Chapter 3

Comparison of different psychoacoustic techniques to measure loudness functions

Abstract

In this chapter different techniques used to measure loudness growth functions are described and discussed. Magnitude estimation and categorical scaling are compared in more detail. Specifically, loudness scaling experiments were performed employing magnitude estimation, restricted magnitude estimation and a categorical scale with many categories. The stimulus was a narrowband noise centered at 1 kHz. The results obtained with these three methods are very similar. With all three techniques, the loudness functions obtained exhibit a steeper increase near threshold than at mid and high levels when plotted on a logarithmic scale. This steeper increase may be partly due to the so–called logarithmic response bias. A less curved loudness function is observed for measurements employing a categorical scale with few categories. Since this method exhibits also a practical advantage, it will be used in subsequent chapters.

3.1 Introduction

For modeling sensorineural hearing impairment in an appropriate way, the determination of functions relating subjective loudness to sound pressure level ("loudness function") plays an important role. Beside diagnostic purposes (i.e., estimating the amount of recruitment), these functions also provide a basis for the selection and fitting of hearing aids that employ multichannel compression. The different methods of measuring loudness functions and their different advantages and problems are briefly reviewed in the following sections.

3.1.1 Loudness matching/balancing

The basic idea of loudness balancing/matching technique is to compare the loudness of two different sounds and to adjust one of them in level to produce equal loudness. This technique is often used to measure equal loudness contours in normal–hearing subjects and to investigate the influence of different experimental parameters such as the range of stimulus levels or the frequency separation between the target and reference tones on the judgement of loudness. Usually, loudness comparisons are performed between two alternating stimuli. One of them is fixed in level (reference) while the other is variable in level (target) and has to be adjusted to produce equal loudness. This is repeated for several levels of the reference sound. The measurements are usually performed twice, once with the reference stimulus in the region of normal hearing (or in the normal ear), and once with the target stimulus in this region. Systematic differences are often observed for these two cases, and taking the average is generally held to reduce bias effects.

In hearing–impaired subjects this method can be employed to measure the amount of recruitment in two ways (Brunt, 1994; Miskolczy–Fodor, 1960): (1) For subjects with near–normal hearing at some frequencies, loudness balances can be obtained between a stimulus at a frequency where the absolute threshold is nearly normal with that of a stimulus at a frequency where the absolute threshold is elevated (monaural loudness balance procedure, MLB); (2) For subjects with unilateral losses (which is rare), loudness balances can be obtained between the two ears, using a single frequency (binaural loudness balance procedure, BLB).

There are several problems associated with the loudness balancing method. Bias effects can occur if the experimental parameters are not chosen properly (Poulton, 1989). Several studies suggest that methodological differences can strongly affect the results obtained with the loudness balancing procedure (Poulton, 1989; Suzuki and Sone, 1993; Gabriel et al., 1994). Also the variability of the matches increases as the frequency separation of the reference and test stimuli increases. This effect probably influences the results not only of the MLB procedure, but also those of the BLB procedure, since many impaired subjects suffer from diplacusis (in which a single frequency evokes different pitches in the two ears). Diplacusis sometimes depends on level, decreasing with increasing level (Burns and Turner, 1986). A second parameter that influences loudness balance judgements is the chosen range of levels of the variable stimulus. The finally adjusted level is shifted towards the center of the presented stimulus range of the variable stimulus (Gabriel et al., 1994). The magnitude of this shift increases with increasing spectral separation between test and reference tone. This creates problems in using loudness matching procedures with hearing–impaired subjects, since the appropriate range of levels is not known in advance. Furthermore, loudness matching cannot be used to determine loudness functions in hearing–impaired people with bilateral losses at all frequencies. Unfortunately, this type of hearing loss is very common. Finally, loudness matching does not provide a direct measure of the loudness sensation. However, a method to relate directly the physical magnitudes of stimuli to their subjective loudness was proposed by Stevens (1957). He advocated the application of scaling proce-

dures, specifically, magnitude estimation and magnitude production. These techniques are discussed in the next section.

