Noise Reduction Schemes for Digital Hearing Aids and their Use for the Hearing Impaired

Vom Fachbereich Physik der Universität Oldenburg
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
angenommene Dissertation.

Mark Marzinzik
geb. am 28. Juni 1970
in Bremen

Erstreferent: Prof. Dr. Dr. Birger Kollmeier
Korreferent: Prof. Dr. Volker Mellert
Tag der Disputation: 19. Dezember 2000

Noise Reduction Schemes for Digital Hearing Aids and their Use for the Hearing Impaired

Vom Fachbereich Physik der Universität Oldenburg
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
angenommene Dissertation.

Mark Marzinzik
geb. am 28. Juni 1970
in Bremen

Erstreferent: Prof. Dr. Dr. Birger Kollmeier
Korreferent: Prof. Dr. Volker Mellert
Tag der Disputation: 19. Dezember 2000

The ear was not fashioned with the prospect of industrial revolution in mind. Its superlative sensitivity and scope of action have made it victim to the culmination of the last few hundred years of industrial and social development. Much of what we now hear is, in one sense or another, unwanted, and it is this element of unwantedness which defines a sound as noise.

Dylan M. Jones

Abstract

The aim of this thesis is to improve both the assessment methods and the available algorithms for noise reduction in hearing aids. In particular, the whole development chain from the construction of algorithms, subjective assessment of algorithmic performance by normal-hearing and hearing-impaired listeners as well as objective assessment methods is considered.

The speech pause detection algorithm proposed in Chapter 2 detects speech pauses by tracking minima in a noisy signal's power envelope, specifically in its low-pass and high-pass power envelopes. It maintains a low false-alarm rate over a wide range of signal-to-noise ratios. This facilitates its application for noise estimation in noise reduction algorithms.

Chapter 3 shows that the musical noise phenomenon, one widely reported artifact of most single-microphone noise reduction schemes based on spectral subtraction, can to a high degree be overcome by the Ephraim-Malah noise reduction algorithms (Ephraim and Malah, 1984, 1985). If combined with the procedure for automatically adjusting the noise spectrum estimate during speech pauses (Chapter 2), a self-adaptive noise reduction scheme is obtained.

Comprehensive evaluations of the Ephraim-Malah noise reduction algorithms with hearing-impaired subjects show that besides better "sound quality" (Chapter 5), most obvious benefits are reductions in the mental effort needed to listen to speech in noise and hence in listener fatigue over longer periods of time. To assess this feature, a new listening effort test is developed (Chapter 4).

Although a significant amount of noise reduction is obtained with the Ephraim-Malah algorithms for various noise conditions, an increase in speech intelligibility measured with a sentence test is not found. Only the binaural directional filter and dereverberation algorithm (Wittkop, 2000) is found to provide speech intelligibility improvements. On the other hand, differences in terms of listening effort are found for different algorithms which did not show up in word recognition scores. These findings indicate that conventional speech recognition tests and tests of listening effort measure different aspects of the effect of noise reduction schemes in speech perception.

The method of paired comparisons in combination with the Bradley-Terry scaling model is suggested for subjective quality assessment of the algorithms in Chapter 5. The results show that noise reduction is worthwhile in all of the different noises that were investigated. The Ephraim-Malah single-microphone noise reduction algorithms can be recommended for use in rather stationary noises. They fail in strongly fluctuating noises where the binaural directional filter and dereverberation algorithm may be used, particularly at lower SNRs.

In Chapter 6, the predictive power of several "objective" speech quality measures is investigated with respect to the subjective noise reduction effect for hearing-impaired listeners. Particularly the PMF and LAR objective quality measures reflect different subjective results.

It is demonstrated how objective measures can be employed to assess the often large parameter space in the development of noise reduction algorithms aiming at a preselection of noise reduction algorithms and parameter settings which are worthwhile a comprehensive subjective evaluation.

Finally, it is hoped that the proposed methods might be used in the future to provide further benefit to hearing-impaired patients from "intelligent" digital hearing aids.

Kurzfassung

Das Ziel dieser Dissertation ist die Entwicklung bzw. Verbesserung von existierenden Störgeräuschunterdrückungsalgorithmen für digitale Hörgeräte sowie von Methoden zur Evaluation derartiger Algorithmen. Dabei wird die gesamte Entwicklungskette von der Entwicklung der Algorithmen über die Erfassung ihrer Fähigkeiten und Unzulänglichkeiten mit normalhörenden und schwerhörigen Versuchspersonen bis hin zur objektiven Qualitätsvorhersage mit technischen und psychoakustischen Maßen berücksichtigt.

