Zur Homepage der Dissertation

Objective assessment of audio quality using an auditory processing model

Vom Institut für Physik an der
Fakultät für Mathematik und Naturwissenschaften
der Universität Oldenburg
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
angenommene Dissertation

Rainer Huber
geb. am 18. Juni 1969
in Oldenburg

Erstreferent: Prof. Dr. Dr. Birger Kollmeier
Korreferent: Prof. Dr. Volker Mellert
Tag der Disputation: 19. Dezember 2003

Contents

CHAPTER 1

General introduction

Acoustics form the most important medium for human communcation. The introduction of technics allowing for acoustical communication over long distances and storage of audio signals had revolutionary effects on human culture. The need for evaluations of systems that transmit, store or process audio signals exists as long as those systems themselves. Before the "digital age", any kind of signal processing was "lossy" in the sense that it inevitably produced alterations of the original signal. In the case of audio signals (including speech), the most relevant criterion for the assessment of signal alterations and therefore of the transmission quality is the auditory perception. Unfortunately, the perception is merely of minor use if the transmission quality of a system needs to be communicated, since perceptions are subjective and can be described precisely and quantitatively only to a very limited extent. As a consequence, objective measures roughly accounting for properties of the auditory system, such as the signal-to-noise ratio (SNR), the (total) harmonic distortion or the frequency transfer function using logarithmic scales have been established for this purpose instead. These measures may impart some idea of the perception of some simple and well known stationary impairments such as additive noise, linear distortions or non-linear distortions that are due to compressive or expansive characteristics. However, they clearly fail to describe the perception of audio distortions produced by modern digital speech and audio coding-decoding algorithms (codecs), which are highly non-linear and non-stationary. An illustration of the failure of the SNR, for example, is the so-called "13-dB miracle"; it describes the phenomenon of a noise that is superimposed on an audio signal and becomes almost inaudible if its tempo-spectrally shape is adapted to that of the audio signal, even when the SNR declines to just 13 dB (Brandenburg and Sporer, 1992). For this reason, the present work is concerned with the development and evaluation of computational audio quality measures that overcome the limitations of conventional measures by incorporating a valid model of auditory perception.

Increasing demands on the transmission of speech over channels of limited bandwidth and progressing computational power led to the development of lossy digital speech compression schemes, which were introduced in the second half of the 1980s. Efforts on the development of audio codecs increased considerably at the end of the 1980s, at first to reduce the storage consumption of digital audio associated with digital video. (The name of the ISO[1] working group that has created one of the most established audio codecs of these days, MPEG (Moving Pictures Experts Group), attests to this circumstance.) Later, the reduction of audio data was exploited to allow for affordable digital home recording using cheap media with rather small storage capacities (Sony's MiniDisc, Philips' DCC - Digital Compact Cassette). Another great push of the audio codec development was caused by the demand of audio transmission via channels of restricted bandwidth, primarily the internet, which extensively spread within the last decade.

Audio compression algorithms are similar in common that they exploit masking properties of the auditory system. They permit a larger amount of quantization noise, which is shaped spectrally and temporally to be (ideally) masked by the signal and thus become inaudible. Coding artifacts produce very complex, time-variant distortions. Those kinds of distortions lead to impairments of the audio quality, which can only to a very limited degree be described by conventional quality measures that do not account for properties of the auditory system in an adequate way. For this reason, subjective listening tests are still the "golden standard" for evaluating audio (transmission) quality. Since formal listening tests are time consuming, expensive and in some cases not applicable (e.g. for online monitoring), efforts were made to develop new computational methods for the objective measurement of the perceived audio quality (degradation).

Promising new approaches incorporated psychoacoustic models. Two main concepts emerged: The masked threshold concept and the concept of comparing internal sound representations. The former concept is characterized by employing a psychoacoustic model to estimate the masking pattern of a given (undistorted) audio signal. The ratio between the energy of the actual distortion noise and the masking threshold is calculated and integrated over frequency and time (cf., e.g., Brandenburg, 1987). The second concept uses an auditory model to transform input and output signals of a considered audio device into corresponding representations in the perceptual domain, which are assumed to be used by subjects in their assessment of the audio quality. The internal representations are compared by mathematical means, typically yielding a distance or similarity measure, which

[1] International Organization for Standardization

estimates the perceived overall audio quality (difference) (e.g., Beerends and Stemerdink, 1992; Hansen and Kollmeier, 2000).

Early psychoacoustically motivated objective quality measures were mainly applied to speech codecs (Schroeder et al., 1979; Karjalainen, 1985). The first measure that was applied to wide-band audio codecs was the Signal-to-Mask Ratio (NMR) introduced by Brandenburg (1987). In the first half of the 1990s, a number of further objective methods for the perceptual measurement of speech and audio quality were developed. Most of them use the concept of comparing internal representations (see, e.g., Paillard et al., 1992; Beerends and Stemerdink, 1992; Wang et al., 1992; Beerends and Stemerdink, 1994; Hollier et al., 1994). Apart from this common basic concept, those methods differ regarding details of the actual realization of the psychophysical transformation and by the measures calculated from the internal representations: While the difference between original and test representation directly represents the quality measure in (Beerends and Stemerdink, 1992) and (Paillard et al., 1992), the overall probability of detection is derived from the representation difference in (Colomes et al., 1995) and (Sporer, 1997). Thiede and Kabot (1996) use the internal representation to calculate the partial loudness of linear and non-linear distortions and alterations in the temporal envelope, which are mapped to a final measure of the overall audio quality.

Although some of the objective quality measurement methods achieved good correlations with subjective ratings of speech quality (e.g. ITU-T, 1996b; Hansen and Kollmeier, 2000), none of the objective audio quality methods proposed as a standard was found to be sufficiently reliable by an ITU[2] committee in 1994. Consequently, a new and improved method was developed jointly by the seven proponents. The resulting method combines features of all of the originally proposed methods and includes an artificial neural network that maps several output variables to a final measure of the overall audio quality. The new method, called PEAQ (Perceptual Evaluation of Audio Quality), became the ITU standard for objective measurement of perceived audio quality in 1998 (ITU-R BS.1387, ITU-R, 1998a).

Due to the combination of different methods (including different auditory models) and the use of an artificial neural network, PEAQ hardly allows for direct conclusions on the actual mechanisms involved in human perception of audio quality. Moreover, its high degree of specialization possibly represents a risk of restricted applicability. For these reasons, the objective of the present thesis was to develop new, reliable methods for the

[2]International Telecommunication Union

objective, perceptual assessment of audio quality, using a psychoacoustically validated model of auditory perception.

In Chapter 2, a new method for the prediction of the perceived difference of the overall audio quality between audio signals is proposed. This method represents an expansion of the speech quality measure q_C of Hansen and Kollmeier (2000), who successfully applied their method to predict subjectively rated speech transmission qualities of mainly low-bit rate speech codecs. The core of both the original and the expanded method is the psychoacoustically validated, quantitative model of the "effective" signal processing in the auditory system of Dau et al. (1996a, 1997a), following the concept of comparing internal sound representations. Free model parameters were not altered, but adopted and kept fixed from psychoacoustical modeling. Apart from replacing the model employed in (Hansen and Kollmeier, 2000) by a more recent version, which uses a modulation filterbank instead of a modulation lowpass filter (cf. Dau et al., 1997a), cognitive effects mainly concerning the relation between time-varying instantaneous and overall audio quality were modeled additionally. Quality prediction results are presented using a large database of subjectively rated audio signals. The influence of the modulation processing within the model on the prediction performance is investigated, as well as the influence of the parameters of the cognitive model parts. The performance of the presented method is compared with that of the current ITU standard BS.1387.

Chapter 3 presents a masking experiment that was carried out to measure thresholds of wide-band audio distortions. This experimental paradigm is proposed as an alternative method for the subjective evaluation of near-transparent audio codecs. In addition, the experiment was simulated, by employing the "complete" auditory models described in (Dau et al., 1996a, 1997a), i.e., including the final detector stage. (Note, that this stage was not used for the prediction of audio quality in Chapter 2.) Measured and simulated data are presented. Again, the influence of the modulation processing modeling on the prediction accuracy is examined.

The applicability of the audio quality measures presented in Chapter 2 for the assessment of noise reduction schemes for speech is investigated in Chapter 4. For this purpose, signals and corresponding subjective ratings obtained from experiments carried out for the evaluation of single and multi-channel noise reduction schemes are used. Based on the presented results, one of the existing audio quality measures was modified to yield an optimized measure for either the prediction of the perceived naturalness of speech or the amount of background noise.

An exemplary application of the new audio quality measure is described in Appendix C: Embedded in the framework of an audio quality test bench, this measure is already being applied within the priority program "Fundamentals and Methods for Low-Power Information Processing (VIVA)" of the Deutsche Forschungsgemeinschaft (DFG) as a tool for ongoing optimizations of implementations of audio processing algorithms.

CHAPTER 2

Objective assessment of audio quality

Abstract

*A new method for the objective assessment and prediction of the perceived audio quality is introduced. It represents an expansion of the speech quality measure q_C, introduced by Hansen and Kollmeier (2000). It is based on a psychoacoustically validated, quantitative model of the "effective" peripheral auditory processing by Dau et al. (1996a, 1997a). To evaluate the audio quality of a given distorted signal relative to a corresponding high quality reference signal, the model is employed to compute "internal representations" of the signals, which are partly assimilated in order to account for cognitive aspects. The linear cross correlation coefficient of the assimilated internal representations represents the new audio quality measure PSM (*Perceptual Similarity Measure*). While PSM shows good correlations with subjective quality ratings if different types of audio signals are considered separately, a better accuracy of signal-independent quality prediction is achieved by another quality measure, PSM_t, represented by the 5%-quantile of the sequence of instantaneous audio quality PSM(t). The new measures were evaluated using a large database of subjective listening tests that were originally carried out on behalf of the ITU[1] and MPEG[2] for the evaluation of various low bit-rate audio codecs. The results support the concept of amplitude modulation processing by a modulation filterbank and suggest a nonlinear relationship between the perceived instantaneous and overall audio quality. The observed good performance of PSM_t in predicting subjective quality ratings is similar to the ITU standard BS.1387 ("PEAQ") for most of the tested data.*

[1] International Telecommunication Union
[2] Moving Pictures Experts Group

2.1 Introduction

In the last decade, efforts were made to develop computational methods for perceptual assessment of transmission quality of lossy wide-band audio compression techniques as an alternative to costly listening tests. Many approaches are similar in that they use a psychoacoustically motivated auditory model which is applied to a pair of reference-test-signals whose quality difference is to be evaluated (e.g. Beerends and Stemerdink (1992); Baillard et al. (1992); Colomes et al. (1995)). The simulated auditory processing transforms the given signals into corresponding "internal representations", i.e. that kind of information that is assumed to be contained in the output of the peripheral auditory system (in terms of neural activity patterns) and serve higher cognitive stages as input. These internal representations are then compared by means of a mathematical distance measure or similarity measure. Ideally, the resulting measure should correspond to the perceived signal differences, which are regarded as audio quality degradations.

Predicting the perceived quality degradation turned out to be more difficult for general wide-band audio signals with rather small distortions than for narrow-band speech signals with greater distortions: Although some objective speech quality schemes performed quite well whereby one was standardized by the ITU (ITU-T, 1996b), none of the generalized audio quality measures fulfilled the requirements of the ITU-R in 1994, when proposed as a standard. Consequently, the seven proponents agreed to jointly develop an improved measurement method. The resulting new method, called PEAQ (*Perceptual Evaluation of Audio Quality*) is a combination and expansion of the best elements of the original methods. It became ITU standard in 1998 (ITU-R, 1998a, ITU-R BS.1387,). PEAQ includes two different ear models and makes use of the masked threshold concept as well as of a comparison of internal representations. An artificial neural network maps several output values to a single final quality value. Optimization and training was done using a set of listening test databases. PEAQ is characterized by a high degree of optimization and adaptation to a single task. Being a composite of rather simple auditory models, refined technical approaches and a costly artificial neural network, PEAQ does not likely represent a realistic, valid model of auditory perception (which the authors do not claim). However, the main purpose of PEAQ is the perceptual evaluation of audio quality, which seems to be served quite well for most of the audio items tested so far (Thiede et al., 2000; Treurniet, 2000).

In contrast, the aim of the present work is two-fold: On the one hand, the perceived audio quality for any kind of distortion and any kind of audio signal should be predicted

as well as possible. On the other hand, the core of the method should be formed by a psychoacoustically validated model of auditory processing, in which the general applicability should not suffer in consequence of possible adaptation to the present task. Ideally, necessary modifications should contradict other psychoacoustical findings, rather yield fruitful contributions to psychoacoustical modeling and understanding principles of auditory perception.

In this study, the speech quality measurement method of Hansen and Kollmeier (2000) was expanded in order to predict not only the perceived quality of distorted narrow-band speech, but of any distorted wide-band audio signal in comparison to the undistorted reference signal. Hansen and Kollmeier showed that their measure performs on average as well as the ITU-T standard P.861, but uses an auditory model that has been validated in a wide variety of psychoacoustical measurements (Münkner, 1993; Dau et al., 1996a,b; Fassel, 1994; Sander, 1994; Verhey, 1998; Derleth, 1999). Its free parameters were taken on from psychoacoustical modeling and kept fixed. The final quality measure, q_C, is built by calculating the normalized cross correlation coefficient of the downsampled, spectral weighted model outputs (internal representations) of the distorted and undistorted (reference) signal.

The present study applies a modified version of the method described above to the prediction of the perceived quality degradation of wide-band audio signals distorted by low bit-rate audio coding-decoding schemes ("codecs").

In the first part of this chapter, a method for signal-dependent quality measurement is presented, followed by a presentation of an expanded version for signal-independent quality measurement in the second part. The methods are evaluated using a large database of audio signals and corresponding subjective quality ratings. Among other aspects, the influence of the modulation processing and the relationship between instantaneous and overall audio quality is examined in particular. The results are presented, compared with results obtained by the ITU standard BS.1387 and discussed.

2.2 Subjective audio quality tests

In order to test and optimize the objective audio quality measurement scheme, a database of subjectively rated audio signals with different types and degrees of quality degradation was used. It consists of material from listening tests conducted between 1990 and 1995 by the International Telecommunication Union (ITU) and the Moving Pictures Experts Group (MPEG). The purpose of these listening tests was to assess the transmission quality of various low-bit rate audio codecs (e.g. ADPCM (ITU-T, 1990), Sony ATRAC (MiniDisc) (Tsutsui et al., 1996), Dolby AC-2 and AC-3 (Fielder et al., 1996), MPEG-1 Audio Layer 2+3 (ISO/MPEG, 1992)). Six data sets emerged from these listening tests, including critical audio signals processed by the codecs and corresponding subjective quality ratings. These data sets are denoted as MPEG90, MPEG91, ITU92DI, ITU92CO, ITU93 and MPEG95. (See Appendix A for a more detailed description of the data sets.) The listening tests were carried out in different countries. 19 to 91 "expert listeners" participated in these tests. Unlike speech codecs used in telephone communications, most of the tested wide-band audio codecs produce considerably smaller, sometimes even imperceptible impairments. Therefore, the subjective assessment of the processed audio signals was performed according to the ITU-R recommendation BS.1116 (ITU-R, 1997), which is intended for the assessment of small impairments in audio systems. It describes a triple stimulus test with a hidden reference: Three signals A, B, and C are presented to the listener, who is free to switch between these signals. Signal A is known to be the unprocessed reference signal, whereas signals B and C are the signals processed by the tested system and once more the reference signal in random order. The listener is asked to rate the degradation of the "basic audio quality" of signals B and C relative to signal A on a continuous five-grade impairment scale defined in ITU-R recommendation BS.562-3 (ITU-R, 1990) (see Figure 2.1).

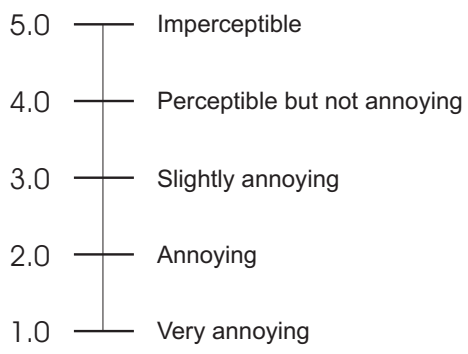

Figure 2.1: ITU-R five-grade impairment scale.

The final quality value is given by the difference of the listener's ratings for signals B and C, called the *Subjective Difference Grade* (SDG), defined as

$$\text{SDG} = \text{grade}_{\text{test signal}} - \text{grade}_{\text{reference signal}}.$$

Unless the listener rates the audio quality degradation of the hidden reference erroneously worse than that of the test signal (which sometimes happens), the SDG has a negative value in the range of -4 (very annoying impairment) to 0 (imperceptible impairment). The SDG values contained in the database described above are mean values over all listeners.

2.3 Objective audio quality measurement

2.3.1 Signal dependent quality prediction

The basic approach of the presented method for objective audio quality measurement is to apply an auditory processing model to a given pair of reference and test signals and take the correlation coefficient of the model outputs as a measure for the perceptual similarity of the signals. Because of the reference signal's high fidelity, it is reasonable to interpret any perceptible deviation from that reference as a degradation of audio quality. The correlation coefficient therefore serves as an objective audio quality (degradation) measure.

Preprocessing

Before the reference signal and the test signal are processed by the auditory model, a possible overall time delay and level difference of the test signal relative to the reference signal has to be eliminated. These deviations are mostly perceptually irrelevant, but could affect the objective quality measure considerably.

If the time delay introduced by the distorting system (e.g. a codec) is not known a priori, it could be estimated by calculating the cross correlation function of the signal envelopes and taking the time lag of its maximum as a delay estimation. The signals are then aligned in time by delaying the reference signal by the known or estimated lag. However, this works only if the system's delay is not varying in time. Otherwise, time alignment has to be done block by block. The signals of the given database are already time aligned.

Level alignment was done by scaling the test signal with a constant factor that was chosen to result in equal overall RMS values. As in the case of time alignment, this way of level alignment will also only work if the system's overall gain has no long term drifts. This was the case for the test material used in this study.

The third step of preprocessing consisted in deleting silent signal intervals. Episodes in the reference signal with levels clearly below hearing threshold (i.e., pauses) were cut out as well as the corresponding intervals of the test signal. It is reasonable to assume that silent intervals do not contribute to the listener's judgement of audio quality as long as they are not filled with audible noise by the distorting system, which is unlikely for audio codecs if the distance between the interval's boundaries and the nearest audible signal segment is larger than the codec's frame length (e.g. up to 24 ms in MPEG-1 Layer III (Brandenburg and Stoll, 1994)). To ensure such a distance and to account for possible temporal masking effects, pauses were not cut out completely, but shortened to a minimum length of 200 ms.

Auditory processing

To simulate the transformation of acoustic stimuli into neural activity patterns by the human ear, a quantitative model of the "effective" auditory signal processing (Dau et al., 1997a) is applied to the preprocessed pair of reference and test signal. This rather psychoacoustically than physiologically motivated model transforms both incoming signals into corresponding "internal representations", i.e. three-dimensional representations of time, frequency and modulation-frequency [3].

Figure 2.2 shows a block diagram of the auditory model. The incoming signal is split up into 33 critical bands by a linear 4th order gammatone filterbank (Patterson et al., 1987), accounting for the basilar membrane's bandpass characteristic. Its center frequencies are equally spaced on an ERB scale (ERB: *equivalent rectangular bandwidth*), with one filter per ERB, ranging from 235 Hz to 14500 Hz. Each filter has a bandwidth of one ERB.

Subsequently, each band is processed independently, beginning with half-wave rectification and lowpass filtering at 1 kHz. This roughly simulates the transformation of

[3] An earlier version of this model used by Hansen and Kollmeier (2000) for speech quality measurement uses a modulation lowpass filter instead of a modulation filterbank (Dau et al., 1996a). That version was also applied in this study for comparison. The version used by Hansen and Kollmeier (2000) and also in this work differs from that described in Dau et al. (1996a) by using a gammatone filterbank instead of Strube's basilar membrane model (Strube, 1985). Moreover, only the preprocessing part of the model is used, while the decision device is omitted.

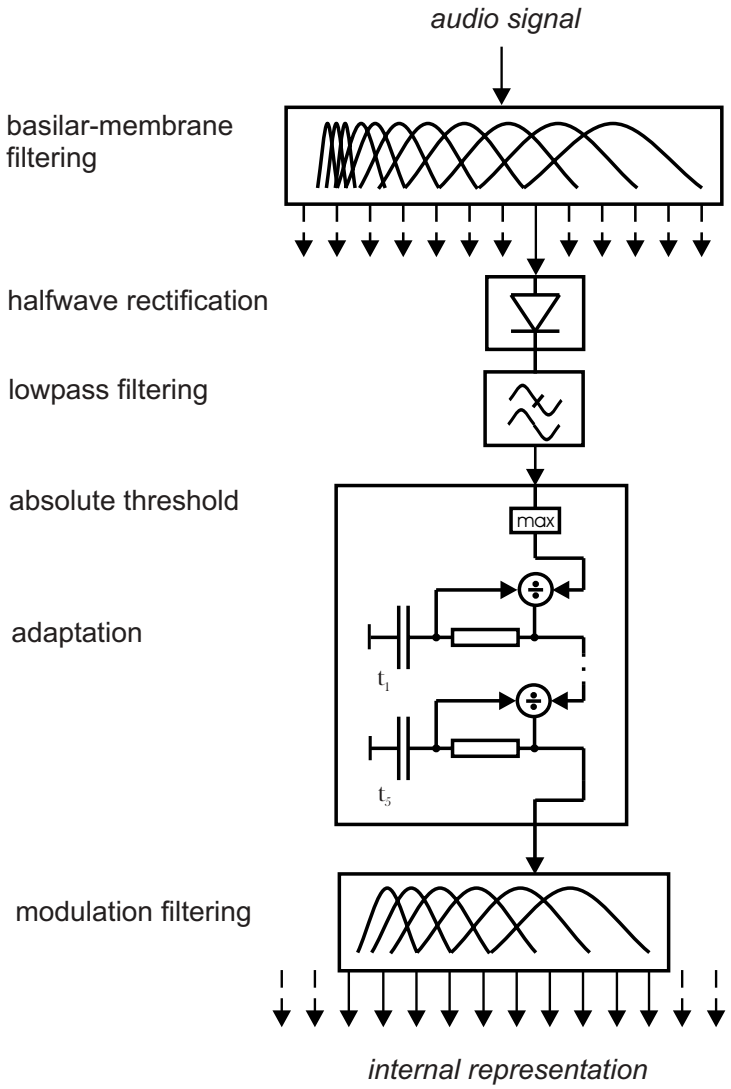

Figure 2.2: Block diagram of the auditory model.

mechanical oscillations to neural firing rates of the inner haircells. This stage essentially preserves the envelope of the signal for high (> 1 kHz) center frequencies, while preserving amplitude and phase for lower center frequencies.

To account for the absolute hearing threshold, the minimal value at the input to the next stage is limited to a lower bound, which depends on the assumed level of the maximum input.

Effects of temporal masking and adaptation are modeled by the subsequent chain of five consecutive nonlinear feedback loops (Püschel, 1988). Each loop consists of a dividing element and a RC-lowpass filter, so that the input is divided by the low-pass filtered output. Thus, for stationary inputs, the output of each loop equals the square

root of its input, and the output of the whole chain equals the 32th root of the input, which approximates a logarithmic compression. Non-stationary inputs are processed less compressively, depending on the rate of envelope fluctuations: Changes in the input signal that are very fast compared to the time constants of the lowpass filters (5 ms to 500 ms) are processed almost linearly. Due to its transformation characteristics, the adaptation stage contrasts signal amplitude fluctuations: Rapid changes (e.g. on and offsets) are emphasized, while stationary parts are compressed. The time constants of the feedback loops cause a kind of "memory" that enables the model to predict temporal effects in auditory perception, such as temporal integration and forward masking.

In the final processing stage, the envelope signal is analyzed by a linear modulation filterbank. This stage of modulation processing in the auditory system is the most substantial difference between the two model versions. The modulation filterbank replaces the 8 Hz lowpass filter of the previous version (Dau et al., 1996a), generalizing the model to account also for psychoacoustic experiments of amplitude modulation detection (Dau et al., 1997a,b). Up to 10 Hz center frequency, the filters have a constant bandwidth of 5 Hz; above 10 Hz, they are scaled logarithmically with a constant Q-value of 2, overlapping at -3 dB. To produce a loss of information for higher center frequencies ($>$ 10 Hz), only the (Hilbert-) envelopes of the output signals are calculated. In this study, eight filters with center frequencies up to 129 Hz were used.

The total output of all 33 x 8 channels forms the "internal representation" of the audio signal. It is a three-dimensional matrix, which can be interpreted as an activity pattern in the frequency and modulation frequency domain, varying in time. To reduce computational effort and storage consumption for subsequent post-processing and analysis steps, the internal representation is downsampled separately for each modulation channel to frequencies of about six times the modulation channel's center frequency.

Postprocessing: modeling cognitive effects

The stages of simulated auditory processing applied so far represent the *preprocessing* part of the complete model by Dau et al. (1996a, 1997a), which was originally designed and optimized to predict detection thresholds from psychoacoustic masking experiments. Originally, noise is added to the internal representation which is then fed into a decision device ("optimal detector"). In this study, instead of a detected vs. not-detected decision, the subjective assessment of perceived similarity/dissimilarity of a pair of audio signals on a continuous scale has to be predicted. For this purpose, the linear cross correlation

coefficient of the internal representations of the reference and the test signal is calculated, following the approach of Hansen and Kollmeier (2000)[4]. But prior to this step, the internal representation of the distorted test signal $Y = (y_{tfm})$[5] is partially equalized to that of the reference signal $X = (x_{tfm})$: Elements of Y having smaller absolute values than the corresponding elements of X are replaced by the mean absolute values of both elements, thus halving the differences:

$$\tilde{y}_{tfm} = \begin{cases} (y_{tfm} + x_{tfm})/2, & |y_{tfm}| < |x_{tfm}| \\ y_{tfm}, & |y_{tfm}| \geq |x_{tfm}| \end{cases}$$

This approach was adopted from Berger (1998) and is based on a sign-dependent difference weighting that was first suggested by Beerends (1994) and also successfully applied by Hauenstein (1997) in speech quality measurement. This approach follows the hypothesis that "missing" components in a distorted signal are less disturbing than "additive" components.

The final cross correlation operation, which yields the quality measure PSM (*Perceptual Similarity Measure*), is performed separately for each modulation channel[6] Each modulation channel is represented by a two-dimensional sub-matrix (i.e., $(x_{tf})|_{m=const}$). The linear cross correlation coefficient of two $N \times M$ matrices is given by

$$r = \frac{\sum_{t,f=1}^{N,M} (x_{tf} - \bar{x})(y_{tf} - \bar{y})}{\sqrt{\sum_{t,f}(x_{tf} - \bar{x})^2 \sum_{t,f}(y_{tf} - \bar{y})^2}}$$

(With N, M representing the number of time samples and frequency channels, respectively, and $\bar{x}$, $\bar{y}$ denoting mean values.) The correlation coefficients per modulation channel r_m are weighted by the normalized mean squared values of the corresponding sub-matrices and summed up to the final quality measure:

[4]Note that the non-uniform band weighting of internal representations suggested by Hansen and Kollmeier (2000) was not adopted in this work.

[5]The matrix indices $t, f$and m refer to time, frequency and modulation frequency.

[6]This independent processing provides a high computational efficiency since the different sampling frequencies within one internal representation do not have to be matched.

$$\mathrm{PSM} = \sum_m w_m r_m, \qquad \text{with} \quad w_m = \frac{\sum_{t,f=1}^{N,M} y_{tfm}^2}{\sum_{t,f,m=1}^{N,M,L} y_{tfm}^2}$$

(With L being the number of modulation channels.) PSM is restricted to the interval [-1, 1], with 1 indicating identity whereas smaller values correspond to larger deviations of the test signal from the reference signal, implying a degradation of the audio quality.

Limitations in predicting audio quality of stereo signals

The auditory model used in the measurement method described so far is a purely monaural model. It does not take into account any binaural effects. If, for example, a codec altered only the interaural phase of a stereo signal, the model would predict a perfect perceptual correlation of the original and the processed signal, if the two channels were assessed independently. In contrast, a human listener would perceive the altered signal quite differently from the original one.

Therefore, in a strict sense, this method should only be applied to predict audio quality degradations of mono signals. However, most of the audio material of the given database is stereo and was presented dichoticly to the assessing listeners. In those cases, objective quality assessment was realized by calculating the quality measure PSM for each signal channel separately and taking the lower PSM value (indicating worse quality) as the final quality measure. Of course, this does not compensate for the lack of a missing binaural model component. But at least it accounts for possible differences between left and right audio channels concerning monaurally detectable quality differences. The influence of the audio channel weighting was examined. The results will be presented in the following section.

2.3.2 Signal dependent audio quality prediction results

The objective audio quality measure PSM was calculated for all test signals (in full length) of the data base described in Section 2.2. In the following figures, subjective quality ratings (SDG) of some of these test signals (see below) are plotted as functions of the corresponding objective quality values PSM[7]. Each symbol represents a different codec. Several appearances of the same symbol within one diagram represent different conditions of that particular codec (e.g. bit rate, cascading). In each panel, the correlation between subjective and objective quality ratings and thus the performance of the objective quality measure is quantified by the linear correlation coefficient r and the Spearman rank correlation coefficient rs. The different types of audio signals are indicated by the abbreviations *pit* (pitch pipe), *spe* (speech), *bag* (bag pipe), *col* (Ornette Coleman), *cas* (castanets) and *glo* ("glockenspiel", chimes). (The speech item consists in fact of three different speech signals that were pooled.) These signals were selected, because they were used in several listening tests and therefore processed by a larger number of codecs than most of the other signals. Moreover, the subjective quality ratings of these particular signals cover almost the whole range of the subjective rating scale, which is not typical for other signals. Altogether these signals represent 48 % of the entire database.

Results with the modulation filterbank

The results obtained by the auditory model with a modulation filterbank are shown in Figure 2.3. All sub-figures exhibit a monotonic relation between subjective and objective quality ratings. No codec-specific clusters aside the mainstream are observed (except to some small extend codec f in the bag-pipe panel). Linear correlation coefficients range from 0.726 to 0.915, rank correlation coefficients from 0.852 to 0.953. The numbers in brackets state the linear correlation coefficients, if the objective quality values are transformed by the shown regression functions (dashed curves). These fitting functions were obtained by a numerical optimization procedure, individually applied for each type of signal. The fitting function is composed of a hyperbola and a linear function:

$$f(x) = \begin{cases} max\{-4,\ \frac{a}{x-b} + c\} & : \quad x < x_0 \\ d \cdot x - d & : \quad x \geq x_0 \end{cases} \tag{2.1}$$

[7] As mentioned in Section 2.2, SDG values are mean values over all listeners. Standard deviations across listeners are considerably large (0.94 scale units on average), but are not plotted for clarity.

Two of the five parameters depend on the others, because of the constraint that $f(x)$ has to be continuously differentiable. Thus, three free parameters are adjusted by the numerical optimization procedure. The transformed objective quality measure shows good correlation with the subjective ratings: Linear correlation values range from 0.903 to 0.953. Fitting the objective data effects linear correlation the most in these cases, where a floor effect of the subjective ratings at the SDG-scale's lower end can be observed. (See, e.g., *glockenspiel* and *castanets*.) This effect, caused by the limited subjective scale, has to be taken into account.

Note that the abscissae are scaled differently; the mapping function $f : \mathrm{PSM} \mapsto \mathrm{SDG}$ depends on the signal type: Its slope is higher for signals with rapid envelope fluctuations such as *castanets* and *glockenspiel* compared to those for rather stationary signals such as *pitch pipe* and *bag pipe*.

Results with modulation lowpass

Figure 2.4 shows the results obtained by the auditory model with the modulation lowpass filter instead of a filterbank. Compared to the preceding results, the overall performance is poorer: linear correlation coefficients for the transformed data range from 0.765 to 0.946, rank correlation coefficients from 0.619 to 0.927. The PSM-ranges of signals that do not contain very rapid amplitude fluctuations (i.e., all except *castanets* and *glockenspiel*) are compressed by factors up to approximately 16 compared to the ranges obtained by the model with a modulation filterbank (cf. Figure 2.3). One possibility for the reduced range of PSM values in the modulation-lowpass version are the larger differences between internal representations of reference and test signals in the high-frequency modulation channels. To test this hypothesis, the left panel of Figure 2.5 shows mean linear correlations[8] between single modulation channels of internal representations of reference and test signals as a function of the modulation center frequency. Linear correlation coefficients of all signals shown in Figure 2.3 except *castanets* and *glockenspiel* were used for the calculation of the mean. Since the quality measure PSM is given by the weighted sum of the correlation coefficients per modulation channel, the mean associated weights are also shown in the left panel. They are proportional to the mean squared amplitude of the internal representation in the corresponding modulation channel and apparently increase for higher modulation center frequencies. As the examined signals are characterized by rather slow envelope

[8] Here and in the following, mean correlation values are obtained by averaging the Fishers-Z transformed correlation coefficients and subsequent inverse transforming of the resulting mean value: $\langle r \rangle = F^{-1}(\langle F(r) \rangle)$. The Fishers-Z transform maps the interval $[-1, 1]$ to $[-\infty, \infty]$ by the following operation: $F(r) = \frac{1}{2} \ln(\frac{1+r}{1-r})$.

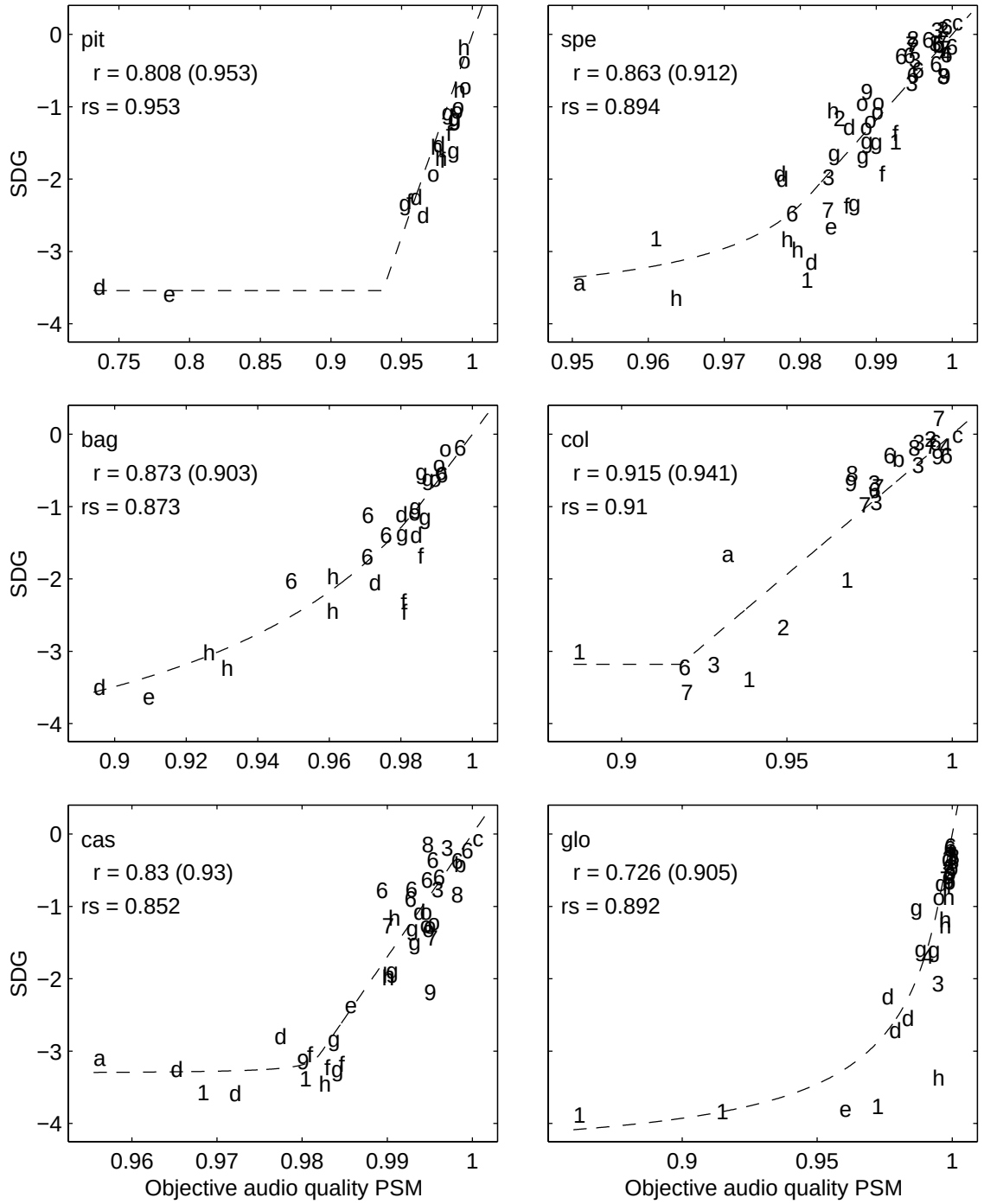

Figure 2.3: Audio quality prediction for six different audio signals (pitch pipe, speech, bag pipe, Ornette Coleman, castanets, glockenspiel). Objective audio qualities were obtained using the modulation filterbank version of the auditory model.

fluctuations, this might appear paradox at first sight. In fact, the power *density* in the modulation frequency domain does decrease for higher modulation frequencies, but this decrease is overbalanced by the increasing bandwidth of the higher modulation filters.