3.1.2 Magnitude estimation and magnitude production

Stevens assumed that listeners judge the loudness of a stimulus on a "ratio scale", e.g., a given sound may be judged twice or three times as loud as another one. This assumption is the basis for the so–called "sone scale" (Stevens, 1957): the loudness of a 1–kHz sinusoid at 40 dB SPL is defined as one sone. Each 10–dB increase in level results approximately in a doubling of loudness, and hence a doubling of the sone value. In magnitude estimation, a stimulus is presented at various levels, and subjects are required to assign to each stimulus a positive number corresponding to the subjective loudness of this stimulus. In early experiments on magnitude estimation, a reference sound was presented to the subjects and labeled as having a certain loudness, e.g., 100 units. The task was to judge each test sound relative to the reference sound (so–called free magnitude estimation). However, these relative judgements are affected by a variety of parameters. For example, the choice of the fixed number assigned to the reference sound appears to bias the results (Hellman and Zwislocki, 1961; Hellman and Zwislocki, 1963; Hellman and Zwislocki, 1964)). Therefore, nowadays the subjects are asked to assign any positive number to the perceived loudness (so–called absolute magnitude estimation, (Hellman and Meiselman, 1993; Hellman, 1993; Gescheider, 1993)). There is a subtle difference between these two techniques, since free magnitude estimation is based on a ratio scale while absolute magnitude estimation is based on an absolute scale of loudness (Hellbrück, 1993). In magnitude production, subjects are asked to adjust the level of the stimulus until its loudness matches a given number. Magnitude production usually yields a slightly steeper loudness function than magnitude estimation. Both magnitude estimation and production require absolute spontaneity and naivete and therefore work best with completely untrained and unexperienced subjects.
A variation of magnitude estimation/production is cross–modality matching. For example, sounds of various levels may be presented and the listener is asked to adjust the length of a line or the brightness of a light so as to match the strength of the subjective impression. The main shortcoming of this method is that it does not yield a direct measure of subjective loudness, i.e., it does not reveal the actual slope of the loudness function. However, the slope derived from cross–modality matching is consistent with that derived from absolute magnitude estimation/production. Hellman and Meiselman (1988, 1990, 1993) showed that the results of the three measurement techniques are internally consistent, i.e., transitivity holds for these three methods even for impaired listeners.
Hellman and Meiselman (1990, 1993) used (absolute) magnitude estimation and production and cross–modality matching to measure loudness functions of hearing–impaired subjects. They found a significant correlation (0.69, $p < 0.01$) between the slopes of the loudness functions and hearing loss: with increasing threshold the slope of the loudness function increases although the interindividual variability in slope also increases. By dividing their subjects into different groups according to age and background and comparing the results of

different scaling experiments, they excluded age and background of the subjects as factors influencing the slopes.
A fundamental assumption of the scaling techniques described, is that humans scale loudness on a ratio scale. One way to overcome these problems related with scaling techniques using a reference stimulus (i.e., a ratio scale) was presented above: the absolute magnitude estimation. Another one, which employs a different scale than numbers for measuring subjective loudness, is discussed in the next section.

3.1.3 Categorical Loudness Scaling

Several researchers (Pascoe, 1978; Heller, 1985) have proposed the measurement of loudness functions using a categorical scale rather than a ratio scale. This scale is based on the assumption that listeners subdivide the dynamic range using verbal categories such as "soft", "comfortable" or "loud". This method is quite often used as a diagnostic tool in audiology and as a tool for the fitting of compression hearing aids (Kollmeier and Hohmann, 1995; Kießling et al., 1994; Hellbrück, 1993; Kießling et al., 1993; Moore et al., 1992; Allen et al., 1990; Pluvinage, 1989).
In this method, stimuli are presented at different frequencies and levels and the task of the listener is to scale loudness using verbal categories like ("not audible") "very soft", "soft", "intermediate" (sometimes called "ok"), "loud", "very loud" ("too loud"). Basically two different scaling methods are used which differ mainly in the fineness of the underlying scale. Heller (1985) and Hellbrück and Moser (1985) proposed a two–step procedure in which two successive judgements of the same stimulus have to be carried out. For the first judgement, all possible verbal categories are presented as response alternatives to the listener, while for the second a fine scale using numbers around the previously chosen category is presented. In the one–step procedure, the stimulus is judged using only the set categories (Moore et al., 1992; Allen et al., 1990). Some researchers allow subjects to make just one response, but on a finer scale using intermediate values between the verbal categories (Hohmann, 1993; Kießling et al., 1993; Kießling et al., 1994; Launer et al., 1994; Kollmeier and Hohmann, 1995) yielding a scale consisting of 10 categories. Hohmann (1993) showed that there is only a small difference in the variability of measured loudness functions between the two–step procedure and the one–step procedure using intermediate values. This scale, shown in Fig. 5.2, has been applied in chapters 4 and 5 to measure loudness functions in normal–hearing and hearing–impaired listeners.
Two factors that influence the results of the categorical scaling technique are the range of stimulus levels and the order of presentation (Heller, 1991; Hellbrück, 1993; Hohmann, 1993; Kollmeier and Hohmann, 1995). In order to produce consistent results across different subjects, the stimuli should be presented at levels covering the entire dynamic range (i.e., ranging from threshold of hearing to uncomfortable loudness level). Subjects rate perceived loudness differently when different stimulus level ranges are employed ("context effect"). They have the tendency to employ all categories of a given scale for the judgement, also for a restricted range of stimulus levels. Thus, they expand the scale if a smaller level range is

