole response scale. The adaptive procedure usually better covers the whole auditory dynamic range. Consequently, it generates loudness function estimates with a wider auditory dynamic range.

The adaptive loudness scaling procedure introduced here appears to be efficient and accurate. However, some possibilities for improvement still exist:
1) The minimum stimulus level should not be limited to 0 dB HL, because the full auditory range of normal–hearing listeners with lower hearing thresholds can only be covered when lower levels are used as well. However, this will only effect estimates of loudness functions of normal–hearing listeners.
2) The number of blocks in the second part of the adaptive procedure can easily be extended to increase measuring accuracy. In this study the number of blocks was set to 2 as a compromise between accuracy and measurement time. However, in order to yield more accuracy, we use 3 blocks in the clinical standard audiometry and in the hearing aid fitting which is performed in our laboratory.
3) The restriction to eleven response alternatives forces some subjects to respond less accurate than they could, especially at soft levels in upwardly concave loudness functions, cf. Chap. 3. Scales with more response alternatives, e.g., the scale of the Würzburg loudness scaling procedure (Heller, 1985) with more than 50 response alternatives and the scale by Keidser *et al.* (1999) with 23 response alternatives, would possibly increase the measuring accuracy. However, some subjects may be puzzled by a too large number of response alternatives.
4) It is recommended to generate uniformly distributed stimulus levels on the loudness scale in order to yield bias–free loudness function estimates (Parducci and Perret, 1971; Poulton, 1989). However, this distribution is often not obtained because the adaptive procedure is optimized to generate equally distributed stimulus levels on a dB–scale and many listeners show upwardly concave loudness functions. To calculate target levels which are equal distributed on the loudness scale, the shape of the whole loudness function (and not only the limits of the auditory dynamic range) has to be estimated during the track. Unfortunately, a stable estimate of the whole loudness function, i.e., of the

parameters L_{50}, m_{lo} and m_{hi}, requires more trials than the estimate of the auditory dynamic range. Lapses in the human response behaviour can also disturb the loudness function estimate at the beginning of the track. Probably, it is possible to estimate the target levels with sufficient accuracy after the first block of the second phase of the adaptive procedure. If the level distribution should be equalized in the remainder of the track, it is necessary to present more trials. The changes in level distribution will probably change the shape of the loudness function, because the local slope of the stimulus–category function increases in parts of the loudness function with higher stimulus density (Parducci, 1963). That means, it can be expected that loudness functions become even more upwardly concave.
5) The average measurement time per trial $\bar{T} \approx 12$ s is very long in the procedure described here compared to other procedures, e.g., $\bar{T} < 8$ s (Allen *et al.*, 1990), $\bar{T} < 6$ s (Ricketts and Bentler, 1996), $\bar{T} \approx 3$ s (Rasmussen *et al.*, 1998). $\bar{T}$ can be reduced about 4 s by reducing the stimulus to a single presentation of a 1 s noise. This will probably have no effect on the result. Further, there is a 1 s time interval after the subject has given his/her rating in which the response can be corrected. This possibility of correction can be omitted because it is only used very rarely. Further, the time interval between response (inclusive correction interval) and next presentation may be reduced, although some subjects need such intervals in order not to be surprised by the subsequent stimulus presentation.
6) Reckhardt *et al.* (1999) found that the influence of the stimulus level range on the estimation of equal–loudness level contours can be reduced dramatically (from 13 dB to 2.3 dB) when adaptive interleaved tracks with different frequencies are used within one run. Therefore, when different loudness functions of different stimuli should be measured using categorical loudness scaling, we propose to present these stimuli interleaved in randomized order to reduce context–effects.

The data of normal–hearing listeners presented here cannot be used as normative data because each subject performed 40 loudness scaling tracks during the study. To derive normative data, it is necessary to perform the reference measurements once with each subject using the same set of stimuli (i.e., different center frequencies, bandwidths, etc.) as in the later diagnostics. Normative data of the constant stimuli procedure were measured by Albani *et al.* (Albani *et al.*, 1997). They derived L_{25} and L_{50} levels which are about 7 dB lower than in this study. Probably, the subjects in this study adapted to the stimulus during the measurement and accepted higher levels.

4.7 CONCLUSIONS

The Oldenburg–ACALOS (Adaptive CAtegorical LOudness Scaling) procedure better approximates the desired equal distribution of stimulus levels than the constant stimuli version of the Oldenburg loudness scaling procedure, i.e., it covers the full auditory

dynamic range of the subject with a minimum number of presentations outside this range. In combination with the model loudness function proposed here, the adaptive version yields a considerably higher accuracy in loudness function estimates than the constant stimuli version, with the same amount of measurement time. Therefore, the adaptive version of the Oldenburg loudness scaling procedure is recommended. The level of each loudness category was estimated with an intraindividual standard deviation of less than 5 dB with a total number of about 20 trials.

ACKNOWLEDGEMENTS

We would like to thank Birgitta Gabriel, Kerstin Sommer and Anita Gorges for performing the measurements. Thanks are also due to Glenis Long for helpful comments on the manuscript.

This study was supported by BMBF 01VJ9305 and by the CEC supporting action NATASHA.

Chapter 5

Parametrization of Loudness Functions in Categorical Loudness Scaling

ABSTRACT

Experimental data were obtained and systematically fitted with a wide selection of different model functions in order to assess the 'optimum' model loudness growth function for categorical loudness scaling measurements. Repeated loudness scaling measurements were performed with 10 normal–hearing and 10 hearing–impaired listeners. Individual reference loudness functions were derived by calculating the median response levels from these measurements. Both the constant stimuli version (Hohmann and Kollmeier, 1995b) and the adaptive version (Brand *et al.*, 1997a; 1997b; Chap. 4) of the Oldenburg loudness scaling procedure were applied. Different model loudness functions were investigated with respect to their reproducibility and systematic errors when fitted to categorical loudness scaling data. Biases (differences between fitted model functions and individual reference functions) and intraindividual standard deviations of response levels were calculated for all model functions. The model function which yielded the lowest bias and the lowest intraindividual standard deviation in response level estimates consists of two straight lines which are connected at the 'medium'–level and smoothed in the transition area. This model function has three free parameters, namely the 'medium'–level and the two slope values of the two straight lines.

INTRODUCTION

The method of categorical loudness scaling (e.g., Pascoe, 1978; Heller, 1985; Allen *et al.*, 1990; Hohmann and Kollmeier, 1995b) assesses the loudness function of a listener by presenting stimuli at various levels which have to be rated on a categorical loudness scale. A shift of the loudness function towards higher levels indicates a hearing loss and an increased slope indicates a loudness recruitment. Assuming a simple linear model loudness function, the threshold and the slope parameter can be used to determine the gain and the compression ratio of a hearing aid. However, there is experimental evidence that a simple linear model function does not fit to many loudness scaling data, because loudness functions often exhibit non–linear shapes (e.g., Allen *et al.*, 1990; Kießling *et al.*, 1993; Kollmeier, 1997). Therefore, the purpose of the current study is to determine an appropriate model loudness function based on empirical data of different listeners.

Such an 'optimal' model loudness function must provide enough degrees of freedom to approximate all kinds of loudness functions in hearing–impaired listeners. Some experimenters use polygon fits having as many free parameters as different stimulus levels were used in the measurement (e.g., Allen *et al.*, 1990). However, less free parameters may result in a more stable fit to the data, especially when the track length should be kept low to reduce measurement time. Many different model loudness functions with fewer degrees of freedom than presentation levels have been proposed by different researchers (e.g., Fechner, 1888; Stevens, 1956; Nowak, 1990; Boretzki *et al.*, 1994; Hohmann and Kollmeier, 1995b; Heller *et al.*, 1997; Brand *et al.*, 1997c).

In this study a purely empirical approach was chosen in order to determine an appropriate parametrization of the loudness function for the Oldenburg loudness scaling procedure (Hohmann and Kollmeier, 1995b; Brand *et al.*, 1997a). Individual loudness functions of 10 normal–hearing and 10 hearing–impaired listeners were derived by calculating the median values of the stimulus levels which were rated with the same loudness category from 10 repeated loudness scaling runs per listener. Thus, the individual reference loudness functions were derived without assuming an underlying model loudness function.

The Oldenburg loudness scaling procedure can be performed in two different ways – a constant stimuli version (Hohmann and Kollmeier, 1995b) and an adaptive version (Brand *et al.*, 1997a; 1997b; Chap. 4). Both versions were applied to investigate whether the measurement procedure effects the shape of the loudness function.

10 different model functions were fitted to each single loudness scaling run and from these data biases (mean differences between fitted model functions and individual reference functions) and the standard deviations of loudness level estimates were calculated.

5.1 EXPERIMENTAL METHOD

5.1.1 Stimulus

A third–octave band of noise with a center frequency of 1 kHz was used as stimulus. The signal was generated from a random noise with Gaussian amplitude statistics, 5 s duration and 44.1 kHz sampling rate. The signal was transformed to the frequency domain by an FFT. All FFT–coefficients outside the desired band were set to zero and the resulting signal was transformed back to the time domain by an inverse FFT. A part of 2 s duration was selected randomly and windowed with 100 ms $\cos^2$ ramps. During each trial, the noise was presented twice with a silent interstimulus interval of 1 s duration. The stimuli were presented to the listeners by Sennheiser HDA 200 headphones. They were calibrated in hearing level (HL) according to the hearing threshold of sinusoidal stimuli (ISO 389 (1991)).

5.1.2 Response scale

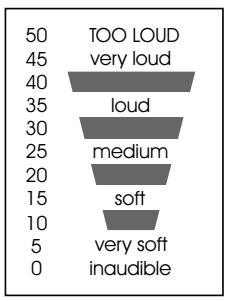

Figure 5.1: *Category scale with 11 response alternatives used by the listeners to rate the loudness. The numbers on the left side indicate the categorical units (cu) which are used for data storage and analysis. They were not visible to the listener.*

The listeners' task was to indicate their ratings on the response scale shown in Fig. 5.1. The scale consists of eleven response alternatives including five named loudness categories, four not named intermediate response alternatives and two named limiting categories. The named response categories are 'sehr leise' ('very soft'), 'leise' ('soft'), 'mittel' ('medium'), 'laut' ('loud') and 'sehr laut' ('very loud') and correspond to 5, 15, 25, 35 and 45 cu (categorical units) as shown on the left side of Fig. 5.1. The not named response alternatives are used to increase the total number of response alternatives. They are indicated with horizontal bars with increasing length for increasing loudness and are placed between the named loudness categories. They correspond to the categorical units

10, 20, 30 and 40 cu, respectively. The two limiting categories are named 'unhörbar' ('inaudible') and 'ZU LAUT' ('TOO LOUD') and correspond to 0 and 50 cu.

Prior to testing, each listener was instructed verbally by the experimenter (cf. A.2). During the instruction, the response box was practically demonstrated and any questions were clarified.

5.1.3 Apparatus

A computer–controlled audiometry workstation was used which was developed within a German joint research project on speech audiometry (Kollmeier *et al.*, 1992). A personal computer with a coprocessor board (Ariel DSP 32C) with 16–bit stereo AD–DA converters was used to control the complete experiment as well as stimulus presentation and recording of the listener's responses. The stimulus levels were adjusted by a computer–controlled custom–designed audiometer comprising attenuators, anti–aliasing filters and headphone amplifiers. The signals were presented monaurally to the listeners with Sennheiser HDA 200 headphones. The headphones were calibrated to free–field sensitivity level according to Richter (1992). The listeners were seated in a sound–insulated booth. Their task was to rate the loudness of each stimulus presented using an Epson EHT 10S handheld computer with an LCD touchscreen showing the response scale. The handheld computer was connected to the personal computer via serial interface. The listeners loudness ratings for each stimulus were stored for later statistical analysis.

5.1.4 Measurement procedures

Two different categorical loudness procedures were used in this study. Both procedures use the same response scale and stimuli but differ in stimulus level placements.

5.1.4.1 Constant stimuli procedure

The constant stimuli version of the Oldenburg loudness scaling procedure (Hohmann and Kollmeier, 1995b) includes two parts. The auditory dynamic range of the individual listener is estimated by presenting an ascending level sequence in the first part of the measurement. The loudness function is assessed by presenting stimuli covering the so determined full auditory dynamic range in the second part.
The first part uses an ascending stimulus level sequence with an initial level of 0 dB HL and a step size of 5 dB. The listener's task is to press the response button as soon as the stimulus is audible. Then the level is further increased in 15 dB steps up to 85 dB and increased in 5 dB steps beyond 85 dB. Now, the listener is asked to press another response button 'too loud' immediately when the stimulus is perceived as too

loud. In case that the listener does not press the response button, the sequence stops at 120 dB HL.
In the second part of the procedure, the loudness function is estimated. Two stimuli are presented at each of 7 different levels which are distributed equidistantly on a dB–scale between the limits of the dynamic range estimated in the first part of the procedure. The listener rates the loudness using the scale described above. The stimuli are presented in pseudo–random order in a way that the maximum difference of subsequent presentation levels is smaller than half of the dynamic range of the sequence in order to avoid context effects which are due to the tendency of many listeners to rate the current stimulus relatively to the previous stimulus. After completion of the track a model function is fitted to the data by a modified least–square fit, cf. Sec. 5.1.5.

5.1.4.2 Adaptive procedure

The adaptive version of the Oldenburg loudness scaling procedure (Brand *et al.*, 1997a; Brand *et al.*, 1997b), cf. Chap. 4, consists of two different phases. However, the listener is not aware that there are two phases, since his/her task is the same – to scale loudness – in both phases. The dynamic range of the listener is roughly estimated in the first phase. More data are collected in the second phase and the estimate of the dynamic range in which the stimuli are presented is updated two times.
The first phase starts with a stimulus at 80 dB HL. When the initial stimulus is inaudible or too loud, it is increased or decreased, respectively, in 15 dB steps until a response between inaudible and too loud is achieved. Thereafter, two interleaved sequences of stimuli begin. The first sequence increases the stimulus level in 5 dB steps until the response 'too loud' is given or the maximum level of 115 dB is reached. The final stimulus level of this sequence is the first estimate of the upper limit of the dynamic range. The second sequence decreases the stimulus level in 15 dB steps until it is inaudible or the limit of 0 dB HL is reached. In the first case the level is increased again with 5 dB steps until it is audible. The final stimulus level of this sequence is used as the first estimate of the lower limit of the dynamic range.
The second phase consists of two blocks according to the following procedure: Five stimulus levels, i.e. L_5 ('very soft'), L_{15} ('soft'), L_{25} ('medium'), L_{35} ('loud') and L_{45} ('very loud'), are estimated by linear interpolation between the estimated limits of the dynamic range. For the first block the limits estimated within the first phase are used. Before the second block starts, the limits are recalculated by fitting a linear function to the results of all previous trials in the track including the first phase with a modified least–squares fit (cf. Sec. 5.1.5). The five levels in each block are presented in pseudo–random order. Subsequent levels do not differ more than half of the dynamic range of the sequence. After the second block, a model function is fitted to all data with a modified least–squares fit.

There are two exceptions to these rules: Firstly, the estimated value of L_5 is not presented in the first block since there have been several trials near L_5 during the first phase. Secondly, the listener is protected against stimuli which are too loud. If the adaptive rule requires a stimulus level which was rated as 'too loud' before, the stimulus level is limited to maximal 5 dB above the level which was rated as 'too loud' (cf. Chap. 4 for more details). This procedure yields approximately a uniform distribution of responses on the categorical scale with a minimal number of 'too loud' responses, cf. Chap. 4.

5.1.5 Fitting

The model function $F(L)$ was fitted to the data $y_i(L_i)$ using a modified least–squares fit, i.e., by minimizing $\sum_i \Delta_i^2 = \sum_i (y_i(L_i) - F(L_i))^2$. To take the limited range of the response scale into account, the difference between model function and data was defined as:

$$\Delta_i = \begin{cases} 0 & \text{for } F(L_i) < 0 \quad \wedge \quad y_i = 0 \\ 0 & \text{for } F(L_i) > 50 \quad \wedge \quad y_i = 50 \\ y_i - F(L_i) & \text{else} \end{cases} \tag{5.1}$$

5.1.6 Subjects and measurement program

10 normal–hearing (5 male, 5 female; aged 24–57 years; median 28 years) and 10 hearing–impaired (6 male, 4 female; aged 22–76 years; median 58 years) listeners participated in the experiment. The hearing threshold of the normal–hearing listeners was smaller than 15 dB HL at the standard audiometric frequencies from 125 Hz to 8 kHz. Three of the normal–hearing listeners were members of the research group. The other listeners had no prior experience in psychoacoustical experiments and were paid on an hourly basis. The hearing–impaired subjects showed sensorineural hearing losses of different degrees. Their audiograms ranged between 15 and 85 dB at 500 Hz and between 15 and 95 dB at 4 kHz They had never performed loudness scaling prior to the experiment. All subjects performed 10 loudness scaling tracks on each ear with both the constant stimuli and the adaptive procedure. The measurements were performed in blocks of 5, i.e., 5 tracks with the same procedure and ear were performed successively followed by a block with a different ear and/or procedure. The order of these blocks was randomized between subjects.