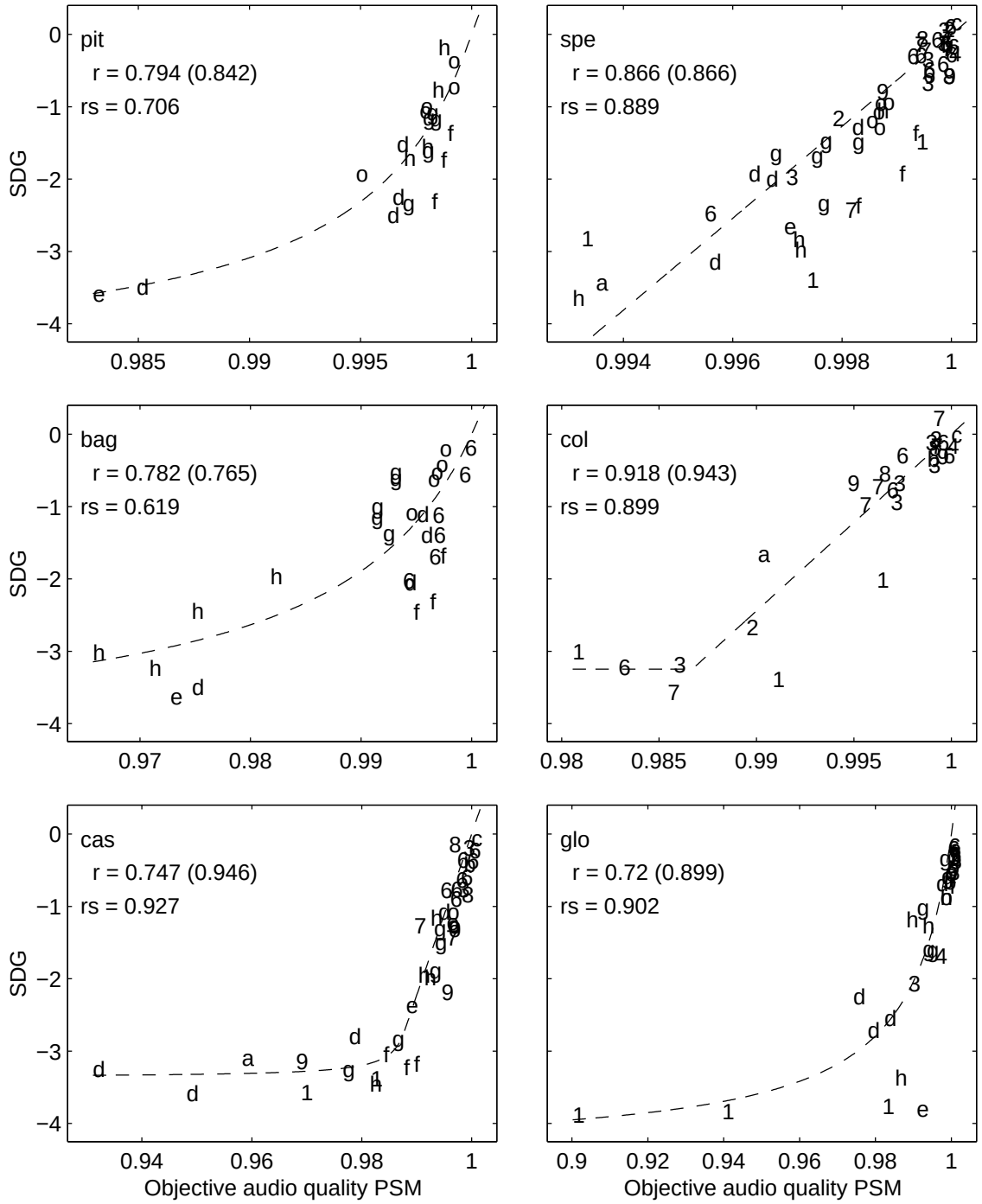

Figure 2.4: Audio quality prediction for the same signals as in Figure 2.3, obtained by the auditory model with a modulation lowpass filter.

Consequently, the PSM values decrease due to the contribution of lower correlation values from the high modulation channels, which confirms the hypothesis stated above.

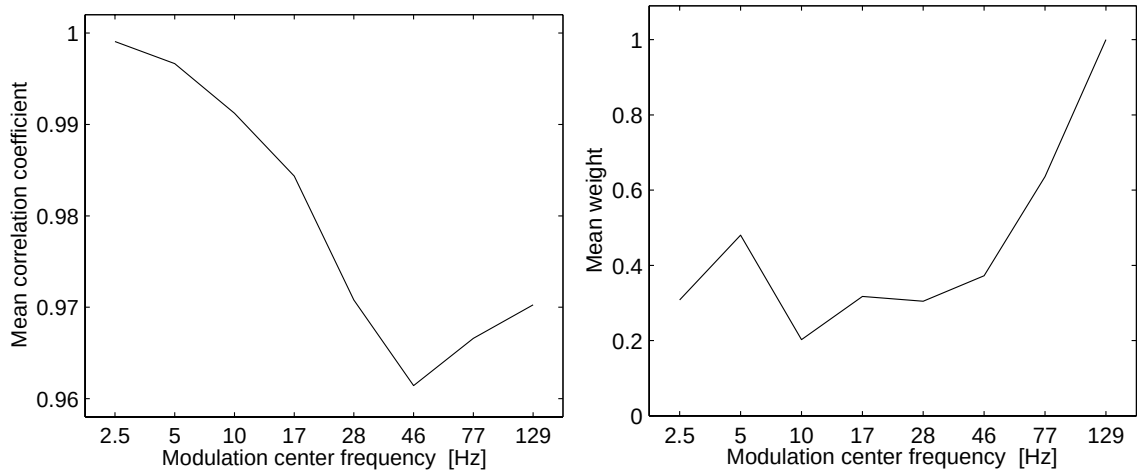

Figure 2.5: Mean correlation coefficients between modulation channels of internal representations of reference and test signals shown in Figure 2.3 (except *castanets* and *glockenspiel*) (left panel) and associated weights (right panel).

Effect of modulation processing

In order to understand the reason for the superior prediction performance of the auditory model version using a modulation filterbank compared to the modulation lowpass version, the influence of the particular way of modulation processing was examined in more detail.

First, the question was addressed how many modulation bandpass filters have to be contained in the filterbank, i.e. what is the highest modulation frequency accounted for, so that the best quality prediction performance is obtained? To answer this question, Figure 2.6 presents the dependency of the prediction performance on the highest modulation center frequency. Again, mean correlation coefficients $\langle r \rangle$ and $\langle rs \rangle$ are obtained for the signals shown in Figure 2.3 as a measure of the average prediction accuracy. They vary non-monotonically with modulation frequency, indicating best prediction performance for modulation filters up to 10 Hz and 129 Hz center frequency, respectively.

The changes of $\langle r \rangle$ and $\langle rs \rangle$ as a function of the modulation frequency range are not striking. However, it should be noted that the prediction performance depends on the modulation frequency range in a more intricate way than is apparent from the correlation coefficients alone. To illustrate this, Figure 2.7 shows quality predictions of the signal *Ornette Coleman*, resulting from two different ways of modulation processing: The results on the left hand side were obtained by applying a modulation filterbank with just three filters (a 2.5 Hz lowpass and two bandpass filters centered around 5 Hz and 10 Hz, respectively), while using eight modulation channels up to 129 Hz center frequency led to

quality predictions displayed on the right hand side. The correlation measures $\langle r \rangle$ and $\langle rs \rangle$ suggest almost identical goodness-of-prediction performance in either case. However, a look at the scatter plots implies that taking more modulation channels into account leads to superior prediction performance: The data in the left panel reveal vertical clusters in the higher quality domain, i.e., the ability of the objective quality measure to resolve small quality differences in that region seems to be reduced. This is not the case if additional information of higher modulation channels is also analyzed, as the results in the right panel show.

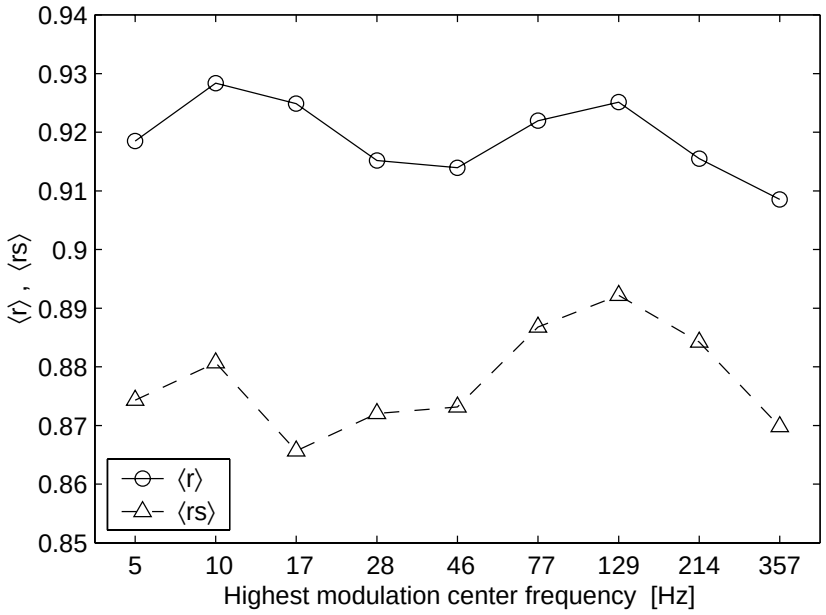

Figure 2.6: Quality prediction performance, quantified by the linear correlation coefficient r and the rank correlation coefficient rs for subjective and predicted quality ratings, as a function of the highest modulation frequency.

As a second question concerning modulation processing, it was investigated whether it is necessary to use a bank of bandpass filters instead of a simple lowpass filter, covering the same frequency range. Figure 2.8 shows the prediction performance for different cutoff frequencies of the modulation lowpass filter. The two solitary items in the upper right corner represent the correlation values that result from applying an eight-channel modulation filterbank covering a frequency range of 160 Hz for comparison. Up to about 30 Hz, the prediction performance is barely influenced by the modulation cutoff frequency, but decreases considerably above that frequency. In none of the cases the prediction performance of the modulation filterbank can be achieved.

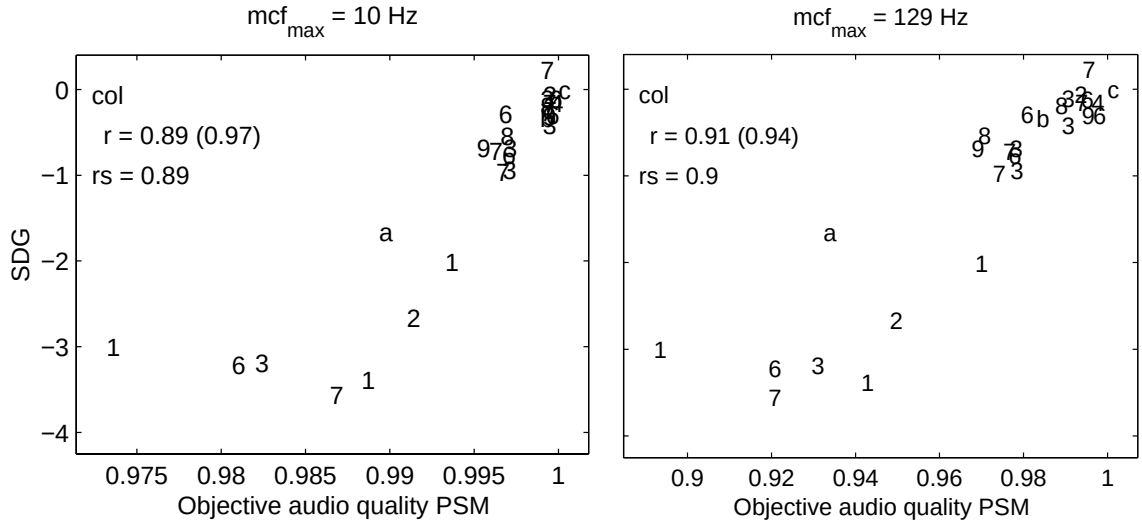

Figure 2.7: Quality predictions for signal *Ornette Coleman*, computed with different modulation filterbanks. Left: three filters, highest cf = 10 Hz; right: eight filters, highest cf = 129 Hz. Note the different scalings of the abscissa.

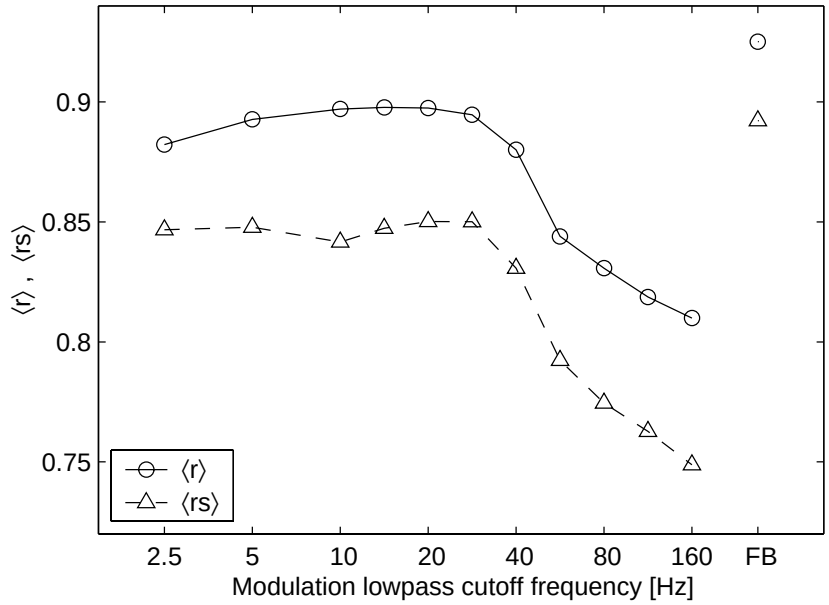

Figure 2.8: Quality prediction performance as a function of the modulation lowpass cutoff frequency. The two solitary items represent results obtained using a modulation filterbank covering a range of 160 Hz.

The main reason for decreasing performance above 30 Hz is found to be the missing ability to distinguish between amplitude modulations and the carrier phase in the low carrier frequency - high modulation frequency domain. This is exemplified by the left panel of Figure 2.9 that shows quality prediction results for speech signals (i.e. three different

speech samples, spoken by a male English speaker, a male German speaker and a female German speaker). One of the items (label "4") is an obvious outlier: The audio quality of the corresponding signal (male speech) is very much underestimated by the objective measure PSM. On the right hand side, a sample time interval of the lowest frequency channel of that signal's internal representation is displayed as well as the corresponding reference signal's internal representation. Both are oscillating with an average period of approximately 4.25 ms, which is the inverse of 235 Hz, the center frequency of that particular frequency channel. Obviously, that channel of the internal representation reflects the carrier phase. Apparently, reference and test internal representation are opposite phase, which dramatically decreases the cross correlation and thus the quality measure PSM, without necessarily impairing the perceived audio quality.

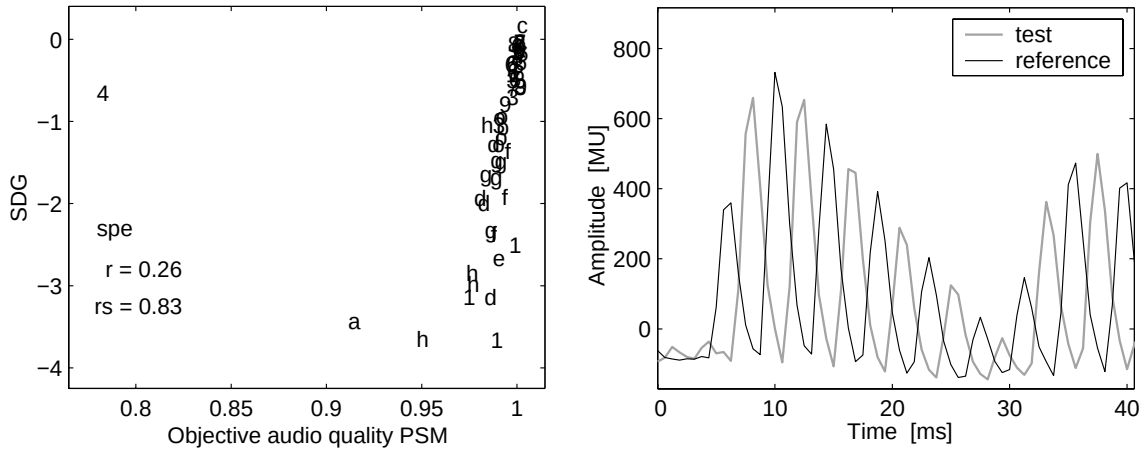

Figure 2.9: Left panel: Quality predictions for speech signals, obtained with a 160 Hz modulation lowpass filter. Right panel: sample time interval of the internal representation's lowest frequency channel of signal "4" from the left plot.

This undesired effect can be avoided if only those carrier frequencies are considered that are clearly higher than the modulation frequency. In case of quality prediction with the multi-modulation-channel model, the quality measure PSM is therefore computed by correlating the internal representations of a given pair of audio signals separately for each modulation channel, while only those frequency channels are considered, in which the center frequencies are greater than four times the respective modulation center frequency. In this way, information from lower frequency channels does not get lost completely, but is just excluded from higher modulation channels. In contrast, this approach can not be realized the same way in the case of a single modulation lowpass channel. There, either information from low frequency channels gets lost or low carrier frequencies are fed through all stages of auditory processing and will be reflected in the internal representation.

Influence of channel weighting for stereo signals

The auditory model used in the presented method is monaural and thus can not account for binaural effects. The lack of a binaural model component was compensated to some extent by calculating PSM for each audio channel independently and taking a weighted sum of the two values to build the final measure:

$$\mathrm{PSM}(final) = \alpha \cdot min\{\mathrm{PSM}(left), \mathrm{PSM}(right)\} + (1-\alpha) \cdot max\{\mathrm{PSM}(left), \mathrm{PSM}(right)\}$$

The effect of the weighting was further investigated. Figure 2.10 presents the mean[9] correlation of subjective and objective quality measures, $\langle r \rangle$ and $\langle rs \rangle$, as a function of the weighting factor α.

The results show that the prediction performance is essentially independent of the weighting factor for the particular sample of audio signals investigated. The mean rank correlation coefficient, $\langle rs \rangle$, indicates a small benefit of taking both audio channels into account. However, the mean linear correlation coefficient, $\langle r \rangle$, does not support this conclusion. In viewing computational effort, this result suggests that evaluating just one audio channel of a stereo signal is reasonable, unless the audio channels have considerable differences.

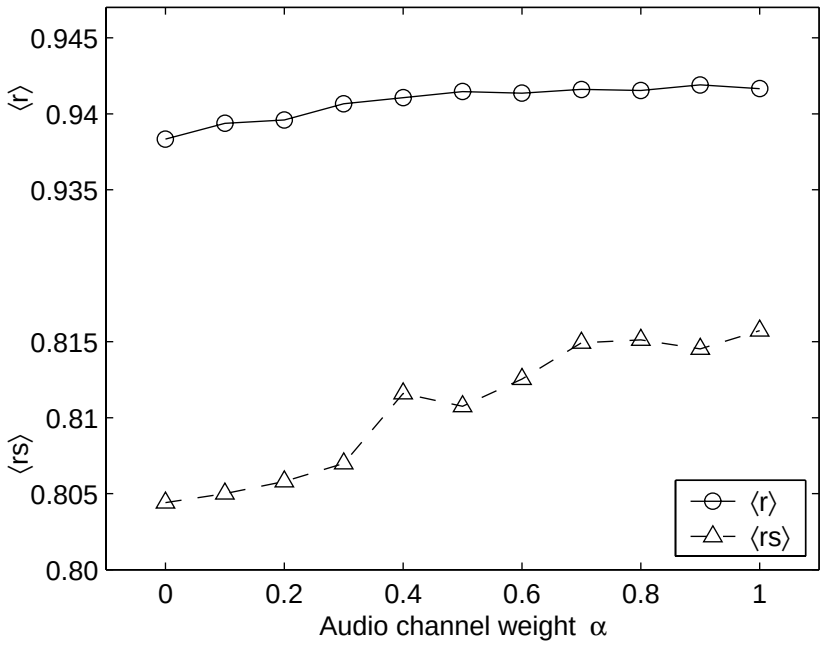

Figure 2.10: Quality prediction performance as a function of the audio channel weighting factor α.

[9] mean over five different audio signals

Effects of cognitive postprocessing

As described in Section 2.3.1, differences between internal representations of reference and test signals are reduced according to their sign. If the difference is defined as $\Delta IR = |IR_{test}| - |IR_{ref}|$, then positive elements of the matrix ΔIR are not altered, while negative elements are reduced by a factor $\beta < 1$. The effect of the choice of β was further investigated. Figure 2.11 shows the prediction performance as a function of the weighting factor β.

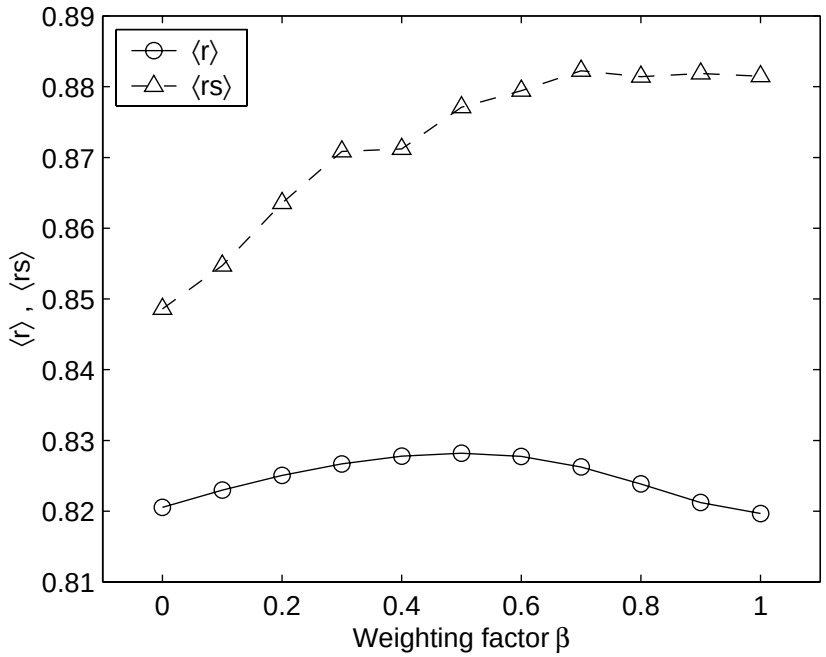

Figure 2.11: Quality prediction performance as a function of the difference weighting factor β.

The results show that reducing negative differences improves the prediction performance. The actual best choice of the weighting factor β is not indicated by $\langle r \rangle$ and $\langle rs \rangle$ consistently: while $\langle r \rangle(\beta)$ becomes maximal at $\beta = 0.5$, $\langle rs \rangle$ increases with β rather asymptotically, suggesting $\beta = 0.7$ to be the best choice. However, due to this asymptotically characteristic, the increase of $\langle rs \rangle$ for $\beta = 0.5$ compared to $\beta = 0$ already amounts to 87% of the maximum increase at $\beta = 0.7$, so choosing any $\beta \geq 0.5$ does not affect $\langle rs \rangle$ very much. Moreover, a choice of $\beta = 0.5$ would be in line with findings of Berger (1998). Resuming the preceding considerations, a choice of $\beta = 0.5$ was finally concluded.

Prediction of average codec quality

The audibility of signal alterations caused by audio codecs depend not only on the particular codec but also on the audio signal. Thus, if the overall transmission quality of a given codec is to be evaluated, it is mandatory to test it using a set of audio signals that are representative for the intended real world application.

Figure 2.12 documents the ability of the PSM to predict the average transmission quality of 22 different codecs or codec conditions, respectively. Each circle represents the mean transmission quality of a different codec (condition), given by the mean value of quality ratings for a set of six audio signals[10]. In the right panel, objective quality ratings were mapped to the subjective quality scale for each signal type individually (using the transforms shown in Figure 2.3) prior to averaging. This procedure improves the prediction performance to some extent.

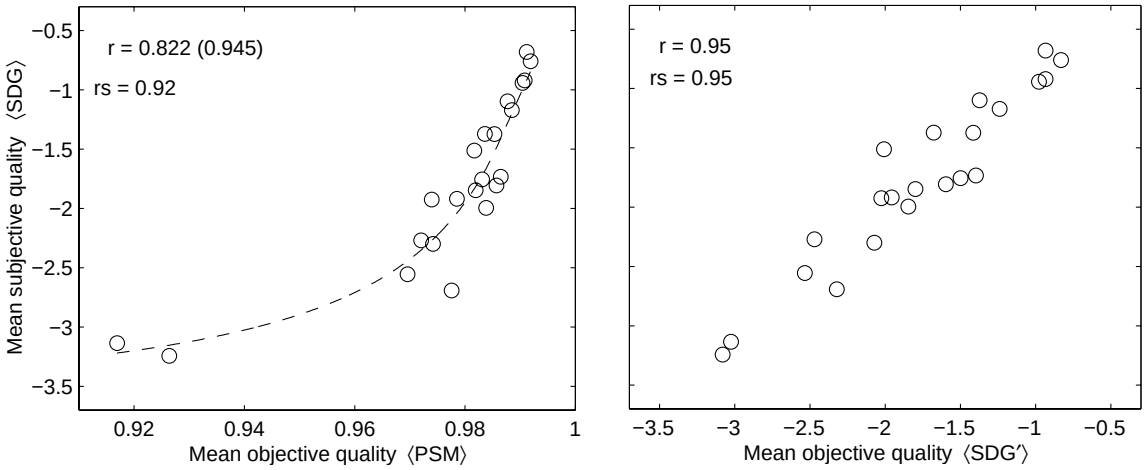

Figure 2.12: Prediction of the average audio quality of six audio signals processed by different codecs. Left: mean objective quality measure ⟨PSM⟩. Right: mean predicted subjective quality measure ⟨SDG'⟩, found by mapping PSM↦SDG' for each signal respectively and then averaging across signals.

Predicting audio quality of different signals

To emphasize the effect of signal dependency on audio quality prediction, Figures 2.13 and 2.14 show results of quality prediction for different audio signals within the same plot.

[10]These 132 items (6 signals processed by 22 codecs) represent the corpus of the database MPEG95. (See appendix A for details about this database.)

In Figure 2.13, quality ratings for the signals *pitch pipe* and *castanets* are depicted[11]. The signal dependency is apparent: ratings cluster along separate lines, both originating from [PSM = 1, SDG = 0], but having different slopes. As mentioned before, signals with rapid envelope fluctuations (such as *castanets*) tend to show steeper slopes in the PSM-SDG-plane than more stationary ones (such as *pitch pipe*). This yields rather poor overall correlations between subjective and objective quality measures (in this case: $r = 0.739$, $rs = 0.746$), while correlations within each signal are much better ($r = 0.93$, $rs = 0.852$ for *castanets*, $r = 0.953$, $rs = 0.953$ for *pitch pipe*, cf. Figure 2.3).

If all audio times of the database used in this study are taken into account to derive an "average" mapping function, the overall correlation of subjective and transformed objective quality ratings improves, but is still not satisfying. Figure 2.14 shows the relationship between the subjective measure SDG and the objective measure PSM for all 439 audio items, covering 28 different types of signals (i.e., 28 different *reference* signals). The overall correlation is quantified by $r = 0.769$ and $rs = 0.737$.

2.3.3 Discussion

The results can be summarized as follows: The relationship between subjective and objective ratings depends on the type of audio signal used. However, it depends rather little on the specific lossy audio processing system tested in the present database. This might indicate that the auditory model weights differences across codec schemes in a similar way as the average normal hearing listener. However, since most of the tested systems are perceptual audio codecs, it might also reflect similar distortion characteristics across audio codecs. Further evaluations using signals with quite different impairments such as linear distortions, reverberation or additive noise seem to be required to decide this question.

Necessity of spectral weighting?

Hansen and Kollmeier (2000) reported a spectral weighting function that improved the prediction performance of their measure q_C for telephone-band-pass filtered speech. This "band importance" weighting function shows a flat characteristic up to 1 kHz and increases with higher center frequencies. It becomes maximal at the highest center frequency (about 3400 Hz), amounting to 10 times its initial value.

[11] Two of the *pitch pipe* items with very small PSM values were omitted in order to allow for a better resolution of the *castanets* data.

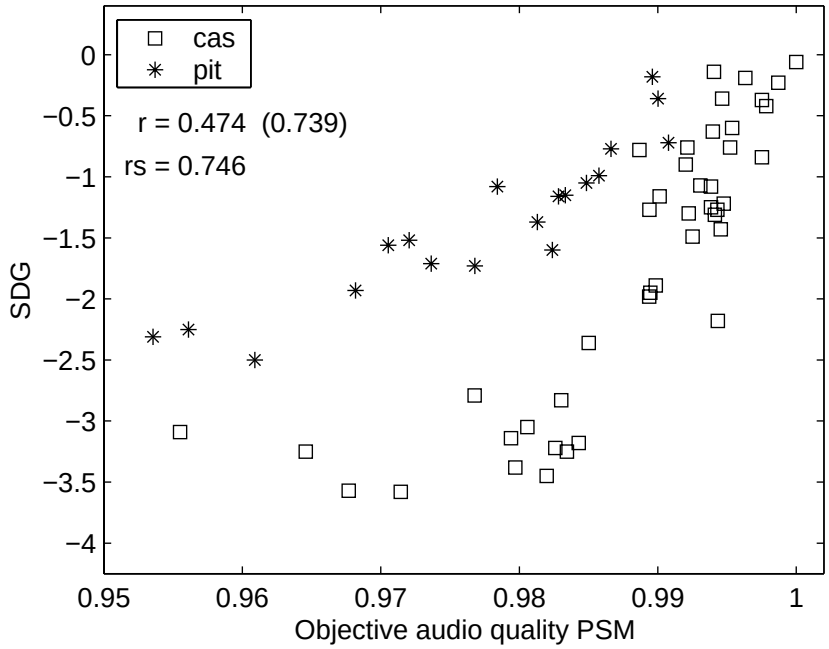

Figure 2.13: Audio quality prediction for the audio signals *pitch pipe* and *castanets*.

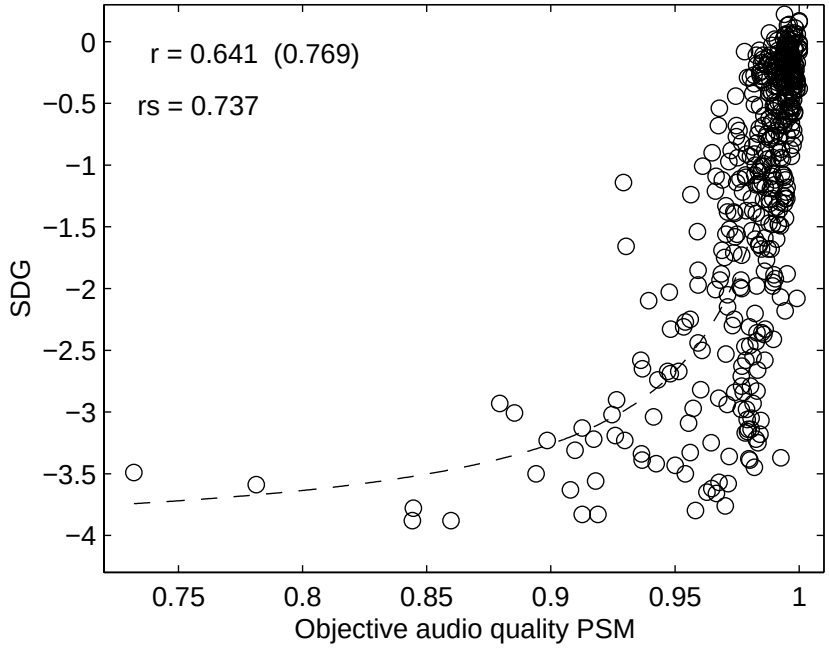

Figure 2.14: Audio quality prediction for all 439 items of the given database.

To account for the higher bandwidth of the high fidelity (reference) audio signals used in the present study, the peripheral filterbank of the auditory model had to be

extended accordingly. As a result, the band importance weighting function could not be adopted directly in the present study. The use of an extended version of the weighting function obtained by extrapolation did not appear reasonable either. Instead, numerical optimizations of band weighting functions with different boundary conditions (including no boundary condition at all) were performed with the aim to maximize the correlation between subjective and objective quality ratings. Optimizations were also performed in the modulation frequency domain. No weighting functions could be found that improved the prediction performance consistently for different data sets, neither in the frequency domain nor in the modulation frequency domain.

Hansen and Kollmeier (2000) argued that emphasizing high frequency bands yields better prediction results, mainly because one of the distorting systems under test was the "Modulated Noise Reference Unit" (MNRU), that produces speech distortions that are essentially spectrally flat. In contrast, spectra of distortions produced by most of the speech codecs are similar to the long term spectrum of speech, so that the distortions are (partially) masked by the speech. Since the latter spectra are not flat but fall with higher frequencies, MNRU produces stronger distortions at high frequencies than typical speech codecs. Without spectral weighting, the quality of subjectively equally rated speech was estimated higher systematically by the speech quality measure, if the speech was distorted by MNRU compared to speech that was distorted by speech codecs. Applying a band weighting function that emphasizes higher frequency bands effects the quality estimates of the MNRU-distorted speech items more than the codec-distorted items. Consequently, the quality estimates of the MNRU-distorted speech are shifted more towards smaller values than the codec-distorted speech, thus shifting the estimates together and improving the overall correlation. They also mentioned an improvement of the correlation between q_C and subjective quality ratings of the codec-processed speech material (i.e., without MNRU distorted speech), if the non-uniform weighting is applied. This could indicate a cognitive effect concerning the importance of the highest frequency bands of band-limited speech for the perception of speech quality. The latter reason for the observed improvement of the overall prediction results, however, was classified as secondary by the authors.

The database used in the present study contained neither MNRU-distorted signals nor band-limited signals. Moreover, only a small fraction of the database consisted of speech signals. This might be the reason why an improving weighting function could not be found in the present case.

Possible reasons for signal dependency

As demonstrated by Figure 2.13 and 2.14, the assessment of audio quality with the objective quality measure PSM depends on the type of signal. Especially signals that differ with respect to the rate of envelope fluctuations are rated differently by PSM. Figure 2.13 indicates that the mapping function f :PSM $\mapsto$ SDG is steeper for signals with rapid fluctuations (here: *castanets*) than for rather stationary signals (here: *pitch pipe*). In other words: The more rapid the fluctuations, the more the quality is overestimated by PSM. To investigate possible reasons, Figure 2.15 compares the internal representations of the audio signals mentioned above on the left hand side. (Only the highest frequency channel of the highest modulation channel is depicted.)

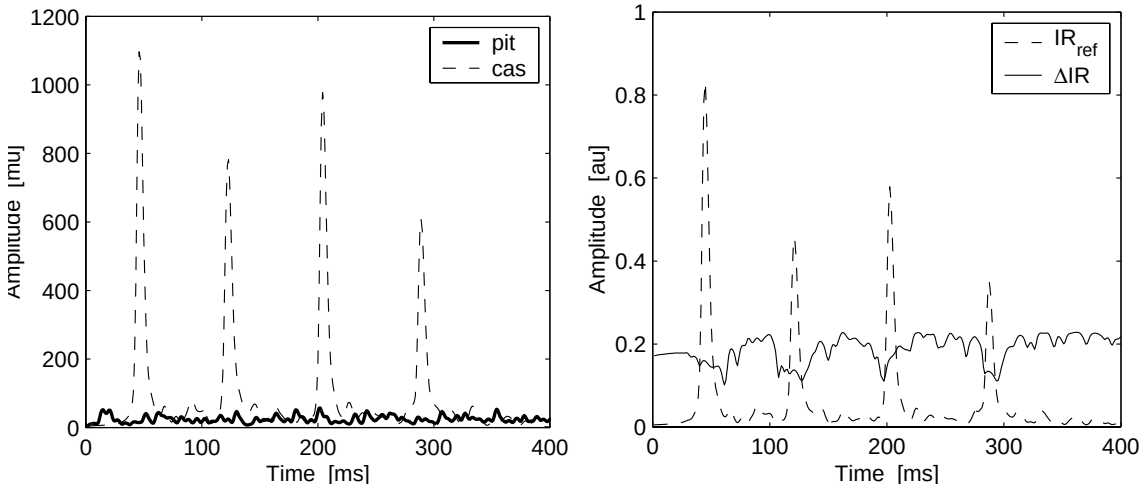

Figure 2.15: Left: Sample interval of the internal representations (highest frequency channel, highest modulation channel) of the signals *pitch pipe* and *castanets*. The audio signals had equal peak levels. Right: Internal representation of the signal *castanets* and the relative difference between internal representations of the reference and a test signal.

The effect of contrasting envelope fluctuations by the adaptation loops of the auditory model becomes very apparent: Although both signals have the same physical peak level, the maximum of the internal representations of the *castanets* signal exceeds that of the *pitch pipe* signal by a factor of about 20. In the calculation of cross-correlation between internal representations like that of the *castanets*, these prominent peaks contribute superproportional with respect to their temporal extent. Possible differences aside these peaks are therefore weighted much less. This will cause higher overall correlation values, if differences aside the peaks are greater than within. The right panel of Figure 2.15 shows that in fact this is the case for the *castanets* signal distorted by one of the audio codecs.

The relative difference between the internal representations of reference and test signal (i.e., $\Delta\mathrm{IR} = |(\mathrm{IR}_{\mathrm{test}} - \mathrm{IR}_{\mathrm{ref}})/\mathrm{IR}_{\mathrm{ref}}|$) is smaller within the peaks than outside.

This finding suggests that if the emphasizing characteristic of the auditory model concerning envelope fluctuations was more moderate, objective quality differences between stationary and more fluctuating audio signals should decrease.

Alternatively (or additionally), the differences of contribution to the overall quality measure between internal representation segments with different amplitudes would be lessened, if the internal representations were not cross-correlated at once, but in short temporal sections. This would yield a sequence of short-time correlation values that could be mapped to a single overall value, e.g. by averaging or a different operation. The contributions of particular segments to the overall value would equalize with a decreasing length of the correlation interval. However, this effect could possibly be accompanied with the drawback of the model losing its ability to account for temporal forward masking, if the interval length falls below 200 ms.