used, i.e., the scale resembles a kind of "rubber scale". Furthermore, the stimuli should be presented completely randomized to prevent subjects from relative judgements.
The categorical scale does not require spontaneity and naivete as does the absolute magnitude estimation technique. Previous experience in loudness scaling has no influence on the measured loudness functions, see chapter 4 and (Kießling et al., 1993; Kießling et al., 1994). It is still unclear which of both methods, absolute magnitude estimation or categorical scaling, yields more reliable results. Elberling and Nielsen (1993) presented data indicating that magnitude estimation techniques could be more reliable. They compared both methods in 10 hearing–impaired subjects and found a strong correlation between audiometric threshold and slope of loudness function for the magnitude estimation technique but not for the categorical scaling technique. In their experiments they used a variation of the absolute magnitude estimation method, the so–called restricted magnitude estimation originally proposed by Geller and Margiolis (1984) and Keller–Knight and Margiolis (1984). In this technique the range of numbers from which subjects may choose is restricted ($0 - 100$). For the categorical scaling they used the scale proposed by Allen et al. (1990) consisting of only seven categories. Sebald (private communication), however, pointed out that using only few categories yields much larger variability in the data. This could provide one explanation for part of the larger variability seen in the categorical data of Elberling and Nielsen. Furthermore, Elberling and Nielsen presented 16 equally spaced stimulus levels between threshold of hearing and uncomfortable loudness level. Thus, much fewer categories for rating the loudness were available to the subjects than stimulus levels employed. This might have strongly influenced the slopes of the loudness functions. According to Poulton (1989) subjects do not scale loudness in this case but simply put those stimuli together which are most easily confused. Therefore, the differences in variability of the measurements may have been due to differences in the task being performed.
However, the question remains whether both scales yield fundamentally different results if a categorical scale with many categories (> 20) is employed instead of a scale with few categories. Thus, it is still unresolved whether both methods yield a similar shape of the loudness functions, especially near threshold. It is well known that using an absolute magnitude estimation technique yields a curvature near threshold well described by a power law. However, no data have been presented in the literature indicating whether or not a similar curvature is observed in loudness functions using a categorical scale with many categories. Therefore, in this study absolute magnitude estimation, restricted magnitude estimation and categorical scaling using many categories (50) were compared with each other. In the following the experiments performed for comparing the different scales are described.

3.2 Method

Apparatus

All experiments in this study were carried out using a computer–controlled setup. An

audiological workstation supports a variety of different audiological tests, like loudness scaling, speech audiometry, and different psychoacoustical experiments.

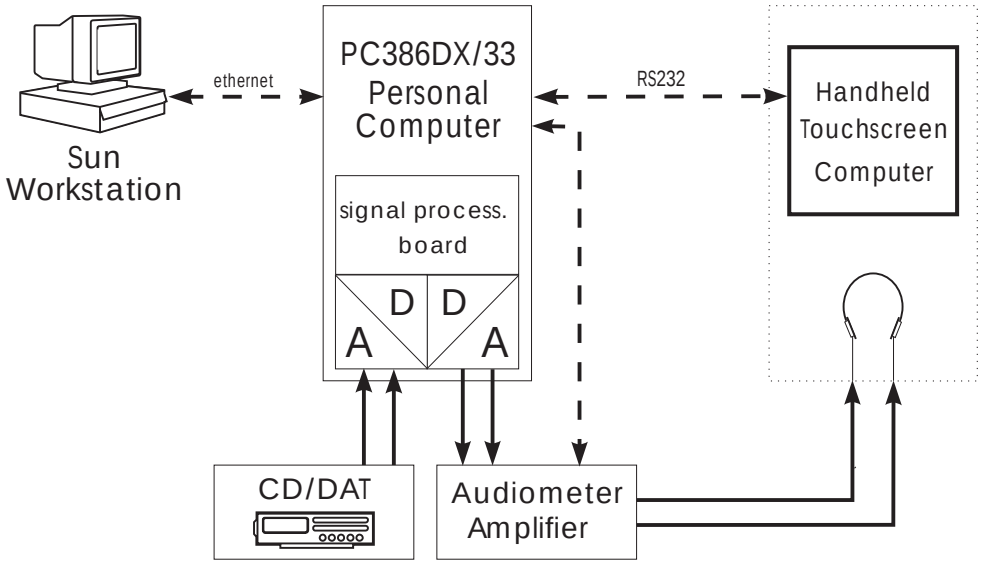

Fig. 3.1: *Schematic diagram of the experimental setup used for performing loudness scaling experiments.*

Figure 3.1 shows the experimental setup. The SUN workstation is used for signal generation and storage of the signals employed. It is connected via ethernet to a PC 386 (Hewlett Packard) which controls the experimental procedure and records subjects' responses. The stimuli are transmitted from the PC via a signal processing board with a 16–bit D/A converter to the computer–controlled audiometer amplifier. This signal processing board is also used to equalize the system response across different headphones or loudspeakers. In this study the signals were always presented monaurally to the subjects via headphones (BeyerDynamik DT48).

Stimuli

In this study a bandpass–filtered frozen noise, centered at 1 kHz with a bandwidth of 200 Hz, was employed as stimulus. It was generated off–line. A Gaussian noise (duration 3 s) was digitally generated first at a sampling rate of 25 kHz. It was Fourier–transformed and bandpass–filtered at a center frequency of 1kHz with a bandwidth of 200 Hz. After transforming the signal back to the time domain, it was windowed with a rectangular window (2 s duration including 50 ms cosine–ramps).