5.2 Model functions

Different authors proposed several different model functions to parameterize loudness functions derived from categorical loudness scalings. In this study, 10 different model

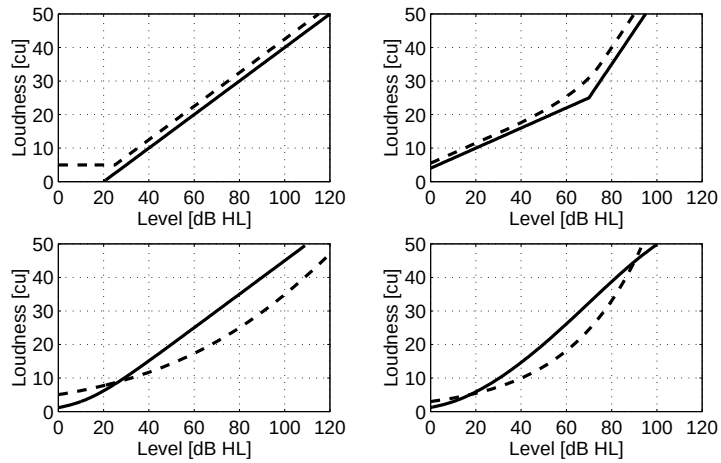

Figure 5.2: *Examples for 8 of the tested model functions. Upper left panel: Linear function (Fechner's law) (solid) and linear function with offset (dashed). Upper right panel: Two linear functions with different slope values, connected in the L_{25} level without smoothing (solid) and with smoothing (dashed). Lower left panel: Modified Fechner function according to Nowak (Heller et al., 1997) (solid) and further modified Fechner function (Brand et al., 1997c) (dashed). Lower right panel: Nowak function (Nowak, 1990) (solid) and Stevens function (Stevens, 1956)(dashed).*

functions were tested with respect to systematic errors and fitting accuracy. Eight examples of the tested model functions are shown in Fig. 5.2.

5.2.1 Linear function (Fechner's law)

The first model function tested in this study was a simple linear loudness function according to Fechner's law (Fechner, 1888) which is based on the Weber fraction. The Weber fraction says that the just noticeable increase in stimulus intensity ΔS is proportional to the intensity S of the stimulus:

$$\frac{\Delta S}{S} = const.$$

Assuming that the loudness perceived can be derived by integration over ΔS and that the categorical loudness is proportional to the loudness perceived, the categorical loudness function in categorical units is linear on a dB–scale. Thus, the model function is:

$$F(L) = 25 + m(L - L_{25}) \tag{5.2}$$

The medium loudness level L_{25} and the slope m are the two free parameters.

5.2.2 Linear function with offset

Since categorical loudness scaling data often show an upwardly concave shape the linear model function was modified by different authors. One modification of Eq. (5.2) was introduced by Hohmann and Kollmeier (1996). They used a third free parameter b which determines a constant offset in the loudness function for soft stimuli:

$$F(L) = \max(b,\, 25 + m(L - L_{25})) \tag{5.3}$$

5.2.3 Modified Fechner function

Another parametrization of loudness functions was proposed by Nowak (Heller *et al.*, 1997) who included a masking noise with intensity Z into the Weber fraction

$$\frac{\Delta S}{S + Z} = const$$

which resulted in a modified Fechner loudness function (Boretzki *et al.*, 1994):

$$F(L) = a \cdot \log(1 + 10^{\frac{L-R}{20}}), \tag{5.4}$$

with the two free Parameters a and R, where R denotes the sound pressure level of the noise in dB.

5.2.4 Further modified Fechner function

Eq. 5.4 was further modified by introducing a third free parameter b controlling the bending of the loudness function (Brand *et al.*, 1997c):

$$F(L) = a \cdot \log(1 + 10^{\frac{L-c}{b}}) \tag{5.5}$$

The free parameters of this function are a, b and c.

5.2.5 Nowak function with 4 parameters

Nowak (1990) developed a further parametrization of loudness functions to achieve a closer approximation to the data obtained with the Würzburg loudness scaling procedure (Heller, 1985):

$$F(L) = A \cdot \left[1 - \exp\left(c\left(10^{\frac{R}{20} \cdot n} - \left(10^{\frac{L}{20}} + 10^{\frac{R}{20}}\right)^n\right)\right)\right] \tag{5.6}$$

The parameter A denotes the maximum scale value which a listener would give at the pain limit. It can be interpreted as the total length of the not restricted loudness scale of the Würzburg loudness scaling procedure.[1] The parameter R denotes the level of the internal noise which is responsible for the hearing threshold, c and n denote further free parameters.
Model functions 5.2 to 5.5 have a slope which remains constant or increases with sound pressure level. However, the Nowak function shows a somewhat unusual behavior, since the slope of the loudness function decreases again for high loudness values.

5.2.6 Nowak function with 3 parameters

Since Eq. (5.6) has 4 free parameters, it is somewhat difficult to fit. Nowak (1990) empirically derived the total length of the not restricted response scale as $A = 60$. Thus, the model function

$$F(L) = 60 \cdot \left[1 - \exp\left(c\left(10^{\frac{R}{20}\cdot n} - \left(10^{\frac{L}{20}} + 10^{\frac{R}{20}}\right)^n\right)\right)\right] \tag{5.7}$$

was also tested in this study.

5.2.7 Stevens' law

Stevens (1956) found empirically in magnitude estimation experiments that subjective loudness is a power function of physical intensity scaled in linear pressure. In a dB–scale the Stevens loudness function corresponds to:

$$F(L) = a \cdot \exp(bL), \tag{5.8}$$

with a and b denoting free parameters. This model function is used by some researchers to fit categorical loudness scaling data as well (e.g., Robinson and Gatehouse, 1996; Keidser *et al.*, 1999).

5.2.8 Polynomials

Further a second grade polynomial with 3 free parameters and a third grade polynomial with 4 free parameters were tested.

$$F(L) = a_0 + a_1 \cdot L + a_2 \cdot L^2 \tag{5.9}$$

$$F(L) = a_0 + a_1 \cdot L + a_2 \cdot L^2 + a_3 \cdot L^3 \tag{5.10}$$

[1] The listener can also respond with categorical loudness values beyond 50 cu in the Würzburg loudness scaling procedure.

5.2.9 Two linear functions

A further model function consisting of two linear parts which have independent positive slope values m_{lo} and m_{hi} and which are connected at the L_{25} level is suggested and tested in this study:

$$F(L) = \begin{cases} 25 + m_{\mathrm{lo}}(L - L_{25}) & \text{for} \quad L \leq L_{25} \\ 25 + m_{\mathrm{hi}}(L - L_{25}) & \text{for} \quad L > L_{25} \end{cases} \tag{5.11}$$

5.2.10 Two linear functions, smoothed

Another model function smoothes the sudden change of the slope at L_{25} in Eq. (5.11) between the categories 15 and 35 using a Bezier fit:

$$F(L) = \begin{cases} 25 + m_{\mathrm{lo}}(L - L_{\mathrm{cut}}) & \text{for} \quad L \leq L_{15} \\ \mathrm{bez}(L, L_{\mathrm{cut}}, L_{15}, L_{35}) & \text{for} \quad L_{15} < L < L_{35} \\ 25 + m_{\mathrm{hi}}(L - L_{\mathrm{cut}}) & \text{for} \quad L \geq L_{35} \end{cases} \tag{5.12}$$

The Bezier smoothing is given in Appendix A.1 in detail. Because of the smoothing the L_{cut} parameter in Eq. (5.12) does not represent the medium loudness level L_{25} but the level where the two linear parts would meet when they were not smoothed. Nevertheless, it is recommended to specify L_{25} rather than L_{cut} for two reasons 1) L_{25} can be calculated from L_{cut}, m_{lo} and 2) m_{hi} and 2) L_{25} can better be interpreted than L_{cut} in clinical diagnostics and hearing aid fitting.

5.3 RESULTS

5.3.1 Individual reference loudness functions

It is necessary to determine the individual reference loudness function for each listener without assuming an underlying model function in order to determine how close the individual loudness functions of the different listeners can be approximated by the different model functions. This was achieved by 10 repeated loudness scaling runs with a 1 kHz narrow–band stimulus presented monaurally to each ear of each listener using both the constant stimuli and the adaptive procedure. The orders of side and procedure were randomized for each listener. From the repeated loudness scaling runs with the same listener, ear and procedure all stimulus levels L_{x} which were rated with the same category x were pooled and their median value $\hat{L}_{\mathrm{x}}$ was calculated. The inverse of the $\hat{L}_{\mathrm{x}}$ values as a function of x was interpreted as the individual reference loudness function.

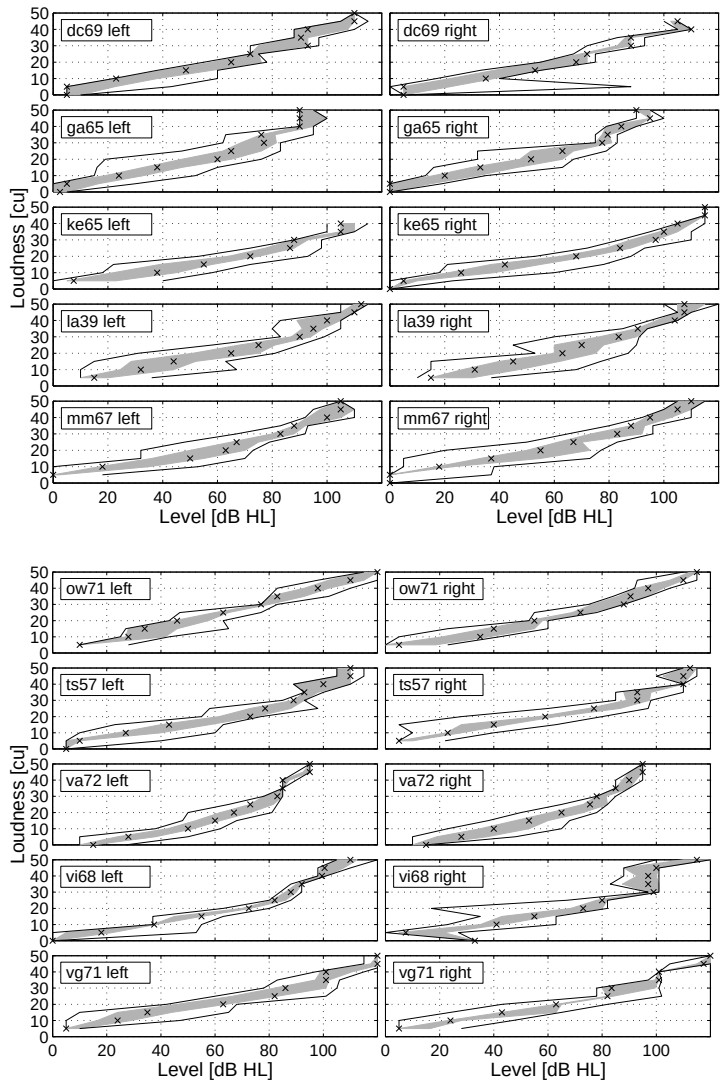

Figure 5.3: *Individual reference loudness curves of normal–hearing listeners obtained with the constant stimuli procedure: Medians (crosses), quartile ranges (grey areas) and extrema (lines) of the stimulus levels L_x which were rated with the same loudness category x.*

Fig. 5.3 shows the individual reference curves and quartile ranges of the L_x–levels measured with the constant stimuli method for all of the 20 normal–hearing ears. Fig. 5.4

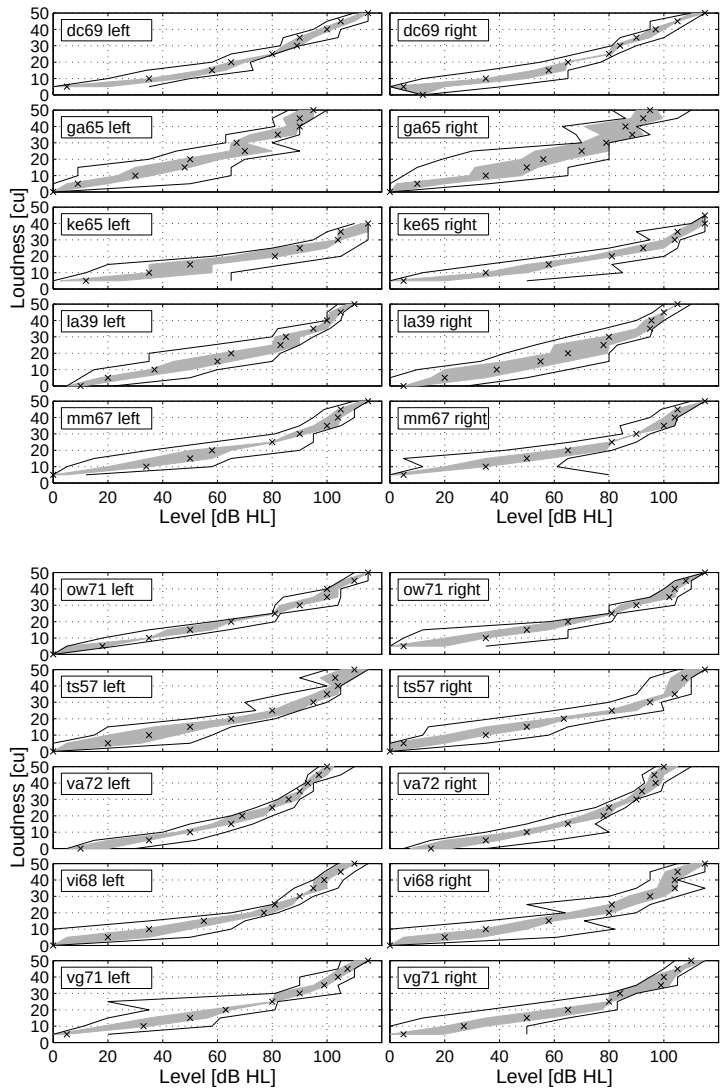

Figure 5.4: *Individual reference loudness curves of normal–hearing listeners obtained with the adaptive procedure: Medians (crosses), quartile ranges (grey areas) and extrema (lines) of the stimulus levels L_x which were rated with the same loudness category x.*

shows the respective data obtained from the adaptive procedure. Data for the hearing–impaired listeners are presented in Figs. 5.5 and 5.6.