The idea of computing a sequence of short-time correlation values instead of one overall correlation coefficient is supported by the consideration of a second possible reason for signal dependency, which might be found in another cognitive aspect of subjective quality assessment: the relationship between instantaneous and overall perceived audio quality. It is known from other fields of psychophysics that this relationship is highly complex. Human observer tend to rather focus on extreme occurrences in the temporal course of a considered psychophysic quantity than to integrate over time linearly (Fastl, 1994; Hamberg and de Ridder, 1999). Additionally, accounting for memory effects might also be essential (Hamberg and de Ridder, 1999). Figure 2.16 exemplarily depicts the instantaneous objective audio quality as a function of time for the two audio items mentioned above. The instantaneous audio quality, PSM(t), was obtained in this case by successive cross correlation of 50 ms frames of the internal representations, so that $t = n \cdot 50$ ms, $n = 1, 2, \ldots$. The overall audio qualities of the depicted items were rated equally by human listeners. In contrast, temporal mean values of PSM(t), denoted by dashed lines, are apparently very different ($\Delta = 0.23$). This suggests that either the interval length of short-time cross-correlation is still too long, or the mean value of PSM(t) is not an appropriate measure for the perceived overall audio quality, or both. (In fact, PSM and not $\langle \mathrm{PSM}(t) \rangle$ serves as the objective overall quality measure so far, and PSM and $\langle \mathrm{PSM}(t) \rangle$ are not mathematically equivalent. However, both measures are highly correlated ($r = 0.96$ for the present database), so the above-mentioned argument concerning $\langle \mathrm{PSM}(t) \rangle$ also applies to PSM.)

The possible qualification of these suggestions for improving the signal-independent audio quality prediction will be investigated in the second part of this chapter.

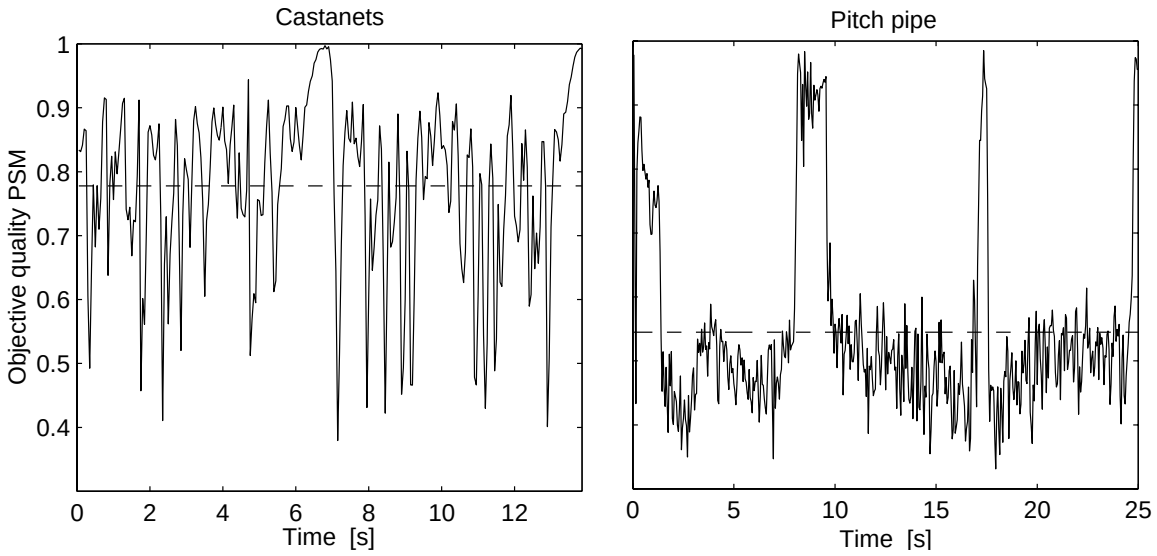

Figure 2.16: Temporal courses of the instantaneous audio quality (solid lines) and corresponding mean values (dashed lines) for the audio signals *castanets* (left) and *pitch pipe* (right). The overall audio qualities of both signals were rated equally by human listeners.

2.3.4 Signal-independent quality prediction

The method for the objective assessment of audio quality differences introduced so far is able to predict subjective ratings well, if the type of audio signal is taken into account. Different types of audio signals may show different relations between subjective and objective measures, especially if they differ with respect to their degree of stationarity. Two possible reasons were suggested in Section 2.3.3. In this section, a modified version of the objective quality assessment method will be presented that accounts for these reasons in order to achieve independency of audio signal characteristics.

Methodical expansions

The findings reported in Section 2.3.3 suggested that a measure with an improved ability to predict audio quality signal-independently should be based on the sequence of short-time correlation values (the instantaneous audio quality), emphasizing low values. For this purpose, the instantaneous objective audio quality, PSM(t), is computed by successive cross correlation of 10-ms frames of the internal representations. Subsequently, PSM(t)

is weighted by the lowpass filtered moving RMS value[12] of the internal representation of the corresponding test signal, in order to account for a slowly varying "loudness". From this weighted time series, the 5%-quantile is extracted and serves as a new measure for the overall audio quality. This quality measure is called PSM_t. In order to be able to predict the subjective rating according to the SDG scale, PSM_t is mapped to an SDG-like scale by a regression function derived from a numerical fitting procedure. The mapping function is of the type presented in Equation (2.1) and reads as

$$\mathrm{ODG}(\mathrm{PSM_t}) = \begin{cases} max\{-4, \ -0.22/(\mathrm{PSM_t} - 0.98) - 4.13\} & : \ \mathrm{PSM_t} < 0.864 \\ 16.4 \cdot \mathrm{PSM_t} - 16.4 & : \ \mathrm{PSM_t} \geq 0.864 \end{cases} \tag{2.2}$$

Following the nomenclature of the subjective scale, the name of the final value was chosen to be ODG (*Objective Difference Grade*).

A block diagram of the expanded method is shown in Figure 2.17.

2.3.5 Signal-independent audio quality prediction results

The new quality measure PSM_t was calculated for all items of the database described in Section 2.2. Its applicability to predict the perceived degradation of audio quality, for individual signals as well as for mixed signals, was tested.

Figure 2.18 shows the prediction results for all 439 audio items contained in the present database. The linear correlation (r) between PSM_t and the subjective ratings (SDG) amounts to 0.7, rank correlation (rs) to 0.86. Again, the discrepancy between these correlation measures is due to the "saturation" effect of the subjective ratings at the bottom of the limited subjective quality scale. If this is taken into account by mapping PSM_t according to Equation (2.2), the linear correlation of the transformed measure and the subjective data increases to 0.895. Figure 2.19 shows the prediction results with the transformed objective quality scale.

Signal-dependent quality prediction with PSM_t - comparison with PSM

As shown in Section 2.3.2, good signal-dependent quality prediction is achieved by PSM, whereas its rather poor performance concerning signal-independent quality prediction led to the development of the measure PSM_t. This new measure is clearly superior to PSM regarding the prediction performance for mixed types of audio signals. In this section the

[12] Lowpass filtered by a 1 Hz, 10. order FIR filter; RMS values are computed for 100 ms segments respectively.

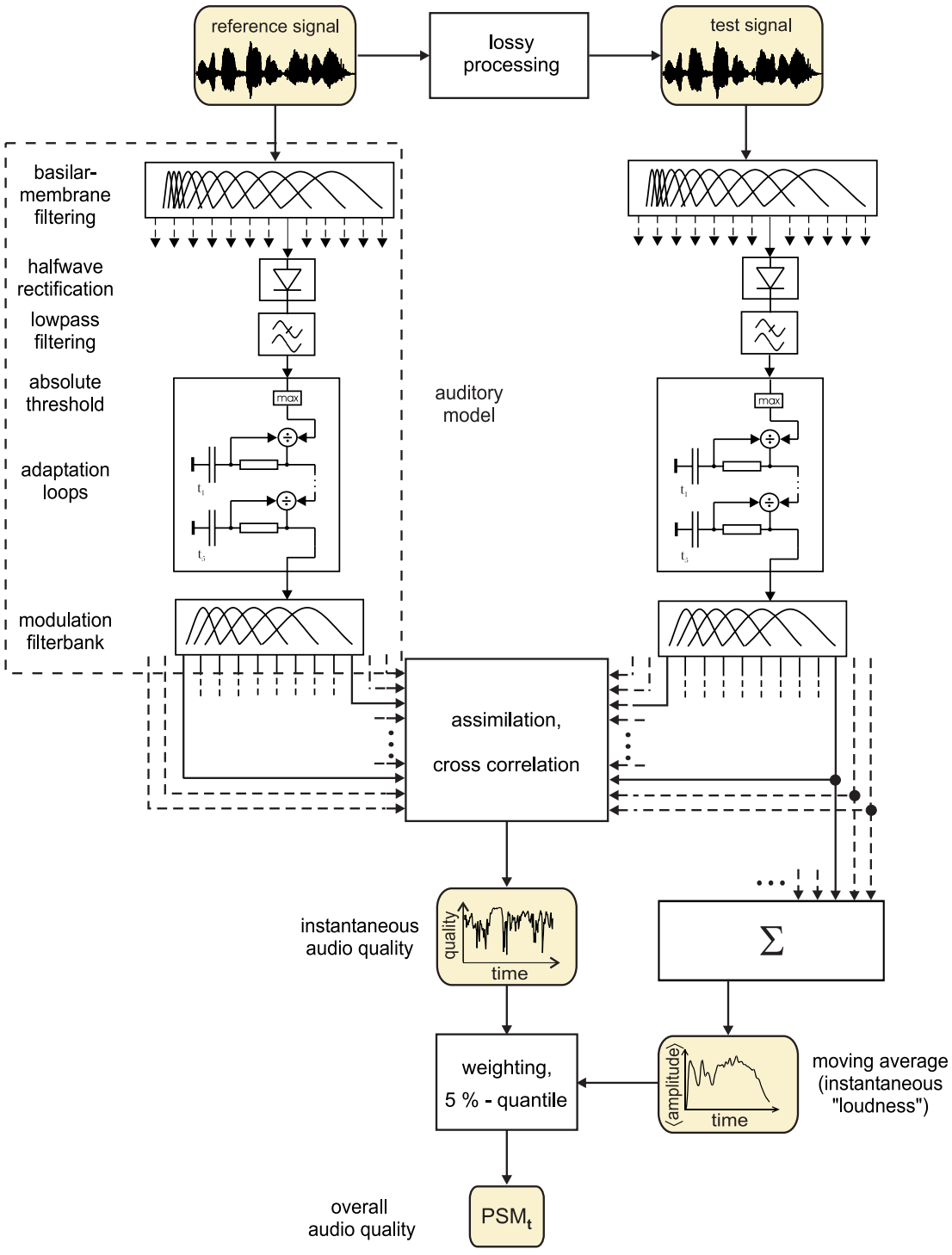

Figure 2.17: Block diagram of the expanded method for objective audio quality estimation.

question is addressed whether the prediction performance of PSM_t for individual signals is comparable to the good performance of PSM.

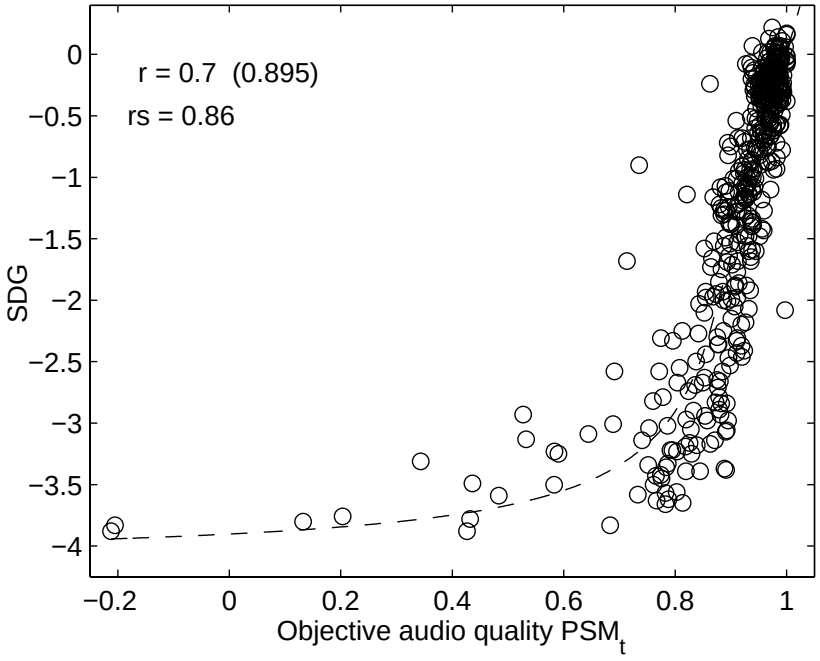

Figure 2.18: Quality prediction performance of PSM_t for all signals.

Results for signal-dependent and -independent quality prediction, expressed as mean correlation values[13], are given in Table 2.1. PSM and PSM_t were mapped to ODG values by either applying individually optimized mapping functions for each signal respectively, or by using a universal mapping function. ("Universal" with respect to different signals, but depending on the respective measure.) The new measure PSM_t is found to be superior in both categories, i.e. signal-independent and signal-dependent quality prediction. It even performs slightly better in the prediction of signal-dependent quality, if the universal mapping function (Equation (2.2)) is applied for all kinds of signals, compared to PSM transformed by signal-dependent optimized mapping functions.

measure	separate signals		mixed signals
	indiv. mapping	univ. mapping	
ODG(PSM_t)	0.930 *0.900*	0.920 *0.900*	0.895 *0.860*
ODG(PSM)	0.921 *0.891*	0.902 *0.891*	0.769 *0.737*

Table 2.1: Mean linear correlation and rank correlation (*italic*) coefficients for transformed quality measures PSM, PSM_t and subjective quality ratings.

[13] As before, correlation values for separate signals were averaged across six signals.

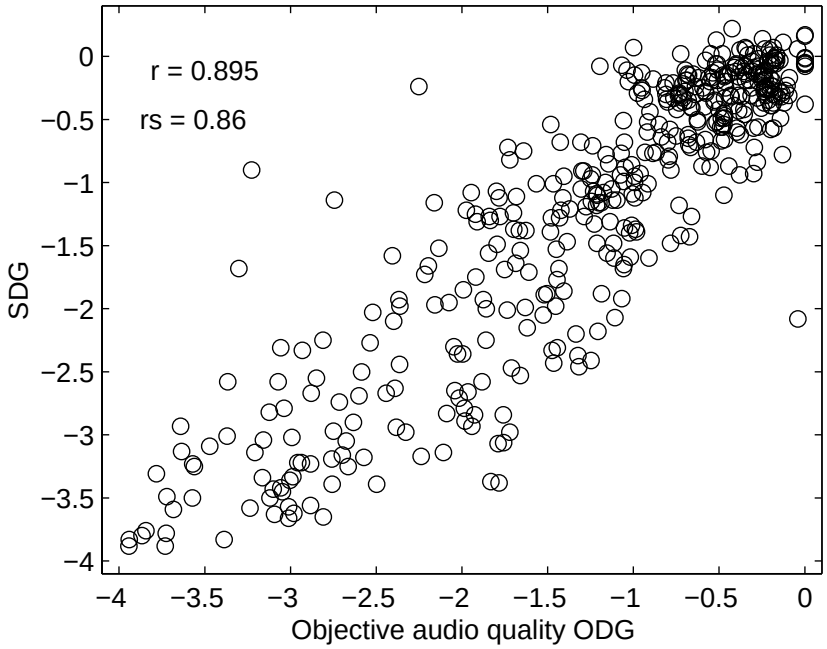

Figure 2.19: Quality prediction performance of the mapped parameter $\mathrm{PSM_t}$, named ODG, for all signals.

Prediction performance of alternative parameters extracted from PSM(t)

In order to evaluate the optimum weighting of the instantaneous audio quality PSM(t) across time, several statistical parameters accentuating low quality values were computed and analyzed with respect to their applicability to predict the subjective overall audio quality ratings. These parameters were: the 0.03-, 0.04-, 0.05- , 0.07- ,0.1-, 0.25- and 0.5-quantiles (denoted as $Q_{0.03}, Q_{0.04}$,..., respectively), the mean value $\langle \mathrm{PSM}(t) \rangle$, the mean values of nonlinear weighted PSM(t), with powers of 1.5 and 2, (in fact, $1-\langle [1-\mathrm{PSM}(t)]^n \rangle^{\frac{1}{n}}$, with $n = 1.5, 2$ has to be taken, because *small* values are to be accentuated) and finally the mean value, the median and the lower quartile of the set of local PSM(t)-minima (i.e., minima of PSM(t)-intervals with lengths of 0.1 s to 2 s, depending on the respective time resolution of PSM(t), see below), denoted as $\langle \{loc.min.\} \rangle$, $Q_{0.5}\{loc.min.\}$ and $Q_{0.25}\{loc.min.\}$, respectively. Moreover, the influence of the time resolution of $\mathrm{PSM}(t) = \mathrm{PSM}(n\tau), n = 1, 2, ...$, i.e., the frame length of short time cross-correlation of the internal representations τ was also examined. For this purpose, τ was varied from 5 ms to 500 ms.

Table 2.2 shows linear and rank correlation coefficients of the listed measures[14] with subjective quality ratings. All measures were fitted to subjective data by nonlinear map-

[14]To reduce the table size, only τ values up to 100 ms are shown. Correlation values of all measures obtained with $\tau > 100$ ms are smaller than obtained with $\tau = 100$ ms and decrease continuously for higher τ values.

ping functions of the type given in Equation (2.1), but with individually optimized parameters. The results show that the best prediction performance is achieved by taking either $Q_{0.05}$, the 5%-quantile of PSM($n\tau$), with $\tau = 10$ ms (according to the values of linear correlation), or $Q_{0.25}\{loc.min.\}$, the lower quartile of the set of local (1.4 s) minima, with $\tau = 25$ ms (according to the values of rank correlation). The correlation differences between these two measures are very small, especially if linear correlations are considered, indicating a slightly better performance of $Q_{0.25}\{loc.min.\}$. Despite this finding, $Q_{0.05}$ was chosen to serve as the final measure for overall quality, $\mathrm{PSM_t}$, because of its simplicity and stronger correspondence to similar measures reported in the literature (Fastl, 1994).

The results also indicate that the benefit of using measures based on a time series of instantaneous audio quality is mainly due to cross-correlating internal representations only over short time intervals instead of calculating the overall correlation coefficient: The mean value of the instantaneous audio quality, $\langle \mathrm{PSM}(t) \rangle$, already shows a considerable better prediction performance than the overall correlation coefficient PSM ($r = 0.870$ vs. $r = 0.767$). Measures that emphasize lower values of the instantaneous audio quality additionally improve the prediction accuracy, because they account for (assumed) cognitive aspects of human audio quality assessment. However, this additional improvement achieved by using such measures is much smaller compared to the former improvement; for example, if the most successful measure, the 5%-quantile, is applied, the correlation coefficient further increases from $r = 0.870$ to $r = 0.897$. (The Fishers-Z-transformations of these correlation values yield $Z(0.767) = 1.013, Z(0.87) = 1.333$ and $Z(0.897) = 1.371$.)

Alternative approaches to signal-independent audio quality prediction - model modifications

If the signal dependency of the quality measure PSM is in fact mainly caused by the strong emphasis of envelope fluctuations by the auditory model as suggested in Section 2.3.3, any reduction in this effect should decrease objective quality differences between slowly and rapidly fluctuating audio signals. (Since the enhancement of rapid envelope fluctuations is a major property of the auditory model being used, any change in this characteristic would also call for a modification of the auditory model.)

In order to test this hypothesis, two alternative versions of the auditory model were implemented and tested with respect to their applicibility for signal dependent and independent audio quality prediction.

measure	$\tau = 5$ ms	$\tau = 10$ ms	$\tau = 25$ ms	$\tau = 50$ ms	$\tau = 100$ ms
$Q_{0.03}$	0.869 *0.839*	0.870 *0.843*	0.867 *0.843*	0.853 *0.833*	0.835 *0.818*
$Q_{0.04}$	0.884 *0.846*	0.884 *0.849*	0.877 *0.847*	0.863 *0.838*	0.854 *0.833*
$Q_{0.05}$	0.894 *0.857*	**0.895** *0.860*	0.893 *0.860*	0.880 *0.853*	0.856 *0.837*
$Q_{0.07}$	0.894 *0.851*	0.894 *0.854*	0.890 *0.854*	0.877 *0.847*	0.850 *0.831*
$Q_{0.1}$	0.888 *0.839*	0.889 *0.842*	0.886 *0.844*	0.870 *0.836*	0.847 *0.823*
$Q_{0.25}$	0.859 *0.805*	0.859 *0.808*	0.857 *0.811*	0.848 *0.808*	0.833 *0.799*
$Q_{0.5}$	0.817 *0.761*	0.816 *0.762*	0.813 *0.763*	0.805 *0.761*	0.774 *0.735*
$Q_{0.25}\{loc.min.\}$	0.877 *0.851*	0.885 *0.862*	0.893 ***0.871***	0.894 *0.859*	0.883 *0.851*
$Q_{0.5}\{loc.min.\}$	0.852 *0.830*	0.867 *0.839*	0.883 *0.850*	0.884 *0.841*	0.857 *0.824*
$\langle\{loc.min.\}\rangle$	0.872 *0.846*	0.880 *0.858*	0.894 *0.868*	0.891 *0.863*	0.876 *0.846*
$1 - \langle[1 - \mathrm{PSM}(t)]^{\frac{3}{2}}\rangle^{\frac{2}{3}}$	0.879 *0.835*	0.879 *0.836*	0.874 *0.833*	0.858 *0.823*	0.832 *0.801*
$1 - \langle[1 - \mathrm{PSM}(t)]^{2}\rangle^{\frac{1}{2}}$	0.863 *0.822*	0.861 *0.822*	0.852 *0.815*	0.831 *0.800*	0.802 *0.776*
$\langle\mathrm{PSM}(t)\rangle$	0.870 *0.819*	0.870 *0.822*	0.868 *0.822*	0.856 *0.816*	0.832 *0.800*
PSM	0.769 *0.737*	0.769 *0.737*	0.769 *0.737*	0.769 *0.737*	0.769 *0.737*

Table 2.2: Linear correlation and rank correlation (*italic*) coefficients for parameters extracted from PSM(t) and subjective quality ratings for all database items, computed for different time resolutions of PSM(t) = PSM($n\tau$). (The objective overall quality measure PSM is also shown for comparison.) Q_x: x-quantile; $\langle\cdot\rangle$: mean value; $\{loc.min.\}$: set of local PSM(t)-minima

The first modification consisted of an additional processing stage subsequent to the haircell stage, that performs an instantaneous compression of the envelope signal. Figure

2.20 shows a block diagram of the modified model. Instantaneous compression was realized by applying the function $f(x) = x^{\gamma}$, with $0 < \gamma < 1$.

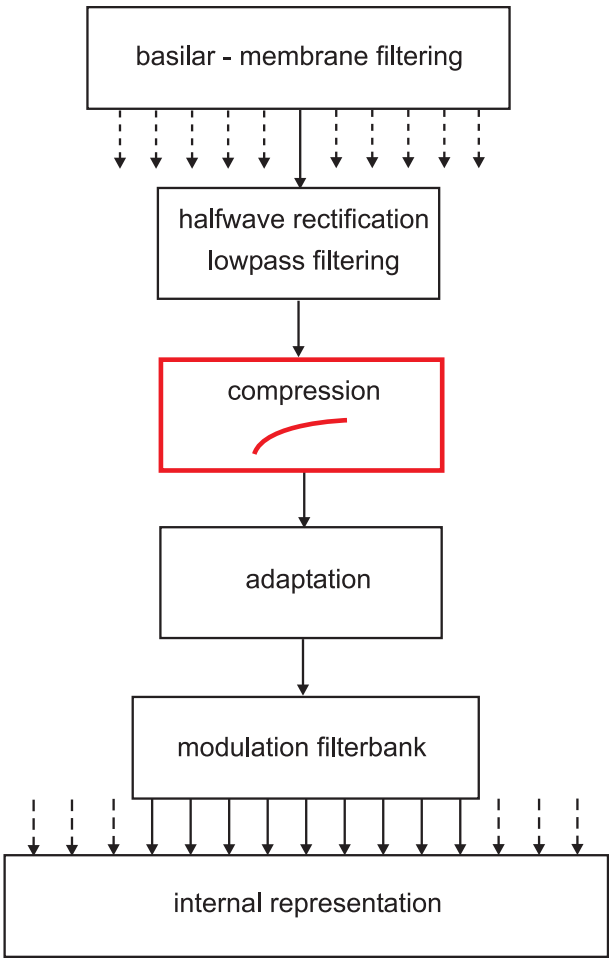

Figure 2.20: Block diagram of the modified model version 1. An additional stage applying an instantaneous compression was inserted between the haircell and the adaptation stage.

Figure 2.21 shows the prediction performance as a function of the compression exponent γ. $\langle r \rangle$ and $\langle rs \rangle$ denote the mean linear and rank correlation coefficient averaged across six different types of audio signals (left panel), while r and rs denote the corresponding correlation values if quality ratings for all types of audio signals are considered together (right panel). As expected, the overall correlation improves with a reduced contrast of envelope fluctuations: r and rs increase with decreasing γ, i.e. stronger compression, unless γ is smaller than 0.1. The maximum prediction performance according to the linear correlation is achieved for $\gamma = 0.15$ ($r = 0.893$). The course of the rank correlation coefficient indicates the best performance for $\gamma = 0.1$ ($rs = 0.893$).

The effect on signal-wise correlation, however, is contrary, but less severe, as shown in the right panel of Figure 2.21: Here, the prediction performance is best, if no instantaneous compression at all ($\gamma = 1$) is applied.

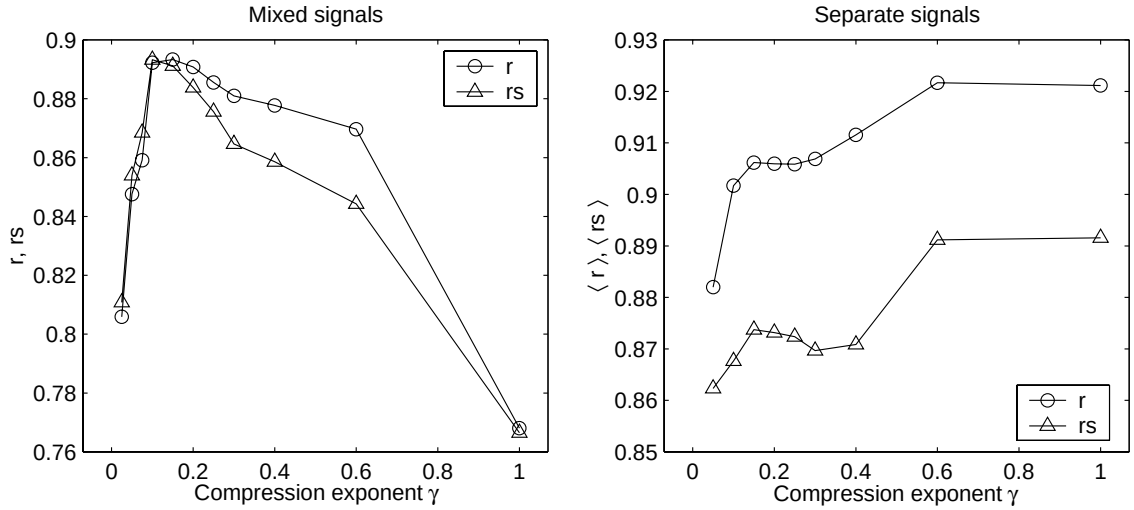

Figure 2.21: Quality prediction performance as a function of the exponent of instantaneous compression γ. Left: correlations between subjective and objective ratings for mixed signal types. Right: mean correlations for separated signal types. (Note the different scalings of the ordinates.)

Compared to the quality measure $\mathrm{PSM_t}$ presented in Section 2.3.5, PSM obtained by the presented modified model shows almost equal prediction performance for mixed signals, if a very small value of 0.15 for the exponent of the additional compression stage is chosen ($\langle r \rangle = 0.892$ vs. $r = 0.897$). On the other hand, predicting the perceived audio quality separately for different signal types is worse with this approach than it is for $\mathrm{PSM_t}$ ($\langle r \rangle = 0.905$ vs. $\langle r \rangle = 0.930$).

With this additional compression, the exponent of the overall compression of the auditory model for stationary input signals amounts to approximately 0.0047 (0.15 introduced by the additional stage, $\frac{1}{32} \approx 0.03$ due to the adaptation loops), which does not appear reasonable. Moreover, Derleth et al. (2001) made a similar modification of the present model by adding a fast-acting compressive adaptation loop subsequent to the haircell stage. They reduced the number of the original adaptation loops from five to three in order to keep the overall compression similar for both the modified and original model. Nevertheless, their modified model failed in predicting modulation-matching experiments, indicating that such a modification does not improve the prediction performance for psychoacoustic experiments.

It is noteworthy that Beerends and Stemerdink (1994) also found an exponent for loudness compression that is much below the value which is found psychoacoustically for loudness of stationary sounds (0.001 instead of 0.23), but maximizes the correlation between subjective and objective speech quality data obtained from their perceptual speech quality measure (PSQM). The authors explained that *"this compression value made the PSQM more robust against implosive noises"*.

In a second modified model version, an alternative approach to reduce the contrast between slowly and rapidly fluctuating signal envelopes was realized by limiting the output of the adaptation loops. This was already proposed by Münkner (1993) in order to improve the model's ability to predict the perceived loudness of non-stationary sounds. Münkner restricted the maximum output of each adaptation loop to ten times the value of its steady state response by introducing an additional stage with a compressive characteristic between the divider and the lowpass element (behind the devider, cf. Figure 2.22). This function reads as

$$f(x) = \begin{cases} x, & x \leq M \\ \frac{2C}{1+e^{-2(x-M)/C}} + M, & x > M, \end{cases}$$

with

$$C = L(M - T) - M.$$

T denotes the amplitude of the input signal at threshold, M the maximum amplitude of the steady state response and L the limit factor ($= 10$ in (Münkner, 1993)) of the output (so that $f(x) \leq L \cdot M$). This function has a linear characteristic for input values within the maximum range for stationary inputs and a sigmoidal characteristic for higher values[15].

The effect of the choice of the limit factor on the quality prediction performance was analyzed. Figure 2.23 shows the prediction results of the modified model as a function of the limit factor L. As expected, the overall correlation is improved by the reduced enhancement of rapid envelope fluctuations due to the limitation (left panel): r and rs increase with decreasing L unless L is smaller than 2.5.

In contrast, the effect on the signal-wise correlation is reverse (right panel): In this case, the correlation deteriorates by decreasing L and the best signal-wise prediction performance is achieved with no limiting at all ($L = \infty$).

[15]The modified adaptation loops were also adopted by Dau et al. (1997a).

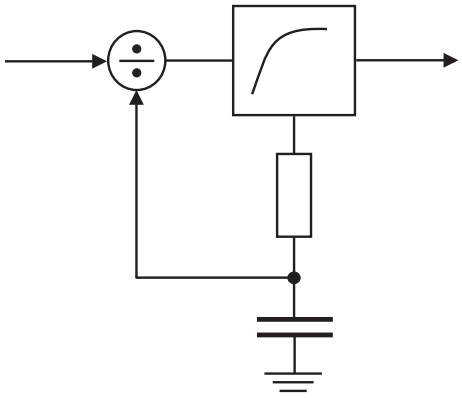

Figure 2.22: Block diagram of the modified adaptation loop proposed by Münkner (1993).

Compared to the measure $\mathrm{PSM_t}$, the quality prediction obtained with PSM using the second modified model version with $L = 2.5$ is worse: the linear correlation coefficient of subjective and objective quality ratings for mixed signals is $r = 0.834$ compared to $r = 0.895$ obtained with $\mathrm{PSM_t}$, and for separated signals on average $\langle r \rangle = 0.896$ compared to $\langle r \rangle = 0.930$.

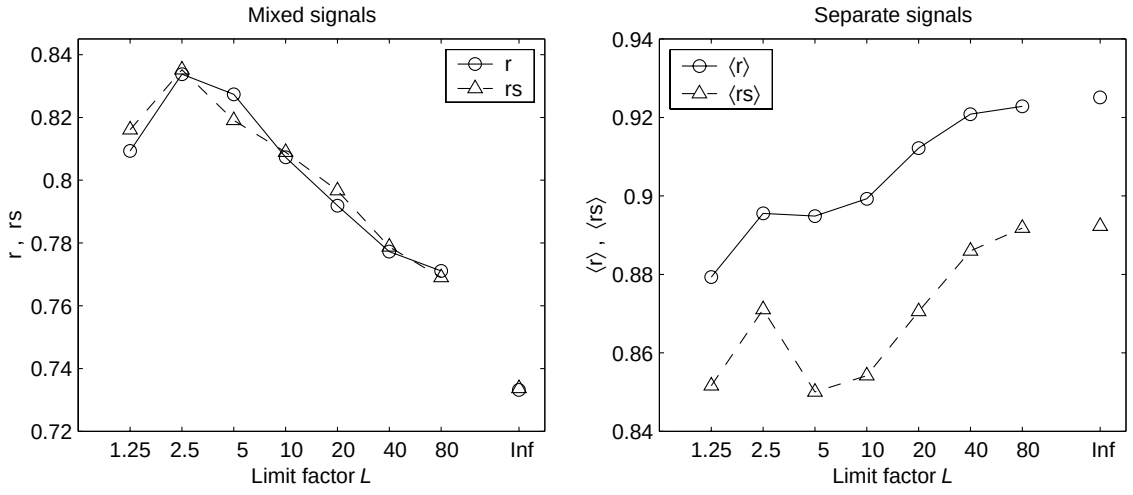

Figure 2.23: Quality prediction performance as a function of the limit of the adaptation loop output. Left: correlations between subjective and objective ratings for mixed signal types. Right: mean correlations for separated signal types. (Note the different scalings of the ordinates.)

2.3.6 Discussion

Possible limitations of $\mathrm{PSM_t}$

The evaluation of internal representations on time scales that are considerably shorter than time constants associated with temporal forward masking (i.e., about 200 ms) is

likely to result in a decreased ability to account for this effect. In view of this, it seems remarkable that the best quality prediction performance is achieved by a measure that is based on short-time cross-correlations of internal representations using a frame length of just 10 ms. Apparently, forward masking merely plays a minor role in the audio quality assessment of everyday audio signals. This might be explained by the fact that most natural sound sources, including non-electronic musical instruments, have decay times of at least some 100 ms. Moreover, most audio signals, music in particular, are either recorded in reverberating environments or reverberated artificially. Thus, sharp offsets are hardly observed in everyday audio signals, such as those contained in the database used in the present study. This explanation is also supported by the fact that the most common artifacts introduced by audio codecs are the so-called pre-echoes, i.e. short bursts of noise preceding sharp signal onsets, in contrast to "real" echoes, which are hardly reported. If this hypothesis is true, $\mathrm{PSM_t}$ should perform worse if subjective ratings of time reversed codec-distorted audio signal are to be predicted.

Another potential limitation concerns a possible dependency of $\mathrm{PSM_t}$ on the kind of the audio distortion and was already mentioned in Section 2.3.3 in connection with the audio quality measure PSM: Based on the present database, it can not be concluded definitely whether the observed rather small dependency on the particular audio codec is due to a similar weighting of differences across codec schemes by human listeners and the auditory model, or if it rather reflects similar distortion characteristics of the concerned audio codecs. Thus, it can not definitely be precluded that different relations between $\mathrm{PSM_t}$ and subjective ratings could possibly be found for quite different impairments of audio signals, such as linear or harmonic distortions, reverberation or additive noise.

Comparison of $\mathrm{PSM_t}$ and PEAQ

In this section the audio quality prediction obtained by the presented measure $\mathrm{PSM_t}$ is compared with the ITU-R standard BS.1387 for "perceptual evaluation of audio quality" (PEAQ) (ITU-R, 1998a).

PEAQ was already introduced in Section 2.1. It *"combines concepts and output variables of most previously known measurement methods of this nature"* (Thiede et al., 2000), namely: DIX (Thiede and Kabot, 1996), NMR (Herre et al., 1992), OASE (Sporer, 1997), PAQM (Beerends and Stemerdink, 1992), PERCEVAL (Paillard et al., 1992; Treurniet, 1998, 1996), POM (Colomes et al., 1995), and Toolbox (unpublished). PEAQ makes use of the masked threshold concept as well of the comparison of internal representations. It

computes measures of nonlinear distortions, linear distortions, harmonic structure, distance to masked treshold, and changes in modulation. These parameters are mapped by an artificial neural network to a single overall measure, the *distortion index* (DI). The DI is linearly related to the estimated perceived basic audio quality, which is denoted as ODG (objective difference grade). The optimization and training of PEAQ was done using a set of listening tests databases, including the six databases that constitute the database used in this study.

A valid comparison of PSM_t and PEAQ would require a database that was not used to develop, optimize or train any of the methods before, neither as a whole nor parts of it. Unfortunately, such a database is not available to the author.

There is one particular database, denoted as DB3 (ITU-R, 1998b), that was not used to optimize the PSM nor the PSM_t measure. Although some (32 out of 84) of its items (i.e. signal/distortion combinations) were not available to the authors of PEAQ during the calibration phase either, all of the signals as well as all of the distorting systems have been used before in different combinations. In contrast, not all kinds of distortions and only 8 of the overall 26 signals contained in DB3 were also contained in the database used for the optimization of PSM and PSM_t.

Figure 2.24 shows the results of quality predictions for the database DB3. In the upper panel, quality predictions by the PEAQ-ODG measure is depicted whereas the lower panel shows the results obtained from the PSM_t-ODG measure (transformed PSM_t). Additionally, Figure 2.25 compares the prediction performance for all available test items, i.e. in the case of PEAQ: all training items plus database DB3 and one validation database[16] (CRC97). In the case of PSM_t: all "training" items plus database DB3.

The comparisons exhibit a clear superior performance of PEAQ regarding database DB3, and a comparable performance if all databases are considered together. (For the latter case, no quantitative measures of the prediction performances were stated in (Thiede et al., 1998), (Thiede et al., 2000) and (ITU-R, 1998a).) The reason for the rather poor performance of PSM_t regarding database DB3 is not clear. This database seems to be more crucial for objective audio quality measures than others. This assumption is supported by the fact that DB3 was originally composed with the intention to serve as a validation database for PEAQ. But after a first performance test of PEAQ using this database, 52 of the overall 84 items were distributed to the authors of PEAQ for further calibration.