Procedure

The subjects were seated in a sound attenuating chamber. Their responses were recorded using a computer keyboard. The instructions and categories were presented using a computer monitor. Loudness scaling experiments were performed using the following techniques:

1. Absolute magnitude estimation (AME), as was described in section 3.1.2
2. Restricted magnitude estimation (RME) using a restricted range of integers between 0 and 50. This can be considered a categorical scale with many categories.
3. Categorical scaling (CS) using the two–step procedure described in section 3.1.3. For the initial judgements, the following five verbal categories were presented to the subjects: inaudible, very soft, soft, intermediate, loud, very loud, too loud. In the subsequent judgement, the listener was allowed to select from ten numbers symmetrically placed around the previously chosen category. Overall, this scale consists of 50 different categories from which subjects can select. Thus, it might be viewed as a two–step RME procedure.

Two different scaling experiments were performed with each method employing two overlapping stimulus level ranges: 0 – 60 dB HL and 30 – 90 dB HL. From both level ranges 21 stimuli, equally spaced in level on a dB scale, were selected respectively and presented in random order. Each stimulus was scaled four times by each subject. The experiment employing the lower levels was always performed first. The experiments were performed in the order AME, RME, CS.

Subjects

Five adult, male, normal–hearing listeners, all staff members, aged 25 –30 years, experienced in other psychoacoustic experiments, participated voluntarily in this study. One of them was the author. Normal hearing was established by routine audiometry. The air conduction threshold of all 5 subjects was below 10 dB HL. Two subjects had no prior experience in loudness scaling experiments.

3.3 Results and discussion

The individual results of the five subjects did not differ markedly. Therefore, these results were averaged arithmetically. The average loudness functions obtained with the three different scaling techniques are shown in Fig. 3.2 and Fig. 3.3. In Fig. 3.2 loudness values and standard deviation are plotted using a linear ordinate versus stimulus level, while in Fig. 3.3 loudness values are plotted using a logarithmic ordinate.
The standard deviation obtained using the CS technique is smaller than that obtained with the AME or RME technique especially at mid and high sound pressure levels. This is probably due to the assignment of fixed values to the verbal categories. This provides a kind of "fixing" of some values to familar verbal categories (Heller, 1991).