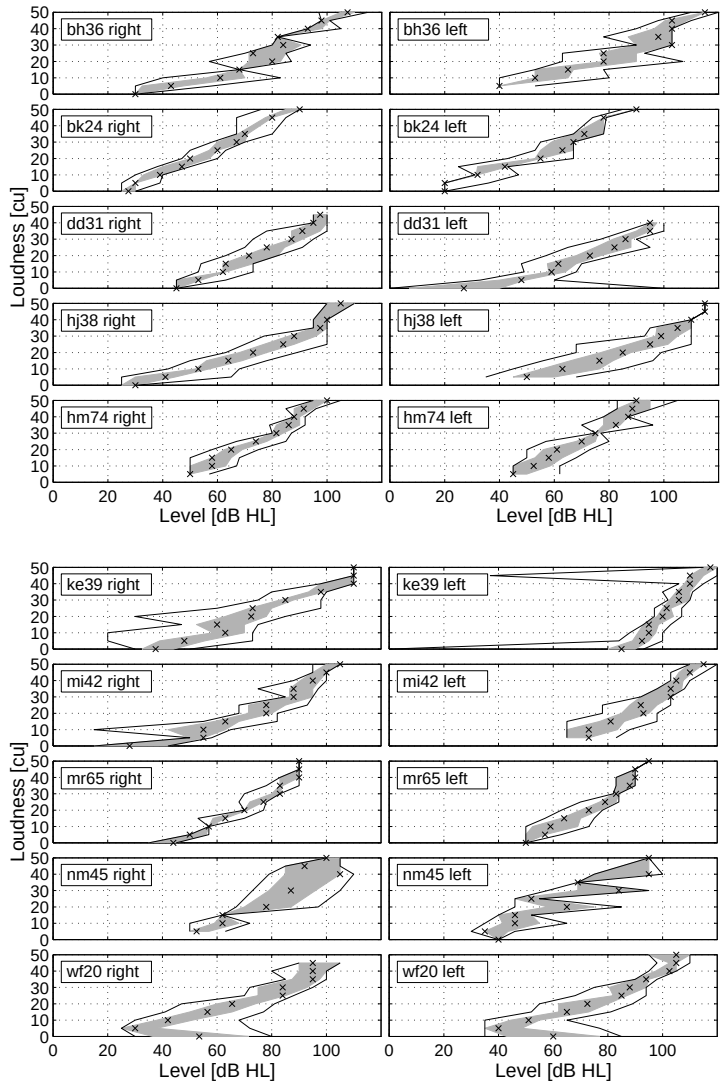

Figure 5.5: *Individual reference loudness curves of hearing–impaired listeners obtained with the constant stimuli procedure: Medians (crosses), quartile ranges (grey areas) and extrema (lines) of the stimulus levels L_x which were rated with the same loudness category x.*

A pronounced variability in range and shape of the individual reference curves can be seen between different listeners even in the normal–hearing group. Most individual

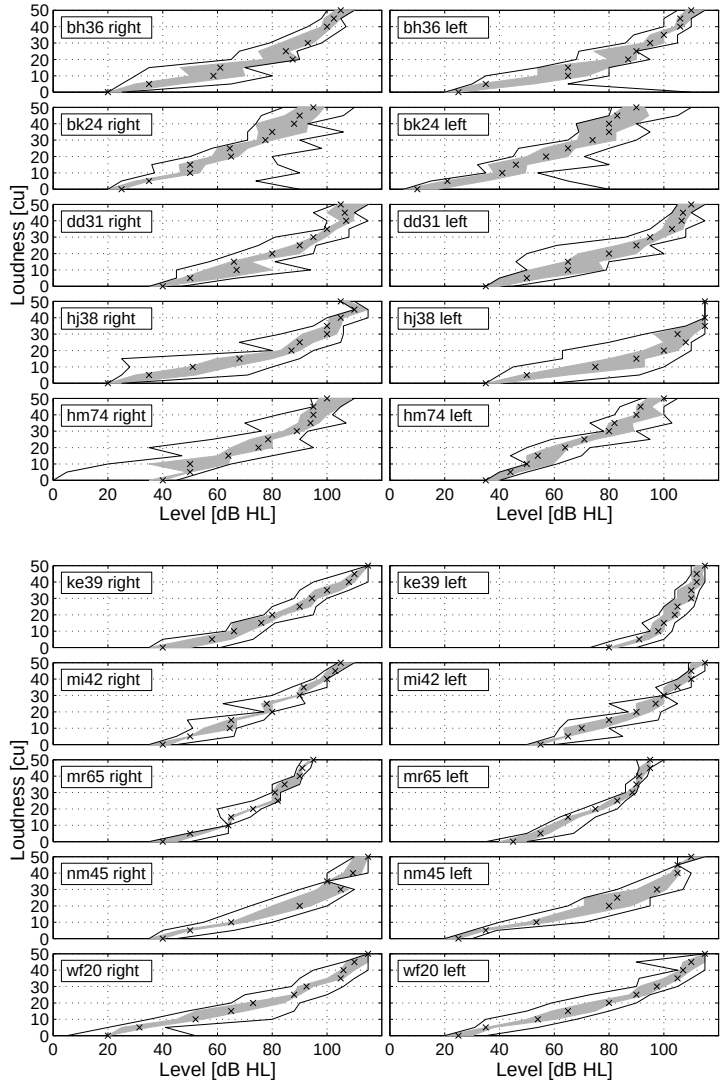

Figure 5.6: *Individual reference loudness curves of hearing–impaired listeners obtained with the adaptive procedure: Medians (crosses), quartile ranges (grey areas) and extrema (lines) of the stimulus levels L_x which were rated with the same loudness category x.*

reference curves of the normal–hearing listeners show a more or less upwardly concave shape. However, some listeners, e.g. ow71 and vg71, show an almost linear loudness

function. The loudness functions of the hearing–impaired listeners are in general more linear, especially for subjects with strong recruitment. At high levels, most hearing–impaired listeners have a loudness curve which is similar to the normal–hearing listeners. At low levels, however, the slope of the loudness function increases according to the increased threshold.

The reference functions derived by the different procedures show slight differences. The adaptive procedure produces slightly more upwardly concave loudness functions compared to the constant stimuli method (cf. Fig. 5.7). This effect is discussed in detail in Chap. 4.

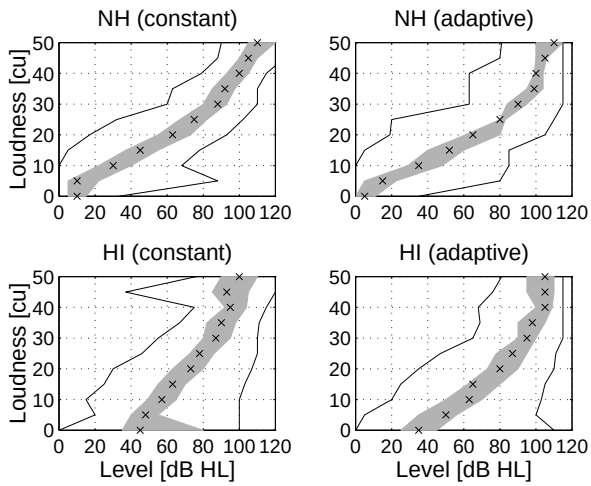

Figure 5.7: Cumulative loudness functions of normal–hearing listeners (upper panels) and hearing–impaired listeners (lower panels) for both constant stimuli (left panels) and adaptive procedure (right panels): Medians (crosses), quartile ranges (grey areas) and extrema (lines) of the stimulus levels L_x which were rated with the same loudness category x.

Fig. 5.7 shows the cumulative loudness functions of the normal–hearing and the hearing–impaired listeners for both procedures. These functions were derived by pooling all responses across all normal–hearing listeners and hearing–impaired listeners, respectively.

5.3.2 Fits to single tracks

In order to assess the differences in experimental adequacy of the different model functions, both biases and standard deviations are calculated for the deviation between the

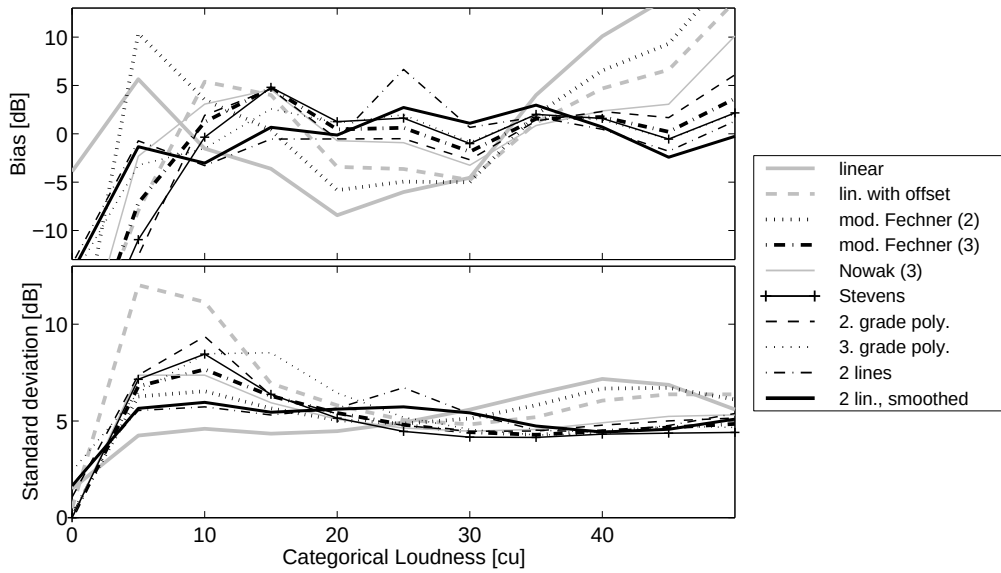

Figure 5.8: *Mean bias (upper panel) and intraindividual standard deviation (lower panel) for different model functions for normal–hearing listeners and the constant stimuli procedure.*

data and the respective model function. The bias of the L_{x} estimate of a given model function $F(L)$ fitted to a given track is given by $\Delta L_{\mathrm{x}} = F^{-1}(x) - \hat{L}_x$, with $F^{-1}(x)$ denoting the inverse of $F(L)$ and $\hat{L}_x$ denoting the reference response levels for the respective listener, ear and procedure. The upper panel of Fig. 5.8 shows the mean of ΔL_{x} averaged across the normal–hearing ears for the constant stimuli procedure for the different model functions. The lower panel of Fig. 5.8 shows the intra–ear standard deviation of the L_{x} estimates. Fig. 5.9 shows the same for the adaptive procedure. Figs. 5.10 to 5.11 show the respective data for the hearing–impaired listeners.

The Nowak function (5.6) with 4 free parameters frequently produced outliers when fitted to the single tracks which caused extremely large biases and standard deviations. This is probably due to its relatively large number of 4 free parameters. Hence, it is not displayed in Figs. 5.8 to 5.11.

The Fechner function (5.2), the Hohmann function (5.3), and the modified Fechner function (5.4) have linear or approximately linear shapes. They overestimate L_{x} for high loudness values and underestimate L_{x} for medium loudness values. The bias is stronger in the normal–hearing listeners because of their more upwardly concave loudness functions. The other model functions give better approximations to the individual loudness curves. The two straight lines with the smoothed transition area (Eq. (5.12)) gener-

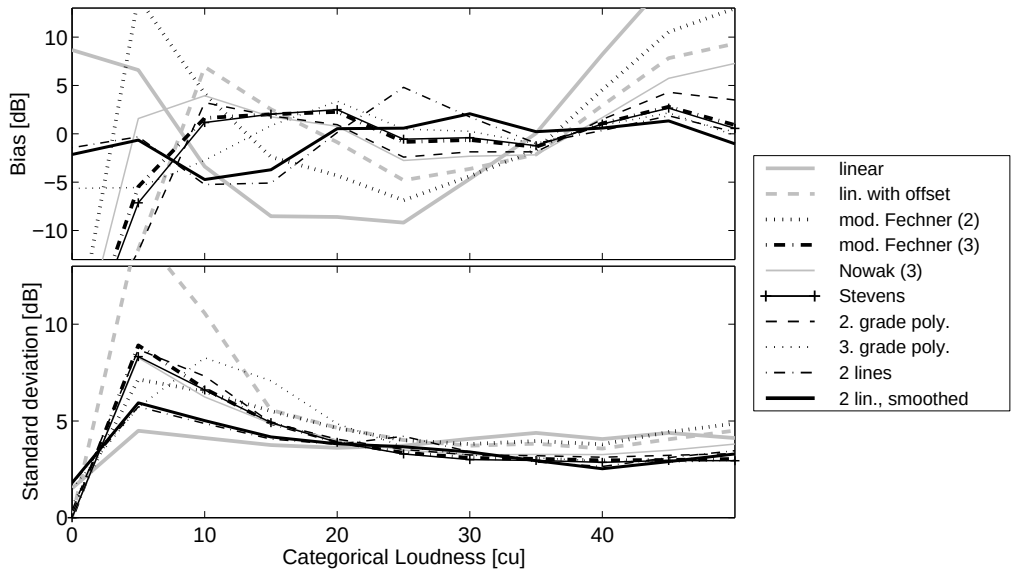

Figure 5.9: *Mean bias (upper panel) and intraindividual standard deviation (lower panel) for different model functions for normal–hearing listeners and the adaptive procedure.*

ate minimal biases and standard deviations averaged across the whole loudness range. The polynomials (Eqs. (5.9) and (5.10)) show large standard deviations in the hearing–impaired listeners when measured with the constant stimuli procedure. As the Nowak function with 4 parameters (5.6), the fits of these two model functions were sometimes not stable. The Stevens function (5.8) yields a good fit under all conditions for loudness values above 10 cu. This model function fits different shapes of individual loudness functions, although it has only two degrees of freedom. For high levels it even yields the best fits in the normal–hearing listeners with the constant stimuli procedure (cf. Fig. 5.8) and hearing–impaired listeners with adaptive procedure (cf. Fig. 5.11). However, it produces significant biases for very low response levels (0 and 5 cu) under all conditions.

The comparison between the two experimental procedures shows that in general the adaptive version yields smaller biases and standard deviations than the constant stimuli version.

Most model functions show large negative biases in L_0 estimates. This was expected for the model functions (5.4), (5.5), (5.6), (5.7) and (5.8) because their loudness value is always positive as they converge asymptotically at zero for low levels. Consequently, they always underestimate L_0 which is the level related to 0 cu.
However, the other model functions also show mostly negative biases in L_0 estimates.

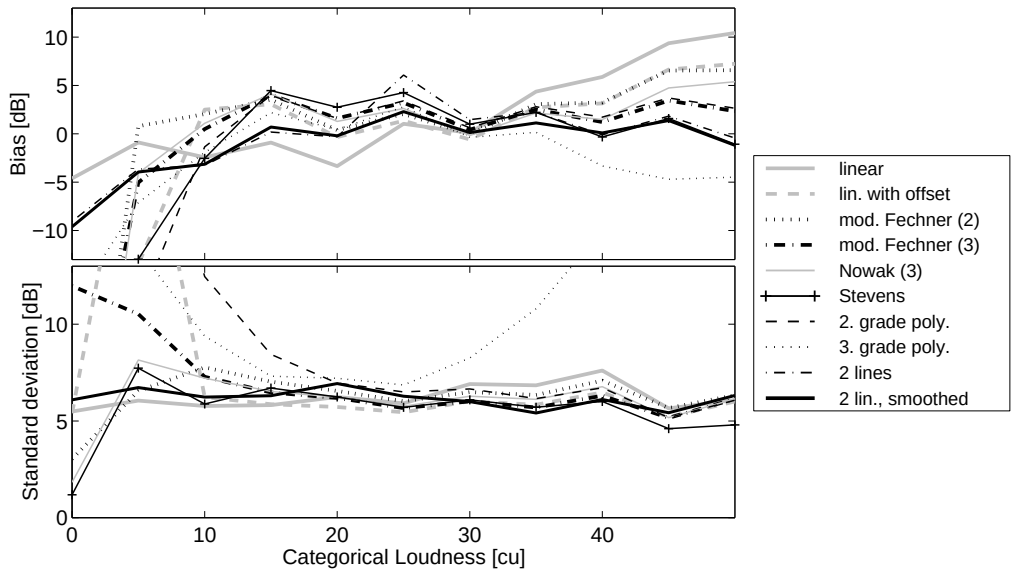

Figure 5.10: *Mean bias (upper panel) and intraindividual standard deviation (lower panel) for different model functions for hearing–impaired listeners and the constant stimuli procedure.*

Especially, in the Fechner function (5.2) this was not expected because this linear model function should generate positive biases at L_0 in upwardly concave loudness functions. This unexpected bias at L_0 is probably due to a bias in the individual reference curves at L_0. In the normal–hearing listeners this bias can be explained as follows: Since only stimulus levels above 0 dB HL were presented in this study, the distribution of stimulus levels which were rated as inaudible was biased because the hearing threshold itself is about 0 dB HL. It can be seen from Figs. 5.3 and 5.4 that in many listeners the median of the L_0 estimates is shifted towards higher levels compared to the linear extrapolation of the lower part of the loudness function. This effect can be seen in the cumulative loudness functions shown in Fig. 5.7 as well. The bias is smaller in the adaptive procedure than in the constant stimuli procedure. This is probably due to the fact that in the adaptive procedure all inaudible responses are used for data analysis whereas the constant stimuli procedure avoids to present inaudible stimuli in the data collection part. Therefore, the data basis was very small at L_0 in the constant stimuli procedure (cf. Fig. 4.5 in Chap. 4).