[16]This database, that was especially composed for the validation of PEAQ, contains eight signals of which seven were already contained in former training data sets (Soulodre et al., 1998).

As mentioned before, these 52 items contained all kinds of signals and distortions (in different combinations, though) that were contained in the remaining part of DB3. The results shown in Figure 2.24 were obtained using the complete database DB3, including the training data. This might be one reason for the considerable superior performance of PEAQ regarding this particular database.

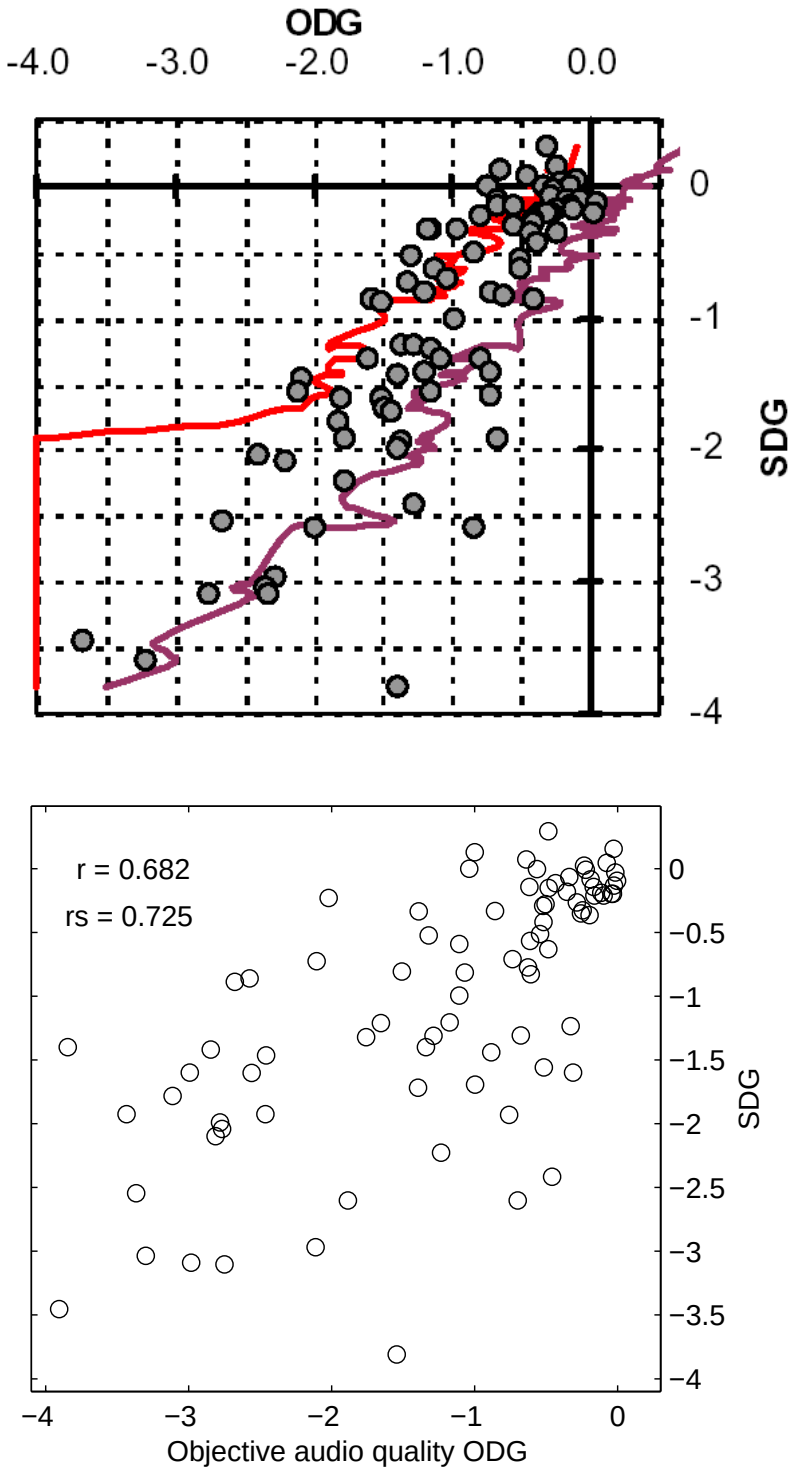

Figure 2.24: Comparison of quality prediction results for database DB3. Upper panel: PEAQ (taken from Thiede et al. (1998)). Lower panel: mapped PSM_t.

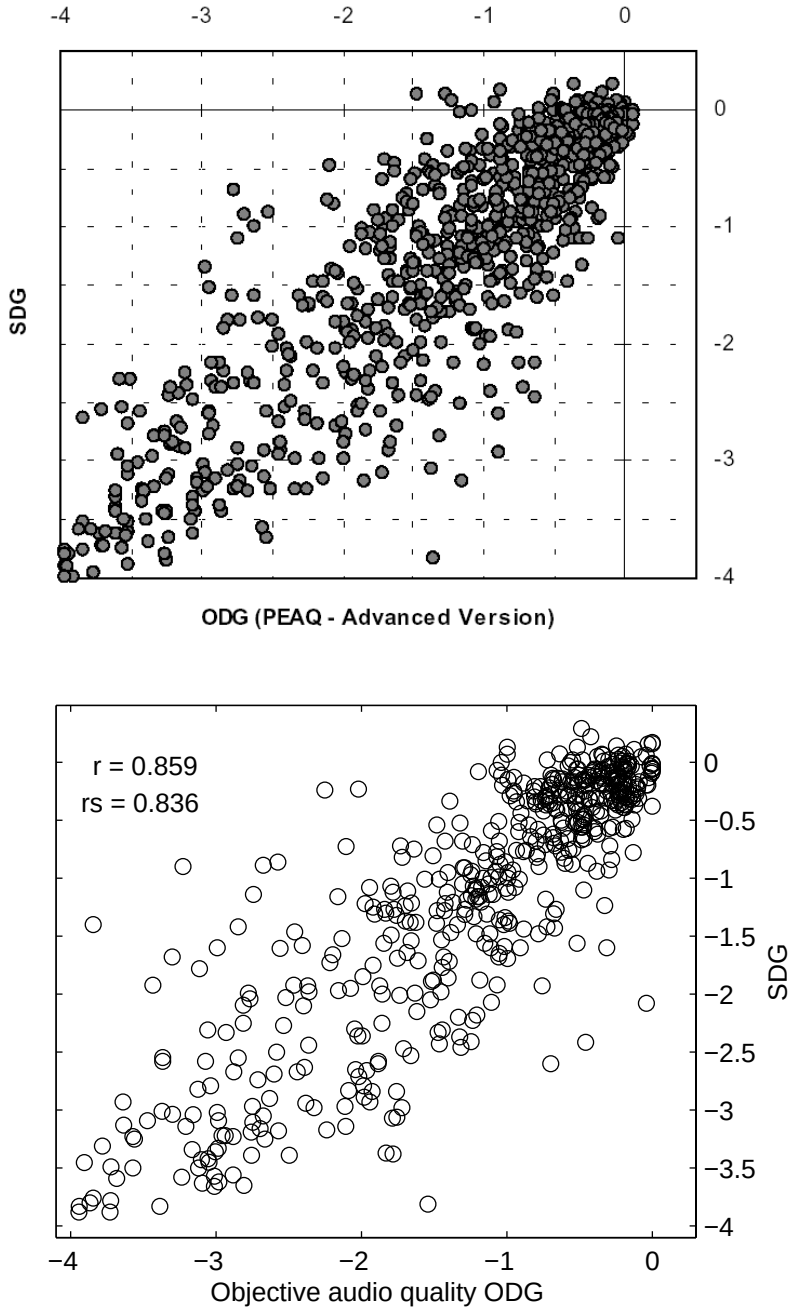

Figure 2.25: Comparison of quality prediction results for all items. Upper panel: PEAQ (taken from Thiede et al. (1998)). Lower panel: mapped PSM_t.

2.4 Summary and conclusions

A new measure for the prediction of the perceived quality difference between audio signals was introduced. It is based on an psychoacoustically validated auditory model by Dau et al. (1997a) and represents an expansion of the speech quality measurement method of Hansen and Kollmeier (2000). The new method is able to predict very small as well as severe quality degradations for different types of audio signals and audio codecs. The predicted audio qualities show good correlations with subjective quality ratings for most of the applied test material. The capability of the presented method seems comparable to that of the ITU recommendation BS.1387 *for objective measurements of perceived audio quality* ("PEAQ"). However, in comparison to PEAQ, because of its use of a psychoacoustically validated auditory model without a special adaptation to the data to be modeled, the presented method better allows for conclusions about actual mechanisms of human perception of audio quality. Moreover, it gives rise to the assumption that it might have a higher ability to generalize and thus being applicable to unknown distortions, audio material and possibly even somewhat different tasks.

The development and investigation of the presented method revealed some aspects of audio quality assessment, from which the following conclusions may be drawn:

- The use of a bank of modulation bandpass filters instead of a modulation lowpass filter improves the prediction performances, especially if very small quality differences are to be resolved. Extending the bandwidth of a single modulation filter to higher modulation frequencies alone is not sufficient.

- For optimum performance of the measure, frequency bands have to be weighted equally. This contrasts to the findings of Hansen and Kollmeier (2000), who used a similar measure for a limited frequency range to predict speech transmission quality.

- Envelope fluctuations seem to be overemphasized by the adaptation loops of the original auditory model. This characteristic led to an overestimation of quality for rapidly fluctuating portions of audio signals. This overestimation can be counteracted by compressing or limiting the amplitude of the adaptation loops output, or by short-time cross-correlation of the model outputs and averaging, instead of computing one overall correlation coefficient.

- The temporal course of the instantaneous audio quality and its relationship to the perceived overall audio quality has to be taken into account. Emphasizing episodes of

particularly low instantaneous quality improves the prediction of the overall quality. This high weighting of low-quality signal portions corresponds well with human behaviour.

CHAPTER 3

Detection of audio distortions

Abstract

This chapter presents a new psychophysical method for the perceptual evaluation of lossy audio processing systems such as low bit-rate audio codecs. This method uses the alternative forced-choice procedure, well established for the psychoacoustical measurement of masked thresholds, for the detection of audio distortions. It focuses on the detectability of signal alterations produced by a lossy processing, thus investigating the transparency *of the system under test. The new method was applied to audio signals that were processed by three lossy audio processing schemes, including two low bit-rate audio codecs. The measured distortion detection thresholds differed largely depending on the respective audio signal and processing scheme.*

For the simulation of the psychoacoustical experiment, two versions of the model of auditory processing of Dau et al. (1996a, 1997a) were employed. The predicted thresholds show high correlations with the measured data for both of the model versions ($r = 0.9$, 0.8) whereby predictions obtained with the modulation filterbank model (Dau et al., 1997a) show distinct deviations from the experimental data in a few cases, leading to a somewhat smaller overall correlation compared to the modulation lowpass filter model (Dau et al., 1996a).

It is concluded that the proposed new subjective method is suitable for the evaluation of the transparency of lossy audio codecs and that the auditory model of Dau et al. (1996a) is capable of predicting detection thresholds of audio distortions with sufficient accuracy and thus represents a tool for the objective evaluation of audio codecs.

3.1 Introduction

To date, the evaluation of transmission quality of lossy audio processing systems is usually realized by performing either subjective listening tests or applying computational methods that try to predict results from listening tests, as presented in the previous chapter. Standard listening tests, as described in ITU-R recommendations BS.562-3 (ITU-R, 1990) and BS.1116, for example, (ITU-R, 1997) aim at assessing the perceived audio quality (difference) on categorical scales that cover a range from "excellent" to "bad" quality (or from "imperceptible" to "very annoying" quality degradation). The output values of computational methods are typically scaled accordingly. However, as efficiency of and demands on new audio codecs have been continuously increasing over the past few years, the aim of an evaluation tends to turn away from addressing the question of *how much* the perceived quality of a processed audio signal is affected by a codec towards the question if quality is affected *at all*, i.e. the "transparency" of a codec is mainly considered. In ITU-R BS.1116, the corresponding category of transparency is only a singular point at the top end of the continuous quality scale. Using this scale, it is hardly possible to interpret the mean quality value over subjects in terms of the detectability of a difference in the signal. For such a purpose, another kind of experimental paradigm is needed.

In the present chapter, an alternative method for the subjective assessment of audio codecs is proposed. With this method, the procedure of determining masked thresholds by an adaptive, 3-AFC ("alternative forced choice") psychacoustical measurement is applied. "Perceptual" audio codecs such as MPEG Layer-3 ("MP3") (ISO/MPEG, 1992) make use of the masked threshold concept by "shaping" the quantization noise, i.e. distributing it in such a way, that it is (ideally) masked by the signal. The question whether distortions introduced by a codec are audible or not is equivalent to the question whether the level of distortions is above or below the masked threshold. In terms of masking, the (original) signal serves as the masker, whereas distortions are the maskee. By comparing the level of distortions with the psychoacoustically measured masked threshold for a given signal and codec, not only can conclusions about the (un-)detectability of the distortions be drawn, but also about the "distance" to the detection threshold. This distance may serve as another measure for the transmission quality of audio processing systems.

A second focus of the present study is the quantitative simulation of the experimental results for the detection of the distortions described above. The applicability of the signal processing model by Dau et al. (1996a, 1997a) is examined for this task. The model has been shown to be particularly suitable for the quantitative prediction of masked thresholds

in simultaneous as well as nonsimultaneous conditions (see, e.g., Dau et al. (1996b)). The use of an "optimal detector" as a decision device enables the model to mimic a human listener in a psychoacoustical measurement[1]. However, primarily artificial signals with relatively simple spectral and temporal properties have been considered so far, such as tones in a bandpass noise masker. In contrast, the signals and distortions usually encountered in audio coding that are investigated here are more complex. It is therefore not clear if the model can account for the results of such an experiment. A successful accomplishment of this task would not only yield a useful method for the computational evaluation of audio processing schemes, but would also represent another validation of the model.

In this chapter, the new method for the subjective evaluation of lossy audio processing schemes is presented in detail. Results of a psychoacoustical measurement with everyday audio signals, distorted by different audio codecs, are presented and discussed. The simulated results for the same signals and codecs are compared to the experimental data.

[1]The method for predicting perceived audio quality differences presented in the previous chapter only uses the "preprocessing" stages of the model; the decision device is replaced by the operation of cross correlating the internal representations of test and reference signal.

3.2 Method

3.2.1 Stimuli and apparatus

Six audio signals were subjected to three types of lossy audio processing schemes. The audio signals were: 1) a 3.8 s excerpt of Bizet's "*Carmen*", presented at an average (peak) level[2] of 62 (76) dB SPL; 2) 2.5 s of "*glockenspiel*" (chimes), presented at 55 (66) dB SPL; 3) 3 s of *castanets*, presented at 54 (78) dB SPL; 4) 4.1 s of German, *male speech*, presented at 60 (74) dB SPL; 5) 3.7 s of *bass guitar*, presented at 66 (80) dB SPL; 6) 2.8 s of German, *female speech*, presented at 60 (72) dB SPL. Distorted test signals were obtained by applying the following signal processing schemes: a) MPEG-1 audio layer III ("MP3") (ISO/MPEG, 1992), b) Windows Media Audio 8 ("WMA") (Microsoft, 2003), c) Modulated Noise Reference Unit (MNRU) (ITU-T, 1996a) [3]. Thus, overall, 18 test signals were investigated. The distortion was "isolated" by taking the difference of the original signal and the processed signal after compensating for the processing delay of the coding scheme:

$$Dist = Proc - Orig$$

A particular test signal ($Test$) was generated by multiplying the distortion ($Dist$) with a constant factor (c) and adding it to the original signal ($Orig$):

$$Test = Orig + c \cdot Dist$$

With $c = 1$, the test signal is equal to the processed signal, i.e. the output signal of the audio codec. Other values of c change the signal-to-distortion ratio according to $\Delta SNR = -20 \cdot log(c)$ dB.[4] Masked thresholds of audio distortions will be expressed by the parameter ΔSNR.

All reference audio signals were taken from collections of audio material that were composed on behalf of the Moving Pictures Experts Group (MPEG, an ISO sub-group), and the International Telecommunication Union (ITU) respectively, for the evaluation of low-bit-rate audio codecs (ISO/MPEG, 1990, 1991; Meares and Kim, 1995; ITU-R, 1992, 1993). The audio material consists of A/D converted recordings, sampled at 48 kHz, 16 bit, and provided as mono or stereo Windows-PCM-wave files. Excerpts of these audio

[2] Levels were adjusted subjectively in order to equate subjective loudness.

[3] The MNRU modulates the input signal $x(t)$ according to $y(t) = x(t) \cdot (1 + m \cdot n(t))$, where $n(t)$ is a white noise with unity variance and m is the modulation depth.

[4] The signal-to-distortion ratio is denoted as SNR, because the distortions are essentially quantization noise.

samples were digitally rescaled, transformed to analog by a 24-bit D/A converter (SEKD 2496), amplified by a headphone amplifier (Beringer "Powerplay") and presented dioticly via headphones (Sennheiser HD580) in a soundproof booth (IAC-1600).

3.2.2 Procedure and subjects

The aim of the psychoacoustical experiments was to determine how much the signal- and codec-specific distortions had to be attenuated (or amplified), just to be masked by the signal. These masked thresholds were measured using an adaptive three-interval forced choice (3IFC) procedure. A trial consisted of three intervals separated by 0.5 s of silence. Two of the intervals contained the original signal (*Orig*), one the distorted test signal (*Test*), in random order. The subject's task was to identify the interval containing the distorted signal. Visual feedback was provided during the measurement. During a threshold run, the level of the distortions was varied according to an 1-up 2-down algorithm (Levitt, 1971), which converges at a level corresponding to 70.7 percent correct answer probability. The experiment was split up into two parts. The first part ("experiment A") served as a pilot experiment in order to check whether the experimental setup was appropriate. The signals 1) to 4) were used. The initial step size of the adaptive procedure of level adjustment was 4 dB and was reduced to 2 dB after the first two reversals. Because of the relatively long signal duration (up to 4 s), this step size was not further reduced during the remainder of the run, and only two more reversals were used to obtain the threshold estimate by taking the mean ΔSNR value at these reversals. In experiment B, the number of reversals used for threshold estimation was increased to four, to increase accuracy of the threshold estimate. Signals 5), 6) and again 1) and 3) were used. The latter signals were used in both experiments in order to check whether the different number of reversals and a possible training effect lead to systematic threshold differences.

Eight male and two female normal-hearing subjects participated in this study. They were aged between 25 and 39 years and had some experience in psychoacoustic measurements.

3.2.3 Simulations

The simulations were performed using the auditory model of Dau et al. (1996a, 1997a). The structure of the model is depicted in Figure 3.1. The model combines several stages of preprocessing with an "optimal detector" as a decision device. First, the incoming signal is processed by a linear gammatone filterbank (Patterson et al., 1987) that simulates the

bandpass characteristic of the basilar membrane. The output of each bandpass filter is halfwave rectified and lowpass filtered at 1 kHz, which approximates the envelope of the signal for higher center frequencies. To simulate effects of temporal adaptation, a subsequent chain of five nonlinear feedback loops transforms the signal depending on its rate of fluctuation. The overall effect of the five consecutive loops is to compress stationary signals with an approximately logarithmic characteristic, while transmitting fast fluctuations almost linearly. Within each loop, the input is divided by the lowpass-filtered fed back output. For stationary signals, this represents a square root operation. The five loops differ with respect to the time constants of the lowpass filter. The final preprocessing stage differs between the two model versions in Dau et al. (1996a) and Dau et al. (1997a): the 8 Hz modulation lowpass filter used in the earlier version was replaced by a modulation filterbank in the more recent version, which generalizes the model such that it also accounts for psychoacoustic experiments of amplitude modulation detection (Dau et al., 1997a,b). An "internal" noise of constant variance is added to the output of each modulation filter to model limitations of resolution. These filter outputs form the "internal representation" of the corresponding input signal. The internal representation is subjected to a decision device (realized as an "optimal detector"), where it is compared with a stored representation of the signal to be detected (template). The comparison is performed by calculating the cross correlation between the two patterns, which can be interpreted as a "matched filtering" process (Dau et al., 1996a).

The simulations presented in the present study were performed using both model versions mentioned above (modulation lowpass vs. modulation filterbank). 33 critical bands in the range from 235 Hz to 14.5 kHz and 8 modulation channels up to 129 Hz (in the modulation filterbank model version) were applied. These settings were adopted from the model configuration used for the prediction of audio quality (cf. previous chapter). Because the value of the variance of the internal noise depends on the overall number of channels, this parameter was adjusted once for each model version to optimize the prediction accuracy for the present experiment. Varying the variance essentially results in a constant threshold shift for all signals (which does not affect the correlation between measured and simulated data). In this study, the variance was adjusted to bring about equal mean values of measured and simulated thresholds. Apart from this adjustment, no alterations of any other model parameters were made, if not stated explicitly.

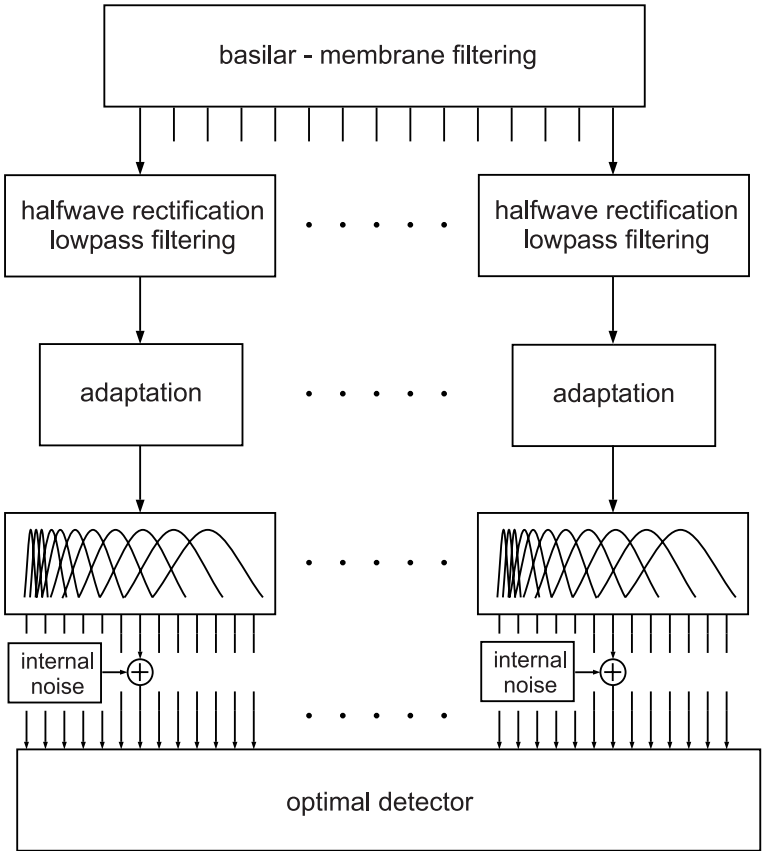

Figure 3.1: Block diagram of the psychoacoustic model as described in Dau et al. (1997a)

3.3 Results

3.3.1 Experimental data

Figure 3.2 shows measured thresholds as mean values and standard deviations over 10 subjects (open symbols) together with corresponding simulated thresholds (filled symbols). Results of experiment A are shown in the upper panel and results of experiment B are represented in the lower panel. The ordinate indicates the SNR-difference at threshold, i.e., the scaling factor applied to the distortions (expressed as $10 \cdot \log(c)$, in dB) that leads to 70.7% correct decision. Positive ΔSNR values mean that the distortions introduced by the audio codec must be attenuated in order to become inaudible, whereas negative ΔSNR values imply transparency of the codec[5]. The higher the threshold, the worse the transmission quality of the audio codec or the more sensitive the ear is. The abscissa shows the different audio signals. Each audio signal was processed by two audio codecs (MP3, WMA) and the MNRU, which is indicated by corresponding symbols (circle, triangle and square, respectively). The correlation between measured and simulated data is quantified by the linear correlation coefficient r.

The measured thresholds show large variations across signal types and codecs, especially for *castanets* and *male speech* signals (cf. upper panel): while the signals processed by MP3 are indistinguishable from the original for all subjects, distortions introduced by WMA are clearly audible. In contrast, both audio codecs perform similarly for the audio items *Carmen* and *glockenspiel*, where quality degradations due to the codecs are, on average, undetectable. In experiment B, two of the signals of the previous experiment (*glockenspiel, male speech*) were replaced by *bass, female speech*, while detection thresholds for the remaining signals (*Carmen, castanets*) were measured again. The results are shown in the lower panel of 3.2. All thresholds for the repeated conditions are shifted towards larger values by 2 to 6 dB, while, on average, the standard deviations decrease. The application of the audio codecs to the new audio signals *bass* and *female speech* lead to audible signal alterations in each case, all thresholds are above 0 dB. A striking difference to the results of experiment A appears in the case of speech, while codec MP3 is transparent when applied to *male speech* in experiment A, its effects on *female speech* in experiment B are clearly detectable. The difference between the detection thresholds amounts to 18.5 dB in this case. In contrast, detection thresholds for distortions introduced by codec WMA are very similar in both experiments.

[5] In the case of the Modulated Noise Reference Unit (MNRU), ΔSNR = 0 dB corresponds to $q = 35$ dB, whereby the MNRU-parameter q and the modulation depth m are related by $q = -20 \cdot \log(m)$.

3.3.2 Simulations using a modulation lowpass

The results of the simulations show good agreement with the measured data: In experiment A, 10 out of 12 simulated threshold lie within the inter-individual standard deviation of the measured thresholds, which are distributed over a wide dynamic range of 24 dB. This results in a high linear correlation value of $r = 0.90$. Because of the smaller inter-individual standard deviations in experiment B, less simulated data points lie within the standard deviation of the data in this case. However, the linear correlation is even slightly higher ($r = 0.91$) than in experiment A.

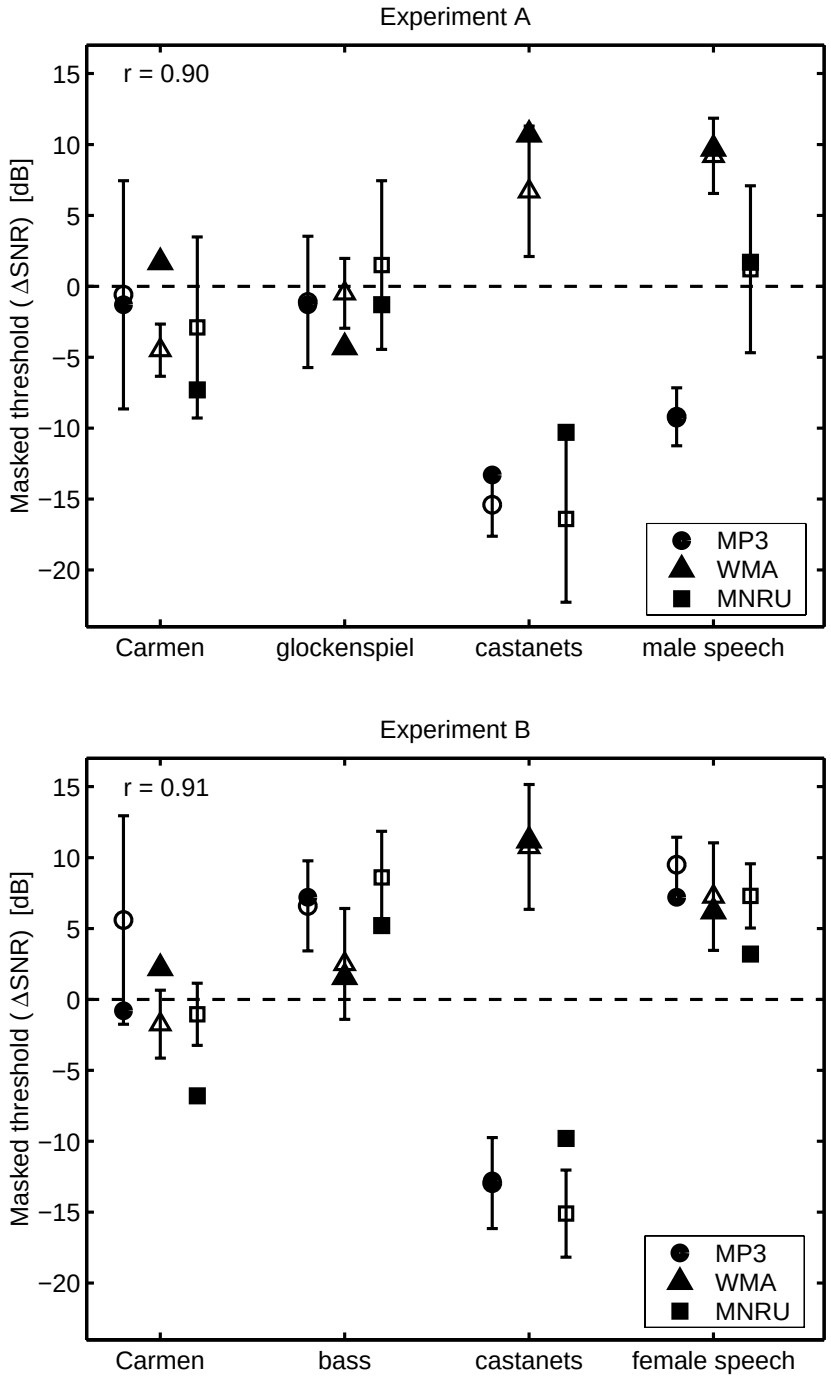

Figure 3.2: Measured (open symbols) and simulated (filled symbols) masked thresholds of distortions introduced by two audio codecs (MP3, WMA) and by amplitude modulation with noise (MNRU) for different audio signals. Thresholds are quantified by signal-to-noise ratio differences (ΔSNR), i.e. the amount of distortion attenuation. Simulated data were obtained with the model using the modulation lowpass approach.

An illustration of the advantage of perceptual coding algorithms like MP3 and WMA in terms of signal-to-noise ratio (SNR) is given in Figure 3.3. It shows the absolute ratios of signal-to-(quantization-)noise energy at detection thresholds as measured in experiment B. Except for the WMA-processed castanets, SNRs at threshold are markedly smaller for the perceptual coders MP3 and WMA than for MNRU (14 - 24 dB) and do not differ from each other by more than 4 dB. A consistent rank order of MP3 and WMA cannot be observed in the present study. However, the threshold for the castanets coded by WMA is particularly high compared to those coded by MP3 and even exceeding the threshold obtained with MNRU. As pointed out earlier, MP3 performs clearly transparent in this case, whereas WMA introduces distinct perceptible distortions [6].

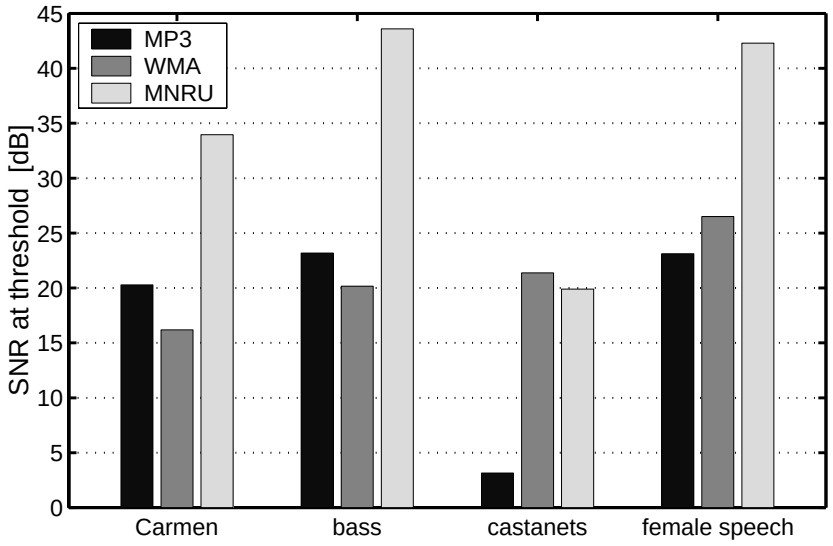

Figure 3.3: Signal-to-(quantization-)noise ratios at threshold for audio signals used in experiment B.

3.3.3 Simulations using a modulation filterbank

In order to examine the effect of the modulation processing stage of the auditory model on the simulated thresholds, the modulation lowpass filter was replaced by a linear bank of bandpass filters as described in 3.2.3, which represents a more recent version of the auditory

[6] Note that both audio codecs were operating at only 48 kbit/s on mono audio signals.

model (Dau et al., 1997a)[7]. Figure 3.4 shows measured data together with simulated data obtained with this model version. The deviations between measured and simulated thresholds are larger than those observed in the previous simulation (cf. Figure 3.2). Linear correlation deteriorates from 0.90 to 0.80. This deterioration is mainly caused by large discrepancies in a few different conditions (most notably: WMA-processed *castanets*, MP3-processed *bass*). In general, the predicted thresholds of those audio signals that have sharp onsets (leading to the excitation of a broad range of modulation frequencies) and that were processed by perceptual audio codecs are increased in most cases. The decreased thresholds for some signals distorted by noise modulation, relative to those obtained with the modulation lowpass (cf. Figure 3.2), are likely due to the process of adjusting the variance of the internal noise for the purpose of aligning simulated and measured data (cf. Section 3.2.3). Thus, this decrease does not necessarily indicate a loss of sensitivity of this model version.

3.4 Summary and discussion

The purpose of this study was two-fold: First, a new method for subjective evaluation of transmission quality of lossy audio processing systems was presented. Second, the ability of an auditory model to predict detection thresholds of audio distortions was examined.

The basic approach of the proposed new method was inspired by the common masked threshold concept of perceptual audio codecs. The basic idea of this concept is to reduce the average resolution of amplitude quantization of audio samples, thus permitting a higher overall level of quantization noise, whereby this noise is spectro-temporally shaped in such a way that it is (ideally) masked by the signal and therefore imperceptible. Hence, it appears reasonable to consider the task of detecting distortions within an audio signal as a masking experiment, where the undistorted original signal represents the masker while the introduced distortions (e.g. quantization noise) represent the maskee. The presented method "isolates" the distortions by simply subtracting the original signal from the distorted signal. Considered as the maskee, the difference signal is scaled independently and added back to the original (the "masker") again to generate particular test signals. This allows for the determination of masked thresholds using well established psychoacoustical standard procedures, which has the advantage of being much more sensitive to small vari-

[7] In fact, the model version applied in the present study differs slightly from that described in Dau et al. (1997a). Dau's modification of the adaptation stage in 1997 was not adopted. Instead, the version of the adaptation stage described in Dau et al. (1996a) was used.

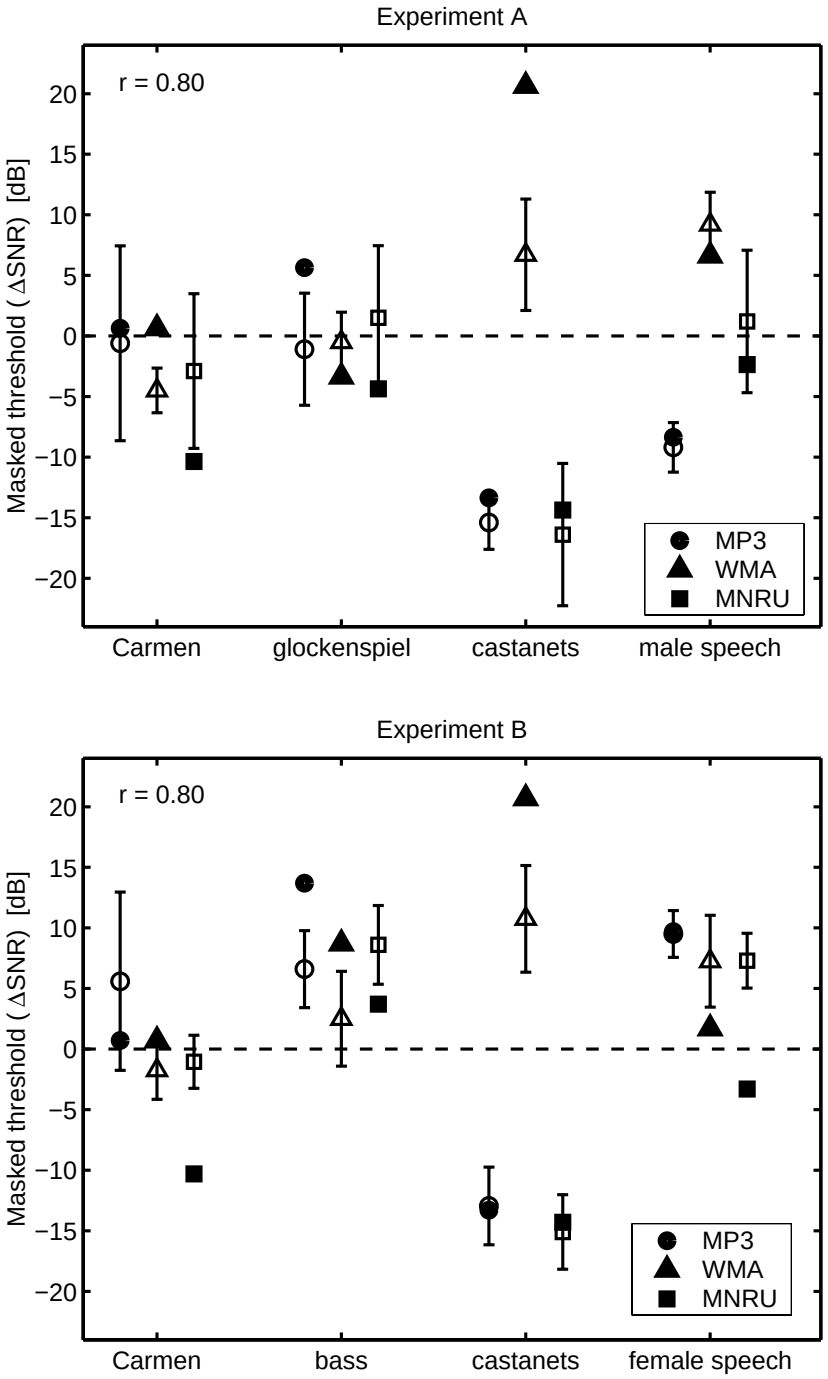

Figure 3.4: As Figure 3.2 but with simulated thresholds obtained using a model version with modulation filterbank.

ations in the signal due to the coding scheme than a category rating procedure. However, this method is only applicable if a simple difference between original and coded signal

yields a reasonable estimate of the quantization noise. This may become difficult if the codec group delay is frequency- or time-dependent.