In Kapitel 2 wird ein Algorithmus zur Sprachpausenerkennung entwickelt. Dieser Algorithmus erkennt Sprachpausen, indem er Minima in den Leistungshüllkurven des Signals sowie des tiefpaß– und hochpaßgefilterten Signals verfolgt. Er zeichnet sich insbesondere durch eine geringe Falsch-Alarm-Rate aus, die er über einen großen Bereich an Signal-Rausch-Verhältnissen bewahrt. Dadurch eignet sich der Algorithmus insbesondere für eine Anwendung zur Schätzung von Störgeräuschspektren, die von vielen Algorithmen zur Störgeräuschunterdrückung benötigt werden. Der Sprachpausenalgorithmus wird kombiniert mit den Störgeräuschunterdrückungsalgorithmen, die von Ephraim und Malah (1984, 1985) vorgeschlagen wurden. Wie in Kapitel 3 gezeigt wird, zeichnen sich diese durch besonders geringe Verarbeitungsartefakte aus.

Neben einer Verbesserung der Klangqualität, die in Kapitel 5 untersucht wird, verringern die Algorithmen insbesondere die mentale Anstrengung, die nötig ist, einem Sprecher in stark störgeräuschbehafteter Umgebung zuzuhören. Zur Erfassung dieses Aspektes wird ein neuartiger Zuhöranstrengungstest in Kapitel 4 entwickelt und angewendet. Obwohl eine starke Störgeräuschunterdrückung mit den monauralen Ephraim-Malah-Algorithmen erreicht wird, schlägt sich dies nicht in verbesserter Sprachverständlichkeit, wie sie mit einem Satztest erfaßt wird, nieder (Kapitel 4). Mit einem binauralen Störgeräuschunterdrückungsalgorithmus (Richtungsfilter und Enthallung; Wittkop, 2000) konnten dagegen Verbesserungen der Sprachverständlichkeit nachgewiesen werden. Andererseits wurden bezüglich der Zuhöranstrengung Unterschiede zwischen Algorithmen gefunden, die sich nicht in den Ergebnissen der Sprachverständlichkeitsmessungen abbilden. Dies kann als ein Hinweis darauf verstanden werden, daß der entwickelte Test auf Zuhöranstrengung tatsächlich andere Aspekte der Störgeräuschunterdrückung erfaßt als konventionelle Sprachverständlichkeitstests.

Zur Erfassung von verschiedenen subjektiv wahrgenommenen Qualitätsaspekten der Algorithmen wird in Kapitel 5 die Paarvergleichsmethode in Verbindung mit dem Bradley-Terry-Skalierungsmodell vorgeschlagen. Die Ergebnisse zeigen, daß die Ephraim-Malah-Algorithmen von den schwerhörigen Versuchspersonen insbesondere in stationären Störgeräuschen bevorzugt werden, während der binaurale Algorithmus in fluktuierenden Geräuschen (besonders bei niedrigen Signal-Rausch-Verhältnissen) besser beurteilt wird.

In Kapitel 6 werden verschiedene "objektive" Sprachqualitätsmaße auf ihre Fähigkeit hin untersucht, die subjektiv erfaßten Qualitätsurteile widerzuspiegeln und damit in gewisser Weise vorhersagen zu können. Insbesondere die Maße PMF und LAR erweisen sich dabei als erfolgreich.

Es bleibt zu hoffen, daß die in dieser Arbeit eingeführten und vorgestellten Methoden zukünftig angewandt werden mögen, um damit schwerhörigen Patienten zu besserer Lebensqualität durch "intelligente" digitale Hörgeräte zu verhelfen.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

It is no secret that noise evokes major communication difficulties in hearing-impaired subjects, even in persons with low to moderate hearing losses (Weiss and Neuman, 1993). The Working Group on Communication Aids for the Hearing-Impaired (1991) points out that it is one of the most common complaints made by hearing-aid users that speech in noise, or speech in a reverberant room, is particularly difficult to understand.

These difficulties are often experienced as a burdensome handicap especially at the working place and during social activities. They are connected with decreased speech intelligibility and with an increased effort to understand speech in noise which is experienced as tiresome and fatiguing.

Noise reduction schemes for digital hearing aids may help to overcome these deficiencies. They aim at increasing the signal-to-noise ratio and thereby increasing the speech intelligibility, lowering the listening effort and improving the perceived quality of the acoustic environment. The current thesis therefore is concerned with several aspects of such schemes.

Many multi-microphone as well as single-microphone noise reduction algorithms have been proposed in the literature so far. However, most multi-microphone schemes will probably not be considered by the hearing aid industry in the *near* future due to cosmetic reasons. The problem of a bi-directional and wireless communication between two hearing aids (left and right ear) is still unsolved. Other propositions, as for example hearing-aid spectacles which enable the placement and wired connection of several microphones, were rejected by the majority of customers in the past.