When the adaptive procedure is used for hearing–impaired listeners there is no bias in the individual reference curves at L_0, i.e. the median of L_0 is equal to the linear extrapolation of the lower part of the loudness function (cf. Figs. 5.6 and 5.7). In this case, the stimulus level distribution which is rated as inaudible is unbiased because the

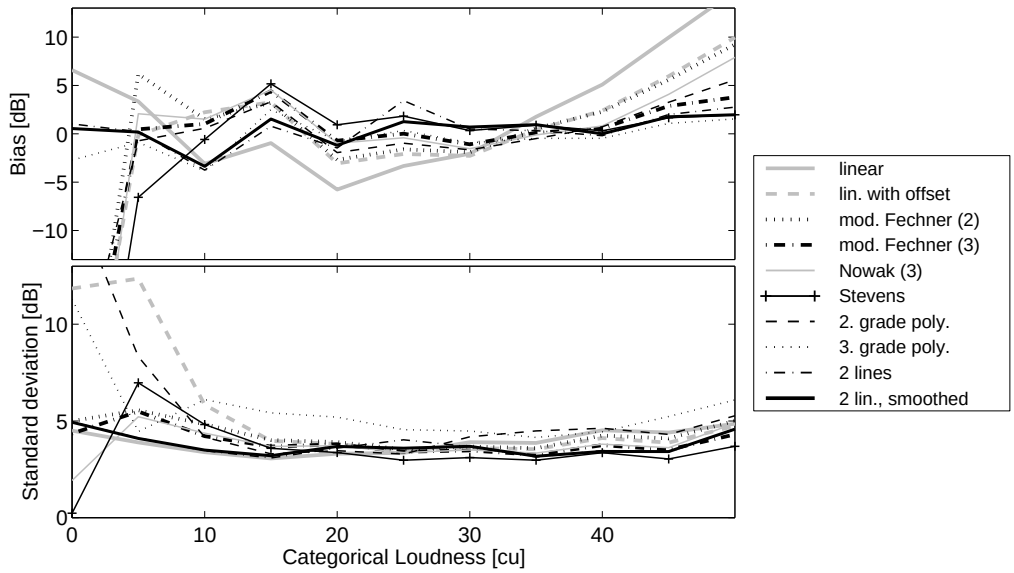

Figure 5.11: *Mean bias (upper panel) and intraindividual standard deviation (lower panel) for the different model functions for hearing–impaired listeners and the adaptive procedure.*

adaptive procedure also presents stimuli below threshold. Since there is no bias at L_0 in the individual reference curves, the biases of the L_0 estimates of the different model loudness functions (Fig. 5.11) have the expected ranking: The Fechner function (5.2) is positively biased. The asymptotical loudness functions (5.4), (5.5), (5.6), (5.7), (5.8) and also the polynomials (5.9) and (5.10) are negatively biased. Only the two straight lines with or without smoothed transition area (5.11) and (5.12) generate unbiased L_0 estimates.
Using the constant stimuli procedure for hearing–impaired listeners there is still a bias in the individual reference curves at L_0 which is probably due to the constant stimuli procedure which also reduces inaudible stimuli in hearing–impaired listeners.

5.4 DISCUSSION

The importance of establishing an appropriate loudness function for categorical loudness scaling originates both from the diagnostics of sensorineural hearing impairment and from the design of dynamic compression in hearing aids. There are different types of hearing loss and loudness growth which require different types of input/output functions in hearing aids (Killion and Fikret-Pasa, 1993). With digital signal processing techniques

available in modern hearing aids arbitrary input/output functions can be implemented: For example, many hearing aids use input/output functions with a linear gain region for low levels, a linear compression region for medium levels and an output limiting region for high levels (e.g., Cornelisse *et al.*, 1995; Hohmann and Kollmeier, 1995a). Other hearing aids use an input/output function which is compressive for medium levels, linear above a completion point and limiting for very high levels (e.g., Killion, 1979; Lunner *et al.*, 1997). Furthermore, different input/output functions can be used in different frequency channels within one hearing aid (e.g., Lunner *et al.*, 1997). Thus, a variety of different input/output characteristics and of parameters like gains, compression ratios, kneepoints and completion points have to be adjusted for the different frequency channels of a hearing aid. Categorical loudness scaling could be used to determine the optimal parameter setting for an individual listener. Therefore, the model loudness function used to fit the data should provide enough variability to approximate a large range of individual loudness functions. In addition, the number of degrees of freedom should be low to allow for a stable fitting even for short track lengths.

In the present study, the individual reference loudness functions were derived by calculating the median value $\hat{L}_x$ of the stimulus levels which generated the same loudness rating. A similar method was formally used by Allen *et al.* (1990) who stated, that it is incorrect to average the levels of different responses since they represent different things to the listener. Heller *et al.* (1997) favored a different approach: They calculated the mean of the scaling values to the same presentation level to derive collective loudness functions for hearing–impaired listeners. Heller and Boretzki (1998) doubt whether averaging the stimulus levels which produced the same rating is adequate. Both, Heller *et al.* and Allen *et al.* used the respective averaging method which was easily to apply to their specific data: Since Allen used only 7 response alternatives, a large number of different stimulus levels correspond to each response value. Heller *et al.* used more than 50 response alternatives. Consequently, a large number of different responses was generated for each level. In the present study there are more levels per response than responses per level which is similar to Allen *et al.*. A detailed discussion of the effect of the number of response alternatives employed is given by Brand *et al.* (1997d) and in Chap. 3.

A comparison of categorical loudness scaling data for narrowband signals of different authors shows that all authors derived loudness functions which are upwardly concave. However, the loudness range at which the loudness function slope shows the largest increase depends on the number of response alternatives:
Allen *et al.* (1990) used 7 response alternatives and fitted polygons to their data. All normal–hearing listeners and most of the hearing–impaired listeners, regularly displayed a more rapid increase in rating at the higher end of the scale above the category 'loud' (35 cu).
Rasmussen *et al.* (1998) used the same scale as Allen *et al.* with pure–tone stimuli. They

derived mean loudness curves for normal–hearing listeners which also showed a linear shape for x below 35 cu ('loud') and a steeper and linear shape above 35 cu.
In the current study, 11 response alternatives have been used and the slope of the loudness function increases at 25 cu ('medium').
Heller *et al.* (1997) used 51 and more response alternatives and fitted the Nowak function with three parameters (5.7) to their data. Their collective loudness functions for normal–hearing listeners and narrow–band frequency–modulated sinusoidals show an upwardly concave shape below the category 15 ('soft') and an almost linear shape with a steeper slope above the category 15.
In the experimental setup of Baumann *et al.* (1998), the listeners had to rate the loudness of 1/3–octave bands of noise–bursts using a non–labeled scale with a length of 270 mm. The scale was sampled with 51 response alternatives which were unknown to the listener. A loudness function consisting of two straight lines connected near the categorical loudness 10 enabled a close fit to the data of Baumann *et al.*.
Taken together, there is a tendency towards lower kneepoints of the fitted loudness function with increasing number of response alternatives. The reason for this behavior is unclear at the moment. However, it has to be concluded that the 'optimal' loudness model function differs across different categorical scales even when similar experimental parameters are used.

One shortcoming of the measuring procedures used in this study is that the minimal stimulus level was limited to 0 dB HL. Since this is the mean hearing threshold level in normal–hearing people, it is likely that some normal–hearing listeners had a hearing threshold below 0 dB HL. Consequently, the limited range of stimulus levels caused a positive bias in $\hat{L}_0$ estimates. This can be seen from the individual reference curves of the normal–hearing listeners in Fig. 5.3 and Fig. 5.4 where the value of $\hat{L}_0$ often deviates from the extrapolation of the remaining data points. This effect is stronger in the constant stimuli method. The effect is smaller in the hearing–impaired group using the constant stimuli method (Fig. 5.5) because the minimum presentation level was below threshold. The remaining bias seems to be due to the constant stimuli method, because in the hearing–impaired group and the adaptive procedure (Fig. 5.6) no bias occurred at $\hat{L}_0$. It is recommended not to limit the minimal stimulus level in order to avoid this bias of L_0 estimate.

A shortcoming of the complete measurement program was that it lasted approximately 2–3 hours per listener including breaks for regeneration. Each listener exhibited a training effect during this period which may have caused a bias. Therefore, the reference curves derived from these measurements cannot be used as normative curves.[2] However, the data basis presented in this study should still allow for conclusions about the optimal parametrization of loudness functions.

[2] Normative curves for the constant stimuli procedure can be found in Albani *et al.* (1997).

Another shortcoming of this study is the limited number of normal–hearing and hearing–impaired listeners. Thus, certain listeners having empirical loudness functions that differ substantially from the general shape observed here might not be included here. However, up to now, such listeners have not been observed in the standard diagnostics performed in our laboratory. Hence, it does not seem likely that a substantial group of hearing–impaired listeners with a totally different loudness function exist.

The model functions investigated in this study can be divided into three groups:
The first group consists of the functions (5.2) to (5.4) which have a linear or nearly linear shape and give insufficient fits in concave loudness functions. All these functions overestimate the L_x levels for categorical loudness values x above 40 cu and underestimate L_x near $x = 25$ cu. Since hearing–impaired listeners with recruitment usually show less upwardly concave loudness functions, this effect is stronger in the normal–hearing listeners. The parameters of the linear funcion with variable offset (Eq. (5.3)) are not linearly independent which causes that the fit is not stable. Small variations in the data can generate large variations of the fitted parameters. The modified Fechner function (5.4) converges asymptotically at zero for low L values. Therefore, the hearing threshold level is not given directly by this model function. However, it is reported to fit reasonably with data measured with the Würzburg loudness scaling procedure (Boretzki *et al.*, 1994). This procedure uses a response scale without an upper limit and at least 51 response alternatives.[3]
The second group of model functions consists of the model functions (5.5) to (5.10). These model functions approximate the individual reference curves clearly better than the first group. The Stevens function (5.8) has only 2 free parameters and gives very good fits under all conditions for loudness values above 10 cu. The functions with 4 free parameters – Nowak(4) Eq. (5.6) and 3rd grade polynomial Eq. (5.10) – fail to extrapolate the loudness function in level ranges where no stimuli were presented. Similar problems occurred in the 2nd grade polynomial (5.9). The Nowak function with 4 parameters (5.6) generated so many outliers that it cannot be used at all.[4] The Nowak functions (Eqs. (5.6) and (5.7)) show a somewhat unusual behavior, because the slope of the loudness function decreases again for very high loudness values. This effect was found by Nowak (1990) but was not observed for the listeners of this study. The polynomials have the disadvantage, that the parameters cannot be interpreted intuitively because different parameter settings can generate similar loudness function shapes. Model functions (5.5) to (5.8) converge asymptotically at zero and the hearing threshold level cannot be extrapolated.
The third group consists of model functions (5.11) and (5.12). Both consist of two straight lines which are connected at the L_{25}. The latter is smoothed in the transition

[3] In the original Würzburg loudness scaling procedure (Heller, 1985) the category 'too loud' is related to the categorical unit 51 and the listeners can also use ratings above 51 cu.

[4] Even though, the initial parameters of each fit were set to the resulting parameters of the Nowak function with 3 parameters, no stable fit resulted for the Nowak function with 4 parameters.

area. They yield the closest approximation to the individual reference curves. Their parameters are linearly independent and can easily be interpreted for diagnostics and hearing aid fitting. Further, the hearing threshold level and the uncomfortable level can always be extrapolated. Model function (5.12) yields the smallest standard deviation and the smallest bias of L_x estimates, when the whole loudness range is considered, compared to all other model functions investigated in this study. For these reasons it is recommended to use model function (5.12) to fit loudness functions to data collected with the Oldenburg loudness scaling procedure.

5.5 CONCLUSIONS

The shape of the empirical loudness functions derived from individual data with categorical loudness scaling depends on procedural factors (such as, e.g., number of response alternatives, averaging across levels or response alternatives), stimuli (such as, e.g., level range and bandwidth) and subjective factors (such as, e.g., hearing loss and recruitment). Therefore, the empirical data obtained in this study differ from those found in other studies (e.g., Allen *et al.*, 1990; Heller *et al.*, 1997; Rasmussen *et al.*, 1998). Moreover, they show a clear interindividual variation even among normal–hearing listeners.
For both, the adaptive and the constant stimuli version of the Oldenburg loudness scaling procedure, the model function (5.12) yielded the closest approximation to the individual reference loudness functions over the whole loudness range in both normal–hearing and hearing–impaired listeners. This model function consists of two straight lines with independent slopes which are connected at the loudness category 25 ('medium'). The transition area is smoothed by a Bezier interpolation between the categories 15 ('soft') and 35 ('loud'). Thus, it is recommended to use this model function to fit data obtained with the Oldenburg loudness scaling procedure.

ACKNOWLEDGMENTS

We would like to thank Birgitta Gabriel for helpful comments on the manuscript.

This study was supported by BMBF, PT-AUG and by the CEC supporting action NATASHA.

Chapter 6

Effect of Center Frequency and Bandwidth on the Shape of Loudness Functions in Categorical Loudness Scaling

ABSTRACT

Loudness functions were measured with 8 normal–hearing and 8 hearing–impaired listeners using a categorical loudness scaling method (Hohmann, 1993). Different stimuli were used, namely three narrowband signals (i.e. 1/3–octave bands of noise with the center frequencies 250, 1,000 and 4,000 Hz) and two broadband signals (i.e. a CCITT speech simulating noise and a German speech sample). In the normal–hearing listeners, the shape of the loudness function hardly depended on the center frequencies of the narrowband stimuli. However, the shapes of the loudness functions differed clearly between the broadband and the narrowband stimuli. Although, there were considerable differences between listeners. In the normal–hearing listeners, generally, narrowband stimuli generated upwardly concave loudness functions, whereas broadband stimuli generated more linear loudness functions. In the hearing–impaired listeners, the narrowband loudness functions generally showed a more linear shape than in the normal–hearing listeners. Theses findings are consistent with other studies (e.g., Zwicker *et al.*, 1957) that applied loudness comparison and found that spectral loudness summation was maximal at medium levels and minimal at the limits of the auditory range. It can be concluded that loudness scaling allows for the quantification of loudness summation effects on an individual basis. The different shapes of loudness functions for narrowband and broadband stimuli have consequences for the restoration of loudness in hearing–impaired listeners

using hearing aids with automatic gain control.

6.1 INTRODUCTION

Categorical loudness scaling (e.g., Pascoe, 1978; Heller, 1985; Allen *et al.*, 1990; Hohmann and Kollmeier, 1995b) is a direct method to investigate the question "How loud is a given stimulus perceived by the listener in categories of 'soft', 'medium' and 'loud'?", because in this method, the listener rates the stimuli directly using these categories. For that reason, categorical loudness scaling is used by many experimenters in clinical audiology (e.g., Allen *et al.*, 1990; Kießling *et al.*, 1993; Hohmann and Kollmeier, 1995b; Launer *et al.*, 1996) and to determine the input/output characteristics of hearing aids with automatic gain control (e.g., Pascoe, 1978; Hellbrück and Moser, 1985; Moore *et al.*, 1992; Kießling, 1995). On the other hand, experimenters who are interested in the question "At which presentation level is stimulus A perceived as loud as stimulus B?", usually use loudness comparison (e.g., Zwicker *et al.*, 1957; Scharf and Hellman, 1966; Verhey, 1989).

The phenomenon of spectral loudness summation has typically been measured using loudness comparison measurements (e.g., Zwicker *et al.*, 1957; Scharf and Hellman, 1966; Hübner and Ellermeier, 1993; Verhey, 1989). It was found that in normal–hearing listeners the loudness increases when the bandwidth of a sound is increased beyond a critical band keeping the overall intensity fixed. This spectral loudness summation effect is most prominent at medium levels and less prominent at very low and high levels. In people suffering from sensorineural hearing loss, the loudness summation is often reduced (e.g., Scharf and Hellman, 1966; Zwicker and Fastl, 1990).

Categorical loudness scaling is only an indirect method to assess spectral loudness summation. Theoretically, however, spectral loudness summation should have an impact on the shapes of loudness functions in categorical loudness scaling experiments, too. Since loudness summation is most prominent at medium levels and less prominent at low and very high levels, the loudness functions derived from categorical loudness scaling measurements are expected to be less upwardly concave in broadband than in narrowband stimuli. Several experimenters investigated this question so far (Launer *et al.*, 1996; Ricketts and Bentler, 1996; Cox *et al.*, 1997; Appell and Hohmann, 1998):

Launer (1996) found spectral loudness summation in normal–hearing and some hearing–impaired listeners. Smaller loudness summation effects were found at high sound pressure levels compared to moderate and low levels which indicates that the loudness functions of narrowband and broadband signals are not equal. In hearing–impaired listeners, loudness summation was reduced in general.

Ricketts and Bentler (1996) examined loudness growth patterns of different narrowband and broadband stimuli in normal–hearing and hearing–impaired listeners using categor-

ical loudness scaling measurements as well. They found that increasing the bandwidth beyond the critical band increased loudness significantly. However, the shape of the loudness growth patterns was reported to be similar for different bandwidths of the test signal.
Cox *et al.* (1997) found different loudness functions for narrowband and for speech signals. The speech stimuli were rated louder than the narrowband stimuli at moderate levels. The difference was less at very soft and very loud levels.
Appell and Hohmann (1998) varied the bandwidth of uniform–exciting noise systematically and found upwardly concave loudness functions in small bandwidths and more linear loudness functions in larger bandwidths. Again, the spectral loudness summation effect was most prominent at intermediate levels.

This study investigates whether categorical loudness scaling measurements show different shapes of loudness functions for stimuli that are typical for audiology and to what extend these loudness functions are consistent with loudness balancing data. For this purpose, measurements using a categorical loudness scaling procedure (Hohmann and Kollmeier, 1995b) were performed with 8 normal–hearing and 8 hearing–impaired listeners. Three narrowband stimuli (i.e. third octave bands of noise with the center frequencies 250 Hz, 1 kHz and 4 kHz) and two broadband stimuli (i.e. a CCITT speech–simulating noise and a German sentence) were used.