The results of the experiment presented in Section 3.3.2 support the masked threshold concept of perceptual audio codecs: SNRs at detection thresholds are generally markedly smaller for the audio codecs MP3 and WMA than for noise modulation (cf. Figure 3.3). While detection thresholds were found to be similar for the two audio codecs for most audio signals and no consistent rank order could be observed, *castanets* turned out to be problematic for the WMA codec: The SNR at threshold for this signal/codec combination was found to be even slightly higher than for *castanets* distorted by the Modulated Noise Reference Unit. The poorer performance obtained with WMA in this case is caused by "pre-echoes", i.e. brief (few ms) bursts of noise preceding sharp signal onsets. This can easily be confirmed by inspecting the waveform of the distorted signal (s. Figure 3.5). A possible reason for this malfunction is the use of too long analysis frames in the encoder, so that rapid transitions in the signal are not taken into account appropriately [8].

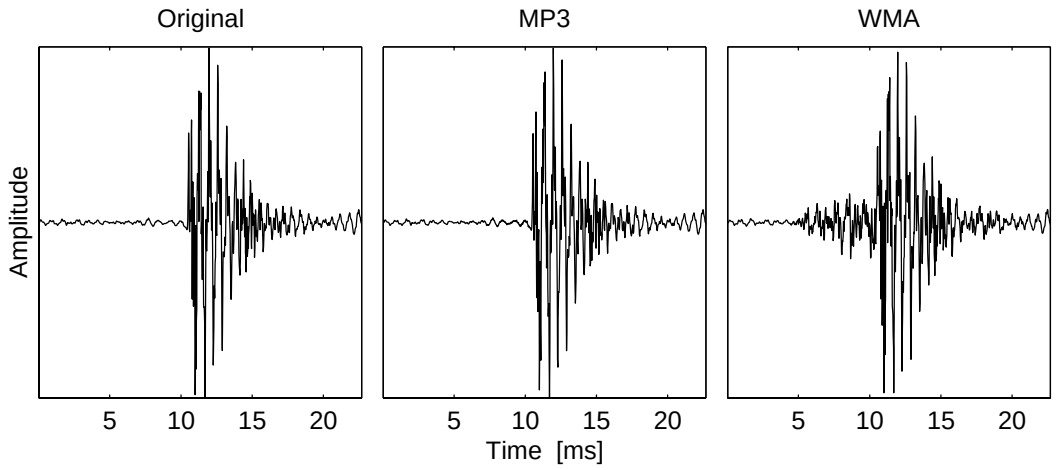

Figure 3.5: Cutout of the audio signal *castanets*: left panel: original, middle panel: processed by audio codec MP3, right panel: processed by audio codec WMA. The latter reveals a distinct "pre-echo".

The comparison of experiment A and B shows that ΔSNR values at threshold for repeatedly used signals increased consistently (i.e. subjects became more sensitive), while inter-individual deviations decreased on average. Both findings might be due to training

[8] Because Windows Media Audio (WMA) is a commercial product of Microsoft and publications of its design are not known to the author, it can only be speculated about this point.

effects (experiment A was conducted before experiment B) as well as to the extended adaptive measurement procedure (two more reversals were added).

3.4.1 Influence of modulation processing on simulations

As shown in Section 3.3.2, the auditory model is able to account for detection thresholds if a modulation lowpass filter with a cutoff frequency of 8 Hz is used at the output of the adaptation stage (Dau et al., 1996a). If a modulation filterbank stage is used instead of the lowpass filter (Dau et al., 1997a), the agreement with the experimental data decreases somewhat. In order to better understand the role of amplitude modulation processing for the predictions investigated here, additional simulations were carried out. In one model version, the cutoff frequency of the modulation lowpass filter was varied, while in another model version, the number of modulation bandpass filters was varied. The accuracy of the predicted thresholds was determined as a function of the lowpass cutoff frequency and the center frequency of the highest modulation filter, respectively, as shown in Figures 3.6 and 3.7. In Figure 3.7, the number of modulation channels ranges from one (in the leftmost condition) to eight (in the right most condition). In the left panels of the two figures, the root mean squared differences between measured and simulated thresholds ("RMS error") is given on the ordinate. The right panels show the corresponding linear correlation coefficient between experimental data and simulation. In the case of the modulation lowpass (Figure 3.6), the results indicate best prediction performance at a cutoff frequency around 7 Hz. This is compatible with findings of Dau et al. (1996a), who determined an optimal value of 8 Hz using data of simultaneous and forward masking experiments. For the modulation filterbank (Figure 3.7), the prediction error is smallest if only the two lowest channels are considered, consisting of a lowpass filter with a cutoff frequency of 2.5 Hz and a bandpass filter centered at 5 Hz (bandwidth = 5 Hz). In this case, prediction performance corresponds to that of the model version using the modulation lowpass. The use of additional higher frequency channels hardly affects the model performance as long as center frequencies up to only about 30 Hz are used. Also, if filters with center frequencies higher than 30 Hz are taken into account, model performance further decreases.

How can these findings be explained? In the case of the modulation lowpass filter, a cutoff frequency much larger than 8 Hz would not account for effects of temporal integration of signal information (e.g. Green, 1985; Moore et al., 1988), and thus a deterioration of the model performance with increasing cutoff frequency is not unexpected. However, in the case of the modulation filterbank, the increasing deviation of the simulated data from

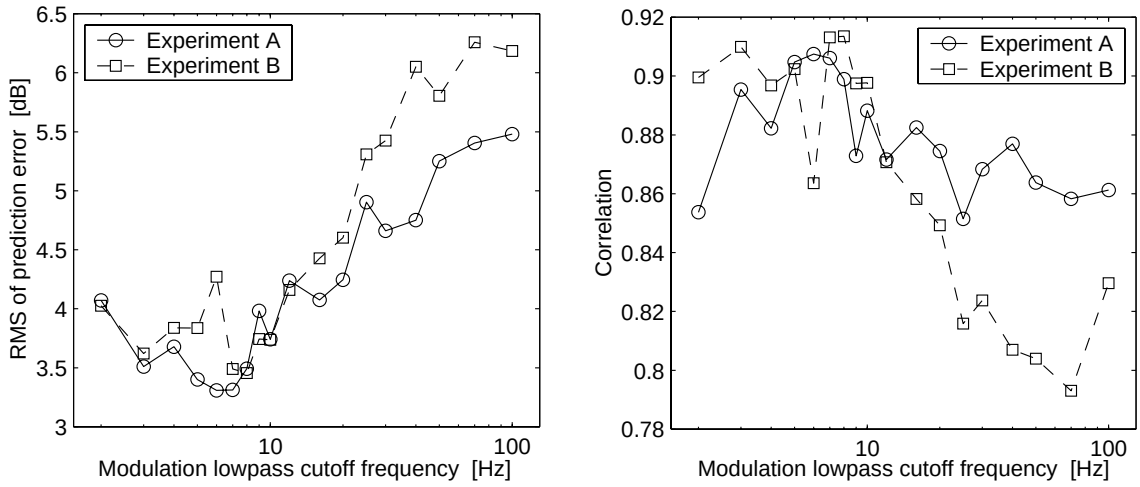

Figure 3.6: Prediction performance of the auditory model with a modulation lowpass filter as a function of lowpass cutoff frequency. Left: root of mean squared differences between simulated and measured thresholds. Right: linear correlation coefficient for the same data.

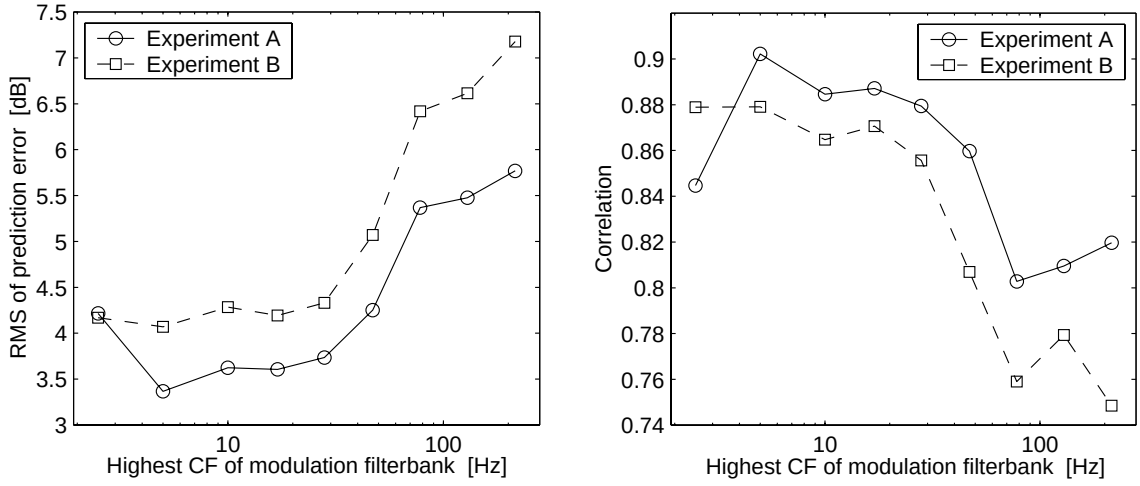

Figure 3.7: As Figure 3.6, but for simulated thresholds obtained from the auditory model with modulation filterbank as a function of the center frequency of the uppermost filter.

the measured data with an increasing number of modulation channels was not expected. Dau and coworkers could show that the modulation filterbank model is able to account for a larger number of psychacoustical data than the earlier (lowpass filter) version, without contradicting findings that were already explained by the original model. The modulation filterbank model might be considered as a generalization of the original model. Indeed, when applied to predict the perceived audio quality, the modulation filterbank model per-

forms better than the lowpass approach, as shown in the previous chapter. Thus, a general overestimation of information contained in higher modulation channels is most likely not the reason for the relatively weak performance of the model in the present task.

It has been reported that the pronounced overshoot of the adaptation loops in response to signal onsets may cause discrepancies between simulated and measured thresholds in specific conditions, and therefore a limitation of the amount of the overshoot response was suggested (Münkner, 1993; Dau et al., 1996b) (see also previous chapter). Overshoots produced by the adaptation loops are damped by the subsequent modulation filtering stage, whereby the damping depends on the cutoff frequency i.e. the bandwidth of the filter. (A larger bandwidth leads to less damping.) The observed decreasing correspondence between measured and simulated data with increasing modulation cutoff frequency in Figures 3.6 and 3.7 might therefore be due to the strong overshoot by the adaptation loops. This hypothesis is supported by the fact that simulated thresholds are shifted upwards most strongly for signals containing distinct onsets like *castanets*, *glockenspiel* and *bass*,[9] when processed by perceptual audio codecs (cf. Figure 3.4). For these signals, pre-echoes, which are a typical artifact of perceptual audio codecs, are more likely to appear than for the other signals. Pre-echoes reduce the sharpness of signal onsets and thus the amount of overshoots considerably. As a consequence, the detectability of signal differences due to pre-echoes is overestimated by the model, where the effect is larger the stronger the overshoots are.

If the decreased accuracy for higher modulation frequencies is in fact caused by too strong of an overshoot at the output of the adaptation loops, then limiting the output of that stage should counteract this effect and reduce the discrepancy between measured and simulated data. For this purpose, a modified version of the adaptation loops, proposed by Münkner (1993) and adopted by Dau et al. (1997a), was applied. The modification is realized by a compressive element[10] between the divider and the lowpass element such that the maximum output of each adaptation loop is restricted to λ times the value of its steady state response. (In Münkner (1993) and Dau et al. (1997a) λ was set to ten, which was motivated by physiological findings of Westermann and Smith (1984).) Figure 3.8 shows simulation accuracy obtained with the modified adaptation loops as a function of the value of the limiting factor λ. As in the previous figures, the left panel shows the RMS

[9] The bass signal used in this study stemmed from an electric bass guitar played in "slap style", which produces sharp onsets.

[10] The characteristic of this element is linear for input values within the maximum range for stationary signals and sigmoidal for higher values.

of the prediction error while the linear correlation coefficient is displayed in the right panel. The results show an improvement of the performance if a limitation is applied[11]. However, the observed effect is rather small (about 1 dB decrease of error-RMS in experiment A, 1.6 dB in experiment B at best) and not consistent for $\lambda < 2.5$. While the best performance is achieved at λ=3.5 in experiment A, the error diminishes continuously with decreasing λ over the entire range tested (λ_{min}=1.25). Prediction errors in these cases still exceed those obtained with the modulation lowpass model version by about 1 - 1.5 dB. Moreover, limitation factors of such small values do not seem reasonable.

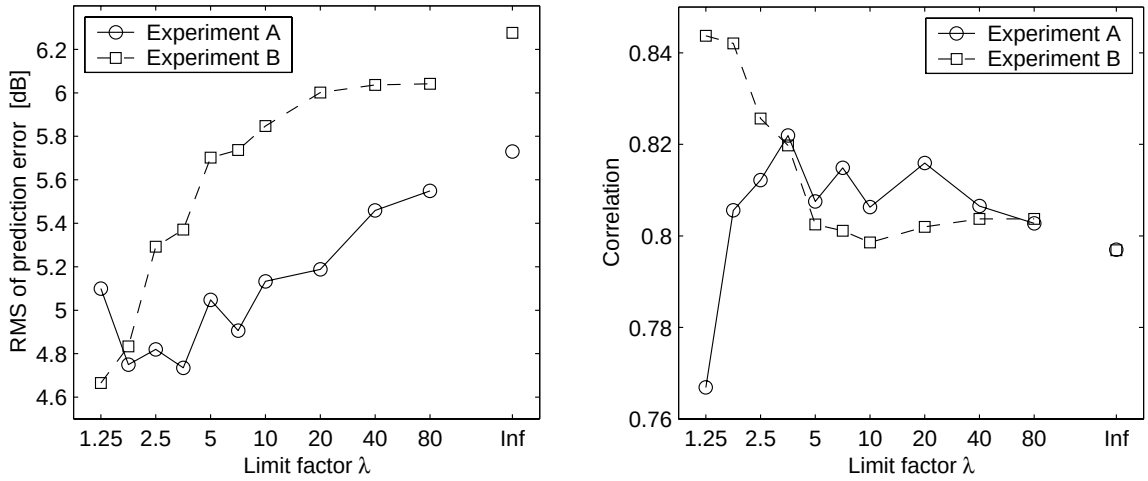

Figure 3.8: Prediction performance of the auditory model with modulation filterbank and modified adaptation loops as a function of the limit factor λ.

In concluding the preceding considerations, the somewhat worse correspondence of measured and simulated data obtained with the modulation filterbank model can at least, to some extent, be explained by too strong of an overshoot of the adaptation loops at signal onsets.

Another possible reason for the remaining discrepancies might be the simplifying model assumption that information is processed independently in the different frequency channels. For many natural stimuli such as, e.g. speech, neural activity in different frequency channels is correlated and can therefore *not* be assumed to be processed as independent information. The maskers and maskees used in the present study were all broadband. If the neural activity in different frequency channels of stimuli are correlated with each other,

[11] Improvements are observed only in association with the modulation filterbank. In case of the modulation lowpass filter, using the modified adaptation loops leads to poorer performance.

but information is processed independently across the channels and finally "optimally" combined by the model, the overall information contained in the stimuli is overestimated. The overestimation increases with increased proportion of correlated channels, degree of correlation and amplitude of the correlated channels. The present results show the largest deviations of simulated data from measured data for those signals that contain sharp onsets in the reference condition (i.e. undistorted signal ≡ masker) and pre-echoes preceeding those onsets (and thus smoothing them) in the corresponding test condition (i.e. distorted signal ≡ masker + maskee). Sharp onsets are broadband and therefore represented in all channels, which thereby become correlated. Moreover, as the bandwidth of the modulation bandpass filter increases with increasing center frequency, the onsets are represented by larger amplitudes in higher modulation channels. Differences between the internal representations of the reference and the test signals are therefore correlated across channels and represented by increasing amplitudes with increasing modulation center frequency. As a consequence, the information provided by these differences and thus detection thresholds could possibly be overestimated by the modulation filterbank model. However, in order to better understand the role of across-channel correlation of excitation (or neural activity) on predicted detection thresholds in the framework of the processing model used here, further studies with well defined and probably more basic spectro-temporal characteristics are needed. This, however, is beyond the scope of the present study.

3.5 Conclusions

- The determination of detection thresholds for audio distortions produced by lossy audio codecs represents a possibility to derive quantitative, statistical conclusions about the transparency of audio codecs, which is becoming the decisive criterion for the evaluation of present and future audio codecs.

- The model of auditory signal processing proposed by Dau et al. (1996a) is capable of quantitatively predicting masked thresholds of distortions in audio signals. This result was not clear in advance since very complex and broadband stimuli were used in the present study and the model has previously been tested primarily in rather simple conditions with well defined and mostly narrowband conditions. Hence, the model can be employed for the objective evaluation of lossy audio processing algorithms or transmission systems that are demanded to be transparent.

- Using a modulation filterbank stage as proposed by Dau et al. (1997a) essentially yields satisfying threshold predictions as well, although this model version performs not quite as well as the modulation lowpass filter model. Discrepancies between simulated and measured data are restricted to a few cases and can partly be explained by a too strong overshoot at the output of the adaptation loops and possibly by the lack of the model to account for across-channel processes. Both of these characteristics become more severe with an increasing number of channels of the modulation processing stage. Thus, limiting the output of the adaptation loops appears reasonable and across-channel processes should be accounted for in future model versions.

CHAPTER 4

Assessment of noise reduction schemes

Abstract

The applicability of computational audio quality measures for the assessment of single-channel and multi-channel noise reduction schemes is examined in this chapter. The speech quality measure q_C of Hansen and Kollmeier (2000), the audio quality measures introduced in Chapter 2 and a modified variant of these measures were employed to predict subjective ratings of the quality of noisy speech signals that were processed by noise reduction schemes. The signals and corresponding subjective ratings were taken from experiments carried out by Marzinzik (2000) and Tontch (2002). The results show dependencies of the prediction performance of the quality measures on the quality aspect (overall preference, speech naturalness, amount of background noise), the group of subjects (normal hearing vs. hearing impaired) and the presence of artifacts. In each condition, at least one of the audio quality measures showed good correlation with the subjective data, but none of the measures performed well in all conditions. The findings suggest that a perceptual evaluation of noise reduction schemes can be achieved using a set of four variants of the presented audio quality measures in order to account for different conditions regarding subjects, quality aspects and artifacts.

4.1 Introduction

Objective methods for the assessment of speech or audio quality are generally intended for predicting the perceived quality of a given test signal relative to a reference signal. Usually, the reference is assumed to be of desired audio quality, whereas any audible differences between the reference and test signal are interpreted as quality impairments of the test signal. The main application of those methods is the assessment of the transmission quality of lossy audio processing systems, such as low-bit rate speech and audio coding-decoding

algorithms (codecs) used for data reduction. Hence, the transmission quality of a codec is directly related to the perceptual similarity of processed and unprocessed signals, where perfect performance is achieved if input and output signals become indistinguishable. In contrast, the purpose of speech enhancement algorithms is to modify noisy speech signals in order to improve the speech intelligibility. In this case, the unprocessed signal is *not* of desired audio quality and audible changes introduced by the processing are intended. Hence, the approach for employing a computational, perceptual distance (or similarity) measure has to be modified in order to become applicable for that specific task. For this purpose, a triple of audio signals is demanded: apart from the noisy input signal and the processed signal, the "clean" speech signal (i.e. speech without noise) has to be provided that serves as a reference, representing the desired quality. The performance of the speech enhancement system is then described by the quality difference between unprocessed and processed signals with respect to the clean speech reference.

One apparent problem for both subjective and objective evaluation of speech enhancement systems is the trade-off between the amount of noise reduction (desired effect) and audible speech distortions (undesired effect). (Distortions are not restricted to the speech but can also produce annoying unnatural background noise. This can degrade the perceived overall signal quality as well, even if the level of noise is reduced.)

These effects are mostly positively correlated and have contrary consequences for the perceptual distance between the processed, noisy speech and the clean speech reference. The correspondence between objective and subjective quality ratings will depend on the subjective weighting of these effects, which itself is influenced by the particular instructions given to the subject. Marzinzik (2000), for example, found different subjective ratings of noise reduction algorithms in paired comparison tests depending on whether subjects were asked to indicate the respective item they "liked more" or the one that "showed less background noise". In contrast, a single objective measure can only account for one subjective dimension, and it is not clear a priori which dimension this will be. Although this problem has been pointed out before repeatedly (Gustafson et al., 1996; Hansen and Pellom, 1998), and a lot of studies applied objective quality measures to the evaluation of speech enhancement algorithms, there are hardly any publications concerning the validation of objective quality measures applied for this particular task using subjective results. For this reason, Marzinzik and Kollmeier (2000) examined the applicability of several objective speech quality measures for predicting the results of listening tests concerning monaural noise reduction schemes. One of the objective measures examined was the speech quality measure q_C proposed by Hansen and Kollmeier (2000). While this measure showed good

correspondence with subjective ratings of the noise reduction, it failed to predict the perceived naturalness of the processed speech and the overall preference of the subjects. In the present study, different modified versions and expansions of this measure, as presented in chapter 2, were applied to the data of Marzinzik and Kollmeier (2000) with the aim to successfully predict the subjective ratings in all of the mentioned categories. Additionally, another quality measure, especially adapted to the prediction of the speech naturalness and the amount of background noise, was developed an applied as well. This new measure and the quality prediction results are presented in the first part of this chapter.

The second part of this chapter is concerned with another validation of the audio quality measures mentioned above. In this case, the noise reduction effect obtained with different multi-channel beamformers used in automobiles was evaluated by Tontch (2002). The objective quality measures were employed to predict subjective assessments of these beamformers, which were obtained from a paired comparison test.

4.2 Experiment I: assessment of single-channel algorithms

In 2000, Marzinzik and Kollmeier presented an evaluation of single-channel noise reduction algorithms based on the concept of Ephraim and Malah (1984). Algorithms were evaluated subjectively as well as by several objective quality measures, including the speech quality measure q_C of Hansen and Kollmeier (2000) (for details see Marzinzik, 2000). This section presents results obtained by the original measure q_C and a number of new measures derived from q_C as introduced in Chapter 2.

4.2.1 Test signals

The target speech signal was a German sentence of approximately 4 s duration, taken from a re-recording of local radio newscasts, spoken by a professional male newscaster in a broadcasting studio. Test signals were generated by adding two kinds of background noise at two different signal-to-noise ratios (SNR): cafeteria noise and drilling machine noise at -5 dB and 5 dB SNR, respectively. These four noisy speech signals were subjected to three noise reduction schemes that consisted of combinations of algorithms proposed by Ephraim and Malah (1984, 1985) and a speech pause detection method introduced by Marzinzik (2000). Including the unprocessed signals, this gave a total number of 12 test signals.

4.2.2 Subjective measurement

Six normal hearing subjects and six subjects with moderate hearing losses participated in a paired comparison experiment. The experiment was divided into four blocks, one per noise condition. In each block, all of the noise reduction algorithms (including no processing) where compared to each other regarding three different criteria: overall preference, speech naturalness and amount of background noise. Test signals were presented diotically via headphones in a sound-attenuating booth. The results of each paired comparison test block were mapped onto a difference scale by fitting a Bradley-Terry model to the data (Bradley and Terry, 1952). Since paired comparisons can only yield relative ratings, the offset of the Bradley-Terry (difference) scale values represents a free parameter. The value of this parameter was chosen to result in a scale value of zero for the unprocessed signal in each noise condition. Thus, the results of the different measurement blocks became comparable. Moreover, the Bradley-Terry scaling allows to apply Pearson's correlation

coefficient to quantify the correspondence with other measurements. (See (Marzinzik, 2000) for details.)

4.2.3 Objective measurements

The objective speech quality measure q_C of Hansen and Kollmeier (2000) as well as several modified versions and expansions of this parameter (cf. Chapter 2) were applied to the same test signals used in the subjective experiment. The particular versions are denoted and characterized as follows:

- **QM1**: speech quality measure q_C according to Hansen and Kollmeier (2000), i.e., center frequencies (CF) of the gammatone filterbank ranging from 350 to 3800 Hz (adapted to telephone band), 8 Hz modulation lowpass filter, downsampling by non-overlapping averaging across 20 ms segments, "band importance weighting" emphasizing high frequencies.
- **QM1$_\mathbf{B}$**: as QM1, but with additional Beerends-Berger assimilation (BBA) of the internal representations (see below).
- **QM2**: the Perceptual Similarity Measure (PSM) introduced in Section 2.3.1 of Chapter 2, i.e. gammatone-filterbank bandwidth according to the actual bandwidth of the input signals (here: 8 kHz), 8 Hz modulation lowpass filter, downsampling to 100 Hz by applying an anti-aliasing filter (40 Hz FIR lowpass) and subsequent resampling, no frequency weighting.
- **QM2$_\mathbf{B}$**: as QM2, but with BBA (cf. QM1$_\mathrm{B}$).
- **QM3**: as QM2, but using a modulation filterbank instead of the lowpass filter (eight channels, highest CF = 129 Hz), downsampling according to the CF of the respective modulation channel.
- **QM3$_\mathbf{B}$**: as QM3, but with BBA.
- **QM4**: audio quality parameter PSM$_\mathrm{t}$ as described in Section 2.3.4 of Chapter 2, i.e. 5%-quantile of the "loudness"-weighted time series of instantaneous audio quality, PSM(t).
- **QM4$_\mathbf{B}$**: as QM4, but with BBA.
- **QM5**: as QM3, but with additional adaptive channel weighting (see below).
- **QM5$_\mathbf{B}$**: as QM5, but with BBA.

Beerends-Berger assimilation (BBA)

The BBA consists of a modification of the internal representation of the test signal (in this case: noisy speech), by partly assimilating it to the internal representation of the reference signal (here: clean speech). In the original approach proposed by Beerends (1994) and adopted in the ITU-T standard P.861 (ITU-T, 1996b), the mean squared weighted difference between internal representations served as a quality measure, where negative differences were weighted less than positive differences. (This approach follows the hypothesis that "missing" components in a distorted signal are less disturbing than "additive" components.) Berger (1998) found that a weighting factor of 0.5 maximizes the correlation between subjective quality ratings and mean squared differences between internal representations for most of the databases examined. Another quality measure developed by Berger is based on the correlation between internal representations, following the approach of Hansen and Kollmeier (2000). In this case, the internal representation of the test signal $T(t, f)$ (as a function of time t and frequency band f) was assimilated to the reference $R(t, f)$ in an asymmetrical way:

$$T(t, f) \mapsto \tilde{T}(t, f) = R(t, f) - \alpha(R(t, f) - T(t, f))$$

with

$$\alpha = \begin{cases} 0.5, & |T(t, f)| < |R(t, f)| \\ 1, & |T(t, f)| \geq |R(t, f)| \end{cases}$$

Again, a weighting factor of $\alpha = 0.5$ for $|T(t, f)| < |R(t, f)|$ was empirically found to be optimal[1]. This approach was adopted in the present study.

The quality measure QM5$_{(B)}$

The new quality measure QM5$_{(B)}$ is a variant of the measure PSM (denoted as QM3$_{(B)}$) presented in Chapter 2 [2], but additionally applies a task-dependent weighting of the frequency and modulation frequency channels. The weighting functions are based on the assumption of different strategies of subjects assessing either the amount of background noise or the naturalness of speech. If the noise is to be assessed, subjects might focus on those frequency bands that exhibit high noise-to-speech-energy ratios, so that they can concentrate better on the noise. In case of broadband noise, this essentially applies to

[1] An optimal choice of $\alpha = 0.5$ was also found in Section 2.3.2 of Chapter 2.

[2] In fact, QM3$_{\mathrm{B}}$ corresponds to PSM as defined in Chapter 2, while QM3 corresponds to a variant of PSM without BBA.

higher frequency bands. If, in contrast, the naturalness of speech is to be rated, subjects are rather likely to focus on those frequency bands containing most of the speech energy, i.e. frequencies up to about 5 kHz. It seems reasonable that these assumed strategies are not restricted to the frequency domain, but also apply to the modulation frequency and time domain.

To test this hypothesis, the audio quality measure PSM was expanded by an adaptive channel weighting of the internal representations and applied to the experimental data. The characteristics of the weighting functions depend on the specific task: If the naturalness of speech is to be assessed, a fixed lowpass characteristic is chosen. In the frequency domain, channels with center frequencies up to about 5300 Hz are equally weighted, while weights for higher channels are set to zero. In the modulation frequency domain, only the lowest three channels are taken into account; the weights of the 2.5 Hz modulation lowpass channel and of the 5 Hz modulation bandpass channel are set to one, the weight of the third channel with a center frequency of 10 Hz is set to 0.5. Instead, if the amount of background noise is to be assessed, the channel weights are set to the inverse of the signal-to-noise ratio, multiplied by the RMS of the noise in the corresponding channel: $w(f) = \mathrm{RMS}^2_{\mathrm{noise}}(f)/\mathrm{RMS}_{\mathrm{speech}}(f)$.

Rescaling of the objective quality ratings

As in the subjective scaling, objective qualities of the unprocessed noisy signals were set to zero by subtracting the objective quality value of the respective unprocessed signal from all quality values for each noise condition. The procedure for comparing subjective and objective quality ratings is outlined in Figure 4.1.

4.2.4 Results

Correlations between measured and predicted results for all categories and objective measures are given in Table 4.1 for hearing impaired and normal hearing subjects. Figures 4.2, 4.3 and 4.4 show subjective versus objective data for each category obtained from hearing impaired subjects and the respective objective measure that shows the best correlation. As reported by Marzinzik and Kollmeier (2000), the speech quality measure q_C of Hansen and Kollmeier ($\equiv$ QM1) correlates very well with measured data of hearing impaired subjects concerning the category "noise reduction" (upper panel), while yielding only poor results in the other categories (cf. Table 4.1). The dependency of the category is opposite for the parameter $\mathrm{QM1_B}$: Emerged from QM1 by applying the BBA addition-

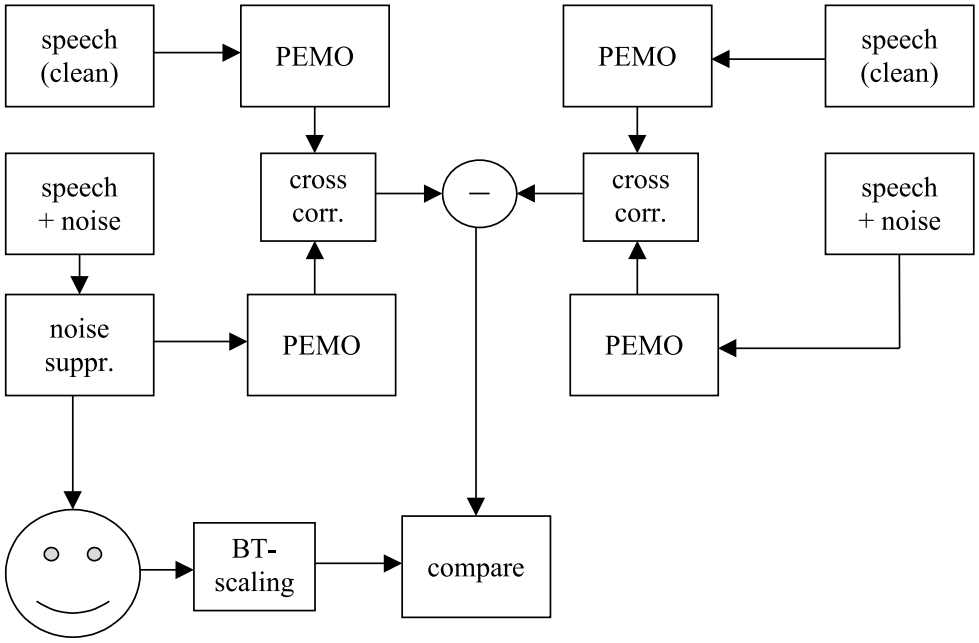

Figure 4.1: Block diagram for comparing subjective and objective quality assessments of noise reduction algorithms. Subjective quality ratings of noisy speech signals processed by different noise reduction schemes (including no processing) are obtained by a paired comparison test with subsequent Bradley-Terry (BT) scaling of the results (lower left corner of the diagram). Objective quality assessments are obtained by correlating the internal representations (= outputs of the model of auditory perception, "PEMO") of clean and noisy speech. The quality of the unprocessed signal serves as a reference and is arbitrarily set to zero within BT scaling, which is accounted in objective assessment by subtracting the corresponding quality value from each quality rating.

ally, this measure shows best prediction performance of all parameters in the categories "speech naturalness" (cf. Figure 4.3) and "overall preference" (cf. Figure 4.4), but only fair correlation with subjective data concerning "noise reduction". Marked improvements of prediction accuracy in the former categories can be observed for all parameters except QM4 ($\equiv \mathrm{PSM_t}$), if BBA is applied.

Correlations of objective parameters with subjective data obtained from normal hearing subjects show similar behaviour concerning "noise suppression" and "overall preference", but very different characteristics in the category of "naturalness". Measured and computed ratings on the latter category are positively correlated for hearing impaired subjects, but negatively correlated for normal hearing subjects. This is caused by converse subjective ratings of the speech items which contain drilling machine noise. The naturalness of the speech contained in these items was rated higher by hearing impaired subjects

if processed by noise reduction algorithms than if unprocessed (see Figure 4.3), whereas normal hearing subjects had a converse impression (see Figure 4.5). Marzinzik concluded that the normal hearing subjects seemed to be able to disregard the background noise in any noise condition and selectively listen to the speech, whereas hearing impaired subjects stated that the "*presence of background noise* was percieved as beeing 'unnatural'" (Marzinzik, 2000).

Best correlation values of objective quality measures with normal hearing's ratings (especially linear correlation values) are generally not as high as in the case of hearing impaired subjects and show a less distinct ranking in the category "noise suppression". According to the rank correlation, the best prediction of subjective data in this category is achieved by QM3 (instead of QM1 for hearing impaired subjects). In contrast the overall preference is still predicted best using the objective parameter $QM1_B$. The best correspondence between subjectively and objectively rated speech naturalness is achieved by QM5 (instead of $QM1_B$ for hearing impaired subjects). Another difference between the groups of subjects observed in the categories speech naturalness and overall preference is a reversal of the the rank order of parameters using or not using the BBA, i.e. applying the BBA *decreases* correlations with ratings of normal hearing subjects. Quality parameters based on PSM_t (QM4 and $QM4_B$) generally show very poor correspondence with subjective data for all subjects, which will be discussed in Section 4.4.

The correlation coefficients obtained from QM5 and $QM5_B$ support the assumption of different strategies applied in the subjective assessment of speech naturalness and background noise. Compared to the results obtained with QM3 (which corresponds to QM5 without channel weighting)and $QM3_B$, correlation coefficients are markedly higher in most cases, except for the speech naturalness rated by hearing impaired listeners, which is likely due to their already mentioned different understanding of speech naturalness. QM5 is the most successful of all measures in predicting the perceived noise suppression (according to linear correlation) and speech naturalness for normal hearing subjects. It also approximates the very good performance of the measure QM1 in the case of noise suppression rated by hearing impaired subjects. (The observed deterioration of correlation from QM3 to QM5 in the case of noise suppression rated by normal hearing subjects, associated with a reverse order of correlations regarding BBA, represents an exception from the general trend, which remains unclear.)

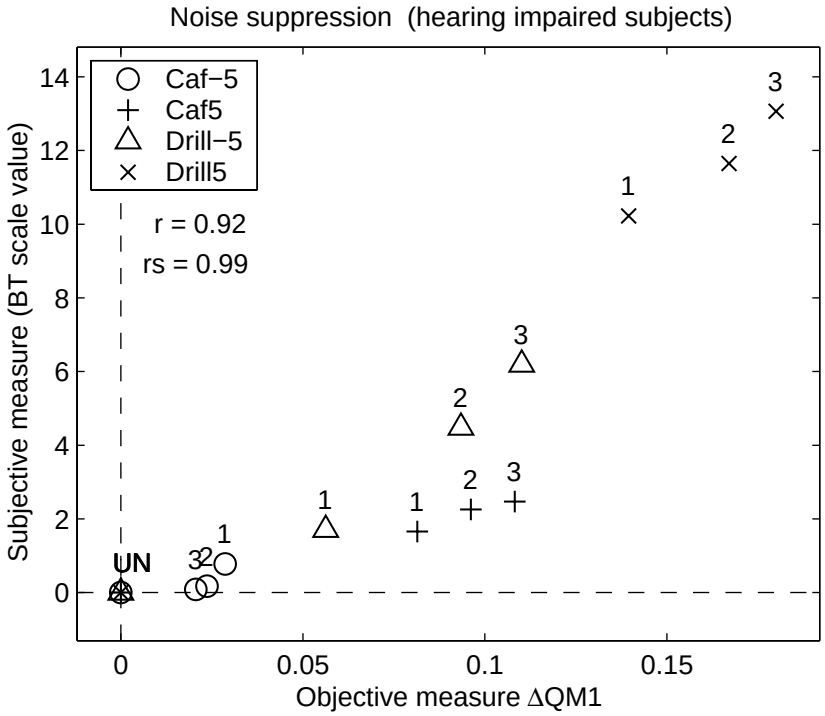

Figure 4.2: Quality prediction results of experiment I. Subjective assessments are plotted versus objective audio quality measures. Subjective measures were obtained by Bradley-Terry (BT) scaling of paired comparison test data. Signals were rated according to the amount of suppression of background noise. BT-scale values of unprocessed signals ("UN") were arbitrarily set to zero. The objective measure was obtained according to Figure 4.1, which reasons the "Δ"-prefix of the objective quality measure, denoting a difference value. Numbers indicate different noise reduction algorithms, symbols represent the kind and amount of background noise: cafeteria and drilling noise at -5 and 5 dB SNR, respectively. r and rs denote linear and rank correlation coefficients, respectively.