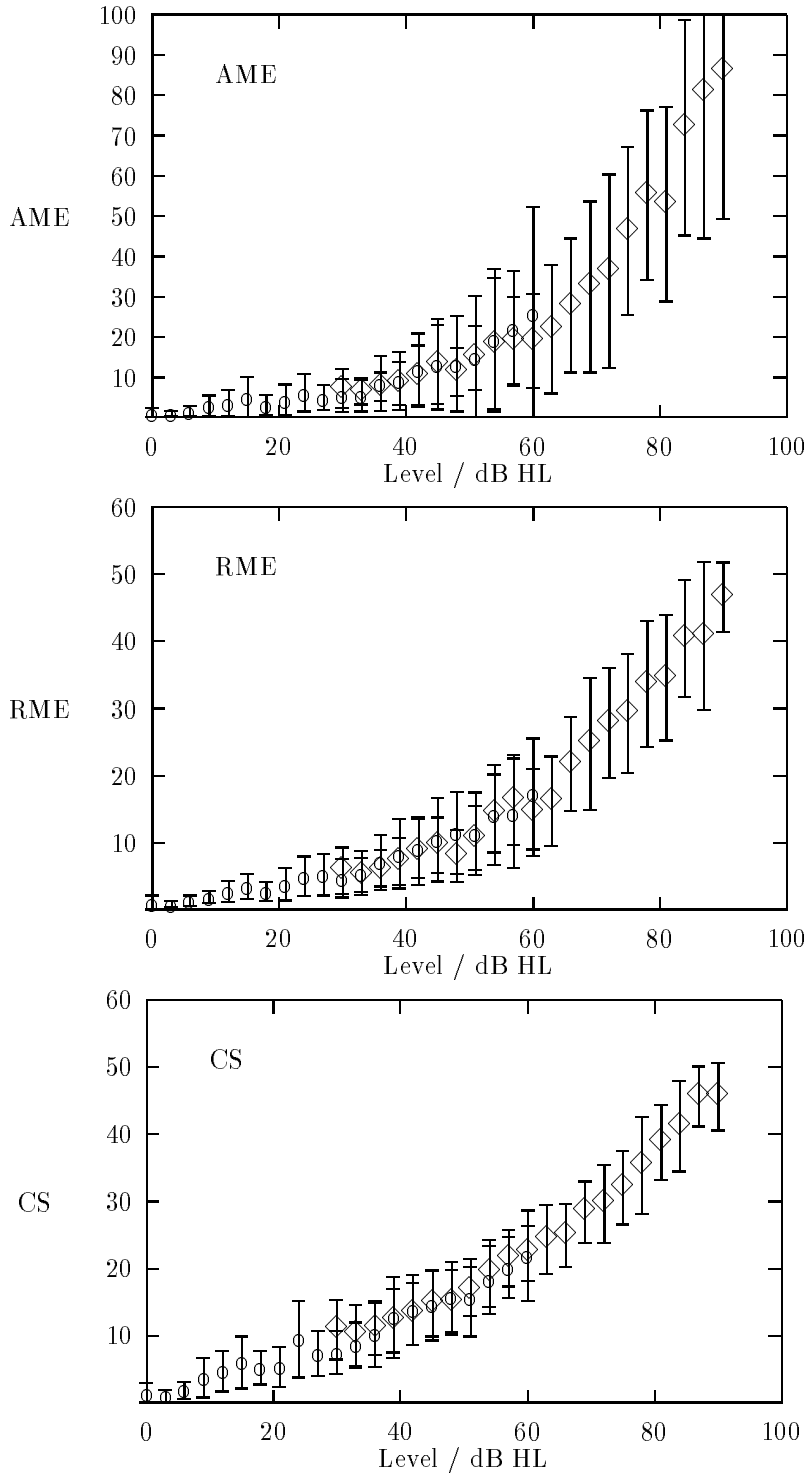

Fig. 3.2: *Loudness functions measured with different scaling techniques averaged over 5 normal–hearing listeners. Loudness in numbers is plotted versus stimulus level. The error bars denote the standard deviation. Circles (○) indicate the responses when employing the level range 0–60 dB HL and diamonds (◇) those with the level range 30–90 dB HL. Upper panel: Absolute magnitude estimation (AME). Mid panel: Restricted magnitude estimation (RME). Lower panel: Categorical scaling (CS) with 50 categories using the two–step procedure. Note that with all three techniques a similar curved loudness function is obtained.*