Since some modern hearing–aids exhibit several separete frequency channels, usually categorical loudness scaling is performed using narrowband stimuli in order to adjust the gain characteristics of the different frequency channels. The over–all loudness, however, produced by broadband signals that are processed by all frequency channels in parallel is influenced by spectral loudness summation. The consequences of spectral loudness summation on the fitting of hearing aids with automatic gain control systems are discussed.

6.2 EXPERIMENTAL METHOD

6.2.1 Procedure

The subjects indicated their loudness ratings on the response scale shown in Fig. 6.1. The scale consists of eleven response alternatives including five named loudness categories, four not named intermediate response alternatives and two named limiting categories. The named response categories are ‘sehr leise’ (‘very soft’), ‘leise’ (‘soft’), ‘mittel’ (‘medium’), ‘laut’ (‘loud’) and ‘sehr laut’ (‘very loud’) and correspond to 5, 15, 25, 35 and 45 cu (categorical units) as shown on the left side of Fig. 6.1. The not named response alternatives are used to increase the total number of response alternatives. They are indicated with horizontal bars with increasing length for increasing loudness and are

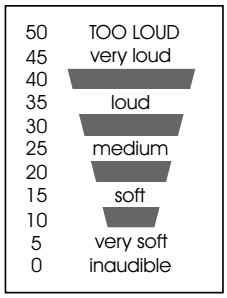

Figure 6.1: *Category scale with 11 response alternatives used by listeners to rate the loudness. The numbers on the left side indicate the categorical units (cu) which are used for data storage and analysis. They were not visible to the subject.*

placed between the named loudness categories. They correspond to the categorical units 10, 20, 30 and 40 cu, respectively. The two limiting categories are named 'unhörbar' ('inaudible') and 'ZU LAUT' ('TOO LOUD') and correspond to 0 and 50 cu.

Prior to testing, each listener was instructed verbally by the experimenter (cf. A.2). During the instruction, the response box was practically demonstrated and any questions were clarified.

The constant stimuli version of the Oldenburg loudness scaling procedure (Hohmann and Kollmeier, 1995b) was used. The procedure includes two parts. The auditory dynamic range of the individual listener is estimated by presenting an ascending level sequence in the first part of the measurement. The loudness function is assessed by presenting stimuli covering the so determined full auditory dynamic range in the second part.
The first part uses an ascending stimulus level sequence with an initial level of 0 dB HL and a step size of 5 dB. The listener's task is to press the response button as soon as the stimulus is audible. Then the level is further increased in 15 dB steps up to 85 dB and increased in 5 dB steps beyond 85 dB. Now, the listener is asked to press another response button 'too loud' immediately when the stimulus is perceived as too loud. In case that the listener does not press the response button, the sequence stops at 120 dB HL.
In the second part of the procedure the loudness function is estimated. Two stimuli are presented at each of 7 diffrent levels which are distributed equidistantly on a dB–scale between the limits of the dynamic range estimated in the first part of the procedure. The listener rates the loudness using the scale described above. The stimuli are presented in pseudo–random order in a way that the maximum difference of subsequent presentation levels is smaller than half of the dynamic range of the sequence. In this way, context effects should be reduced which are due to the tendency of many listeners to rate the current stimulus relatively to the previous stimulus. After completion of the track a

model function is fitted to the data by a modified least–squares fit, cf. Sec. 6.2.2.

6.2.2 Model function and fitting

$$F(L) = \begin{cases} 25 + m_{\mathrm{lo}}(L - L_{\mathrm{cut}}) & \text{for} \quad L \leq L_{15} \\ \mathrm{bez}(L, L_{\mathrm{cut}}, L_{15}, L_{35}) & \text{for} \quad L_{15} < L < L_{35} \\ 25 + m_{\mathrm{hi}}(L - L_{\mathrm{cut}}) & \text{for} \quad L \geq L_{35} \end{cases} \tag{6.1}$$

The model function used in this study consists of two linear parts with independent slope values m_{lo} and m_{hi}. The two parts are connected at the level L_{cut}. The transition area between the loudness categories L_{15} ('soft') and L_{35} ('loud') is smoothed with a Bezier fit denoted with $\mathrm{bez}(L, L_{\mathrm{cut}}, L_{15}, L_{35})$. The exact form of the Bezier smoothing is given in Appendix A.1.[1] This model function provided the best fit to experimental data (cf. Chap. 5).

The model function $F(L)$ was fitted to the data $y_i(L_i)$ using a modified least–squares fit, i.e., by minimizing $\sum_i \Delta_i^2 = \sum_i (y_i(L_i) - F(L_i))^2$. To account for the limited range of the response scale, the difference between model function and data was defined as:

$$\Delta_i = \begin{cases} 0 & \text{for } F(L_i) < 0 \quad \wedge \quad y_i = 0 \\ 0 & \text{for } F(L_i) > 50 \quad \wedge \quad y_i = 50 \\ y_i - F(L_i) & \text{else} \end{cases} \tag{6.2}$$

6.2.3 Stimuli

Third–octave bands of noise with center frequencies of 250 Hz, 1 kHz and 4 kHz were used as narrowband stimuli. They were generated from a random noise with Gaussian amplitude statistics, 5 s duration and 44.1 kHz sampling rate. The signal was transformed to the frequency domain by an FFT. All FFT–coefficients outside the desired band were set to zero and the resulting signal was transformed back to the time domain by an inverse FFT. A segment of 2 s duration was selected randomly and windowed with 100 ms $\cos^2$ ramps. During each trial, the noise was presented twice with a silent interstimulus interval of 1 s duration. The narrowband stimuli were calibrated to free–field sensitivity level of sinusoidals (ISO 389 (1991)) according to Richter (1992).

A telephone band pass filtered speech simulating noise (CCITT G.227, 1964) and the German sentence 'Der Bahnhof liegt sieben Minuten entfernt.' ('The railway–station is seven minutes away.') which is a sample of the Göttingen sentence test (Kollmeier

[1] Because of the smoothing in the medium range, the L_{cut} parameter in Eq. (6.1) does not represent the medium loudness level L_{25} but the level where the two linear parts would meet if they were not smoothed.

and Wesselkamp, 1997) were used as broadband stimuli. Both signals were sampled with 25 kHz. The CCITT noise had a duration of 2 s and was windowed with 100 ms $\cos^2$ ramps. During each trial, the noise was presented twice with a silent interstimulus interval of 1 s duration. The sentence had a duration of 2.3 s and was presented once per trial.
Since no hearing–level correction factors were available for the broadband stimuli, the hearing thresholds of both CCITT noise and speech sample were measured with the normal hearing listeners who participated in this study using a standard audiometry procedure and the same equipment as in the latter loudness measurements. The mean hearing thresholds are 15.6 dB SPL for the CCITT noise and 12.2 dB SPL for the speech sample as measured using the B & K 4153 artificial ear.

6.2.4 Apparatus

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

The broadband signals were presented without free–field equalization. The maximum presentation level possible was at least 120 dB HL for the narrowband stimuli and at least 120 dB SPL for the broadband stimuli. The subjects were seated in a sound–insulated booth. Their task was to rate the loudness of each stimulus presented using an Epson EHT 10S handheld computer with a LCD touchscreen showing the response scale. The handheld computer was connected to the personal computer via serial interface. The loudness ratings for each stimulus were stored for later statistical analysis.

6.3 Subjects and measurement program

8 normal–hearing listeners (3 male, 5 female; aged 24–39 years; median 27 years) and 8 hearing–impaired listeners (5 male, 3 female; aged 16–72 years; median 55 years) participated in the experiment. The hearing threshold of the normal–hearing listeners was lower than 15 dB HL at the standard audiometric frequencies from 125 Hz to 8 kHz. The hearing–impaired subjects showed sensorineural hearing losses of different degrees. Their audiograms ranged between 15 and 85 dB at 500 Hz and between 15 and 95 dB at 4 kHz. Two of the normal–hearing listeners were members of the research group.

All subjects (except hearing–impaired subject ej80) participated in an earlier study in which repeated categorical loudness scaling measurements of the 1 kHz 1/3–octave band of noise were performed, cf. Chaps. 3 and 4. The data for the 1 kHz stimulus shown in this paper were taken from this earlier study. All subjects performed five categorical loudness scaling runs for each stimulus at each ear. After each run the stimulus was changed. Those subjects who participated in the earlier study performed ten runs using the 1 kHz stimulus at each ear.

6.4 RESULTS

6.4.1 Individual loudness functions

Fig. 6.2 shows the individual loudness functions (separately for each ear) for the normal–hearing listeners and the different stimuli. These loudness functions were fitted to all data points collected during the five loudness scaling runs (70 trials) per subject, stimulus and ear.[2] Because of this large data basis per loudness function, these loudness functions can be assumed as being very accurate. Most listeners show upwardly concave loudness functions for the narrowband stimuli and more linear loudness functions for the broadband stimuli (CCITT noise and speech). Generally, the broadband stimuli were rated louder than the narrowband stimuli, especially at moderate levels. These findings can be explained by spectral loudness summation which is known to be most prominent at medium levels and less prominent at high and low levels. However, there are also subjects (vi68 and vg71) who show nearly equal loudness functions for narrowband and broadband stimuli.[3] Thus, in these subjects no spectral loudness summation effect can be observed.
Subject ow71 exhibits unusual loudness functions: The loudness functions for the narrowband stimuli are almost linear, whereas the loudness functions for the broadband stimuli have a downwardly concave shape. However, subject ow71 shows a regular amount of loudness summation with a typical maximum at moderate levels.

Fig. 6.3 shows the individual loudness functions (separately for each ear) for the hearing–impaired listeners. The functions were fitted to all 70 trials performed per subject, stimulus and ear.[4] Most loudness functions show an increased hearing threshold due to hearing loss and an increased slope due to loudness recruitment. Most listeners show

[2] The fits for the 1 kHz stimulus are based on ten scaling runs (140 trials).

[3] In subject vg71 the 1 kHz loudness curve deviates strongly from the remaining narrowband loudness curves. Possibly, this is due to the fact that the 1 kHz loudness curve was measured in another session than the other loudness curves. In this session only the 1 kHz stimulus was used for a large number of loudness scaling runs. This might have caused an adaptation effect which caused a bias. The remaining stimuli were measured in interleaved runs, i.e. after each run a different stimulus was used.

[4] Except for subject ej80 the fits for the 1 kHz stimulus are based on 140 trials.

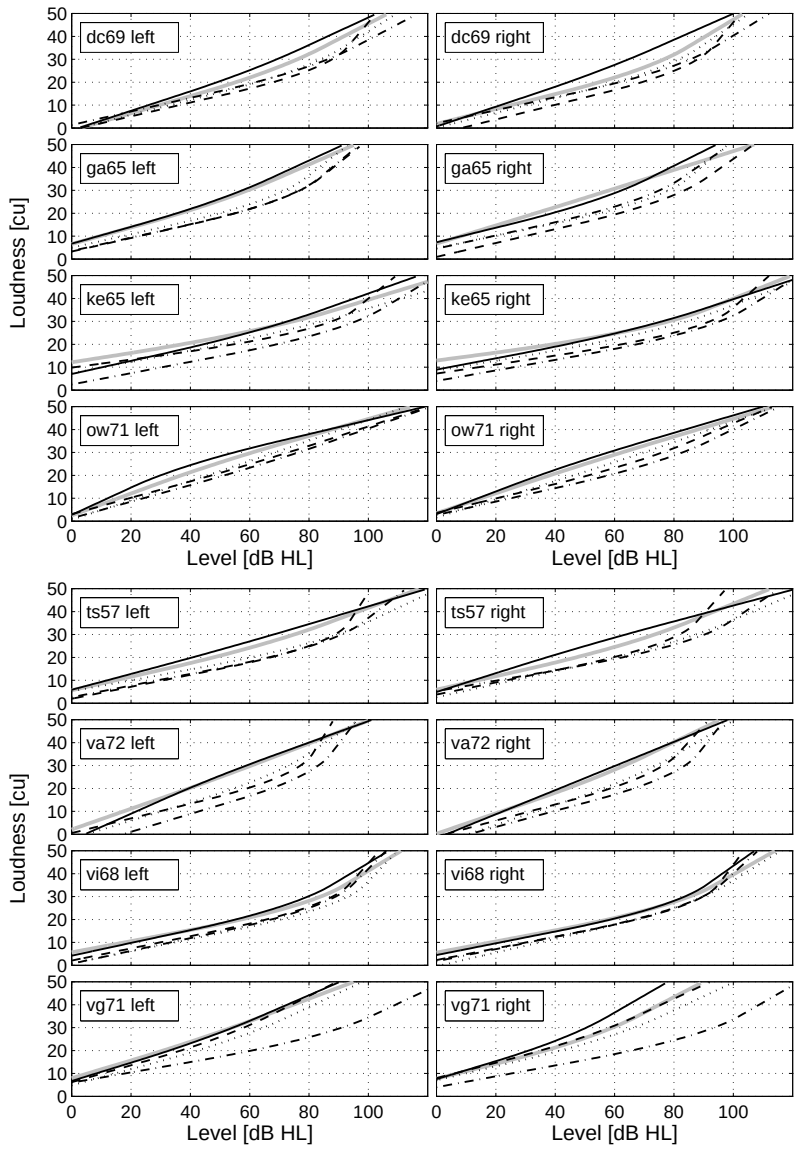

Figure 6.2: *Individual loudness curves of 8 normal–hearing listeners for different stimuli: 250 Hz narrowband noise (dotted), 1 kHz narrowband noise (dash–dotted), 4 kHz narrowband noise (dashed), CCITT speech–simulating noise (solid black) and speech sample (solid grey).*

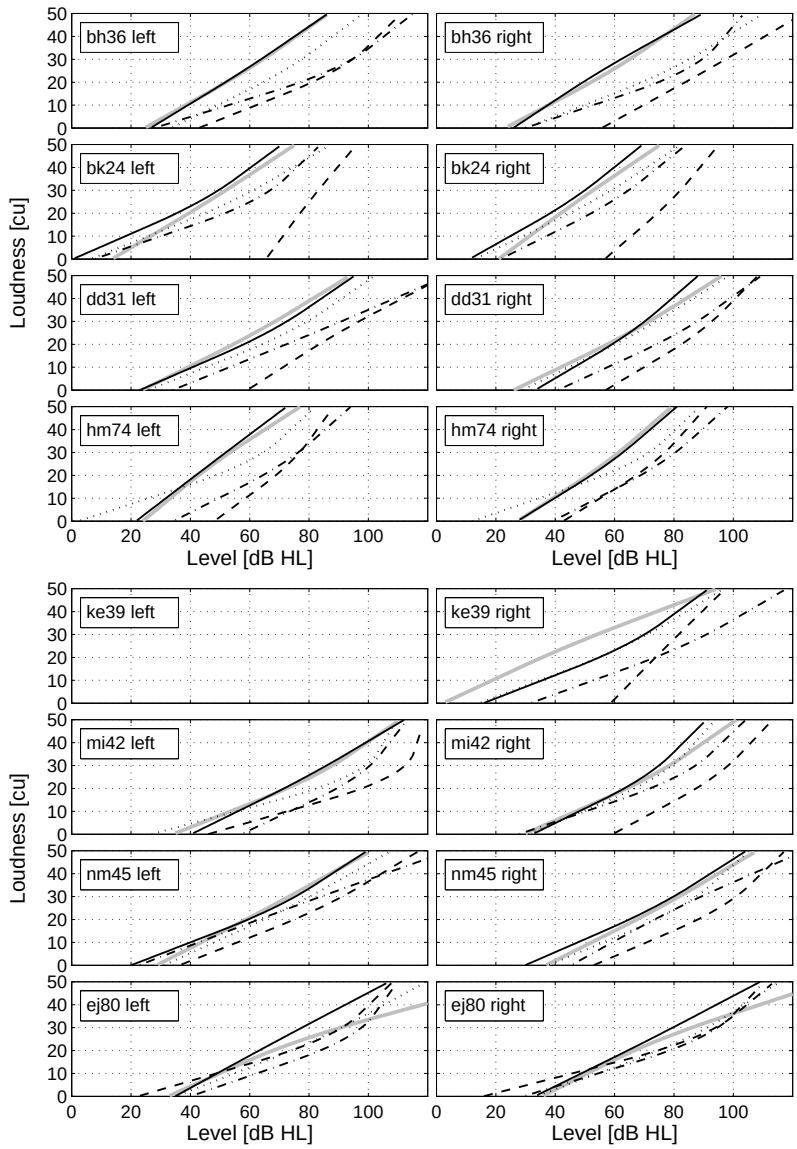

Figure 6.3: *Individual loudness curves of 8 hearing–impaired listeners for different stimuli: 250 Hz narrowband noise (dotted), 1 kHz narrowband noise (dash–dotted), 4 kHz narrowband noise (dashed), CCITT speech–simulating noise (solid black) and speech sample (solid grey).*

less upwardly concave loudness functions in narrowband stimuli than the normal–hearing listeners. That means, the increase in the loudness function slope due to recruitment is most prominent in the lower loudness range. Altogether, the loudness functions of the hearing–impaired listeners show a more linear shape than in the normal–hearing listeners. That means, that at least the level dependency of the loudness summation cannot be found in these subjects.