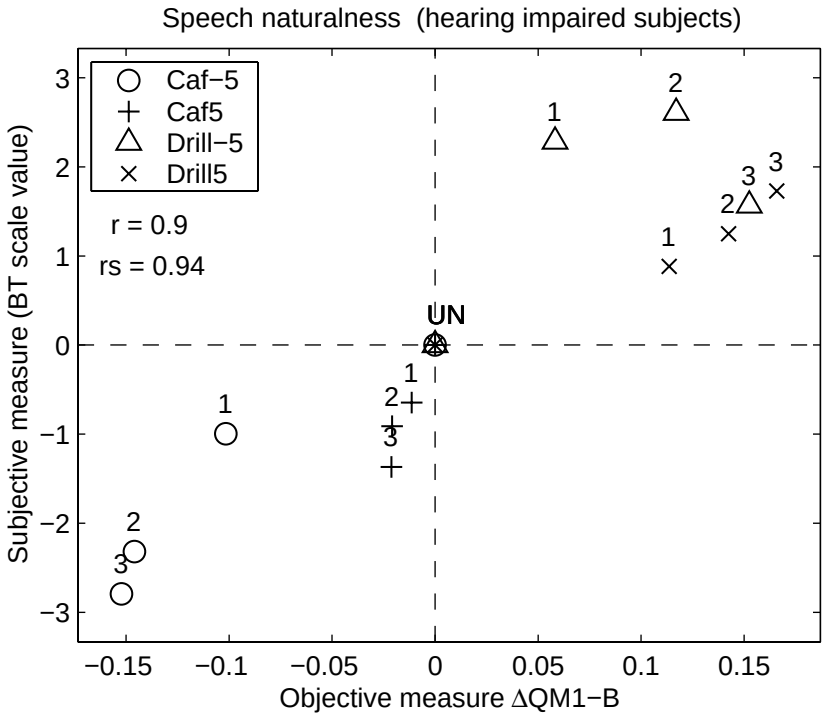

Figure 4.3: As Figure 4.2, but for quality criterion "speech naturalness" and different objective quality measure (ΔQM1 with BBA).

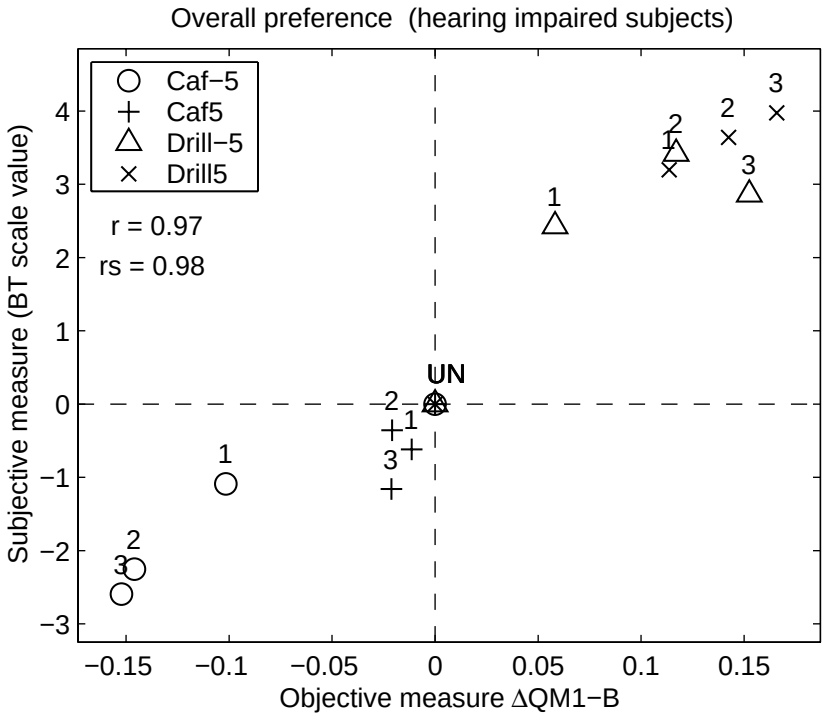

Figure 4.4: As Figure 4.3, but for quality criterion "overall preference".

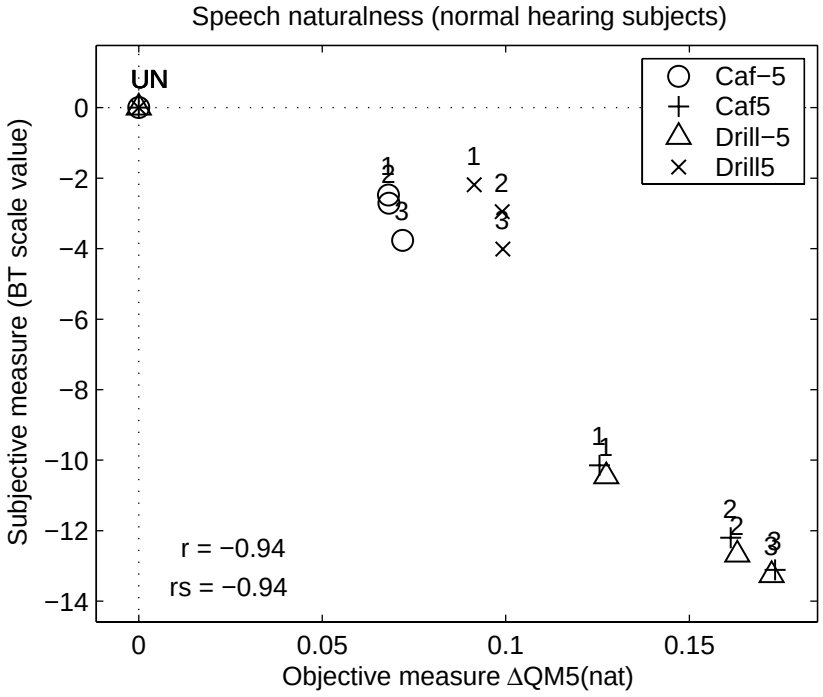

Figure 4.5: Ratings of the speech naturalness obtained from normal hearing subjects and corresponding predictions obtained by differences of the quality parameter QM5(nat).

	hearing impaired			normal hearing		
measure	suppression	naturalness	overall	suppression	naturalness	overall
ΔQM1	**0.918**	0.462	0.694	0.722	-0.444	0.361
	0.988	*0.385*	*0.509*	*0.885*	*-0.615*	*0.350*
$\Delta QM1_B$	0.788	**0.904**	**0.965**	0.749	-0.240	**0.833**
	0.582	***0.941***	***0.976***	*0.550*	*-0.124*	***0.918***
ΔQM2	0.439	0.231	0.331	0.512	-0.876	0.034
	0.794	*0.365*	*0.300*	*0.868*	*-0.932*	*0.312*
$\Delta QM2_B$	0.662	0.825	0.853	0.609	-0.420	0.741
	0.835	*0.797*	*0.815*	*0.797*	*-0.553*	*0.744*
ΔQM3	0.645	0.434	0.564	0.696	-0.772	0.242
	0.900	*0.538*	*0.518*	***0.959***	*-0.826*	*0.503*
$\Delta QM3_B$	0.725	0.765	0.829	0.752	-0.664	0.625
	0.841	*0.776*	*0.747*	*0.853*	*-0.632*	*0.735*
ΔQM4	0.288	0.424	0.419	0.297	-0.621	0.302
	0.585	*0.279*	*0.124*	*0.615*	*-0.721*	*0.253*
$\Delta QM4_B$	0.731	0.047	0.310	0.563	-0.109	-0.107
	0.750	*-0.088*	*0.088*	*0.665*	*-0.403*	*-0.115*
ΔQM5(sup)	0.887	0.291	0.570	0.654	-0.411	0.128
	0.950	*0.256*	0.415	*0.835*	*-0.591*	*0.221*
$\Delta QM5_B$(sup)	0.807	0.738	0.856	**0.808**	-0.567	0.589
	0.938	*0.574*	0.579	*0.947*	*-0.750*	*0.544*
ΔQM5(nat)	0.378	0.238	0.307	0.449	**-0.935**	0.046
	0.741	*0.256*	0.179	*0.791*	***-0.944***	0.197
$\Delta QM5_B$(nat)	0.644	0.790	0.817	0.639	-0.597	0.674
	0.826	*0.765*	0.782	*0.788*	*-0.579*	0.712

Table 4.1: Linear correlation and rank correlation (*italic*) between predicted and measured data obtained from hearing impaired and normal hearing subjects. Three categories were rated: noise reduction, speech naturalness and overall preference. The highest correlation values per category are emphasized (**bold** face). ΔQM5(sup) and ΔQM5(nat) denote the quality measure ΔQM5 with channel weighting adapted to the prediction of noise suppression and speech naturalness, respectively.

Influence of the BBA

The quality parameters based on the speech quality parameter q_C of Hansen and Kollmeier (2000) (i.e. ΔQM1, $\Delta QM1_B$) show the highest correlations with quality ratings of the hearing impaired subjects in all categories. Applying the BBA to the internal representation of the test signal leads to considerable differences of the prediction accuracy. Whether it improves or deteriorates the prediction depends on the category: While it worsens the correlation between measured and predicted quality ratings concerning "noise suppression", much better correspondence is achieved in the categories "speech naturalness" and "overall preference" (cf. Table 4.1). The reason for this contrary effect are the different subjective ratings of the processed (i.e., noise reduced) speech-plus-cafeteria-noise items relative to the unprocessed items in different categories on the one hand. Figure 4.2 shows that the processed speech-plus-cafeteria-noise items are subjectively rated higher than the unprocessed signals concerning "noise suppression", i.e. they are assigned positive BT-scale values. In contrast, these items are assigned negative BT-scale values in the other categories (see Figures 4.3 and 4.4). On the other hand, the objective quality differences ΔQM1 of these signals are affected much more by the BBA than those containing drilling machine noise. (Compare Figure 4.2 (without BBA) with Figure 4.3 (with BBA).)

Apparently, the magnitude of the BBA bias on the quality measures is related to certain signal properties. Possible relevant signal properties will be discussed in Section 4.4.

Comparison with further measures

In (Marzinzik, 2000; Marzinzik and Kollmeier, 2000), the capability of a number of objective speech quality measures for the assessment of noise reduction algorithms was investigated. The following measures were applied to the same signals and subjective ratings as used in this study: The speech quality measure q_C[3] (Hansen and Kollmeier, 2000) (denoted as PMF in the Marzinzik study), the Perceptual Speech Quality Measure PSQM (Beerends and Stemerdink, 1994), the Itakura-Saito Distortion IS (Itakura and Saito, 1979), the Log-Likelihood Ratio LLR (Itakura, 1975), the Log-Area Ratio LAR (Quackenbush et al., 1988), the Segmental Signal-to-Noise Ratio SSNR (Quackenbush et al., 1988) and the Weighted Spectral Slope WSS (Klatt, 1982).

[3]This quality measure was also applied in the present study (QM1), although a slightly different implementation was used.

In comparing the results obtained in the present study, the linear correlation coefficients between ratings of the hearing impaired subjects and the stated objective quality measures are reprinted from (Marzinzik, 2000) in Table 4.2. The comparison shows that the best correlations are achieved by the quality measure $\Delta QM1_{(B)}$ in the categories "noise suppression" and "overall preference", whereas the Log-Area Ratio (LAR) shows the highest linear correlation with the subjective ratings of the speech naturalness.

	subjective criterion		
objective measure	noise suppression	speech naturalness	overall preference
q_C	**0.90**	0.43	0.67
PSQM	0.65	0.85	0.87
SSNR	0.71	0.68	0.77
LAR	0.48	**0.92**	**0.87**
LLR	0.56	0.88	0.86
IS	-0.50	-0.06	-0.21
WSS	0.28	0.78	0.68

Table 4.2: Linear correlation coefficients between objective speech quality measures and mean quality ratings of hearing impaired subjects as reported in (Marzinzik, 2000).

4.3 Experiment II: assessment of multi-channel algorithms

The second set of subjective data that were predicted by objective audio quality measures originates from a subjective evaluation of speech enhancement systems used in automobiles. Tontch (2002) examined the speech enhancement by spatial filtering using different arrangements of microphone arrays in the cabin. In this section, the results of objective assessments of these speech enhancement schemes are compared with the corresponding subjective data presented in (Tontch, 2002).

4.3.1 Test signals

Two target speech signals were used: A German sentence of 2 s duration, read by a male and female speaker in an anechoic chamber. The target signals were played back in the cabin of a mid-size automobile via head and torso simulator (HEAD Acoustics) at three different speeds: 100, 130 and 160 km/h. The superposition of speech and automobile

noise was recorded by using different microphone arrays. The multi-channel recordings were processed by delay-and-sum beamforming algorithms yielding a single-channel output signal with enhanced speech. Eight different variants of beamformers were implemented. Including the unprocessed signal (i.e. single microphone recording), nine test signals per speed and target speech were generated. While the female target speech signal was used in each speed condition (≡ noise condition), the male speech was only used in the 100 km/h condition. This gave a total number of $4 \times 9 = 36$ test signals.

4.3.2 Subjective measurement

The experiment was carried out by the Hörzentrum Oldenburg. 10 normal hearing subjects participated in a paired comparison experiment. The experiment was divided into four blocks, one per noise (speed) condition and target signal. In each block, all of the beamforming algorithms (including no processing) where compared to each other regarding the overall preference. Test signals were presented dioticly via headphones in a sound-attenuating booth. Bradley-Terry scaling was applied to the results of each paired comparison test block.

4.3.3 Objective measurements

All of the objective quality measures described in Section 4.2 were applied to all test signals. Additionally, two rather simple measures, the signal-to-noise ratio (SNR) and the segmental SNR (segSNR)[4], were applied as well. These alternative measures were selected, because the background (car) noise in experiment II was of the same type in each condition and rather stationary. Moreover, noise reduction was achieved using multi-channel beamformers, which generally produce merely mild artifacts (comb filter effects), compared to those typically occurring if single channel algorithms are applied ("musical noise"). Thus, one could assume that the overall preference of human listeners is mainly determined by the (segmental) SNR and should therefore be predictable by this measure.

4.3.4 Results

Correlations between measured and predicted results for all objective measures are given in Table 4.3. It shows that subjective quality ratings according to the overall preference are predicted best by the objective audio quality measure ΔQM2. In contrast to experiment I,

[4]The segmental SNR is the mean SNR (in dB) over short (20 ms in the present study) segments.

the prediction performance deteriorates if the BBA is applied. The quality measure ΔQM5 does not yield the highest correlation with subjective ratings, because it was especially designed to predict the perceived speech naturalness or the amount of background noise. In contrast, the subjective quality criterion applied in this experiment was the overall preference. It is noteworthy that if a channel weighting adapted to the prediction of speech naturalness is applied (i.e., ΔQM5(nat)), the prediction accuracy of the overall preference concerning all beamformers except beamformers 3 and 4 is higher than that obtained with a channel weighting optimized for the assessment of background noise (i.e., ΔQM5(sup)). This contradicts the above stated assumption that the overall preference might essentially be determined by the perceived amount of background noise.

The results also show that the overall SNR and the segmental SNR do in fact correlate with the subjective ratings of the overall preference. However, the observed correlations are considerably smaller than those obtained by the psychoacoustical quality measure QM2. This finding emphasizes the advantage of psychoacoustically motivated computational measures, if perceived acoustical qualities are to be described or predicted by technical means, even in the case of assumed rather simple conditions.

Figure 4.6 shows subjective versus objective data obtained with QM2. Signals processed by two of the eight tested beamformers (denoted as number 3 and 4 in Figure 4.6) were rated worse by the subjects than the unprocessed signals ("UN"). This was due to an artifact of the corresponding beamforming algorithm: Its automatic source-locating procedure estimated wrong, discontinuous speaker positions during speech segments of low energy. As a consequence, the SNR did not improve and the processed signals contained discontinuities which were clearly audible and very annoying, especially at higher car speeds and/or male target speech. The annoyance in those cases was mostly underestimated by the objective quality measure. For comparison, Table 4.3 also shows correlation values if signals containing artifacts were discarded.

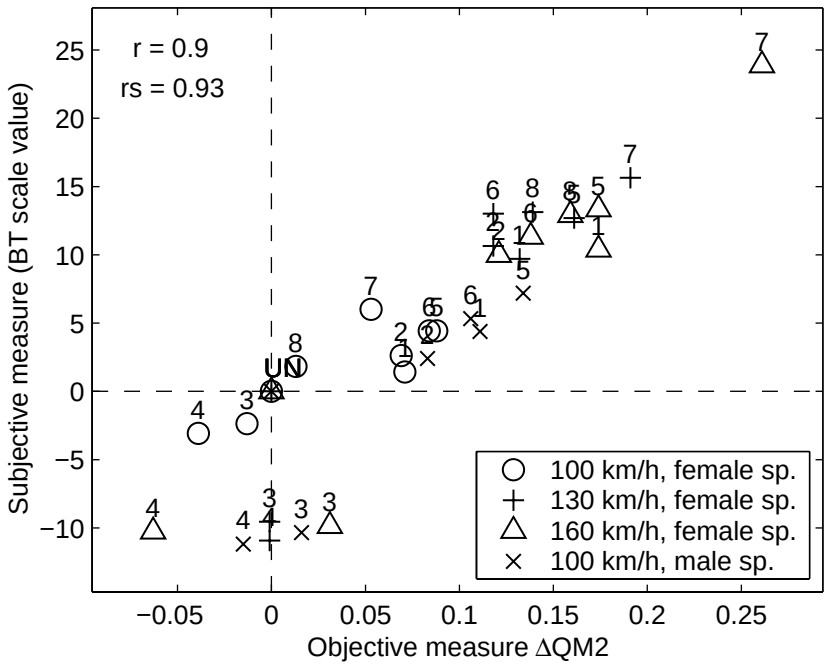

Figure 4.6: Quality prediction results of experiment II. Signals were rated according to the overall preference. Symbols refer to car speed and speaker, numbers indicate different beamforming algorithms. "UN" denotes unprocessed signals, whose quality values were arbitrarily set to zero.

measure	all beamformers	BMF 3, 4 excluded
ΔQM1	0.881 *0.893*	0.897 *0.877*
$\Delta QM1_B$	0.720 *0.744*	0.849 *0.828*
ΔQM2	**0.903** ***0.930***	**0.927** ***0.926***
$\Delta QM2_B$	0.826 *0.876*	0.904 *0.891*
ΔQM3	0.884 *0.853*	0.770 *0.697*
$\Delta QM3_B$	0.857 *0.837*	0.773 *0.694*
ΔQM4	0.825 *0.775*	0.657 *0.530*
$\Delta QM4_B$	0.782 *0.734*	0.527 *0.431*
ΔQM5(sup)	0.892 *0.861*	0.801 *0.726*
$\Delta QM5_B$(sup)	0.870 *0.844*	0.818 *0.714*
ΔQM5(nat)	0.882 *0.894*	0.907 *0.908*
$\Delta QM5_B$(nat)	0.782 *0.854*	0.905 *0.905*
SNR	0.711 *0.680*	0.650 *0.492*
segSNR	0.603 *0.610*	0.715 *0.625*

Table 4.3: Linear correlation and rank correlation (*italic*) coefficients for predicted and measured quality ratings. The highest correlation values are emphasized (**bold** face). The right column shows correlation values obtained if signals processed by beamformers 3 and 4 are discarded.

4.4 Discussion

The results of experiment I and II show that an optimal prediction of the perceived quality can not be achieved by a single audio quality measure, if different subjective quality criterions or groups of subjects (normal hearing or hearing impaired) are considered. Moreover, the presence or absence of artifacts does also affect the correlation between subjective and objective assessments. However, at least one measure can be found in each condition that achieves good correspondence with the subjective ratings.

As far as normal hearing subjects are considered, the quality measure QM5 is most suitable to predict the perceived naturalness of speech or the amount of background noise, depending on the selected mode (i.e., channel weighting function) of this measure.

If the overall preference is to be predicted instead, best results are achieved without channel weighting. Which of the quality measures QM1, QM2 or QM3 is the most adequate for this task in general can not be finally concluded from the somewhat inconsistent results obtained from the two presented experiments.

QM4, representing the audio quality measure $\mathrm{PSM_t}$ (cf. Chapter 2), generally corresponds poorly with subjective data. This may be explained by the fact that this measure was especially designed for the prediction of small audio quality differences, focusing on the largest short-time perceptual differences between test and reference signals. Due to the presence of continuous background noise in the present test signals, very large short-time perceptual deviations from the clean speech reference can occur for any test signal. Thus, the approach of the quality measure $\mathrm{PSM_t}$ does not seem appropriate for the present task.

In the case of hearing impaired subjects, the quality measure QM1 ($\mathrm{QM1_B}$, respectively) shows the highest correlations with subjective ratings in all categories, while only the overall preference is predicted best by this measure if normal hearing subjects are considered. Possible reasons for this finding will be discussed in Section 4.4.2.

Applying the Beerends-Berger-assimilation (BBA) to the internal representations improves the prediction performance in the presence of artifacts introduced by the noise reduction schemes. This might be explained by the trade-off between the degree of noise reduction and the annoyance of artifacts, which are typically positively correlated. The BBA appears to model this trade-off. The effect of the BBA on the quality estimates will be investigated in more detail in the following section.

4.4.1 Factors influencing the BBA bias on the quality estimates

The results of experiment I and experiment II show that the effect of the BBA on the quality prediction performance varies with respect to the signal, the quality aspect and the objective measure. In order to understand the basic underlying mechanisms, this section investigates the relation between the bias of the BBA on the objective quality estimates and signal properties.

The BBA assimilates internal representations (IR) by replacing elements of the test IR that are smaller than corresponding elements of the reference IR by the mean value of both elements, thus halving the former differences. For a given correlation between the original test and reference IR, the effect of the assimilation on the correlation increases with larger differences. In the extreme case of an all-zero test IR for example, the assimilation would result in a perfect correlation, i.e., $\mathrm{QM}_\mathrm{B} = 1$. Therefore, the bias of the BBA on the quality measures should be directly related to the ratio of the test and reference IR magnitudes.

To test this assumption, Figure 4.7 shows the BBA bias in terms of differences between Fishers-Z transformed correlation values before and after applying the BBA to test signals (left panel) and to the test IR of experiment I (right panel). The test signals were generated by adding normal distributed noise to a sine wave, while the latter served as the reference signal. A range of magnitude ratios was realized by different scalings of the test signals. Additionally, the SNR was varied in order to obtain different correlation values. The results displayed in Figure 4.7 confirm the assumed positive correlation between magnitude ratio and BBA bias. The influence of the correlation between test and reference signal before the BBA on the bias is indicated by the two curves in the left panel, which represent very high and very low correlations between reference and test signals. (Average correlation values of $1 - 10^{-6}$ are represented by the dashed line, mean correlations of 0.28 by the solid line.) Intermediate correlations would be represented between these lines, which converge with an increased magnitude ratio. In the right panel, the BBA bias on the quality measure QM2 applied to the speech signals of experiment I is plotted versus the magnitude ratios of the corresponding IR. The bias is given by $F(\mathrm{QM2_B}) - F(\mathrm{QM2})$ (with F denoting the Fishers-Z transform). Signals and noise reduction schemes are indicated by the same symbols and numbers as before (cf. Figures 4.2 - 4.4). Obviously, the unprocessed signals containing cafeteria noise (circles and crosses in Figure 4.7) are much more affected by the BBA than the corresponding noise-reduced signals. In contrast, unprocessed signals with drilling-machine noise are affected similarly to processed signals. This reflects the observed marked shifts of the ΔQM1 (=QM1(noise reduced)-QM1(unprocessed)) values of

the cafeteria-noise items toward negative values, whereas the drilling-machine-noise items are hardly affected (compare Figure 4.2 with 4.4)[5].

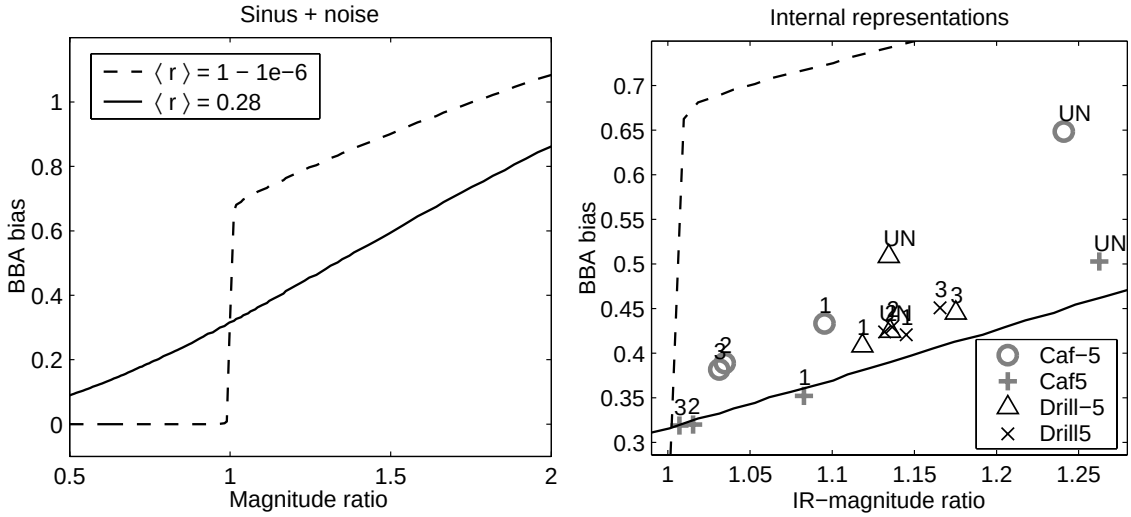

Figure 4.7: Relation between the bias of the BBA on the correlation between signals and the average magnitude ratio of these signals. The results shown on the left hand side were obtained using sine waves with additive gaussian noise. The internal representations of the noisy speech signals used in experiment I yielded the results shown in the right panel. The BBA bias is expressed in terms of differences between Fishers-Z transforms of correlation coefficients (QM2, respectively) for pairs of signals before and after the BBA. The two curves shown in the left panel indicate the influence of the correlation between signals before the BBA. The dashed line corresponds to a high correlation coefficient ($r = 1 - 10^{-6}$), whereas low correlations ($r = 0.28$ on average) are represented by a solid line. Digits indicate different noise reduction algorithms, "UN" denotes unprocessed signals. (Note the different scaling of the axes.)

The reason why IR differences between clean speech and speech plus noise are larger in the case of cafeteria noise than for drilling machine noise is found in the different spectral distributions of these noises. While the power spectrum of the cafeteria noise is similar to the long-term spectrum of speech, the spectrum of the drill noise shows a rather highpass characteristic (see left panel of Figure 4.8). Thus, the SNR in the frequency channels that carry most of the speech energy is worse if cafeteria noise is added than if drill noise is added at the same overall SNR. Consequently, the IR of the speech-plus-drill-noise signal reflects the speech stronger than the IR of the speech-plus-cafeteria-noise signal. This is shown in the right panel of Figure 4.8, where sample intervals of the IR (averaged across frequency channels) of these signals are compared with that of clean speech. (The overall SNR of the speech-plus-noise signals is -5 dB in each case.) The average IR magnitude of

[5]Shifts towards smaller values are also observed using ΔQM2 and ΔQM3, but less distinct.

the signal containing cafeteria noise is smaller and less correlated with the reference IR than in the case of the signal containing drill noise. As a result, the correlation between the IR of the former signal and the reference and thus the quality measure will be affected stronger by the BBA. Therefore, the assumed relation between the BBA bias and IR ratios can be concluded.

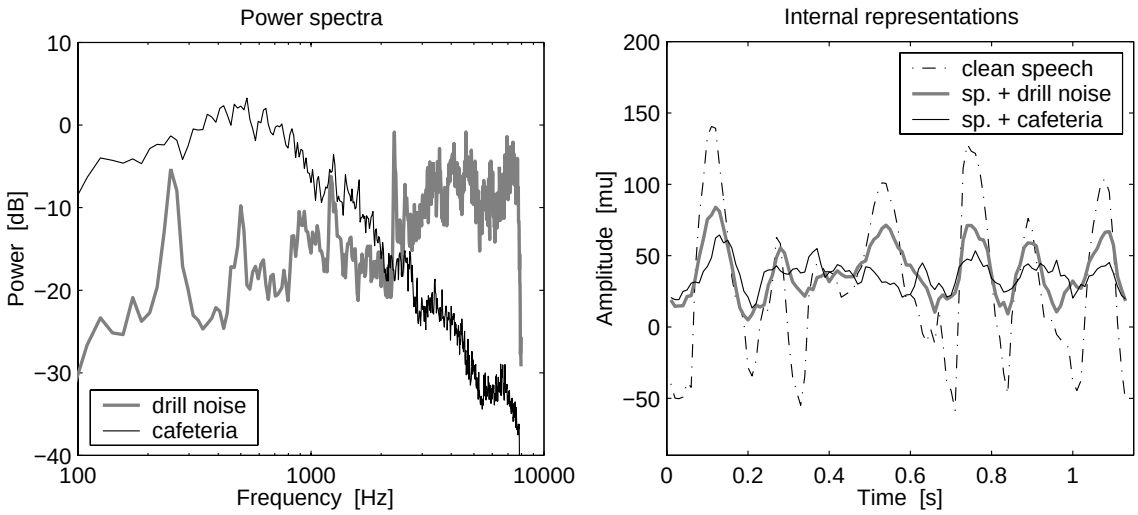

Figure 4.8: Left panel: Power spectra of the cafeteria noise and the drilling machine noise used in experiment I. Right panel: sample interval of internal representations, averaged across frequency channels, of clean speech (dash-dotted line), speech plus drilling machine noise at -5 dB SNR (thick gray line) and speech plus cafeteria noise at -5 dB SNR (solid black line).

In experiment II, the best prediction of the overall preference was achieved with the parameter QM2. In contrast to experiment I, the BBA does not improve but even somewhat deteriorates correlations between measured and computed ratings in this experiment, because its effect on those signals that were subjectively rated worse than the unprocessed reference is reverse with respect to those of experiment I (see Figure 4.9). The negatively rated signals were produced by two beamforming algorithms (beamformers 3 and 4), which did not reduce background noise but produced annoying discontinuities due to the erratic behaviour of the inter-channel delays determined by a malfunctioning automatic speaker localization procedure. All of those signals were rated worse by the subjects than if unprocessed and also worse by the objective audio quality parameter QM2 in most cases (cf. Figure 4.6). In this case, the BBA decreases the absolute quality differences (i.e., $|\Delta\mathrm{QM2_B}| < |\Delta\mathrm{QM2}|$) for most of the items, as shown in Figure 4.9. The degradation of the overall correlation due to the BBA in experiment II is caused by sign reversals and also

by increasing quality differences of some of the negatively rated signals with very small ΔQM2 values.

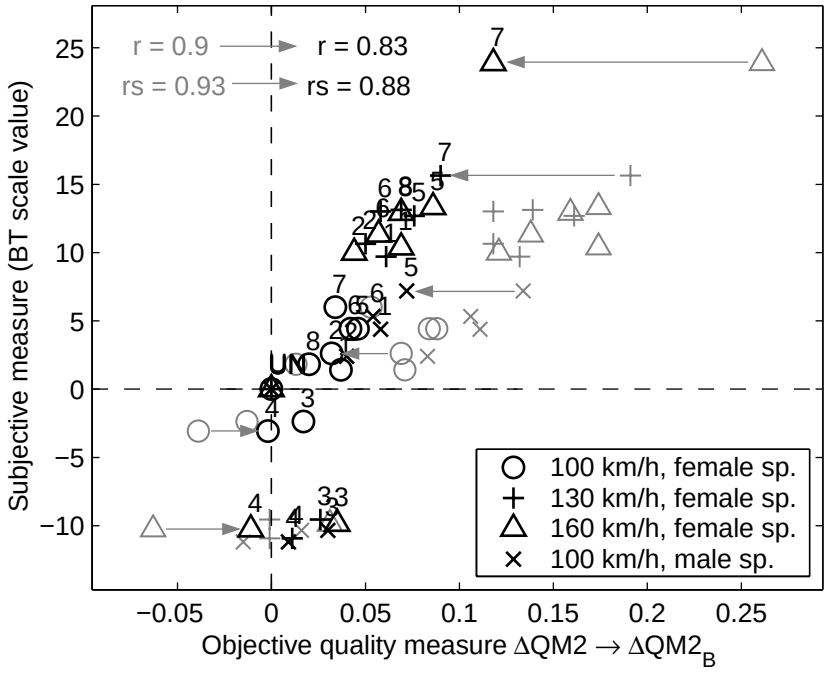

Figure 4.9: Effect of additional BBA on the predicted overall preference in experiment II. The quality estimates obtained with ΔQM2 (i.e., without BBA) are indicated by gray symbols, while black symbols and digits represent $\Delta QM2_B$ values. (The digits indicate different beamformers.) Most of the quality estimates are shifted towards smaller absolute values, if the BBA is applied. Note that some of the negatively rated signals deviate from this trend, which causes a degradation of the overall correlation.

4.4.2 Differences between quality measures

Several variants of audio quality measures were investigated with regard to the perceptual assessment of noise reduction schemes. These variants were derived from the objective speech quality q_C of Hansen and Kollmeier and introduced in Chapter 2. A set of different measures were used, because it has been reported in the literature that the qualities of noise reduction schemes are usually subjectively rated differently, if different subjective quality criterions are applied (e.g. Marzinzik and Kollmeier, 2000). Thus, predicting different quality categories requires different measures.

The objective measures investigated in this study were selected with respect to different assumed target properties: The speech quality measure QM1 ($\equiv q_C$) and the more generalized audio quality measures $QM3_B$ (≡PSM) and $QM4_B$ (≡PSM_t) have been shown

to be suitable for the prediction of perceived overall quality differences between speech and audio signals, respectively (Hansen and Kollmeier, 2000, and Chapter 2 of this thesis). Therefore, these measures might also be qualified to predict the subjective overall preference. $\text{QM2}_{(\text{B})}$ was also tested, because it represents a kind of intermediate speech/audio quality measure; while it uses the same bandwidth of the peripheral filterbank as PSM and PSM_t, amplitude modulations are processed by a modulation lowpass filter instead of a modulation filterbank. q_C applies a modulation lowpass filter as well, but uses a peripheral filterbank of restricted bandwidth (adapted to telephone band). Quality measure QM5 was especially designed to estimate the perceived naturalness of speech, or, alternatively, the amount of background noise, depending on the selected operation mode of this measure.

As expected, the results reveal differences in the predictive ability of the employed quality measures across quality criterions, but also across subjects, experiments and concerning the influence of the BBA. In the following, the question of possible reasons for the observed differences is addressed.

Relation between QM1 and hearing impaired subjects

In experiment I, measure $\text{QM1}_{(\text{B})}$ shows the highest correlations with quality ratings of the hearing impaired subjects in all categories, whereas only the overall preference is predicted best by QM1_B in the case of normal hearing subjects. Two possible reasons for this are to be considered: 1) the limited bandwidth of the peripheral filterbank used in measure QM1, 2) the non-uniform frequency weighting applied in QM1 (see Figure 4.10).

Measure QM1 represents the speech quality measure q_C of Hansen and Kollmeier (2000), originally optimized for the prediction of the quality of distorted telephone-band speech. For this purpose, a gammatone filterbank is used that covers the frequency range of the telephone band, ranging from 320 Hz to 4 kHz. The median hearing loss of the hearing impaired subjects that participated in experiment I exceeds 60 dB HL at frequencies greater than 4 kHz (up to nearly 80 dB at 8 kHz), which was partially compensated by a third-octave band equalization (range of $\pm$ 16 dB in each band), following the half-gain rule (cf. Marzinzik, 2000). The bandwidth of the used test signals was 8 kHz. Thus, the quality measure QM1 does not account for frequencies higher than 4 kHz and might thus roughly account for the reduced sensitivity of the subjects in the high frequency range. If the reduced bandwidth of QM1 compared to QM2 would be the only reason for the superior prediction performance of this measure for hearing impaired subjects, then

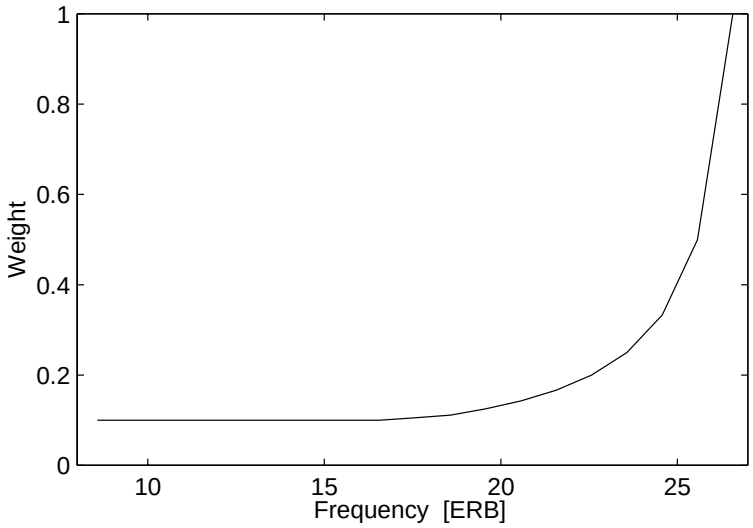

Figure 4.10: Band weighting characteristic of QM1.

disabling the frequency weighting in QM1 should not deteriorate the correlation with the subjective data. In particular, the performance of $\mathrm{QM1}_{-w}$ (i.e. QM1 without frequency weighting) should still beat the performance of QM2. However, this can not be confirmed by the data. Table 4.4 shows correlation values for quality measures QM1, with and without frequency weighting, and QM2. The stated correlation coefficients obtained with QM1 deteriorate markedly, if the frequency weighting is disabled, becoming smaller than the corresponding values obtained with QM2.