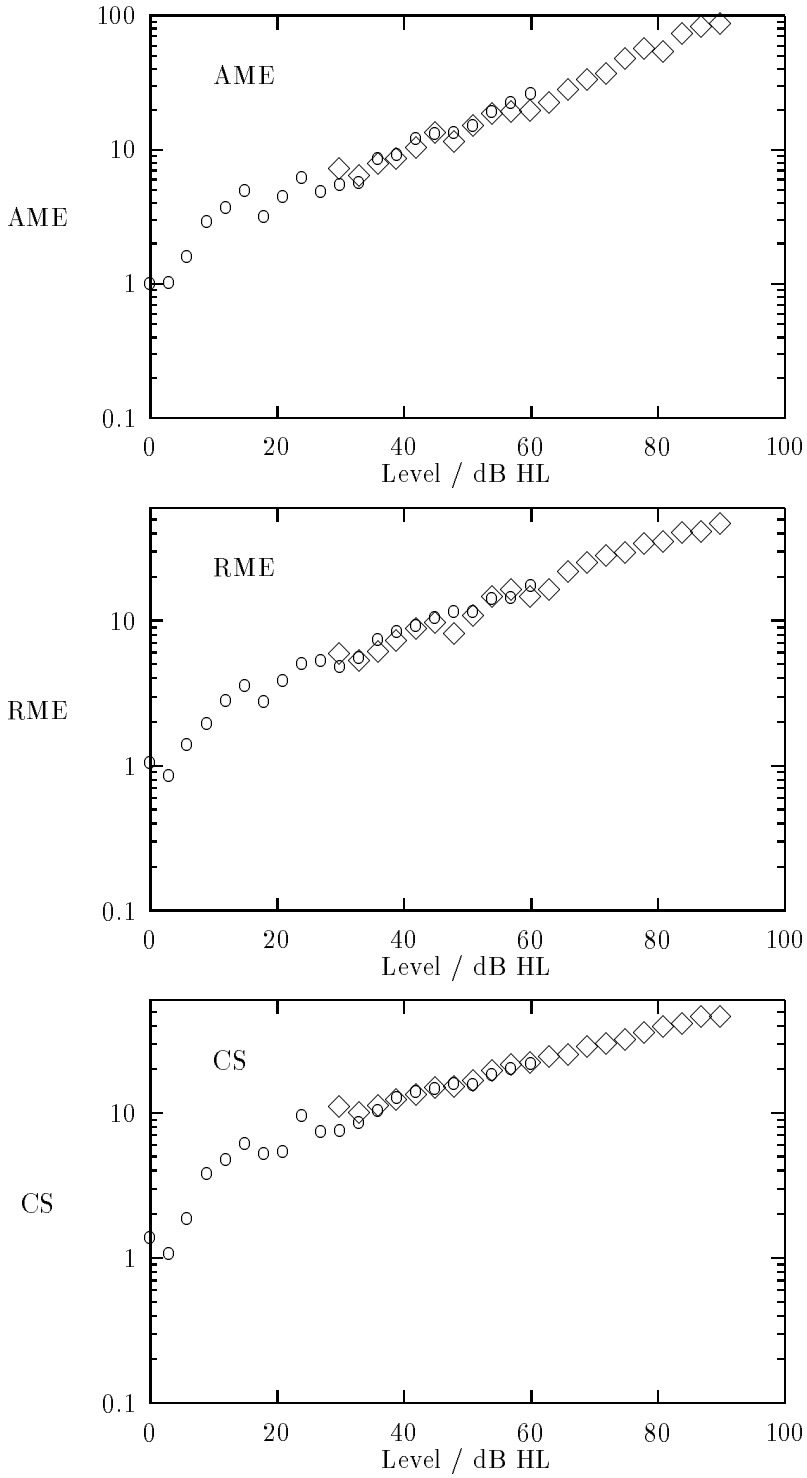

Fig. 3.3: *Same loudness functions as in Fig. 3.2 but using a logarithmic ordinate. Note that with all three techniques a steeper increase of loudness with level is observed near threshold.*

The categories may provide a guideline for the judgement of perceived loudness. Thus, these verbal categories appear to represent a common internal standard accross subjects, or a common "natural scale" (Zwislocki, 1991).
Furthermore, it is evident from these figures that in all applied scaling techniques the results of the two experiments employing split level ranges are nearly identical. A similar finding was reported by Gescheider and Hughson, who measured loudness functions in normal–hearing subjects using the AME technique (Gescheider and Hughson, 1991; Gescheider, 1993). They also showed that the sequence of experiments with different level ranges had no significant influence on the shape and location of the loudness function. Surprisingly, loudness functions measured by means of categorical scaling are also not affected by the splitting of the level range. This contrasts with findings in the literature since it has been reported that this technique strongly depends on stimulus context. Obviously, such a context dependence was avoided in the present experiment by orientation of the subjects about the level range employed. Although no explicit orientation was provided, the subjects were oriented about the employed level range since they performed the AME scaling using the two different level ranges before the categorical scaling. Furthermore, the dependence on context, e.g., an expansion of the scale, might not occur if a sufficiently large number of different categories are available to the subjects for the judgement of perceived loudness. Using many categories resembles using a continous scale. Thus, using 50 categories might also have contributed to the reduced dependence on context.
The values our subjects used for judging loudness in the AME technique differ from those generally reported in the literature, yielding shallower loudness functions than reported by, e.g., Hellman and Zwislocki (1963) and Gescheider and Hughson (1991). On average the subjects in this study applied a number range of 1 to 100 for rating the loudness compared to a range of 0.05 to 100 reported in the literature. This might be caused by two factors. Firstly, subjects were all members of the research group well trained in psychoacoustic experiments and not unexperienced in loudness scaling using a categorical scale. Although two of them never performed a loudness scaling experiment before, they might have been biased to using numbers between 0 and 50 for scaling loudness. Secondly, a computer keyboard was used for recording their responses. Hellman (personal communication) pointed out that using a keyboard could lead subjects to use only integers rather than an expanded range of numbers including fractions, decimals, and small numbers between 0 and 1.
It is evident from the data that all three functions exhibit the same curved shape described by the power function typical of magnitude estimation. For comparison, the results obtained with the different techniques are plotted in Fig. 3.4. In the upper panel the results of AME ($\diamond$) and CS (+) are plotted and in the lower panel those of RME ($\diamond$) and CS (+). The RME and CS (lower panel) mainly differ at low levels where RME yields slightly lower values than CS. This could be due to assigning too large numbers to the low categories. However this could also be due to a nonlinear mapping between numbers and perceived loudness. This is discussed further below. At high levels both methods yield the same values. The AME and CS results differ less at lower levels than those of RME and CS. The AME results lie between those of RME and CS. In the RME task, subjects might