6.4.2 Cumulative loudness functions

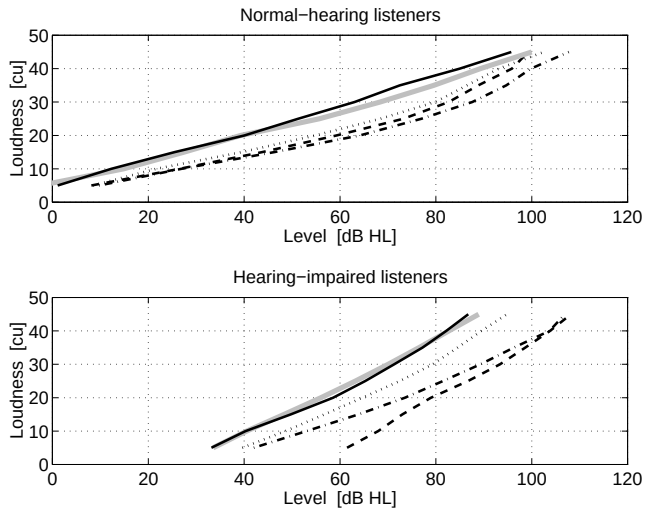

Figure 6.4: *Cumulative loudness functions of 8 normal–hearing listeners (upper panel) and of 8 hearing–impaired listeners (lower panel). 250 Hz narrowband noise (dotted), 1 kHz narrowband noise (dash–dotted), 4 kHz narrowband noise (dashed), CCITT speech–simulating noise (solid black) and speech sample (solid grey).*

Fig. 6.4 shows the cumulative loudness functions across all normal–hearing listeners (upper panel) and across all hearing–impaired listeners (lower panel). For each loudness category the median values of the respective presentation levels (according to the loudness function estimates shown in Figs. 6.2 and 6.3) was calculated across all listeners and ears. In the normal–hearing listeners (upper panel) the cumulative loudness functions of the broadband stimuli are very similar and show a linear shape. The cumulative loudness functions of the narrowband stimuli are upwardly concave. The loudness functions of the 250 Hz and the 4 kHz narrowband noise are very similar. The loudness functions of the 1 kHz narrowband noise, however, is somewhat flatter, i.e. it shows a larger dynamic range. This is consistent with the equal–loudness level contours which also show a larger dynamic range for 1 kHz as compared to 250 Hz and 4 kHz (ISO/R 226, 1961). However,

it has to be considered that the 1 kHz loudness curves were measured in another session 7 months earlier than the other loudness curves. In this session only the 1 kHz stimulus was used for a large number of loudness scaling runs whereas the remaining stimuli were changed between runs. It is possible that adaptation effects – that caused a certain bias – might have occurred in the 1 kHz session.
In the hearing–impaired listeners the cumulative loudness functions are more linear than in the normal–hearing listeners. As in the normal–hearing listeners, the broadband stimuli generated very similar loudness functions. The narrowband loudness functions differ from each other because on average the listeners showed in increasing hearing–loss with increasing frequency.

6.4.3 Level dependency of loudness summation

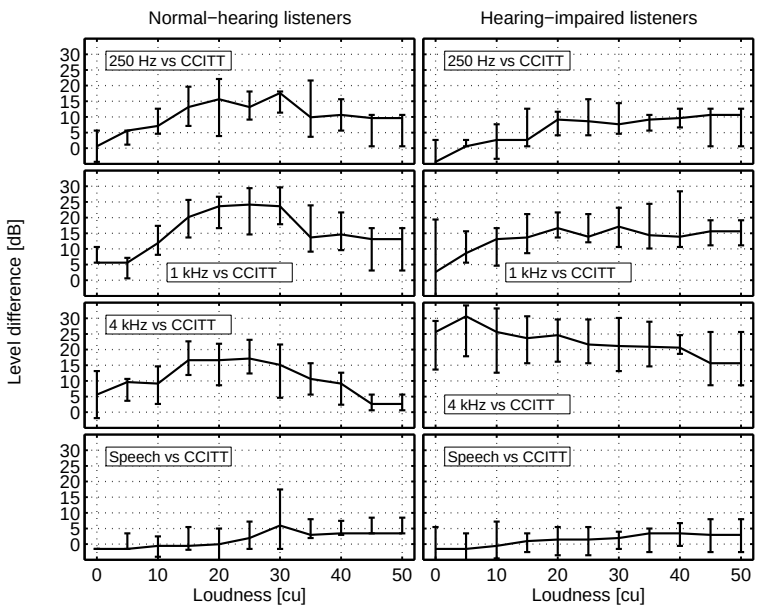

Figure 6.5: *Median level increments (across listeners) and quartile ranges that have to be applied to the different stimuli to produce equal loudness as the CCITT noise. The stimuli are: 250 Hz narrowband noise (upper panels), 1 kHz narrowband noise (2nd row), 4 kHz narrowband noise (3rd row) and speech sample (lower panels). The level increments are given as a functions of categorical loudness. The data of the normal–hearing listeners are shown in the left panels. The data of the hearing–impaired listeners are shown in the right panels.*

Fig. 6.5 shows the median values (across listeners/ears) and the quartile ranges of the level increment by which a certain stimulus (narrowband noise or speech sample, respectively) has to be increased to produce equal loudness as the CCITT noise. The calculation of these median level increments was not based on the fitted loudness functions shown in Figs. 6.2 and 6.3 but on the median levels which were rated with the same loudness category by each listener/ear. In this way, possible biases which might be due to the fitting of the loudness function were excluded. In the normal–hearing listeners (left column), there is a clear spectral loudness summation effect at moderate loudness values, i.e., the levels of all narrowband stimuli have to be increased by approx. 15 dB to produce the same loudness as the CCITT noise. At very low and very high loudness values, the loudness summation effect is much smaller.
The hearing–impaired listeners (right column) show higher level differences for the narrowband stimuli as compared to the equally loud broadband stimuli. This may be due to both loudness summation and hearing loss. However, the effect of hearing impairment seems to dominate this difference for the following reasons: First, the level difference curves between narrowband signals and CCITT noise do not show a maximum at moderate loudness values. This indicates that at least the characteristic level dependency of loudness summation does not occur here. Second, the '4 kHz vs. CCITT noise' curve is increased at most and shows an decreasing tendency with loudness. This can be explained from the fact that the CCITT noise has its maximum spectral energy at 1 kHz and that most hearing–impaired listeners exhibit better hearing thresholds in this frequency range than at 4 kHz. Consequently, the CCITT noise produces more loudness than the 4 kHz stimuli due to audibility, especially at low levels and high frequencies.

The speech sample and the CCITT noise produce almost equal loudness functions in both normal–hearing and hearing–impaired listeners, cf. lower panels in Fig. 6.5. Since these two stimuli have similar spectra, no spectral loudness summation effect was expected here. However, the speech stimulus is much more modulated than the CCITT noise. Obviously, in this case, the modulation has only a small impact on loudness. However, there is a small tendency that the level difference between both stimuli at equal loudness increases with increasing level for both normal–hearing and hearing–impaired listeners. It is not clear whether this is an effect of the modulation. Possibly, the more synthetical CCITT noise gets more annoying at high levels than the more natural speech sample which may influence the loudness ratings of many listeners.

6.5 DISCUSSION

In this study, most normal–hearing listeners showed upwardly concave loudness functions for narrowband stimuli and more linear loudness functions in broadband stimuli. However, there were large differences in the individual shapes of the loudness function as

some normal–hearing listeners show linear loudness functions in narrowband stimuli and downwardly concave loudness functions in broadband stimuli. In loudness comparison experiments, these interindividual differences do not occur because the listeners adjust the level of the stimulus to match the loudness of the standard stimulus. This makes loudness comparison a more direct method than categorical loudness scaling to measure the amount of loudness summation. However, the amount of loudness summation derived from the categorical loudness scaling data of this study (cf. Fig. 6.5) are on average very consistent with loudness comparison data presented by, e.g., Zwicker and Fastl (1990, Fig. 16.24a,b) and Scharf (1961): The spectral loudness summation has an amount of approximately 15 dB at medium levels. At very low and very high levels, the amount of spectral loudness summation is nearly zero.
Most studies which employed categorical loudness scaling found a level dependent loudness summation effect as well (e.g., Launer *et al.*, 1996; Cox *et al.*, 1997; Appell and Hohmann, 1998). The study of Ricketts and Bentler (1996), however, contradicts all other studies including the current study. They used passband signals (150–1500 Hz and 1500–6500 Hz) of a temporally inverted sentence as broadband stimuli. The moderate bandwidths of these signals do probably not generate the full loudness summation effect which is typical for regular speech. Nevertheless, Ricketts and Bentler concluded from their results that level–dependent corrections of loudness estimates measured with narrowband stimuli are not necessary. A consequence of this proposal is that hearing aids that aim at restoring an increased loudness growth to normal have to apply the same gain functions to narrowband and to broadband signals (apart from a level–independent correction factor). The current study, however, gives a clearly different result:

If the loudness function F_{imp} of a hearing–impaired listener is known, the input/output function of an hearing aid that aims at restoring this loudness function to the normal loudness function F_{norm} is given by $out = F_{\text{imp}}^{-1}(F_{\text{norm}}(in))$ with in denoting the level of the input signal and out denoting the level of the output signal. As an example, Fig. 6.6 shows the loudness functions of the 4 kHz stimulus (upper left panel) and the speech stimulus (upper right panel) for the hearing–impaired subject dd31 (right ear) and the normal–hearing subject dc69 (left ear). The lower panels of Fig. 6.6 show the resulting gain functions. The gain functions differ clearly between the two stimuli. The gain function for the 4 kHz stimulus shows a compressive characteristic with a ratio of 2.4 : 1 in the input level range from 0 to 85 dB SPL and a linear amplification of 7 dB beyond 85 dB SPL.[5] The gain function for the speech sample, however, shows a

[5] In a real hearing aid, one would introduce a compression/limiting stage for very high input levels (e.g. above about 100 dB SPL) in order to protect the listener against dangerous sounds. Furthermore, one would introduce a noise gate or expanding stage, respectively, for very low levels (e.g. below about 30 dB SPL) because of technical limitations and because it is not adequate to make each sound audible to a hearing–impaired listener. With these modifications the input/output characteristic shown in the lower left panel of Fig. 6.6 is equivalent to the input/output characteristics proposed by Barfod (1978) and Killion (1993).

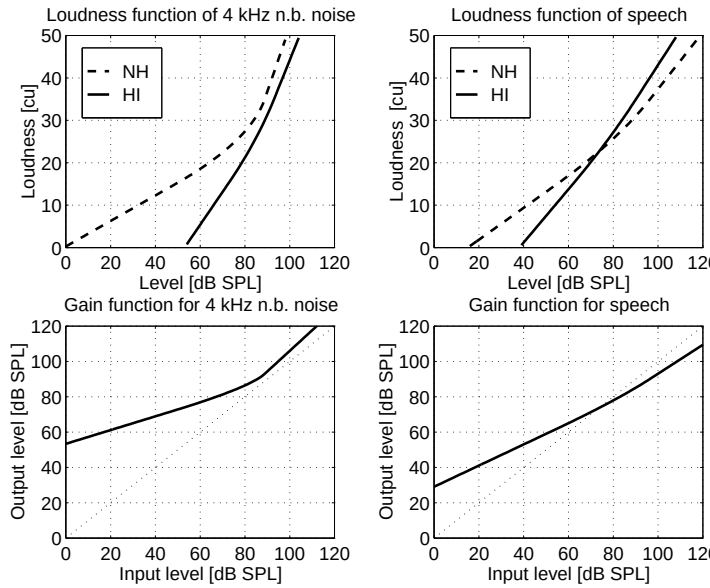

Figure 6.6: *Examples for hearing aid gain functions: The upper left panel shows the loudness functions for the 1 kHz narrowband stimulus for the normal–hearing listener dc69 (left ear) (dashed line) and the hearing–impaired listener dd31 (right ear) (solid line). The hearing aid gain function that results if the loudness perception of the hearing–impaired listener should be restored to that of the normal–hearing listener is shown in the lower left panel. The upper right panel shows the loudness functions of the speech stimulus for the same subjects. The lower rigth panel shows the resulting gain function for the speech stimulus.*

compression ratio of about 1.5 : 1 over the full dynamic range. Note, that this example is not representative for all listeners. Since loudness functions differ between listeners, completely different input/output characteristics result for different hearing–impaired listeners.

From these findings it can be concluded that an automatic gain control system of a hearing aid has to consider the frequency spectrum of the input signal if it aims at restoring the loudness of both narrowband and broadband signals. This has to be done in two ways: 1) The automatic gain control of the hearing aid has to be fitted separately for several frequency bands in order to compensate for loudness recruitment that differs across frequency. 2) Since the bandwidth of the input signal influences the loudness, the different frequency channels have to interact with each other. Certainly, the degree of interaction between channels has to be increased with increasing number of channels: Since narrow frequency channels consist of only a very small number of critical bands,

there is virtually no within–channel loudness summation. Consequently, the amount of between–channel gain correction that accounts for loudness summation has to be larger when a broadband signal is processed by a large number of narrowband channels in parallel.

While many modern multichannel hearing aids fulfill item 1), i.e., they offer several frequency channels that can be fitted separately, item 2) requires some additional research. Since it is not possible to measure the loudness function of every possible input signal, the hearing aid has to apply a loudness model that is able to predict the loudness of arbitrary signals. Such loudness models have been proposed by several authors, e.g. Zwicker and Scharf (1965), Launer (1996), Moore and Glasberg (1996) and Marzinzik *et al.* (1996). However, these models have some limitations because they sometimes produce insufficient estimates of the loudness of certain stimuli. Furthermore, the fitting of these models to the individual loudness functions of hearing–impaired listeners is problematic because all of these loudness models use the sone–scale whereas loudness functions of hearing–impaired listeners are mostly measured using categorical loudness scaling (cu–scale) in clinical diagnostics. Unfortunately, the transformation between sone–scale and cu–scale is not trivial (Hohmann, 1993; Blum *et al.*, 1998).

Additional to the spectral loudness summation effects investigated in this study, an advanced loudness model has to consider temporal loudness summation effects as well. Garnier *et al.* (1999) showed that temporal loudness summation in normal–hearing and hearing–impaired listeners can be measured using categorical loudness scaling as well. Furthermore, it has to be noted that loudness restoration is not the only goal in hearing aid fitting. As a hearing aid that yields an optimal loudness restoration usually does not yield an optimal speech intelligibility, hearing aid fitting always means a compromise.

A further problem in determining the optimal gain functions of hearing aids is the fact that there are large interindividual differences in the loudness functions across normal–hearing listeners. Some listeners show different loudness functions for narrowband and broadband signals and a clear loudness summation effect. Others show equal loudness function for stimuli with different bandwidths, i.e., no loudness summation effect. Unfortunately, most studies about loudness summation cited so far do not present individual data. Therefore, it remains unclear to us if normal–hearing listeners without spectral loudness summation do occur in categorical loudness scaling measurements only or if they also occur in other measuring procedures like loudness comparison or magnitude estimation. Some studies, however, that presented individual data, found large interindividual differences: Boone (1972) measured spectral loudness summation using a loudness production method and found an unacceptable spread in the results of the test subjects. Boone concluded, that the increase of bandwidth calls forth qualitatively different reactions, as far as loudness is concerned, in different tests subjects. Verhey (1989) found interindividual differences up to 10 dB in spectral loudness summation in

normal–hearing listeners using a loudness comparison method with a fixed 55 dB reference stimulus.
Because of the large interindividual differences in normal–hearing listeners, the target loudness function which should be applied to the hearing–impaired individual is not clear. One possibility to solve this problem is to use the cumulative normal–hearing loudness functions as shown in Fig. 6.4. However, this might cause inadequate fits in hearing–impaired listeners whose individual 'normal' loudness functions deviate strongly from the cumulative loudness function. Another approach is to estimate the individual target function based on loudness scaling data obtained from stimuli which are perceived (nearly) normally by the hearing–impaired listener. As it can be seen in the individual loudness functions of narrowband stimuli in normal–hearing listeners shown in Fig. 6.2, the loudness functions across different center frequencies as well as across the two ears do hardly differ. That means, that if a hearing–impaired listener exhibits (nearly) normal hearing in a certain frequency range or at one ear, his/her individual target loudness function might be estimated by loudness measurements in this frequency range or at the normal–hearing ear.