Thus, it can be concluded that the reduced bandwidth can not be the only or major reason for the superior performance of QM1. In contrast, the frequency weighting seems to play an important role in the prediction of ratings of hearing impaired subjects. Hearing impaired subjects possibly put more weight on those of the intact frequency bands neighboring lost frequencies regions. In case of a high-frequency hearing loss (which was characteristic for the hearing impaired subjects that participated in experiment I), this applies to the frequency bands emphasized by the quality measure QM1.

Predicting the overall preference of normal hearing subjects

The fact that the overall preference of the normal hearing subjects is predicted best by $\mathrm{QM1}_{\mathrm{B}}$ (although not as good as in the case of the hearing impaired subjects) is explained by the similar ratings of the normal hearing compared to the hearing impaired subjects in this category, preferring unprocessed speech with cafeteria noise to the processed versions

measure	hearing impaired			normal hearing		
	suppression	naturalness	overall	suppression	naturalness	overall
ΔQM1	0.918	0.462	0.694	0.722	-0.444	0.361
	0.988	*0.385*	*0.509*	*0.885*	*-0.615*	*0.350*
ΔQM1$_{-w}$	**0.408**	**0.153**	**0.259**	**0.445**	**-0.902**	**-0.049**
	0.744	***0.171***	***0.153***	***0.759***	***-0.924***	***0.100***
ΔQM2	0.439	0.231	0.331	0.512	-0.876	0.034
	0.794	*0.365*	*0.300*	*0.868*	*-0.932*	*0.312*

Table 4.4: Correlations between predicted and measured data obtained with quality measure QM1 without band importance weighting (**bold** face) compared to the original QM1 and QM2.

and vice versa for speech mixed with drilling machine noise. As a result, negative BT-scale values were assigned to the processed speech-plus-cafeteria-noise items and positive to the remaining. The required corresponding negative objective quality measure differences ΔQM are only obtained if the BBA is applied. (See Appendix D, Figure D.1.) Moreover, the effect of the BBA shifting the quality estimates towards negative values is restricted to those signals containing cafeteria noise and continuously weakens for parameters $QM2_B$, $QM3_B$ and $QM4_B$, in the sense that just some of the items in question are assigned negative quality values. (This is also shown and explained in Appendix D.) Hence, the prediction performance of these measures deteriorates.

In experiment II, the prediction performance of the quality measures $\Delta QM3_{(B)}$ and $\Delta QM4_{(B)}$ is lower than that of $\Delta QM2_{(B)}$, because the perceptually most relevant differences between unprocessed and noise-reduced speech signals primarily occur in the modulation frequency band of speech (i.e. around 4 Hz). As QM3 and QM4 are based on an auditory model version that also accounts for considerably higher modulation frequencies, it appears that taking additional information from this higher modulation frequency range into account likely impairs the prediction accuracy, because the importance of this information is overestimated.

4.4.3 Is there a task-dependent weighting of channels?

The frequency weighting introduced by Hansen and Kollmeier (2000) as "band importance weighting", strongly emphasizes the uppermost frequency bands of the used peripheral filterbank with restricted bandwidth (see Figure 4.10). Interestingly, it also improves the prediction accuracy for the ratings of normal hearing subjects concerning the amount of background noise, while it deteriorates the performance in the category of speech natural-

ness (cf. Table 4.4). This finding might indicate a task-dependent weighting of channels and inspired the development of a modified quality measure that was especially adapted to the prediction of speech naturalness and background noise. This measure was introduced in Section 4.2 as QM5 and represents a straight-forward realization of employing an adaptive channel weighting. Assuming that subjects focus on the frequency and modulation frequency bands that carry most of the speech energy if they are assessing the naturalness of speech, the new measure puts high weights on these channels. If, in contrast, the amount of interfering background noise is to be estimated, it appears reasonable to listen into channels where the noise is least "disturbed" by interfering speech, so one can concentrate better on the noise. This approach is implemented in the quality measure QM5 by weighting each channel with the product of the noise-to-speech ratio and the noise energy in that band.

The comparatively high predictive ability obtained with this measure presented in Section 4.2.4 (Table 4.1) supports the assumption of different, task-dependent weightings of auditory channels.

4.4.4 Outlook: Further possible extensions of current quality measures

The results presented in this study revealed strengths and weaknesses of the 10 investigated variants of objective quality measures in predicting the different quality categories, subjects and noise/artifact conditions. Although promising results could already be achieved, further work is required to validate and possibly further improve the current quality measures. Two main approaches should be followed:

a) The "data driven approach": More independent data should be gathered, i.e. further subjective measurements should be carried out to validate the measures that were optimized with respect to a certain data set by using different, independent data sets.

b) The "model driven approach": New measures should be developed or existing measures should be modified to become even stronger related to models of human perception.

For example:

- Apart from the reduction of the noise and the distortion of the speech, the *distortion of the noise* due to processing artifacts, which might also influence the subjective overall preference, has not been taken explicitly into account yet. One could think of a combined measure, "symmetrically" treating the speech naturalness, the changes of the noise quality and the amount of the noise reduction.

- The use of alternative perceptual reference situations: If possible changes of the quality of the background noise needs to be assessed as well, it appears reasonable to replace the clean speech target by a speech signal with original noise at an increased SNR.

- Special adaptations of the measures/models to hearing impaired listeners. Apart from accounting for known different properties of the impaired auditory system, such as attenuations or losses of certain frequency ranges and reduced dynamic compression, there are indications of different internal weightings of frequency bands by hearing impaired listeners, which should be further investigated and accounted for by new quality measures for hearing impaired subjects.

4.5 Summary and conclusion

The capability of speech and audio quality measures for the perceptual evaluation of noise reduction schemes for speech was examined in the present study. Because of the sparse underlying experimental database available, this work rather serves as a pilot study to derive first hypotheses. The following conclusions drawn from the results of the present study represent such first hypotheses, which need to be tested by further experimental data.

The findings of the presented experiments indicate that no single audio quality measure is able to properly predict subjective quality ratings for all conditions and subjects. Instead, a set of several variants are required, accounting for different criterions applied to the evaluation of noise reductions schemes and also for different groups of subjects (normal hearing vs. hearing impaired). With such a set of quality measures, good correspondences between subjective and computed assessments of noise reduction schemes were achieved in each condition. If the naturalness of speech is to be assessed, the quality measure should focus on those frequencies and modulation frequencies that contain most of the speech energy. If the amount of the perceived noise (reduction) needs to be predicted instead, channels with rather low speech-to-noise ratios should be emphasized. This concept is realized by the modified audio quality measure QM5, which represents the measure PSM (without BBA) introduced in Chapter 2 of the present work, expanded by an adaptive channel weighting of the internal representations.

If the overall preference is chosen as quality criterion, a well-founded recommendation of the most appropriate measure can hardly be made in face of the somewhat inconsistent results obtained from the few experimental data available so far. The quality measure QM2 appears to be potentially suitable for this task. If there are speech distortions due to artifacts of the noise reduction scheme, the application of BBA in the quality measurement appears to model the trade-off between noise reduction and speech naturalness. Hence, a higher correlation with subjective ratings seems to be found for measure $\mathrm{QM2}_{(\mathrm{B})}$ only in those conditions where speech distortions occur in addition to the noise reduction.

Subjective quality ratings of hearing impaired subjects were predicted best by the quality measure QM1 (i.e. the speech quality measure q_C of Hansen and Kollmeier (2000)), at least partly caused by the non-uniform weighting of frequency bands applied within this measure. This is somewhat surprising, since this measure was originally adapted to telephone-band filtered speech but not to hearing impaired listeners. Further evaluations are required to check whether features of this quality measure possibly account for special

properties of hearing impaired listeners. Moreover, the application of modified quality measures specially designed to properly account for impairments of the auditory system seems promising.

Future work should be invested to validate the presented quality measures using independent data, but also to develop further extensions of the current quality measures in order to become even stronger related to models of human perception.

CHAPTER 5

Summary and conclusions

In this thesis, the applicability of the auditory processing model of Dau et al. (1997a) for the purpose of audio quality prediction was shown. In Chapter 2, a novel method for the objective, perceptual assessment of quality differences between audio signals was introduced. It represents an expansion of the speech quality measure q_C of Hansen and Kollmeier (2000), who successfully applied their method to predict the transmission quality of low bit rate telephone speech codecs. The basic approach was adopted in the present work: The auditory model of Dau et al. (1996a, 1997a, respectively) is employed to transform a pair of reference and test signals into corresponding internal representations. The linear cross-correlation coefficient of these internal representations serves as a measure for the perceptual similarity between test and reference signals, which is interpreted as the perceived audio quality of the test signal relative to the quality of reference signal. However, the extension of the method from narrow-band speech to any kind of broad-band audio signal and from clearly audible to just perceptible distortions required some methodical modifications and expansions: The bandwidth of the peripheral filterbank was extended and the modulation lowpass filter was replaced by a modulation filterbank (cf. Dau et al., 1997a). Apart from these modifications concerning the modeling of the auditory signal processing, the original method was expanded by further stages that model more cognitive aspects of audio quality perception. In fact, the "band importance weights" applied in the speech quality measure q_C likely model a cognitive aspect as well. However, the necessity of a non-uniform frequency weighting could not be confirmed by the results of the present work. Instead, it was found that a sign-dependent weighting of differences between internal representations of test and reference signals somewhat improves the accuracy of the quality prediction. The most substantial expansion of the method, however, is represented by the modeling of the relation between instantaneous and overall audio quality. This was realized by computing a sequence of short-time cross correlation coefficients, weighting this sequence by the moving average of the internal representation of the test signal and finally calculating the 5%-quantile of the weighted sequence. Without accounting for this

relation, good correlations between measured and predicted audio qualities could only be obtained if different signal types were considered separately.

The new method was developed and tested using a large database of subjectively rated audio signals. The results showed that the performance of the presented method and of the ITU standard BS.1387 ("PEAQ") are on average comparable. However, comparisons using the given database are not quite equatable, because PEAQ uses an artificial neural network that was trained using this database. Apart from the restricted applicability for different tasks, another drawback of using a neural network (from a modelers point of view) is its "black box" nature, which makes it difficult to draw conclusions and learn about the actual mechanisms involved in human perception of audio quality. This fact constitutes a major advantage of the presented new method over the ITU standard.

While only the "preprocessing" part of the auditory model described in (Dau et al., 1997a) was used for the computation of the audio quality measure $\mathrm{PSM_t}$, Chapter 3 describes the application of the entire model, including the final detector stage, for the prediction of detection thresholds of audio distortions. The simulated data showed a good correlation with the measured data, which were obtained applying an experimental setup well established for "classical" masking experiments. Because of the use of complex, broadband stimuli, such good prediction performance could not be expected a priori, since the model does not account for across channel processing. However, the somewhat worse prediction accuracy of the modulation filterbank model compared to the lowpass version indicates that across-channel processes do play a role and should therefore be considered in future model versions.

The measurement of detection thresholds of audio distortions represents an alternative approach to evaluate the transmission quality of near transparent audio processing schemes.

In the final chapter of this thesis, the capability of several variants of the audio quality measures presented in Chapter 2 for the perceptual evaluation of noise reduction schemes for speech was investigated. Subjective data obtained from measurements carried out by Marzinzik (2000) and the Hörzentrum Oldenburg on behalf of K. Tontch (published in Tontch, 2002) were compared with corresponding objective data obtained with the quality measures. Since the subjective assessments differed with regard to the considered quality aspect (speech naturalness, amount of background noise, overall preference) and group of

subjects (normal hearing, hearing impaired), no single objective measure was sufficient to predict all of the subjective ratings. However, at least one of the tested quality measure variants could be found in each condition that showed good correlation with the measured data.

Based on these findings, a new quality measure was derived that is especially adapted to the prediction of the speech naturalness and the amount of noise. This measure essentially corresponds to the quality measure PSM presented in Chapter 2, but additionally applies a task-dependent band weighting: If the naturalness of the speech is to be evaluated, those frequency and modulation frequency bands are weighted stronger that contain considerable portions of the speech energy. If the amount of the perceived noise (reduction) is to be predicted instead, bands with rather low speech-to-noise ratios are emphasized. The modified measure achieved high correlations with the subjective data in the stated categories.

The question of the best measure for predicting the overall preference could not be answered consistently by the results obtained so far. Further experiments are required to conclude that question.

Finally, the presented results suggested that subjective quality ratings obtained from hearing impaired subjects should be predicted using a quality measure that is adapted to account for special properties of the impaired auditory system.

In conclusion, this thesis provides a theoretical framework for the perceptual evaluation of audio processing systems that is based on practical applications. In addition, it demonstrates the wide applicability and validity of the used auditory processing model: This model does not only explain effects of signal detection at threshold successfully, but also provides a perceptual representation of acoustic stimuli that appears to be a suitable input for modeling higher processes in auditory perception.

A further application in the field of audio quality assessment, for example, could be the identification and measurement of certain qualities of audio impairments in addition to the one-dimensional impairment of the overall audio quality (such as roughness, fluctuation, sharpness or other psychophysical attributes). However, this extension would require a better knowledge and model of human perception and interpretation of acoustical information than currently available. Even though this thesis has provided some progress in this area, more work will have to be invested.

APPENDIX A

DESCRIPTION OF THE DATABASE

The database used for the development of the audio quality measures presented in this work consists of six data sets that evolved from listening tests performed for the evaluation of various low-bit rate audio codecs. An additional data set (DB3), which was was especially created for the validation of the ITU standard BS.1387, was set aside until the end of the development phase and finally applied for validation and comparison with the ITU standard BS.1387. The data sets consist of audio files[1] and corresponding subjective ratings averaged across listeners and will be summarized briefly in the following.

A.1 Data set name: MPEG90

This data set evolved from a listening test described in (ISO/MPEG, 1990).

The audio reference material consists of ten stereo sequences: bass guitar, bass synth, castanets, Tracy Chapman, Ornette Coleman, glockenspiel (chimes), fireworks, English male speech and trumpet (Haydn).

The reference signals were processed by two codecs (MUSICAM (Stoll and Dehéry, 1990) and ADPCM (ITU-T, 1990)) at three bit rates (64, 96 and 128 kbit/s/channel), yielding 50 test signals (only five reference signals were processed at the lowest bit rate).

While only five subjects rated the audio quality degradation in the test condition with the lowest bit rate, 59 to 70 subjects participated in the remaining test conditions of this listening test.[2]

[1] The lengths of the audio fragments contained in these files are in the range from 10 to 32 s; the average length is 23 s.

[2] The subjects that participated in all of the listening tests described in this appendix were characterized as "expert listeners" (ITU-R, 1998a).

A.2 Data set name: MPEG91

This data set evolved from a listening test described in (ISO/MPEG, 1991).

The audio reference material consists of ten stereo sequences: accordion/triangle, electric bass guitar, Carmen, Ornette Coleman, George Duke, glockenspiel, English male speech, percussion, tambourine and Suzanne Vega.

The reference signals were processed by six codecs (MPEG layer 1-3 (ISO/MPEG, 1992), MUSICAM (Stoll and Dehéry, 1990), ASPEC (Brandenburg et al., 1991) and NICAM (ETSI, 1998)) at three bit rates (64, 96 and 128 kbit/s/channel), yielding 105 test signals. (Some of the codecs only operated at the highest bit rate and not all of the reference signals were processed by each codec condition.)

Depending on the test condition, 40 to 91 subjects participated in this listening test.

A.3 Data set name: MPEG95

This data set evolved from a listening test described in (Meares and Kim, 1995).

The audio reference material consists of six mono sequences: bag pipe, castanets, glockenspiel, harpsichord, pitch pipe and English female speech.

The reference signals were processed by 22 encoding variations of six audio codecs provided by AT&T, Fraunhofer, Sony, GCL, RAI/Alcatel and Philips, yielding 132 test signals.

63 subjects participated in this listening test.

A.4 Data set name: ITU92DI

This data set evolved from a listening test described in (ITU-R, 1992).

The audio reference material consists of twelve stereo sequences: Asa Jinder, electric bass guitar, castanets, Ornette Coleman, Dalarnas Spelmarsforbund (Swedish Folk), Dire Straits, Ravel ("Feria"), harpsichord, triangels and Stravinsky ("Wind Octet").

The reference signals were processed by five codecs (MPEG layer 2 + 3 (ISO/MPEG, 1992), Dolby AC-2 (Fielder et al., 1996), AWARE and NHK) at 120 kbit/s/channel, yielding 60 test signals.

23 subjects participated in this listening test.

A.5 Data set name: ITU92CO

This data set evolved from a listening test described in (ITU-R, 1992).

The audio reference material consists of ten stereo sequences: Asa Jinder, electric bass guitar, castanets, Ornette Coleman, Dalarnas Spelmarsforbund (Swedish Folk), harpsichord, triangels and Stravinsky ("Wind Octet").

The reference signals were processed by six codecs (MPEG layer 2 + 3 (ISO/MPEG, 1992), Dolby AC-2 (Fielder et al., 1996), Dolby Low-Delay, AWARE and AT&T DSQ 5TR620) at 180 kbit/s/channel, yielding 60 test signals. Each item was processed by the same codec three times in tandem.

19 subjects participated in this listening test.

A.6 Data set name: ITU93

This data set evolved from a listening test described in (ITU-R, 1993).

The audio reference material consists of seven stereo sequences: Asa Jinder, bagpipe, bass clarinet, castanets, harpsichord, German male speech and violin.

The reference signals were processed by different tandem code configurations of MPEG layer 2 (ISO/MPEG, 1992) at 256 and 192 kbit/s/channel, yielding 42 test signals.

33 subjects participated in this listening test.

A.7 Data set name: ITU92CO

This data set evolved from a listening test described in (ITU-R, 1992).

The audio reference material consists of ten stereo sequences: Asa Jinder, electric bass guitar, castanets, Ornette Coleman, Dalarnas Spelmarsforbund (Swedish Folk), harpsichord, triangels and Stravinsky ("Wind Octet").

The reference signals were processed by six codecs (MPEG layer 2 + 3 (ISO/MPEG, 1992), Dolby AC-2 (Fielder et al., 1996), Dolby Low-Delay, AWARE and AT&T DSQ 5TR620) at 180 kbit/s/channel, yielding 60 test signals. Each item was processed by the same codec three times in tandem.

19 subjects participated in this listening test.

A.8 Data set name: DB3

This data set evolved from three listening tests described in (ITU-R, 1998a).

The audio reference material consists of 27 stereo sequences: bagpipe, castanets, clarinet, claves, drum, English and German female speech, English and German male speech, flute, glockenspiel, harpsichord (2×), kettle drum, marimba, piano, pitch pipe, Ry Cooder, saxophone, snare, soprano, strings, Suzanne Vega , tambourine, triangle, trumpet, tuba and xylophone.

The reference signals were processed by six codecs (ATRAC (MiniDisc) (Tsutsui et al., 1996), MPEG layer 2 + 3 (ISO/MPEG, 1992), Dolby AC-2 + AC-3 (Fielder et al., 1996) and MPEG AAC (Bosi et al., 1997)) alone and in tandem (including cascadings of different codecs) at bit rates from 128 to 256 kbit/s/2 channels, and by adding quantization distortions, harmonic distortions and additive noise. A selection of 84 test signals in total were used in the listening tests. (39 to 44 items were used per listening test.)

28 to 33 subjects participated in the listening tests.

APPENDIX B

Quality prediction results per data set

The following figures show the quality prediction results for the six data sets that constitute the overall database used in the present work. In addition, the results obtained with the data set DB3 that were used for the comparison with the ITU standard BS.1387 are shown as well. In each figure, subjective quality ratings (mean over subjects) in terms of the *Subjective Difference Grade* (SDG) are plotted versus corresponding objective quality ratings obtained with the audio quality measure aq_{ct} for various test items (audio signal fragments impaired by lossy audio codecs). Audio codecs are discriminated by different colors, while the types of the audio signals are indicated by different symbols. (Because of the very large number of types of audio signals and distortions in data set DB3, the items of this data set were not broke down by different symbols and colors.) The symbols are explained in the legends using the following abbreviations: *acc* = accordion, *bag* = bag pipe, *bas* = electric bass guitar, *car* = Carmen, *cas* = castanets, *cha* = Tracy Chapman, *cla* = clarinet, *col* = Ornette Coleman, *dal* = Dalarnas Spelmarsforbund, *dir* = Dire Straits, *duk* = George Duke, *fir* = fireworks, *glo* = glockenspiel, *har* = harpsichord, *jin* = Asa Jinder, *per* = percussion, *pit* = pitch pipe, *rav* = Ravel ("Feira"), *spe* = speech, *str* = Stravinsky ("Wind Octet"), *syn* = bass synth, *tam* = tambourine, *tri* = triangle, *tru* = trumpet (Haydn), *veg* = Suzanne Vega, *vio* = violin. The linear and rank correlation coefficients (r, rs) are stated in the upper left corner of each figure. The bracketed values correspond to linear correlation coefficients that are obtained if the objective ratings are transformed by the regression functions indicated by the dashed lines.

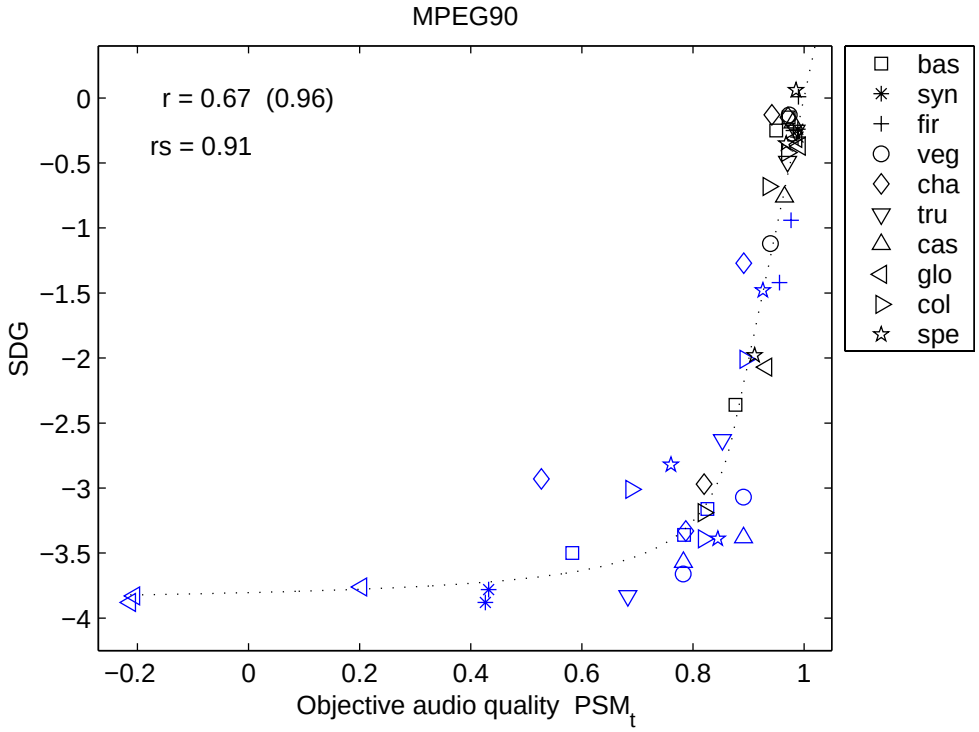

Figure B.1: Quality prediction results for data set MPEG90.

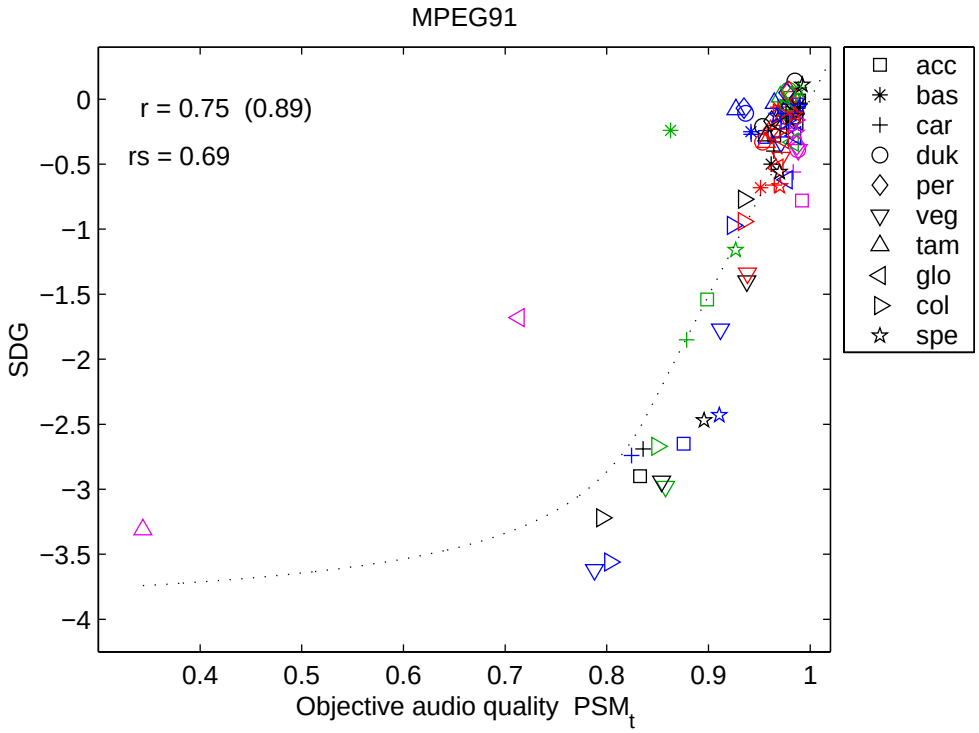

Figure B.2: Quality prediction results for data set MPEG91.

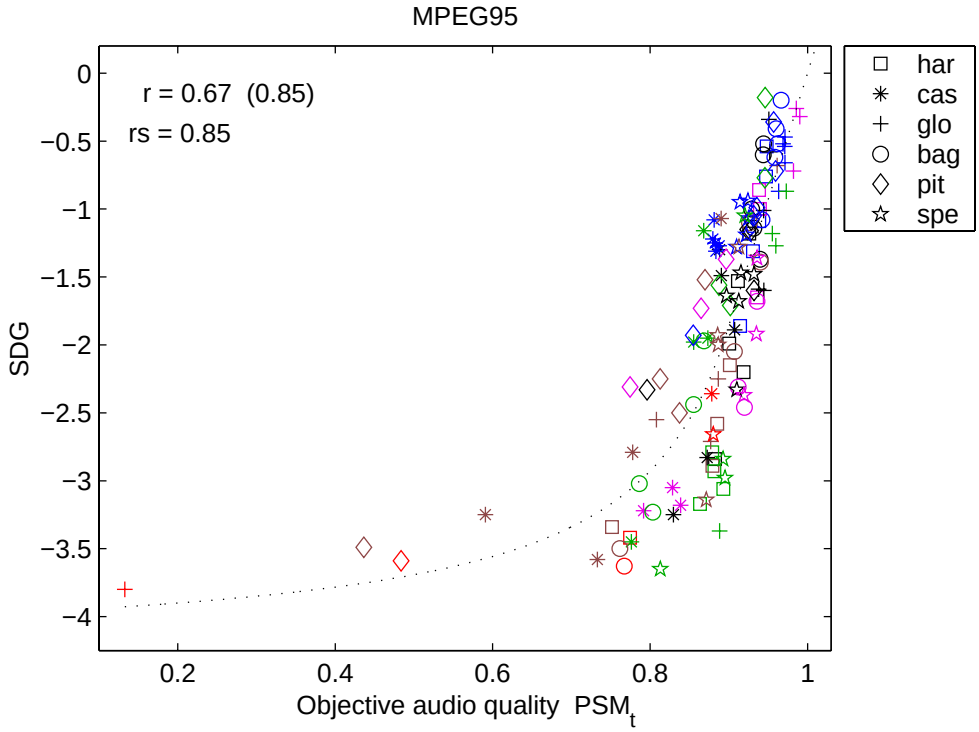

Figure B.3: Quality prediction results for data set MPEG95.

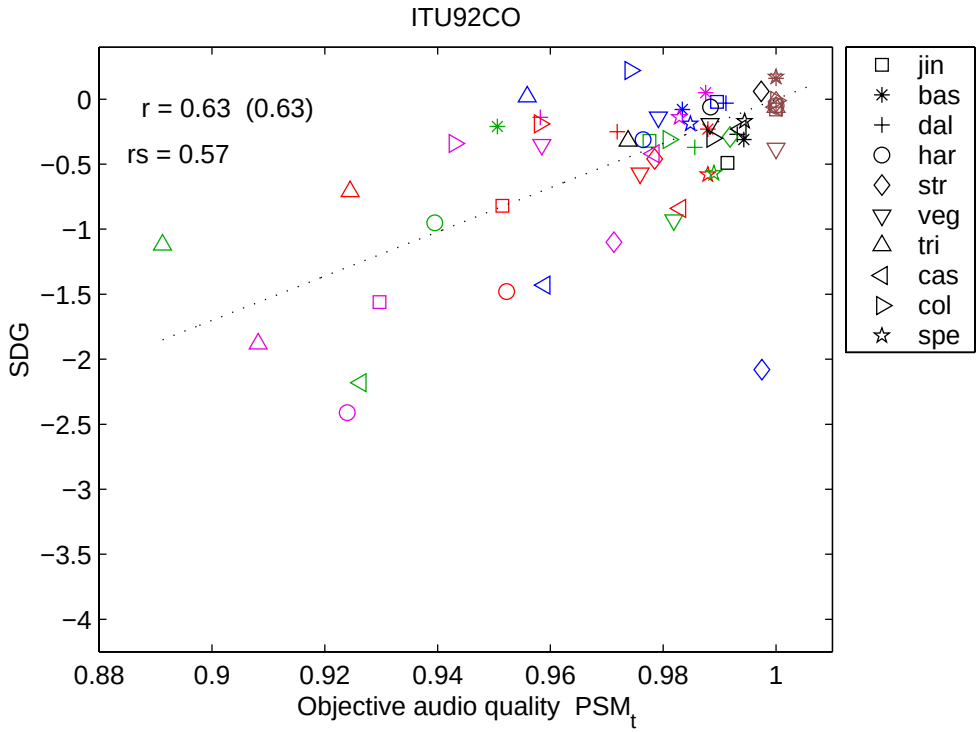

Figure B.4: Quality prediction results for data set ITU92CO.

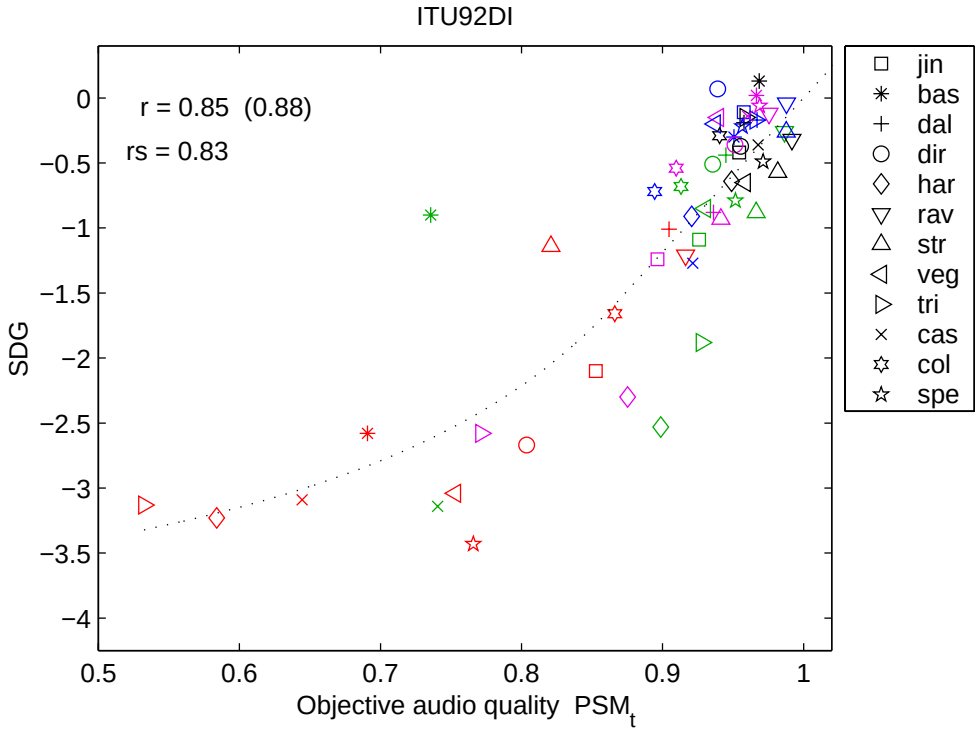

Figure B.5: Quality prediction results for data set ITU92DI.

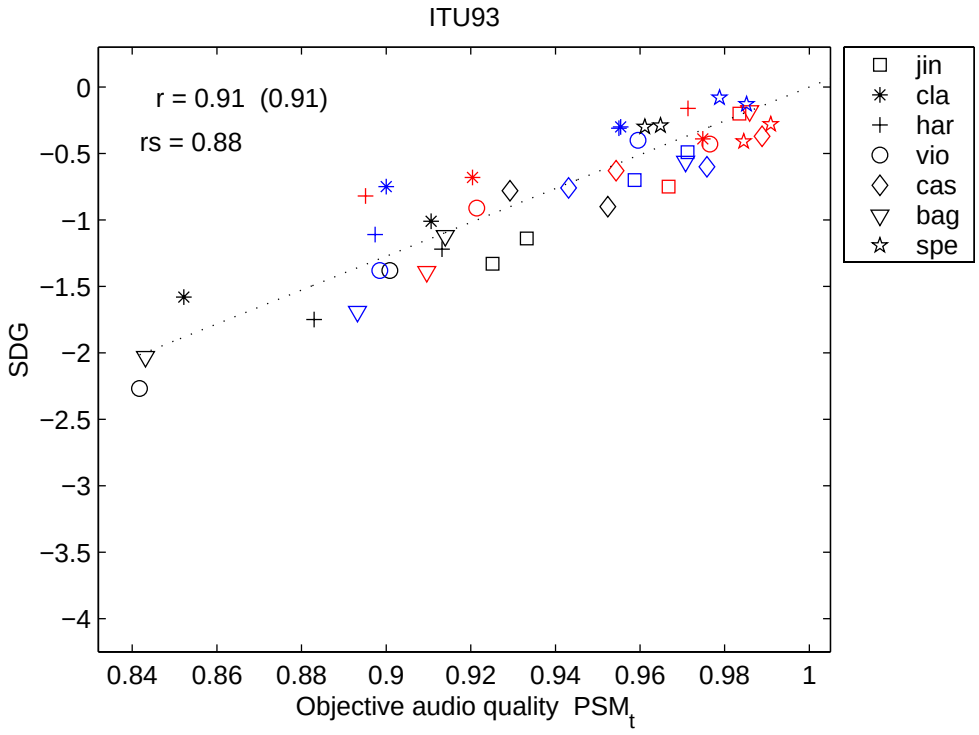

Figure B.6: Quality prediction results for data set ITU93.

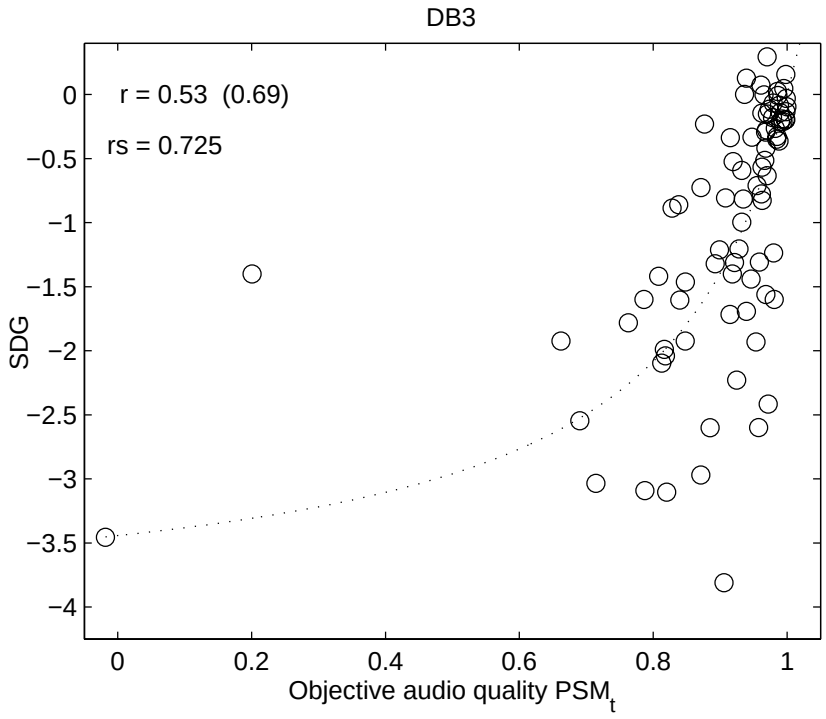

Figure B.7: Quality prediction results for data set DB3.

APPENDIX C

APPLICATION OF PSM_T FOR THE OPTIMIZATION OF AUDIO PROCESSING ALGORITHMS

The objective audio quality measure PSM_t presented in this thesis is especially qualified to serve as a tool for the optimization of audio processing algorithms. This appendix exemplarily describes an actual application of PSM_t for this purpose.

One of the projects that is supported by the priority program "Fundamentals and Methods for Low-Power Information Processing (VIVA)" of the Deutsche Forschungsgemeinschaft (DFG) is the PRO-DASP project (Power Reduction for Digital Audio Signal Processing[1]). This project deals with low power optimized design of algorithms and architectures for audio and speech signal processing. One way to reduce the power consumption of a signal processing hardware is to reduce its computational accuracy, thus producing losses of the processed signal. In order to trade the power consumption with the quality impairment due to the lossy processing, an audio quality test bench embedding the audio quality measures PSM and PSM_t is utilized within this project (cf. Damaschke et al., 2002). Using this test bench, parameters influencing both power consumption and audio quality are adjusted to produce "acceptable" quality impairments of the processed audio signal. In terms of the transformed objective quality measure ODG (Objective Difference Grade), "acceptable" corresponds to quality impairments that are not worse than "perceptible, but not annoying", i.e. ODG $\geq$ -1.