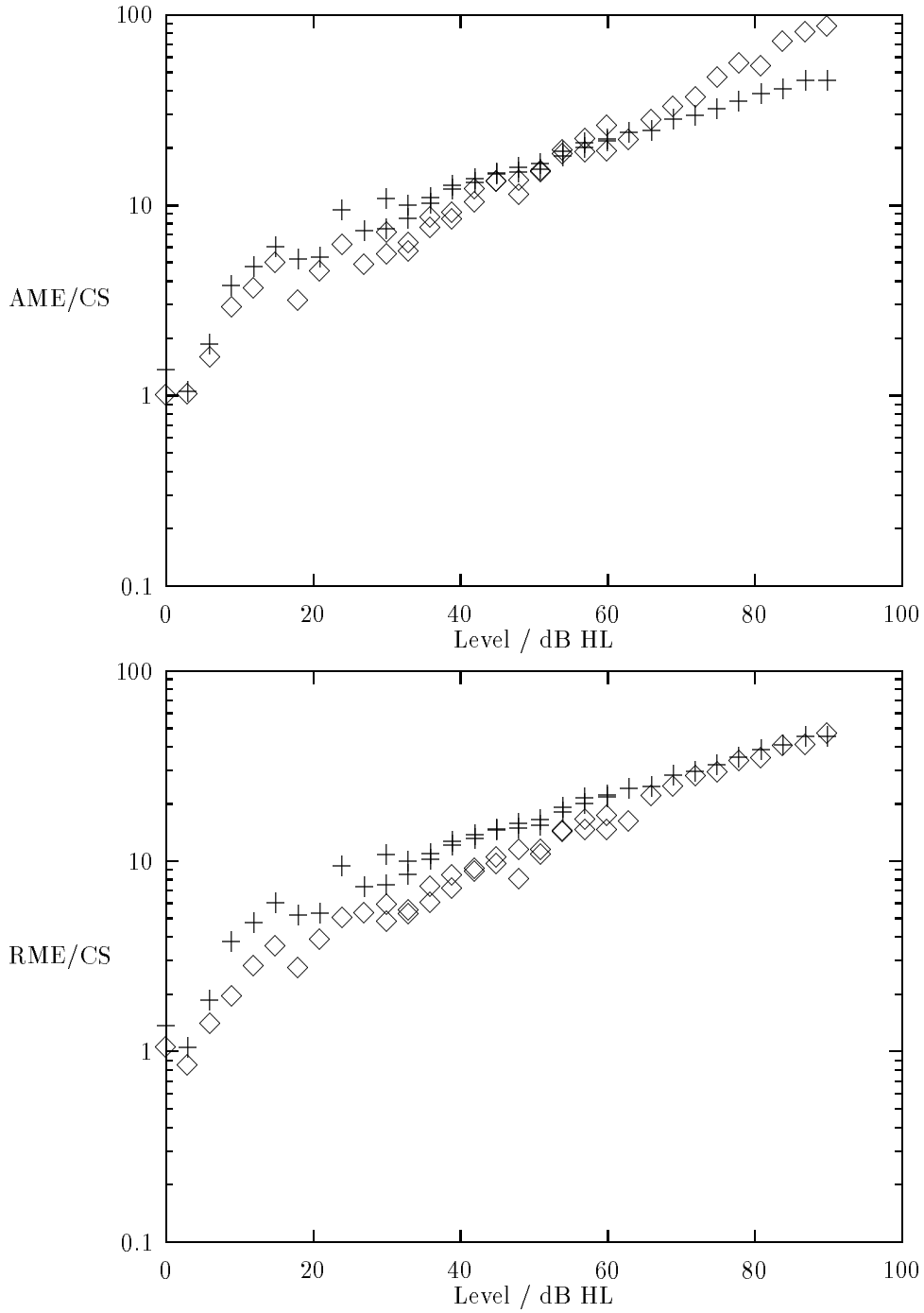

Fig. 3.4: *Same representation as in Fig. 3.3 to facilitate the comparison of measured loudness functions using AME (◇) with those of CS (+) (upper panel), and RME (◇) with those of CS (+) (lower panel).*

have scaled loudness conservatively at low levels, in order to maintain a sufficient range of

values for scaling at higher levels. At mid to high sound pressure levels the results of the AME differ markedly from those of RME and CS. Obviously, the restriction of values in the latter two methods yields a shallower increase in loudness than using a free numerical scale. However, the overall shape of the loudness functions obtained with these three methods is very similar. The loudness functions exhibit a concave shape when loudness is plotted using a linear ordinate, while they show a convex shape when plotted using a logarithmic ordinate. Specifically, the loudness function obtained using the CS technique also exhibits a steep increase with level near threshold.

Thus, it can be concluded that using a categorical scale with many categories (much more than the number of stimuli) yields results similar to those using a free numerical scale. A similar conclusion was drawn by Poulton (1989) but for a different sensory modality (electric shock). Furthermore, the CS (= two–step RME) and the RME do not differ markedly. According to Poulton the curved shape of the loudness function near threshold is partly due to the so–called logarithmic response bias. This logarithmic response bias reflects the fact that subjects (on average) apply a nonlinear scale to map stimulus intensity to numbers. In other words numerical estimates might not scale linearly with sensation but instead $S = N^{\alpha}$, where S is the "real" sensation, N is the numerical estimate produced by the listener, and α is an exponent with a value less than one (Krueger, 1989; Poulton, 1989). Furthermore, the CS curve seems to be less curved than the RME curve. This could indicate that subjects employ different strategies when using the CS scale and the RME scales. It has been suggested that using a two–step procedure with few categories in the first step might reduce the logarithmic response bias. However, a logarithmic response bias might still occur if listeners use the scale in the same way as the RME, i.e., more or less ignoring the first step. Indeed, two subjects reported to have performed in this way. This can also be observed in the individual data, where some subjects show a stronger curvature while others show an almost linear increase. Thus, the logarithmic response bias appears to more strongly affect the results of the RME than the CS method, causing a stronger curvature of the obtained loudness function at low levels. However, different authors argue that the curved shape near threshold reflects properties of signal processing in the auditory system (Hellman, 1991) rather than properties of the applied scaling technique. However, the "true" origin of this curved shape can not be discovered from the data presented here. The reader is referred to Krueger (1989), Poulton (1989) and Hellman (1991).
In Fig. 3.5 the results of the CS technique are compared to those using a one–step categorical scaling technique in which only 10 categories are used. These data are taken from Hohmann (1993). Hohmann performed loudness scaling experiments also with 5 normal–hearing subjects and with a 200–Hz wide noise band centered at 1 kHz. According to Poulton (1989) a linear relationship between stimulus level and perceived loudness should be obtained when using such few categories. Actually, using few categories yields a less curved, although not perfectly linear, loudness function. Linear functions were fitted to both curves. For fitting the data, a variation of a least–squares technique, the so–called chi–square technique was applied. This technique takes the variances of the results into account. The quality of fit