In some normal–listeners the fitted loudness functions indicate extremely low hearing thresholds (cf. subjects ga65, ke65, ts57 and vg71) which are probably due to a negative bias in loudness level estimates at low loudness values. (This bias of the constant stimuli version of the Oldenburg loudness scaling procedure has been discussed in Chap. 5.) The bias had no influence on the examination of loudness summation because the cumulative loudness summation data shown in Fig. 6.5 were based on the median values of the loudness level estimates of each listener/ear instead of the fitted loudness functions shown in Figs. 6.2 and 6.3.
However, the use of a model loudness function was necessary in the individual data because it stabilized the estimate: Although 5 repeated loudness scaling runs were performed per subject/ear, the individual medians of the loudness level estimates themselves showed too much statistical noise to see any systematic trends. The fitted model loudness function, however, showed clear differences between the loudness functions of different stimuli which are consistent with spectral loudness summation effects described in literature. These effects are even more visible for averaged data across listeners.

6.6 SUMMARY AND CONCLUSIONS

A clear spectral loudness summation effect can be observed in many normal–hearing and some hearing–impaired listeners using a categorical loudness scaling procedure with stimuli typically used in audiometry (i.e. 1/3–octave bands of noise, broadband noises and speech samples). This loudness summation effect is most prominent at medium levels and less prominent at very low and high levels. As a consequence, the loudness

functions of normal–hearing listeners are usually upwardly concave for narrowband stimuli and more linear for broadband stimuli. This has an impact on hearing aid fitting and loudness restoration strategies for hearing aids. If a hearing aid is fitted based on categorical loudness scaling data of narrowband stimuli at different center frequencies, loudness summation effects have to be taken into account, especially if many different frequency channels are involved.

ACKNOWLEDGEMENTS

We would like to thank Birgitta Gabriel, Kerstin Sommer and Anita Gorges for performing the measurements.

This study was supported by BMBF 01VJ9305.

Chapter 7

Summary and conclusion

Several approaches to analyze and improve psychophysical measurement methods in audiology were introduced in this thesis. A special feature was that simulations and actual measurements with humans were interactively used to derive and test new adaptive procedures.

The first part of this thesis is concerned with measurements of speech intelligibility using sentence tests. The influence of the stimulus level placement on the accuracy of psychometric function estimates was calculated on the basis of the binomial theory. In contrast to other studies (e.g., Wetherhill, 1963; Levitt, 1971; O'Regan and Humbert, 1989), this calculation was performed not only for two presentation levels placed symmetrically around the midpoint of the psychometric function but also for asymmetrically placed stimulus levels. Two different approaches to adaptive procedures for sentence intelligibility tests were proposed. The adaptive procedures converge either at the optimal target for efficient threshold estimates (*sweetpoint*) or at the optimal targets for concurrent threshold and slope estimates (*pair of compromise*). The first approach is a generalization of an adaptive procedure proposed by Hagerman and Kinnefors (1995) which calculates the presentation level of the next trial according to the response of the previous sentence. The second approach is a modification of the transformed up/down procedures as described by Levitt (1971). Both approaches utilize the fact that in each sentence trial more than one Bernoulli trial is performed. The number of statistically independent perceived elements per sentence is given by the j factor according to Boothroyd and Nittrouer (1988). For $j \geq 2$, the procedures converged considerably faster on their targets than usual adaptive procedures in psychophysical tasks with only one Bernoulli trial per trial. Furthermore, they produce reliable discrimination function estimates using much smaller track lengths. The approach according to Hagerman and Kinnefors (1995) gave better results in Monte–Carlo simulations than the approach according to Levitt (1971).

The step sizes of both approaches were optimized using Monte–Carlo simulations in order to make the different adaptive strategies comparable. The optimal final step sizes found in these simulations are considerably smaller than those used of other sentence tests.
It is recommended to use at least 20 sentence trials and an adaptive procedure which converges at a discrimination value of $p = 0.5$ with decreasing step size to obtain reliable, bias–free SRT estimates with an intraindividual standard deviation of less than 1 dB. A different adaptive procedure should be used to assess SRT and discrimination function slope concurrently. This procedure converges at the discrimination values $p_1 = 0.2$ and $p_2 = 0.8$ in randomly interleaved order and yields a relative standard deviation of slope estimates of about 30 % and a relative bias of about 10 % within 30 sentence trials.
The Göttingen and the Oldenburg sentence test were used (Kollmeier and Wesselkamp, 1997; Wagener *et al.*, 1999c) to evaluate the adaptive procedures. The Göttingen sentence test yielded the same accuracy in the measurements as predicted by the simulations. Because of its higher j factor, the Oldenburg sentence test was expected to yield a higher accuracy compared to the Göttingen sentence test. However, the Oldenburg sentence test produced less accurate results than predicted on the basis of measurements with constant presentation levels. This was especially the case in procedure A2 which converges interleaved at two different target intelligibilities. This is probably due to the fact that the j factor decreases when the presentation level varies from trial to trial due to the adaptive level placement.
Taken together, speech reception thresholds and discrimination function slopes can be estimated using the adaptive procedures for sentence intelligibility tests proposed here much more efficient than using conventional psychophysical methods. Unfortunately, clinical diagnostics require measurement times that are much shorter than those typical for psychophysics. Therefore, it remains unclear whether discrimination function slopes will be assessed in future audiological standard diagnostics, particularly because slope values are highly negatively correlated with threshold values. Hence, they might not bear additional diagnostic meaning. However, the discrimination function slope is an important parameter in audiological research. It measures the increased spread of intelligibility differences across speech elements encountered in hearing–impaired listeners. At least 30 sentence trials are necessary in measurements which require very accurate estimates of speech reception thresholds. The slope estimates is an useful add–on information in such track lengths.

The second part of this thesis dealt with categorical loudness scaling.
A simple statistical model was developed which enables to predict the accuracy of arbitrary categorical loudness scaling procedures using Monte-Carlo simulations. The model requires the knowledge of the subject's loudness function and of his/her response statistics, i.e. the reproducibility of the loudness ratings. The underlying response statistics were derived from test–retest measurements by Hohmann *et al.* (1993). This response

statistics can be characterized by a standard deviation of categorical loudness ratings which is equal to 4 cu (categorical units) in normal–hearing listeners. In the hearing–impaired subjects, this standard deviation ranges between 3 and 10 cu. The simulations gave the following results:
When the incidence of outliers in the subject's responses exceeds 7 %, the robust Lorentzian fit is favorable as compared to the Gaussian fit (least-squares fit).
The standard deviations of the estimates of the loudness function parameters L_{25} (level corresponding to the loudness category 'medium') and m (slope of the loudness function) decrease proportional to $\frac{1}{\sqrt{n}}$, with n denoting the track length. The actual standard deviations of the estimates depend on the respective reproducibility of the loudness ratings of the subject. It is recommended to use at least 14 trials to yield a reliable loudness function estimate. This track length yields intraindividual standard deviations of about 2 cu divided by m in L_{25} estimates and a relative intraindividual standard deviation of about 0.1 in m estimates in normal–hearing and many–hearing impaired listeners. However, in subjects with response statistics which are more variable than normal, the standard deviation of L_{25} estimates increases up to 5 cu divided by m and the standard deviation of m estimates increases up to 0.3 relative to m.
At least 10 response alternatives are necessary to avoid losses in accuracy. Otherwise, standard deviations increase and in subjects with outliers in their responses the m estimates are strongly biased positively. In some subjects, the accuracy can be further improved if the number of response alternatives is increased up to 15. A further increase of the number of response alternatives generates no further improvement of accuracy.

The statistical model of categorical loudness scaling was also used to optimize and to evaluate the Oldenburg–ACALOS (Adaptive CAtegorical LOudness Scaling) procedure which was introduced in this thesis. This procedure better approximates the desired uniform distribution of stimulus levels than the constant stimuli version of the Oldenburg loudness scaling procedure, i.e., it covers the full auditory dynamic range of the subject with a minimum number of presentations outside this range. When combined with the model loudness function proposed in this thesis, the adaptive version yields a considerably higher efficiency in loudness function estimates than other procedures (Pascoe, 1978; Heller, 1985; Allen *et al.*, 1990; Elberling and Nielsen, 1993; Hohmann and Kollmeier, 1995b; Ricketts and Bentler, 1996; Cox *et al.*, 1997; Rasmussen *et al.*, 1998; Keidser *et al.*, 1999). With a total number of about 20 trials the level of each loudness category was estimated with an intraindividual standard deviation of less than 5 dB.

For both the adaptive and the constant stimuli version of the Oldenburg loudness scaling procedure, a non–linear model function yielded the closest approximation to the individual reference loudness functions considered over the whole loudness range in both normal–hearing and hearing–impaired listeners. This model function consists of two straight lines with independent slopes which are connected at the loudness category

25 ('medium'). The transition area is smoothed by a Bezier interpolation between the categories 15 ('soft') and 35 ('loud'). This model function parameterizes not only hearing threshold and slope of the loudness function but also the bending of the loudness function. It allows for a more accurate assessment of the shape of loudness functions in individuals than the previously used target functions.

This fact was utilized in a study that investigated the influence of the stimulus bandwidth on the shape of the loudness function. Most normal–hearing listeners show upwardly concave loudness functions in narrowband stimuli and more linear loudness functions in broadband signals. In many hearing–impaired listeners, loudness function of both narrowband and broadband signals show a more linear shape. These findings are consistent with loudness matching measurements (e.g., Zwicker *et al.*, 1957) which showed that spectral loudness summation is most prominent at moderate levels and less prominent at very low and high levels. For hearing aids that aim at restoring the loudness perception of hearing–impaired listeners to normal, these spectral loudness summation effects have to be considered, i.e. the input/output characteristic of the hearing aid has to depend on the bandwidth of the input signal. This requires the use of a loudness model which is able to calculate the loudness of arbitrary sounds perceived by the listener. The modern digital signal processing technologies which are more and more applied in hearing aids will enable to implement such loudness models in the near future. Therefore, the open questions concerning loudness models and their fitting to individual data should be addressed by basic audiological research.

In summary, the procedures developed in this thesis appear to yield the highest practically achievable precision and validity in a given amount of time. Since they cover important, until now overlooked auditory functions in normal–hearing and hearing–impaired listeners, they might be of clinical significance in the future.

Appendix A

A.1 Bezier smoothing

The Bezier interpolation of model function (5.12) between the categorical loudness values 15 and 35 cu is given by

$$\mathrm{bez}(L, L_{\mathrm{cut}}, L_{15}, L_{35}) = y_0 + y_1 \cdot t + y_2 \cdot t^2 \tag{A.1}$$

with

$$t(L) = \begin{cases} -\frac{x_1}{2x_2} - \sqrt{\frac{L - x_0}{x_2} + \frac{x_1^2}{4x_2^2}} & \text{for} \quad \frac{y_k - y_a}{x_k - x_a} < \frac{y_k - y_b}{x_k - x_b} \\ -\frac{x_1}{2x_2} + \sqrt{\frac{L - x_0}{x_2} + \frac{x_1^2}{4x_2^2}} & \text{for} \quad \frac{y_k - y_a}{x_k - x_a} > \frac{y_k - y_b}{x_k - x_b} \end{cases}$$

and with

$$\begin{array}{llll} x_0 = L_{15} & , & y_0 = 15, \\ x_1 = 2 \cdot L_{\mathrm{cut}} - 2 \cdot L_{15} & , & y_1 = 2 \cdot 25 - 2 \cdot 15, \\ x_2 = L_{15} - 2 \cdot L_{\mathrm{cut}} + L_{35} & \text{and} & y_2 = 15 - 2 \cdot 25 + 35. \end{array}$$

The inverse of Eq. (A.1) is given by:

$$\mathrm{bez}^{-1}(F, L_{\mathrm{cut}}, L_{15}, L_{35}) = x_2 \cdot \left(t + \frac{x_1}{2x_2} \right)^2 - \frac{x_1^2}{4x_2^2} + x_0 \tag{A.2}$$

with

$$t = \frac{y}{y_1} - \frac{y_0}{y_1}$$

and with x_0, x_1, x_2, y_0, y_1 and y_2 as defined above.

Eq. A.2 can be used to calculate, e.g., the L_{25} parameter.
($L_{25} = \mathrm{bez}^{-1}(25, L_{\mathrm{cut}}, L_{15}, L_{35})$.)

A.2 Instruction

German version

Bei diesem Experiment bitten wir Sie, die Lautstärke von Geräuschen zu beurteilen.

Ihnen werden über Kopfhörer Geräusche vorgespielt. Nach jedem Geräusch erscheint eine Skala auf der Antwortbox. Bitte geben Sie auf dieser Skala an, wie laut Sie das Geräusch empfunden haben. Ihnen stehen die Antworten "unhörbar", "sehr leise", "leise", "mittel", "laut", "sehr laut" und "zu laut" zur Verfügung. Sie können auch die als schwarze Balken dargestellten Zwischenstufen wählen. Bitte geben Sie Ihre Antwort durch Berühren der entsprechenden Taste. Wenn Sie das Geräusch z.B. als "laut" empfunden haben, drücken Sie bitte auf das Wort "laut". Wenn Sie das Geräusch z.B. zwischen "sehr leise" und "leise" empfunden haben, drücken Sie bitte auf den Balken zwischen "sehr leise" und "leise". Sollte die Antwortskala erscheinen, ohne daß Sie etwas gehört haben, berühren Sie bitte das Feld "unhörbar".

Es gibt keine richtigen und falschen Antworten. Entscheidend ist nur, wie laut Sie das Geräusch empfunden haben.

Erst wenn Sie Ihre Antwort gegeben haben, verschwindet die Skala und es wird Ihnen ein weiteres Geräusch dargeboten.

Haben Sie noch Fragen?

English version

In this experiment, we ask you for judging the loudness of sounds.

You will be presented with some sounds via headphones. After each sound presentation, a response scale will appear on the response box. Please indicate how loud you have perceived the sound on this scale after each presentation. You have the response alternatives 'inaudible', 'very soft', 'medium', 'loud', 'very loud' and 'too loud'. You can even choose the intermediate steps indicated by black bars.
Please indicate your response by touching the corresponding field. If you have perceived the sound, e.g. as 'loud', please touch the word 'loud'. If you have perceived the sound, e.g. in between 'very soft' and 'soft', please touch the bar between 'very soft' and 'soft'. If the response scale appears and you have not heard anything, please touch the field 'inaudible'.

There are no right or wrong answers. All that matters is how you perceived the sound.

Not before you have given your response, the response scale will disappear and the next sound will be presented.

Do you have any questions?

References

Albani, S., Brand, T., Gabriel, B., Hohmann, V. and Kollmeier, B. (**1997**). Referenzdaten der Oldenburger Hörflächenskalierung. In B. Kollmeier (Ed.), *Hörflächenskalierung, Buchreihe Audiologische Akustik*, pp. 18–22. Heidelberg: median–verlag.

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**:745–753.

Appell, J. and Hohmann, V. (**1998**). Messungen der Lautheitssummation bei Normal– und Schwerhörenden. In A. Sill (Ed.), *Fortschritte der Akustik – DAGA 98*, pp. 306–307, DEGA e. V. Oldenburg.

Barfod, J. (**1978**). Multichannel compression hearing aids: Experiments and considerations on clinical applicability. *Scand. Audiol. Suppl.*, **6**:315–340.

Berger, K. W., Hagberg, E. N. and Rane, R. L. (**1988**). *Prescription of hearing aids: Rationale, procedure and results.* (5th ed.). Herald Kent.

Blum, R. (**1998**). *Experimente und Modellvorstellungen zur Lautheitswahrnehmung bei Normal- und Schwerhörenden*, (Master's thesis). Universität Oldenburg.

Blum, R., Hohmann, V., Dau, T. and Kollmeier, B. (**1998**). Vergleich verschiedener Meßmethoden zur Lautheit. In A. Sill (Ed.), *Fortschritte der Akustik – DAGA 98*, pp. 302–303, DEGA e. V. Oldenburg.

Boone, M. M. (**1972**). Loudness measurements on pure tone and broad band impulsive sounds. *Acustica*, **29**:198–204.

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

Boretzki, M., Heller, O., Knoblach, W., Fichtl, E., Stock, A. and Optitz, M. (**1994**). Untersuchungen zur Reliabilität und Sensitivität der Hörfeldaudiometrie. In *Fortschritte der Akustik – DAGA 94*, DPG Kongreßgesellschaft mbH. Bad Honnef.