To give an example, Figure C.1 shows the estimated audio quality degradations of several test signals that were processed by a fixed-point implementation of an audio processing algorithm (Voss and Mertsching, 2002). The quality degradation is given in terms of the ODG as a function of the word length in bits.

[1] See `http://getwww.uni-paderborn.de/research/prodasp`

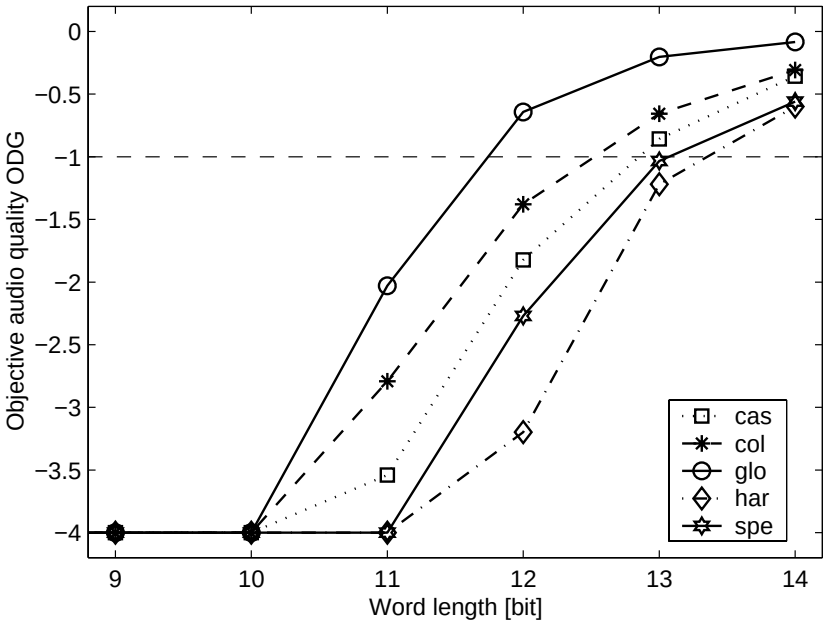

Figure C.1: Estimated quality impairments introduced by a fixed-point implementation of an audio processing algorithm for different word lengths and audio signals. The objective quality of ODG = -1 is indicated by the dashed line and corresponds to a perceived impairment of the "basic audio quality" that is "perceptible, but not annoying". (*cas* = castanets, *glo* = glockenspiel (chimes), *col* = Ornette Coleman, *spe* = speech, *har* = harpsichord.)

APPENDIX D

Bias of the Beerends-Berger-assimilation on different quality measures and signals

In Chapter 4 of this thesis, two experiments dealing with the prediction of subjective quality assessments of noise reduction schemes were presented. The results of the first experiment showed that the prediction of the overall preference improves, if the Beerends-Berger-assimilation[1] (BBA) is additionally applied. The degree of the improvement due to the BBA was found to differ between different audio quality measures.

In this appendix, possible reasons for the observed improvement and for the dependence on the quality measure are investigated.

At first, the effect of the BBA on the quality estimates obtained with different quality measures is demonstrated. Figure D.1 shows subjective vs. objective quality assessments of several single-channel noise reduction algorithms concerning the overall preference. The subjective ratings were obtained from normal hearing subjects, while the objective quality estimates stem from three different kinds of quality parameters (QM1, QM2, QM3)[2]. Each row of panels represents another kind of quality measure. The panels in the left column of Figure D.1 show quality prediction results without BBA, the right column with BBA. Applying the BBA shifts the quality estimates of the speech items containing cafeteria noise towards negative values more than the remaining items. If BBA is applied to the ΔQM1 measure (uppermost row), all items that were subjectively rated negative (i.e., the noise reduced speech was rated worse than the unprocessed version) are shifted from positive to negative ΔQM1 values, thus improving the overall correlation. The effect of the BBA on the quality estimates of the speech-plus-cafeteria-noise items continuously weakens for parameters $QM2_B$, $QM3_B$ and $QM4_B$, in the sense that less of the items in question are assigned negative qual-

[1] See Chapter 2, Section 2.3.1, or Chapter 4, Section 4.2.3 for a description of the Beerends-Berger-assimilation.

[2] See Chapter 4, Section 4.2.3, for a description of these quality measures.

ity values. Consequently, the improvement of the overall correlation due to BBA decreases.

The reason why the bias of the BBA on the quality estimates levels off for parameters $QM2_B$, $QM3_B$ and $QM4_B$ is found in the dependency of this effect on the frequency and modulation frequency in the case of speech enhancement. Those channels, where noise reduction leads to significant SNR improvements, will show greater speech-like amplitude modulations than before processing so that the magnitude of the internal representation will increase here. On the other hand, the improvement of the correlation with the clean speech reference achieved by the BBA becomes smaller with increasing magnitude of the test signal's internal representation. This is why the quality measure of the unprocessed, noisy speech signal is increased more by the BBA than that of the processed, noise-reduced speech signal (cf. Chapter 4, Section 4.4.1). To evaluate the effect of noise reduction on the magnitude of the internal representation for different frequencies and modulation frequencies, Figure D.2 shows the mean ratio of the absolute amplitudes after and before BBA ($\langle|IR_B|/|IR|\rangle$) for noise-reduced and unprocessed noisy speech signals as functions of modulation frequency (upper left panel) and frequency (upper right panel). The ratios were averaged across all signals containing cafeteria noise and across frequency or modulation frequency, depending on which parameter represented the variable. The lower panels of Figure D.2 show mean correlations between Beerends-Berger-assimilated IR of the noisy signals (unprocessed and processed) with that of clean speech.

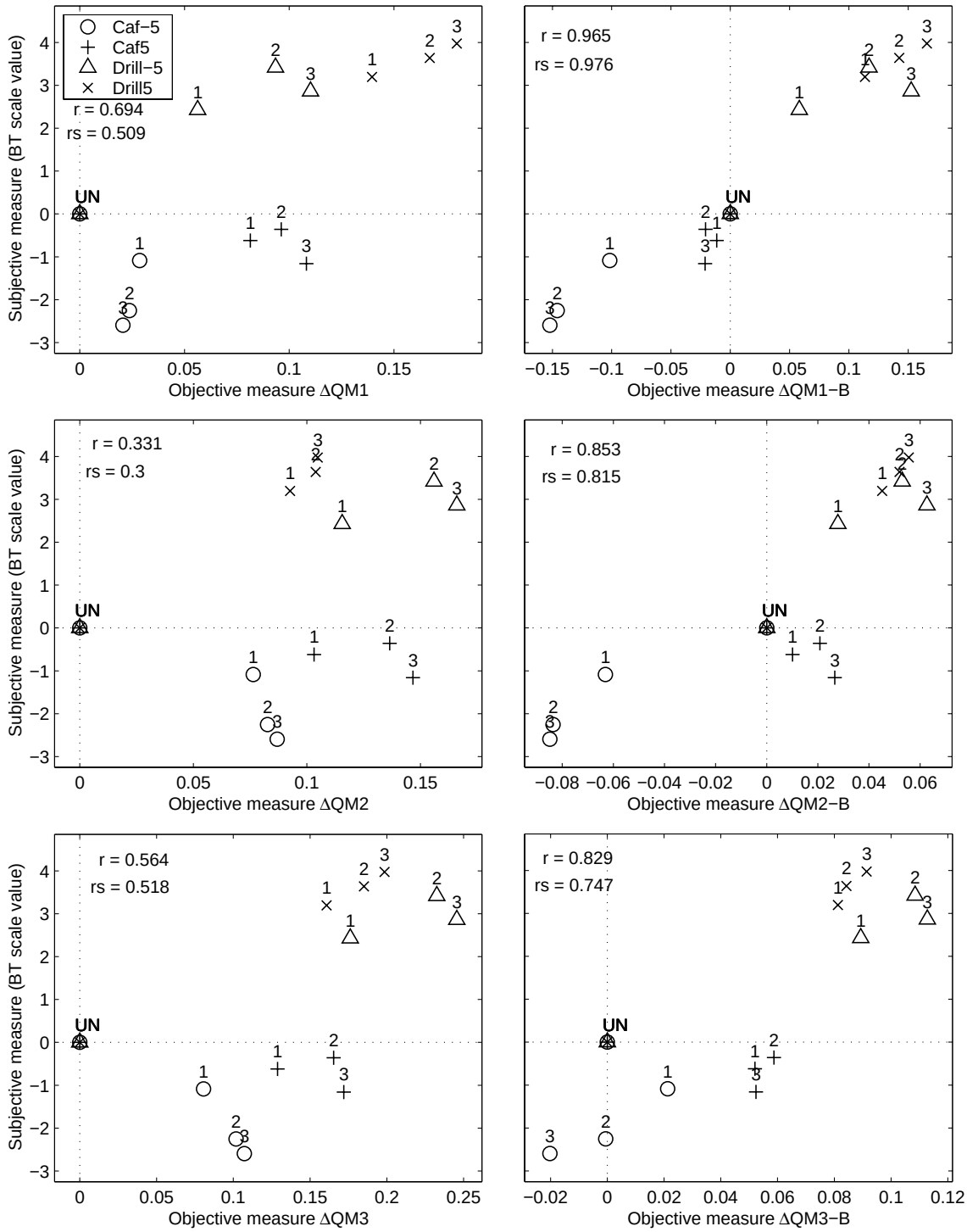

Figure D.1: Effect of the BBA on different quality measures. Measured quality ratings obtained from normal hearing subjects concerning the overall preference of noisy speech signals are plotted versus corresponding objective quality estimates obtained with quality measure QM1 (upper row), QM2 (middle row) and QM3 (lower row). The panels in the left (right) hand side show the results obtained without (with) BBA. The speech signals were mixed with cafeteria noise (circles and crosses) or drilling machine noise (triangles and x-marks) at -5 dB or 5 dB SNR.

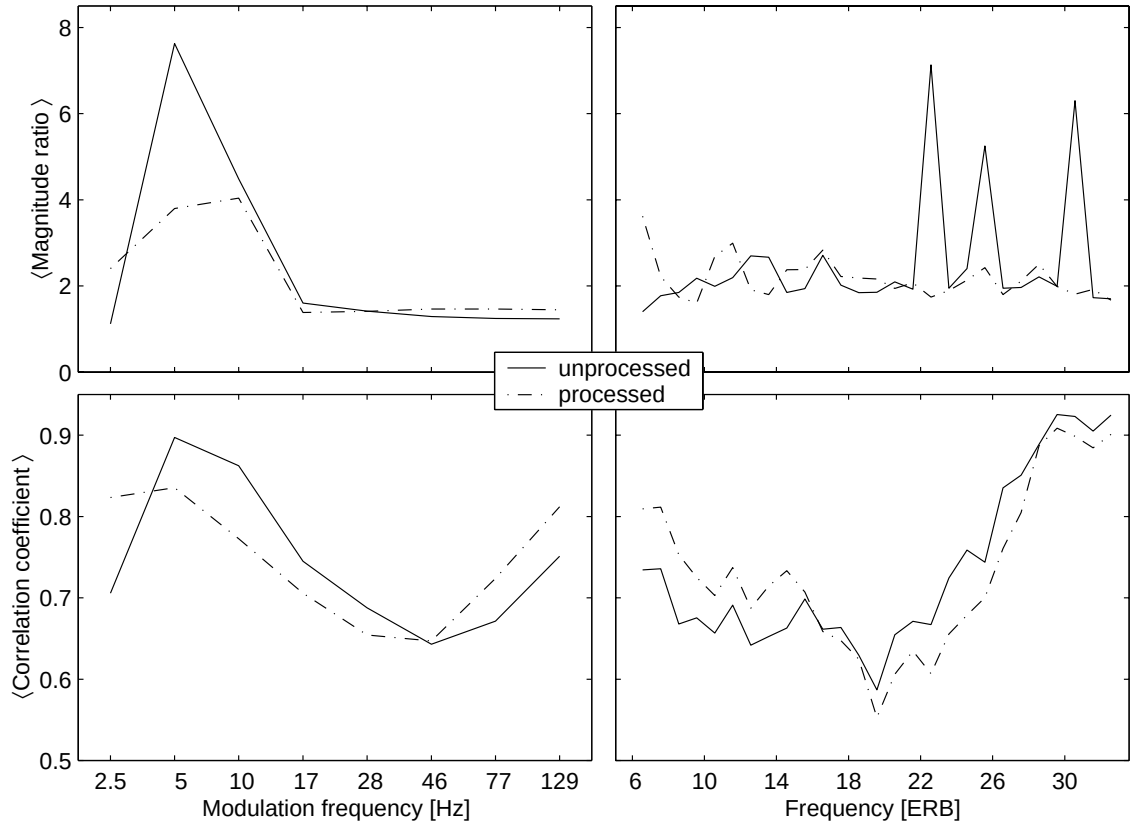

Figure D.2: Upper panels: Mean magnitude ratios of internal representations after and before BBA ($\langle |\mathrm{IR_B}|/|\mathrm{IR}| \rangle$) for noise-reduced and unprocessed noisy speech signals as functions of the modulation frequency (left) and frequency (right). Lower panels: Mean linear correlations between single channels of Beerends-Berger-assimilated IR of clean speech and unprocessed noisy speech (solid curves) and noise-reduced speech (dash-dotted curves), respectively, as functions of the modulation frequency (left) and frequency (right).

As expected, the magnitude ratios and correlation differences are related in such a way that large magnitude ratios (i.e. the average magnitude of the internal representation of the noise-reduced speech signal is considerably higher than that of the unprocessed signal) are associated with higher correlation coefficients for the unprocessed signal if BBA is applied. This leads to negative quality differences ΔQM_B. The highest magnitude ratios are observed for modulation frequencies around 5 Hz, which approximately correspond to the frequency of maximum amplitude modulation of speech. Thus it appears that those quality parameters are effected stronger by the BBA that focus on this modulation frequency region. This is the case for the measures $QM1_B$ and $QM2_B$, which employ versions of the auditory model that use a modulation lowpass filter with a cutoff frequency of 8 Hz (Dau et al., 1996a). Instead, the quality measures $QM3_B$ and $QM4_B$ are based on a model version that uses a modulation filterbank according to Dau et al. (1997a), accounting for modulation frequencies up to about 160 Hz (in the present study). As shown in the lower left panel of Figure D.2, correlations between single modulation channels of internal representations of clean speech and speech plus cafeteria noise are higher, if the noisy signal is unprocessed and the modulation center frequency does not exceed 28 Hz. (The lowest indicated modulation frequency of 2.5 Hz corresponds to the cutoff frequency of a lowpass filter. In this channel, the processed signal shows a higher correlation.) Consequently, the quality measure differences $\Delta QM3_B$ and $\Delta QM4_B$ are less effected by the BBA than $\Delta QM1_B$ and $\Delta QM2_B$.

The reason why $\Delta QM2_B$ is less affected by the BBA than $\Delta QM1_B$ is found in its dependency on frequency, as indicated in the right panels of Figure D.2: The highest magnitude ratios and thus higher correlation values for the unprocessed signal are observed for frequencies higher than about 20 ERB ($\approx$ 1750 Hz). In contrast, lower correlations are assigned to the unprocessed signal if frequency bands below 15 ERB (= 924 Hz) are considered. Because the parameter QM1 represents the speech quality measure q_C of Hansen and Kollmeier (2000), higher frequencies are emphasized (cf. Chapter 4, Figure 4.10). Thus, the overall correlation (which builds the speech quality measure) is dominated by the very frequency region that reveals the largest positive differences between correlation values of unprocessed and processed speech signals. As a consequence, the parameter $\Delta QM1_B$ is affected more by the BBA than $\Delta QM2_B$. This increases the correlation between subjective and objective ratings and thereby apparently overbalances the inappropriate bandwidth of the peripheral filterbank used in $QM1_B$, which is in fact too small to properly account for broad-band audio signals and normal hearing subjects.

The latter might reason the markedly smaller correlation value for $\Delta \mathrm{QM1_B}$ in the case of normal hearing subjects compared to hearing impaired subjects.

Bibliography

Baillard, P., Mabilleau, B., Morisette, S., Soumagne, J., 1992. PERCEVAL: Perceptual evaluation of the quality of audio signals. J. Audio Eng. Soc. 40(1): 21–31. 8

Beerends, J. G., 1994. Modelling cognitive effects that play a role in the perception of speech quality. Workshop 'Speech quality assessment' at Ruhr-Universität Bochum. 15, 76

Beerends, J. G., Stemerdink, J. A., 1992. A perceptual audio quality measure based on a psychoacoustic sound perception. J. Audio Eng. Soc. 40(12): 963–978. 3, 8, 44

Beerends, J. G., Stemerdink, J. A., 1994. A perceptual speech quality measure based on a psychoacoustic sound perception. J. Audio Eng. Soc. 42(3): 115–123. 3, 42, 84

Berger, J., 1998. Instrumentelle Verfahren zur Sprachqualitätsschätzung - Modelle auditiver Tests. Shaker ISBN 3-8265-4091-3. 15, 26, 76

Bosi, M., Brandenburg, K., Quackenbush, S., Fielder, L., Akagiri, K., Fuchs, H., Dietz, M., Herre, J., Davidson, G., Oikawa, Y., 1997. ISO/IEC MPEG-2 advanced audio coding. J. Audio Eng. Soc. 45(10): 789 – 814. 110

Bradley, R. A., Terry, M. E., 1952. Rank analysis of incomplete block designs, I. The method of paired comparisons. Biometrika 39: 324–345. 74

Brandenburg, K., 1987. Evaluation of quality for audio encoding at low bit rates. presented at the 82th convention of the Audio Engineering Society; J. Audio Eng. Soc. (Abstracts), Preprint 2433. 2, 3

Brandenburg, K., Herre, J., Johnston, J. D., Mahieux, Y., Schroeder, E. F., 1991. ASPEC: adaptive spectral entropy coding of high quality music signals. presented at the 90th convention of the Audio Engineering Society; J. Audio Eng. Soc. (Abstracts), Preprint 3011 A-4. 108

Brandenburg, K., Sporer, T., 1992. NMR and Masking Flag: Evaluation of Quality Using Perceptual Criteria. In: Proc. of the AES 11th International Conference (Portland, Oregon, USA), pages 169 – 179. 1

Brandenburg, K., Stoll, G., 1994. ISO-MPEG-1 Audio: A generic standard for coding of high-quality digital audio. J. Audio Eng. Soc. 42(10): 780–792. 12

Colomes, C., Lever, M., Dehery, Y. F., 1995. A perceptual model applied to audio bit-rate reduction. J. Audio Eng. Soc. 43(1): 233–240. 3, 8, 44

Damaschke, J., Huber, R., Hohmann, V., Kollmeier, B., 2002. PRO-DASP: An audio quality testbench for optimizing low-power chip designs of speech processing algorithms. In: Müller, D., Kretzschmar, C., Siegmund, R., eds., 3. Kolloquium des Schwerpunktprogramms der Deutschen Forschungsgemeinschaft - VIVA - Grundlagen und Verfahren verlustarmer Informationsverarbeitung, pages 50 – 54. TU Chemnitz, ISBN 3-00-008995-0, Chemnitz. 117

Dau, T., Kollmeier, B., Kohlrausch, A., 1997a. Modeling auditory processing of amplitude modulation: I. Modulation Detection and masking with narrowband carriers. J. Acoust. Soc. Am. 102(5): 2892–2905. 4, 7, 12, 14, 42, 49, 51, 52, 55, 56, 57, 62, 65, 67, 70, 103, 104, 123

Dau, T., Kollmeier, B., Kohlrausch, A., 1997b. Modeling auditory processing of amplitude modulation: II. Spectral and temporal integration. J. Acoust. Soc. Am. 102(5): 2906–2919. 14, 56

Dau, T., Püschel, D., Kohlrausch, A., 1996a. A quantitative model of the 'effective' signal processing in the auditory system: I. Model structure. J. Acoust. Soc. Am. 99: 3615–3622. 4, 7, 9, 12, 14, 51, 52, 55, 56, 62, 65, 70, 103, 123

Dau, T., Püschel, D., Kohlrausch, A., 1996b. A quantitative model of the 'effective' signal processing in the auditory system: II. Simulations and measurements. J. Acoust. Soc. Am. 99: 3623–3631. 9, 53, 67

Derleth, R. P., 1999. Temporal and compressive properties of the normal and impaired auditory system. BIS ISBN 3-8142-0695-9, Oldenburg. 9

Derleth, R. P., Dau, T., Kollmeier, B., 2001. Modeling temporal and compressive properties of the normal and impaired auditory system. Hearing Research 159: 132–149. 41

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing 32 (6): 1109–1121. 74

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error short-time log-spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing 33 (2): 443–445. 74

ETSI, 1998. NICAM 728: transmission of two-channel digital sound with terrestrial television systems B, G, H, I, K1 and L. Tech. Rep. EN 300 163 V1.2.1 (1998-03), European Telecommunications Standards Institute. 108

Fassel, R., 1994. Experimente und Simulationsrechnungen zur Wahrnehmung von Amplitudenmodulationen im menschlichen Gehör. Ph.D. thesis, Universität Göttingen. 9

Fastl, H., 1994. Evaluation and measurement of the perceived average loudness. In: Schick, A., ed., Contributions to Psychological Acoustics, vol. 5, pages 205–216. BIS Oldenburg. 32, 38

Fielder, L., Bosi, M., Davidson, G., Davis, M., Todd, C., Vernon, S., 1996. AC-2 and AC-3: low-complexity transform-based audio coding. Collected papers on digital audio bit-rate reduction, Audio Engineering Society. 10, 108, 109, 110

Green, D. M., 1985. Temporal factors in psychoacoustics. In: Michelsen, A., ed., Time resolution in auditory systems. Springer Verlag, New York. 65

Gustafson, S., Martin, R., Vary, P., 1996. On the optimization of speech enhancement systems using instrumental measures. In: Workshop on quality assessment in speech, audio and image communication; Darmstadt, Germany, pages 36–40. ITH Informationstechnische Gesellschaft, EURASIP European Association for Signal Processing. 72

Hamberg, R., de Ridder, H., 1999. Time-varying image quality: Modelling the relation between instantaneous an overall quality. SMPTE Journal pages 802–811. 32

Hansen, J., Pellom, B., 1998. An effective quality evaluation protocoll for speech enhancement algorithms. In: Proceedings ICSLP '98, Sydney, Australia. 72

Hansen, M., Kollmeier, B., 2000. Objective modelling of speech quality with a psychoacoustically validated auditory model. J. Audio Eng. Soc. 48: 395–409. 3, 4, 7, 9, 12, 15, 28, 30, 49, 71, 72, 74, 75, 76, 84, 95, 97, 100, 103, 123

Hauenstein, M., 1997. Psychoakustisch motivierte Maße zur instrumentellen Sprachgütebeurteilung. Ph.D. thesis, Universität Kiel. 15

Herre, J., Eberlein, E., Schott, H., Schmidmer, C., 1992. Analysis tool for real time measurement using perceptual criteria. In: Proc. AES 11th Int. Conf. (Portland, OR, 1992), pages 180–190. 44

Hollier, M., Hawksford, M., Guard, D., 1994. Objective perceptual analysis: comparing the audible performance of data reduction scheme. presented at the 96th convention of the Audio Engineering Society, Amsterdam; J. Audio Eng. Soc. (Abstracts), Preprint 3797 (P3.6). 3

ISO/MPEG, 1990. Audio test report. ISO/IEC/JTC 1/SC 2/WG 11 MPEG MPEG90/N0030, International Organization for Standardization. 54, 107

ISO/MPEG, 1991. Audio test report. ISO/IEC/JTC 1/SC 2/WG 11 MPEG MPEG91/N0010, International Organization for Standardization. 54, 108

ISO/MPEG, 1992. Coding of moving pictures and associated audio for digital storage media up to about 1.5 Mbit/s, part 3: audio. ISO/IEC JTC1/SC29/WG11 MPEG International Standard IS 11172-3, International Organization for Standardization. 10, 52, 54, 108, 109, 110

Itakura, F., 1975. Minimum prediction residual principle applied to speech recognition. IEEE Trans. Acoust., Speech and Signal Processing ASSP-23(1): 67 – 72. 84

Itakura, F., Saito, S., 1979. A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies. In: Schafer, R. W., Markel, J. D., eds., Speech Analysis. IEEE Press, New York. 84

ITU-R, 1990. Subjective assessment of sound quality. Series BS: Broadcast Services Recommendation BS.562-3, International Telecommunications Union. 10, 52

ITU-R, 1992. CCIR Listening Tests - Basic Audio Quality of Distribution and Contribution Codecs, Sweden. CCIR Listening Tests CCIR-Doc. 10-2/24, International Telecommunications Union. 54, 108, 109

ITU-R, 1993. CCIR Listening Test - Network Verification Tests without Commentary Codecs, Canada and Italy. CCIR Listening Tests CCIR-Doc. 10-2/43, International Telecommunications Union. 54, 109

ITU-R, 1997. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Series BS: Broadcast Services Recommendation BS.1116, International Telecommunications Union. 10, 52

ITU-R, 1998a. Method for objective measurement of perceived audio quality. Series BS: Broadcast Services Recommendation BS.1387, International Telecommunications Union. 3, 8, 44, 45, 107, 110

ITU-R, 1998b. Report on the sixth meeting of ITU-R Task Group 10/4. Tech. Rep. Doc. 10-4/21, International Telecommunications Union. 45

ITU-T, 1990. 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM). Series G: Transmission systems and media, digital systems and networks Recommendation G.726, International Telecommunications Union. 10, 107

ITU-T, 1996a. Modulated noise reference unit (MNRU). Series P: Telephone transmission quality, telephone installations, local line networks Recommendation P.810, International Telecommunications Union. 54

ITU-T, 1996b. Objective quality measurement of telephone-band (300-3400 Hz) speech codecs. Series P: Telephone Transmission Quality Recommendation P.861, International Telecommunications Union. 3, 8, 76

Karjalainen, M., 1985. A new auditory model for the evaluation of sound quality of audio systems. In: Proc. Int. Conf. on Acoust., Speech and Signal Processing (ICASSP), pages 608 – 611. IEEE Press, New York. 3

Klatt, D. H., 1982. Prediction of perceived phonetic distance from critical-band spectra : a first step. In: Proc. Int. Conf. on Acoust., Speech and Signal Processing (ICASSP). IEEE Press, New York. 84

Levitt, H., 1971. Transformed up-down procedures in psychoacoustics. J. Acoust. Soc. Am. 49: 467–477. 55

Marzinzik, M., 2000. Noise reduction schemes for digital hearing aids and their use for the hearing impaired. Shaker ISBN 3-8265-8513-5. 71, 72, 74, 75, 79, 84, 85, 95, 104

Marzinzik, M., Kollmeier, B., 2000. Quality assessment of noise reduction for digital hearing aids. In: Fortschritte der Akustik – DAGA 2000, pages 280–281. DEGA, Oldenburg. 72, 73, 77, 84, 94

Meares, D., Kim, S.-W., 1995. NBC time/frequency module subjective tests: overall results. ISO/IEC JTC 1/SC 29/WG 11 MPEG N0973 MPEG95/208, International Standardisation Organistation. 54, 108

Microsoft, 2003. Windows Media Homepage. www.microsoft.com/windows/windowsmedia/. 54

Münkner, S., 1993. A psychoacoustical model for the perception of non-stationary sounds. In: Schick, A., ed., Contributions to Psychological Acoustics, vol. 6, pages 121–134. BIS Oldenburg. 42, 43, 67

Moore, B. C. J., Glasberg, B., Plack, C. J., Biswas, A. K., 1988. The shape of the ear's temporal window. J. Acoust. Soc. Am. 83: 1102–1116. 65

Münkner, S., 1993. Modellentwicklung und Messungen zur Wahrnehmung nichtstationärer akustischer Signale. Shaker ISBN 3-86111-850-5, Aachen. 9

Paillard, B., Mabilleau, P., S., M., Soumagne, J., 1992. PERCEVAL: perceptual evaluation of the quality of audio signals. J. Audio Eng. Soc. 40: 21–31. 3, 44

Patterson, R. D., Nimmo-Smith, J., Holdsworth, J., Rice, P., 1987. An efficient auditory filterbank based on the gammatone function. Paper presented at a meeting of the IOC Speech Group on Auditory Modelling at RSRE. 12, 55

Püschel, D., 1988. Prinzipien der zeitlichen Analyse beim Hören. Ph.D. thesis, Universität Göttingen. 13

Quackenbush, S. R., III, T. P. B., Clements, M. A., 1988. Objective Measures of Speech Quality. Prentice Hall, Englewood Cliffs. 84

Sander, A., 1994. Psychoakustische Aspekte der subjektiven Trennbarkeit von Klängen. Ph.D. thesis, Universität Oldenburg. 9

Schroeder, M. R., Atal, B. S., Hall, J. L., 1979. Optimizing digital speech coders by exploiting masking properties of the human ear. J. Acoust. Soc. Am. 66: 1647 – 1652. 3

Soulodre, G., Grusek, T., Lavoie, M., Thibault, L., 1998. Subjective evaluation of state-of-the-art 2-channel audio codecs. J. Audio Eng. Soc. 46: 164–177. 45

Sporer, S., 1997. Objective audio signal evaluation - applied psychoacoustics for modelling the perceived audio quality of digital audio. presented at the 103rd convention of the Audio Engineering Society, J. Audio Eng. Soc. (Abstracts). 3, 44

Stoll, G., Dehéry, 1990. High quality bit-rate reduction system family for different applications. In: Proc. IEEE ICC Supercom 1990, vol. 3, pages 937 – 941. GA, Atlanta. MUSICAM. 107, 108

Strube, H. W., 1985. A computationally efficient basilar-membrane model. Acustica 58: 207–214. 12

Thiede, S., Kabot, E., 1996. A new perceptual quality measure for the bit rate reduced audio. presented at the 100th convention of the Audio Engineering Society, J. Audio Eng. Soc. (Abstracts). 3, 44

Thiede, T., Treuniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., Colomes, C., Keyhl, M., Stoll, G., Brandenburg, K., Feiten, B., 2000. PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality. J. Audio Eng. Soc. 48(1): 3–29. 8, 44, 45

Thiede, T., Treuniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., Colomes, C., Keyhl, M., Stoll, G., Brandenburg, K., Feiten, B., 1998. PEAQ - der künftige ITU-Standard zur objektiven Messung der wahrgenommenen Audioqualität. Bericht der 20. Tonmeistertagung in Karlsruhe. 45, 47, 48

Tontch, K., 2002. Instrumentelle und auditive Beurteilung der Qualität des Sprachsignals nach räumlicher Filterung des Sprechschalls in Personenkraftwagen. Shaker ISBN 3-8322-0411-3. 71, 73, 85, 104

Treurniet, W. C., 1996. Simulation of individual listeners with an auditory model. presented at the 100th convention of the Audio Engineering Society, J. Audio Eng. Soc. (Abstracts). 44

Treurniet, W. C., 1998. Objective measurement of perceived audio quality. CRC Tech. Note CRC-TN-98-007, Communication Research Center, Ottawa, ON, Canada. 44

Treurniet, W. C., 2000. Evaluation of the ITU-R objective audio quality measurement method. J. Audio Eng. Soc. 48(3): 164–173. 8

Tsutsui, K., Suzuki, H., Shimoyoshi, O., Sonohara, M., Akagiri, K., Heddle, R., 1996. ATRAC: Adaptive transform coding for MiniDisc. Collected Papers on Digital Audio Bit-rate Reduction, Audio Engineering Society. 10, 110

Verhey, J. L., 1998. Psychoacoustics of spectro-temporal effects in masking and loudness patterns. BIS ISBN 3-8142-0622-2, Oldenburg. 9

Voss, N., Mertsching, B., 2002. PRO-DASP: Using transformations to implement hardware-macros for a low power design methodology. In: Müller, D., Kretzschmar, C., Siegmund, R., eds., 3. Kolloquium des Schwerpunktprogramms der Deutschen Forschungsgemeinschaft - VIVA - Grundlagen und Verfahren verlustarmer Informationsverarbeitung, pages 55 – 60. TU Chemnitz, ISBN 3-00-008995-0, Chemnitz. 117

Wang, S., Sekey, A., Gersho, A., 1992. An objective measure for predicting subjective quality of speech coders. IEEE Journal on Selected Areas in Communications 10(5): 819 – 829. 3

Westermann, L. A., Smith, R. L., 1984. Rapid and short-term adaptation in auditory nerve response. Hearing Research 15: 249–260. 67

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbstständig verfasst habe und nur die angegebenen Hilfsmittel verwendet habe.

Oldenburg, den 28. Oktober 2003

Rainer Huber

Danksagung

Nicht nur einer guten Tradition, sondern auch einem "ehrlichen" Bedürfnis folgend, möchte ich mich an dieser Stelle bei all denen bedanken, die mich bei der Anfertigung dieser Dissertation unterstützt bzw. sie überhaupt ermöglicht haben.

Dies betrifft an erster Stelle meinen Doktorvater, Prof. Dr. Dr. Birger Kollmeier, der mir die Gelegenheit gab, ein interessantes Thema unter guten Arbeitsbedingungen bearbeiten zu können. Sowohl als wissenschaftlicher Kopf als auch als "Manager" unserer Arbeitsgruppe "Medizinische Physik" ist er eine Ausnahmeerscheinung.

Herrn Prof. Dr. Volker Mellert danke ich für die freundliche Übernahme des Korreferats sowie für seinen federführenden Einsatz im Graduiertenkolleg "Psychoakustik", das mich drei Jahre als Stipendiat gefördert hat.

Prof. Dr. Torsten Dau unterstützte mich vor allem durch seine unübertroffene Kompetenz in Sachen Perzeptionsmodell und allgemeinen Fragen zur Psychoakustik. Insbesondere danke ich ihm für seine wertvollen Korrekturvorschläge zum dritten Kapitel dieser Dissertation. Sein Weggang schmerzt aus wissenschaftlicher und noch mehr aus menschlicher Sicht.

Eine Hauptstütze dieser Arbeitsgruppe ist Dr. Volker Hohmann, von dessen allseitiger Kompetenz, vor allem auf dem Gebiet der digitalen Signalverarbeitung, ich jederzeit und ungefragt profitieren durfte.

Dr. Michael Kleinschmidt hat sich heldenhaft durch die Rohversion zweier Kapitel gekämpft, um mir sprachliche Tipps, aber auch nützliche inhaltliche Anmerkungen zu geben. Ihm und Dr. Jörn Otten danke ich darüber hinaus für die angenehme und unterhaltsame Atmosphäre unserer langjährigen Bürogemeinschaft. Schade, dass auch diese beiden Kollegen und Freunde mittlerweile "in den Süden gezogen" sind.

Dr. Thomas Brandt hat ein komplettes Semester meiner Lehrverpflichtung übernommen, um mir dadurch einen großen Schritt in Richtung Vollendung der Dissertation zu ermöglichen. Dafür schulde ich ihm großen Dank. In gleicher Weise hat mir Matthias Müller-Wehlau zu einem "unvermeidbaren" Traumurlaub während der Vorlesungszeit verholfen. Dafür und für die stets locker-amüsante Arbeitsatmosphäre im neuen Domizil danke ich auch ihm.

Meinen zehn Versuchspersonen danke ich für ihre einsame Zeit in stahlummantelter Enge.

Dass sich der Anteil vergeudeter Arbeitszeit durch nicht-funktionierende Computer und -Netzwerke in engen Grenzen hielt, verdanke ich Dr. Volker Kühnel, Dr. Oliver Fobel (ehem. Wegner), Dr. Thomas Wittkop, Johannes Nix, Frank Grunau und evtl. noch weiteren Wärtern des Rechner-Zoos, die möglicherweise von mir vergessen oder unbemerkt ihre pflegenden Finger im Spiel hatten.

Weitere technische Hilfe und Rat in allen Lebenslagen gab es von unserer technischen Assistentin und "Labor-Maus" Anita Gorges.

Unsere Sekretärinnen Karin Bramstedt, Susanne Garre und Ingrid Wusowski waren stets hilfsbereit und kompetent in Sachen Verwaltung und Organisation.

Als mindestens ebenso förderlich wie die guten materiellen und (inzwischen) räumlichen Arbeitsbedingungen hat sich die überaus gesunde "Chemie" zwischen den KollegInnen dieser Arbeitsgruppe erwiesen. Ich danke daher allen derzeitigen und ehemaligen "Medis" für das erstklassige Miteinander der letzten acht(!) Jahre.

Nach meiner Stipendiatenzeit erfuhr ich weitere finanzielle Unterstützung durch das Programm der Universtät Oldenburg zur "Förderung des wissenschaftlichen Nachwuchses".

Für ein ausgleichendes, sorgenfreies Privatleben danke ich meiner Familie und vor allem meiner geduldigen Frau Svenja.

Lebenslauf

Am 18.6.1969 wurde ich als Sohn von Gerd und Ursel Huber, geb. Bunjes, in Oldenburg geboren. Auf der Nordseeinsel Wangerooge besuchte ich von 1975 – 1985 die Grundschule, die Orientierungsstufe und schließlich das Gymnasium. Anschließend wechselte ich auf das Niedersächsische Internatsgymnasium in Esens, um dort im Juni 1988 das Abitur zu absolvieren.

Nach der Grundwehrdienstzeit vom Januar 1989 bis zum März 1990 in Hamburg und Jever nahm ich zum Wintersemester 1990/91 an der Carl-von-Ossietzky-Universität Oldenburg mein Physikstudium auf. Mein Vordiplom absolvierte ich im August 1992. Meine Diplomarbeit über akustische Verfahren zur objektiven Klassifizierung heiserer Stimmen schrieb ich in der Arbeitsgruppe Medizinische Physik bei Prof. Dr. Dr. Birger Kollmeier, um im Januar 1998 mein Studium mit dem Diplom abzuschließen.

Im Mai 1998 erhielt ich ein Doktorandenstipendium des Graduiertenkollegs Psychoakustik und begann mit der Arbeit an der vorliegenden Dissertation. Seit April 2001 besetze ich als wissenschaftlicher Angestellter eine Stelle zur "Förderug des wissenschaftlichen Nachwuchses" des Fachbereichs Physik bzw. des Instituts für Physik.