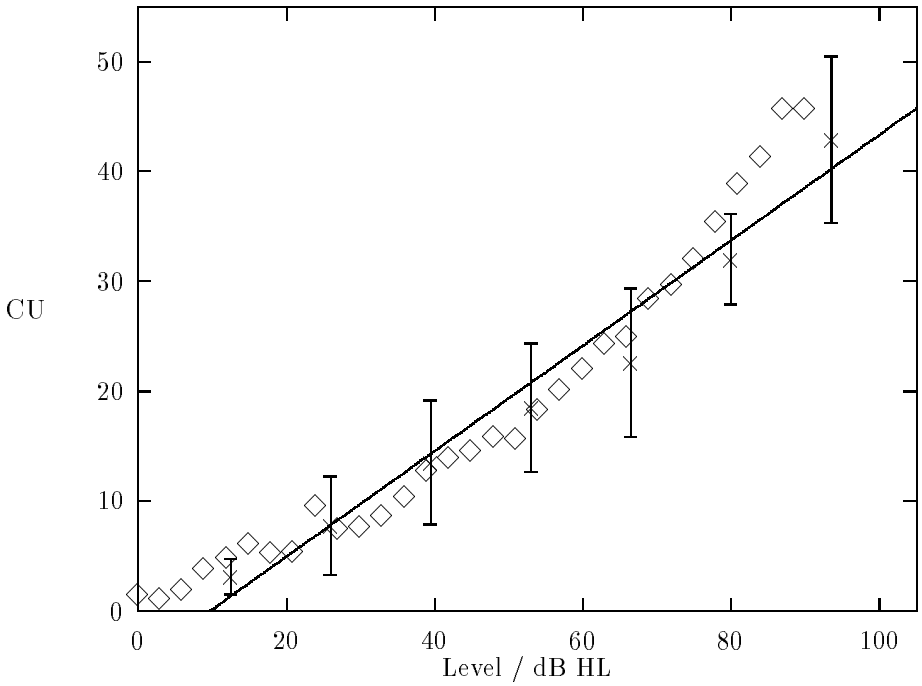

Fig. 3.5: *Comparison of measured loudness functions using either a two–step (◇) or a one–step (x) categorical scaling technique using 10 categories.*

was 0.995 for the one–step categorical scaling while it was 0.984 for the two–step technique. Thus, the loudness function obtained with the one–step technique is slightly better described by a linear function than the one described by the two–step technique. The differences between the results of the one–step and the two–step technique in Fig. 3.5 at high levels might be due to the different subjects. Hohmann presented his stimuli at 7 equally spaced levels between individual audiometric threshold and uncomfortable loudness level. On average these levels fell in the range between 10 and 95 dB HL. Thus, his subjects show a small shift in dynamic range of hearing compared to the subjects of this study. This causes different levels L_{25} of the category "intermediate", i.e., 25 categorical units indicating a shift of the loudness functions to higher values. The L_{25} are 65.1 dB HL in the present study and 71.0 dB HL in Hohmann's study. In summary, if a categorical scale with only few categories is applied for measuring loudness growth functions, an approximately linear relationship between perceived loudness and stimulus level is obtained. A similar result was obtained by Poulton (1989) but for a variety of different sensory modalities. Poulton concluded that using few categories avoids the above mentioned logarithmic response bias, since a less curved shape near threshold is observed.

3.4 Conclusions

The comparison of categorical scaling, absolute magnitude estimation and restricted magnitude estimation revealed that in none of the applied techniques did the splitting of the level range 0 – 90 dB HL in two overlapping intervals have a significant influence on the shape and location of the loudness function. In addition, applying a categorical scale with many categories (i.e., 50 categories) yields results similar to those obtained with both numerical scales. All three of the applied scales show a steeper increase of loudness with level near threshold when subjective loudness is plotted on a logarithmic scale versus level. Plotting the loudness on a linear ordinate yields the opposite trend. The loudness function obtained with CS shows a less curved, i.e., more linear, shape than those measured using AME and RME.
The curved shape of loudness functions obtained with numerical scales as well as with categorical scales with many categories, might partly be due to the logarithmic response bias. This logarithmic response bias is less pronounced in categorical scales with few categories only. Therefore, in the subsequent experiments in this work, a one–step categorical scaling technique will be applied for measuring loudness functions.