Brand, T., Hohmann, V. and Kollmeier, B. (**1997a**). Adaptive categorical loudness scaling. In A. Schick and M. Klatte (Eds.), *Contributions to Psychological Acoustics – 7th Oldenburg Symposium on Psychological Acoustics*, University of Oldenburg. Oldenburg.

Brand, T., Hohmann, V. and Kollmeier, B. (**1997b**). Die adaptive Hörflächenskalierung. In B. Kollmeier (Ed.), *Hörflächenskalierung, Buchreihe Audiologische Akustik*, pp. 146–166. Heidelberg: median–verlag.

Brand, T., Hohmann, V. and Kollmeier, B. (**1997c**). Meßgenauigkeit der (adaptiven) kategorialen Hörflächenskalierung. In *Fortschritte der Akustik – DAGA 97*, DEGA e. V. Oldenburg.

Brand, T., Hohmann, V. and Kollmeier, B. (**1997d**). Wie genau ist die kategoriale Lautheitsskalierung. In B. Kollmeier (Ed.), *Hörflächenskalierung, Buchreihe Audiologische Akustik*, pp. 121–145. Heidelberg: median–verlag.

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

Bronkhorst, A.W. and Plomp, R. (**1988**). The effect of head–induced interaural time and level differences on speech intelligibility in noise. *J. Acoust. Soc. Am.*, **83**:1508–1516.

Bronkhorst, A.W. and Plomp, R. (**1989**). Binaural speech intelligibility in noise for hearing–impaired listeners. *J. Acoust. Soc. Am.*, **86**:1374–1383.

Buus, S., Florentine, M. and T., Poulsen (**1999**). Temporal integration of loudness in listeners with hearing losses of primarily cochlear origin. *J. Acoust. Soc. Am.*, **105**:3464–3480.

Byrne, D. and Dillon, H. (**1986**). The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid. *Ear & Hearing*, **7**:257–265.

Campbell, R. A. (**1974**). Modifications to the BUDTIF procedure. *J. Acoust. Soc. Am. Suppl. 1*, **56**:S57.

CCITT G.227 (**1964**). **C**omitée **C**onsultatif **I**nternationale de **T**élégraphique et **T**éléphonique (CCITT).

Cornelisse, L. E., Seewald, R. C. and Jamieson, D. G. (**1995**). The input/output formula: A theoretical approach to the fitting of personal amplification devices. *J. Acoust. Soc. Am.*, **93**:1854–1864.

Cox, R. M., Alexander, G. C., Taylor, I. M. and Gray, G. A. (**1997**). The contour test of loudness perception. *Ear & Hearing*, **18**(5):389–400.

Elberling, C. and Nielsen, C. (**1993**). The dynamics of speech and the auditory dynamic range in sensorineural hearing impairment. In J. Beilin and G.R. Jensen (Eds.), *Recent developments in hearing instrument technology – 15th Danavox Symposium*, pp. 99–133.

Fechner, G. Th. (**1888**). Über die psychischen Maßprinzipien und das Webersche Gesetz. In W. Wundt (Ed.), *Philosophische Studien. Vol. 4*. Leipzig.

Gabriel, B. (**1996**). *Equal-loudness Level Contours: Procedures, Factors and Models*, (Ph.D. thesis). Universität Oldenburg.

Garner, W. R. (**1954**). Context effects and the validity of loudness scales. *J. Exp. Psychol.*, **48**:218–224.

Garnier, S., Micheyl, C., Berger-Vachon, C. and Collet, L. (**1999**). Effect of signal duration on categorical loudness scaling in normal and in hearing–impaired listeners. *Audiology*, **38**:196–201.

Geller, D. and Margiolis, R. H. (**1984**). Magnitude estimation of loudness I: Application to hearing aid selection. *J. Speech. Hear. Res.*, **27**:20–27.

Green, D. M. (**1989**). Stimulus step size and heterogeneous stimulus conditions in adaptive psychophysics. *J. Acoust. Soc. Am.*, **86**:629–636.

Green, D. M. (**1990**). Stimulus selection in adaptive psychophysical procedures. *J. Acoust. Soc. Am.*, **87**:2662–2674.

Green, D. M. and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics*. Wiley, New York. reprinted in 1988 by Peninsula Publishing, Los Altos, CA.

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

Hagerman, B. (**1996**). Personal communication.

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

Hall, J. L. (**1981**). Hybrid adaptive procedure for estimation of psychometric functions. *J. Acoust. Soc. Am.*, **69**:1763–1769.

Hellbrück, J. and Moser, L. M. (**1985**). Hörgeräte–Audiometrie: Ein computerunterstütztes psychologisches Verfahren zur Hörgeräteanpassung. *Psycholog. Beiträge*, **27**:494–508.

Heller, O. (**1985**). Hörfeldaudiometrie mit dem Verfahren der Kategorienunterteilung (KU). *Psycholog. Beiträge*, **27**:478–493.

Heller, O. (**1990**). Scaling and orientation. In F. Müller (Ed.), *Fechner Day 90, Proc. of the 6th annual meeting of the International Society of Psychophysics. Würzburg*, pp. 52–57.

Heller, O. and Boretzki, M. (**1998**). personal communication.

Heller, O., Boretzki, M., Fichtl, E., Knoblach, W., May, B., Nowak, T. and Stock, A. (**1997**). *Projektbereich: Hilfen für Hörgeschädigte.* Psychologisches Institut der Universität Würzburg. Abschlußbericht.

Hellman, R. P. and Zwislocki, J. (**1963**). Monaural loudness function at 1000 cps and interaural summation. *J. Acoust. Soc. Am.*, **35**:856–865.

Hellmann, R. P. and Meiselman, C. H. (**1993**). Rate of loudness growth for pure tones in normal and impaired hearing. *J. Acoust. Soc. Am.*, **93**:966–975.

Hohmann, V. and Kollmeier, B. (**1995a**). The effect of multichannel dynamic compression on speech intelligibility. *J. Acoust. Soc. Am.*, **97**(2):1191–1195.

Hohmann, V. and Kollmeier, B. (**1995b**). Weiterentwicklung und klinischer Einsatz der Hörfeldskalierung. *Audiologische Akustik*, **34**:48–59.

Hohmann, Volker (**1993**). *Dynamikkompression für Hörgeräte – Psychoakustische Grundlagen und Algorithmen*, (Ph.D. thesis). Universität Göttingen.

Hübner, R. and Ellermeier, W. (**1993**). Additivity of loudness across critical bands: A critical test. *Perception & Psychophysics*, **54**:185–189.

ISO 389 (**1991**). *Acoustics-Standard reference zero for the calibration of pure-tone air conducting audiometers.* International Organization for Standardization, Geneva.

ISO/R 226 (**1961**). *Normal equal–loudness contours for pure tones and normal threshold of hearing under free field listening conditions.* International Organization for Standardization, Geneva.

Jenstad, L. M., Cornelisse, L. E. and Seewald, R. C. (**1997**). Effects of test procedure on individual loudness functions. *Ear & Hearing*, **18**:401–408.

Jesteadt, W. (**1980**). An adaptive procedure for subjective judgements. *Perception & Psychophysics*, **28**:85–88.

Keidser, G., Seymour, J., Dillon, H., Grant, F. and D., Byrne (**1999**). An efficient, adaptive method of measuring loudness growth functions. *Scand. Audiol.*, **28**:3–14.

Kießling, J. (**1995**). Loudness growth in sensorineural hearing loss – consequences for hearing aid design and fitting. *Audiol. Acoust.*, **34**:2, 82–89. (in German).

Kießling, J., Schubert, M. and Archut, A. (**1995**). Adaptive fitting of hearing instruments by category loudness scaling. *Scand. Audiol.*, **32**:153–160.

Kießling, J., Steffens, T. and Wagner, I. (**1993**). On the clinical application of loudness scaling. *Audiol. Acoust.*, **32**(4):100–115. (in German).

Killion, M. C. (**1979**). *Design and evaluation of high fidelity hearing aids*, (Ph.D. thesis). Northwestern University. Ann Arbor, MI, University Microfilms.

Killion, M. C. (**1993**). The K–amp hearing aid: An attempt to present high fidelity for the hearing impaired. In J. Beilin and G. R. Jensen (Eds.), *Recent developments in hearing instrument technology – 15th Danavox Symposium*, pp. 167–229.

Killion, M. C. and Fikret-Pasa, S. (**1993**). The three types of sensorineural hearing loss: Loudness and intelligibility considerations. *Hearing Journal*, **46**(4):31–36.

Kollmeier, B. (Ed.) (**1997**). *Hörflächenskalierung – Grundlagen und Anwendung der kategorialen Lautheitsskalierung für Hördiagnostik und Hörgeräteversorgung.* median–verlag.

Kollmeier, B., Gilkey, R. H. and Sieben, U. (**1988**). Adaptive staircase techniques in psychoacoustics: A comparison between theoretical results and empirical data. *J. Acoust. Soc. Am.*, **83**:1852–1862.

Kollmeier, B., Müller, C., Wesselkamp, M. and Kliem, K. (**1992**). Weiterentwicklung des Reimtests nach Sotscheck. In B. Kollmeier (Ed.), *Moderne Verfahren der Sprachaudiometrie, Buchreihe Audiologische Akustik*, pp. 216–237. Heidelberg: median–verlag.

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

Kollmeier, Birger (**1990**). *Meßmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache.* Universität Göttingen. Habilitationsschrift.

Lam, C. F., Mills, J. H. and Dubno, J. R. (**1996**). Placement of observations for the efficient estimation of a psychometric function. *J. Acoust. Soc. Am.*, **99**:3689–3693.

Laming, D. and Marsh, D. (**1988**). Some performance tests of QUEST on measurements of vibrotactile thresholds. *Perception & Psychophysics*, **44**:99–107.

Launer, S. (**1995**). *Loudness perception in Listeners with Sensorineural Hearing Impairment*, (Ph.D. thesis). Universität Oldenburg.

Launer, S., Hohmann, V. and Kollmeier, B. (**1996**). Modeling loudness growth and loudness summation in hearing–impaired listeners. In W. Jestaedt (Ed.), *Modeling sensorineural hearing loss*. Hillsdale: Erlbaum.

Leek, M. R., Hanna, T. E. and Marshall, L. (**1992**). Estimation of psychometric functions from adaptive tracking procedures. *Perception & Psychophysics*, **51**:247–256.

Levitt, H. (**1971**). Transformed up–down methods in psychoacoustics. *J. Acoust. Soc. Am.*, **49**:467–477.

Lunner, T., Hellgren, J., Arlinger, S. and C., Elberling (**1997**). A digital filterbank hearing aid: Three digital signal processing algorithms – user preference and performance. *Ear & Hearing*, **18**:373–387.

Margiolis, R. H. (**1985**). Magnitude estimation of loudness III: Performance of selected hearing aid users. *J. Speech. Hear. Res.*, **28**:411–420.

Marks, L. E. and Warner, E. (**1991**). Slippery context effect and critical bands. *J. Exp. Psychol.*, **17**:986–996.

Marzinzik, M., Hohmann, V., Appell, J. E. and Kollmeier, B. (**1996**). Zur Modellierung der Lautheitswahrnehmung bei Normalhörenden und Innenohrschwerhörigen. *Audiologische Akustik*, **35**:136–144.

McCandless, G. A. and Lyregaard, P. E. (**1983**). Prescription of gain/output (POGO) for hearing aids. *Hear. & Instr.*, **34**(1):16–21.

Montgogomery, H. (**1975**). Direct estimation: Effect of methodological factors on scale type. *Scand. J. Psychol.*, **16**:19–29.

Moore, B. C. J. and Glasberg, B. R. (**1996**). A revision of Zwicker's loudness model. *Acustica – acta acustica*, **82**:335–345.

Moore, B. C. J., Seloover Johnson, J., Clark, T. M. and Pluvinage, V. (**1992**). Evaluation of a dual-channel full dynamic range compression system for people with sensorineural hearing loss. *Ear & Hearing*, **13**(5):349–370.

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

Nowak, T. (**1990**). Loudness scaling based on Fechner's idea. In F. Müller (Ed.), *Proceedings of the sixth Annual Meeting of the International Society for Psychophysics*. Würzburg, Germany: Universität Würzburg.

O'Regan, J. K. and Humbert, R. (**1989**). Estimating psychometric functions in forced–choice situations: Significant biases found in threshold and slope estimations when small samples are used. *Perception & Psychophysics*, **46**:434–442.

Parducci, A. (**1963**). Range–frequency compromise in judgement. *Psychological Monographs*, **77**. (2, Whole No. 565).

Parducci, A. (**1965**). Category judgement: A range–frequency model. *Psychological Review*, **72**:407–418.

Parducci, A. and Perret, L. F. (**1971**). Category rating scales: Effect of relative spacing and frequency of stimulus values. *J. Experi. Psychol. Monograph*, **89**:427–452.

Parducci, A. and Wedell, D. H. (**1989**). The category effect with rating scales: Number of categories, number of stimuli and method of presentation. *J. Experi. Psychol.: Human Perception and Performance*, **12**(4):496–516.

Pascoe, D. P. (**1978**). An approach to hearing aid selection. *Hear. Instrum.*, **29**:12–16.

Peissig, J. and Kollmeier, B. (**1997**). Directivity of binaural noise reduction in spatial multiple noise–source arrangements for normal and impaired listeners. *J. Acoust. Soc. Am.*, **101**:1660–1670.

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

Poulton, E. C. (**1968**). The new psychophysics: Six models for magnitude estimation. *Psychological Bulletin*, **69**:1–19.

Poulton, E. C. (**1989**). *Bias in Quantifying Judgements*. Hillscale, New Jersy: Lawrence Erlbaum.

Rasmussen, A. N., Olsen, S. Ø., Borgkvist, B. V. and Nielsen, L. H. (**1998**). Long-term test-retest reliability of category loudness scaling in normal-hearing subjects using pure-tone stimuli. *Scand. Audiol.*, **27**:161–167.

Reckhardt, C., Mellert, V. and Kollmeier, B. (**1999**). Factors influencing equal–loudness–level contours. *J. Acoust. Soc. Am.*, **105**:1083.

Richter, U. (**1992**). *Kenndaten von Schallwandlern der Audiometrie*. PTB-MA-27.

Ricketts, T.A. and Bentler, R.A. (**1996**). The effect of test signal type and bandwidth on the categorical scaling of loudness. *J. Acoust. Soc. Am.*, **99**:2281–2287.

Robinson, K. and Gatehouse, S. (**1996**). Test–retest reliability of loudness scaling. *Ear & Hearing*, **17**:120–123.

Scharf, B. (**1961**). Loudness summation and spectrum shape. *J. Acoust. Soc. Am.*, **34**:228–233.

Scharf, B. and Hellman, R. P. (**1966**). Model of loudness summation applied to impaired ears. *J. Acoust. Soc. Am.*, **40**:71–78.

Stevens, S. S. (**1956**). The direct estimation of sensory magnitudes – loudness. *Am. J. Psychol.*, **69**:1–25.

Stevens, S. S. (**1957**). On the psychophysical law. *Psychol. Rev.*, **64**:153–181.

Taylor, M. M. and Creelman, C. D. (**1967**). PEST: Efficient estimates on probability functions. *J. Acoust. Soc. Am.*, **41**:782–787.

Teightsoonian, R. (**1973**). Range effects in psychophysical scaling and a review of Stevens' law. *Am. J. Psychol.*, **86**:3–27.

Verhey, J. L. (**1989**). *Psychoacoustics of spectro–temporal effects in masking and loudness perception*, (Ph.D. thesis). Universität Oldenburg.

Wagener, K., Brand, T. and Kollmeier, B. (**1999a**). Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests. *Zeitschrift f. Audiologie*, **38**:86–95.

Wagener, K., Brand, T. and Kollmeier, B. (**1999b**). Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests. *Zeitschrift f. Audiologie*, **38**:44–56.

Wagener, K., Kühnel, V. and Kollmeier, B. (**1999c**). Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. *Zeitschrift f. Audiologie*, **38**:4–15.

Wetherhill, G. B. (**1963**). Sequential estimation of quantal response curves. *J. Roy. Statist. Soc.*, **B25**:1–48.

Zwicker, E. and Fastl (**1990**). *Psychoacoustics – Facts and Models.* Springer.

Zwicker, E., Flottorp, G. and Stevens, S. S. (**1957**). Critical band width in loudness summation. *J. Acoust. Soc. Am.*, **29**:548–557.

Zwicker, E. and Scharf, B. (**1965**). A model of loudness summation. *Psychol. Review*, **72**:3–26.