A crucial requirement of most single-microphone noise reduction algorithms is the estimation of the noise spectrum. Since most realistic noisy environments are characterized by non-stationarity, it is necessary to frequently adjust the noise spectrum estimate to maintain an effective noise reduction processing. A natural possibility is to perform this adjustment whenever target speech is absent, which means that the input signal consists of noise only. A speech pause detection scheme which especially meets

the constraints for hearing aid applications is developed and evaluated in Chapter 2.

Chapter 3 gives a brief overview of single-microphone noise reduction algorithms that were developed in the last decades. Specifically, the algorithms proposed by Ephraim and Malah (1984, 1985) are reviewed and some of their outstanding features which favor them for hearing aid applications are discussed in more detail.

In order to construct a self-adaptive noise reduction scheme for digital hearing aids, the Ephraim-Malah noise reduction algorithms are combined with the speech pause detection algorithm developed in Chapter 2. A comprehensive subjective evaluation of the algorithms is presented in Chapters 4 and 5.

Since there is a need for the development of reliable measurement tools to reflect the benefits of noise reduction circuits (Kuk *et al.*, 1990), a listening effort test is proposed and developed in Chapter 4. This new test is employed to assess the Ephraim-Malah single-microphone noise reduction schemes and – as competitor – a multi-microphone noise reduction algorithm proposed by Wittkop (2000).

Besides speech intelligibility and listening effort (Chapter 4) the assessment of the subjective processing quality of the algorithms is considered in Chapter 5 since sound quality is, in general, a major feature for the acceptance of a hearing aid. Paired comparisons are applied to assess different aspects of subjective quality. The Bradley-Terry model (Bradley and Terry, 1952) is employed to obtain scale values for the algorithms.

Finally, different "objective quality measures" are investigated in Chapter 6 with regard to their applicability to *predict* the perceived sound quality of the noise reduction algorithms. The evaluation results obtained in Chapter 5 are used to determine correlations between subjective data and objective predictions. The results are compared with others given in the literature. The "best" objective measure is finally applied to assess different modifications of the Ephraim-Malah noise reduction algorithm.

Chapter 2

Speech pause detection for noise spectrum estimation by tracking envelope minima

A speech pause detection algorithm is an important and sensitive part of most single-microphone noise reduction schemes for enhancing speech signals corrupted by additive noise as an estimate of the background noise is usually determined when speech is absent. An algorithm is proposed which detects speech pauses by adaptively tracking minima in a noisy signal's power envelope both for the broadband signal and for the high-pass and low-pass filtered signal. In poor signal-to-noise ratios, the proposed algorithm maintains a low false-alarm rate in the detection of speech pauses while the standardized algorithm of ITU G.729 shows an increasing false-alarm rate in unfavorable situations. These characteristics are found with different types of noise and indicate that the proposed algorithm is better suited to be used for noise estimation in noise reduction algorithms, as speech deteriorations may thus be kept at a low level. It is shown that in connection with the Ephraim-Malah noise reduction scheme (Ephraim and Malah, 1984), the speech pause detection performance can even be further increased by using the noise-reduced signal instead of the noisy signal as input for the speech pause decision unit.

2.1 Introduction

New technologies in mobile telecommunication, robust speech recognition and digital hearing aids are a strongly driving force in the development of real-time noise reduction algorithms. The number of publications on single-microphone noise reduction algorithms indicates an unbroken interest in this research field over the past two or three decades. A crucial point for these kind of algorithms is the concurrent estimate of the target speech spectrum

and the interfering noise spectrum in particular. Since most realistic noisy environments are characterized by non-stationarity, it is necessary to update the noise spectrum estimate as often as possible to maintain an effective noise reduction. This can for example be done whenever target speech is absent, which means that the input signal consists of noise only. Another constraint is the limited complexity of the algorithm when it is supposed to become implemented in digital circuits. Hence, computational and memory requirements should be as low as possible.

Different algorithms have been proposed which *continuously* update the noise estimate and hence avoid the need for explicit speech pause detection. Martin (1993, 1994) uses the minimum of the sub-band signal power within a time window of about 1 s as an estimate of the noise power in the respective sub-band. This idea was already formulated by Paul (1981). Doblinger (1995) proposed a continuous noise estimation scheme similar to Martin's which is computationally more efficient. This scheme was, however, not systematically tested. Hirsch (1993) and Hirsch and Ehrlicher (1995) proposed an algorithm which is based on the observation that the most commonly occurring spectral magnitude value in clean speech is zero. Hence, having noisy speech their algorithm measures the distribution density function of the spectral magnitude and determines the maxima which are then used as an estimate of the respective noise magnitude. These kind of algorithms which avoid speech pause detection for noise estimation are supposed to cope better with non-stationary (i.e., fluctuating) noise, since they are generally faster in their adaptation to changing noise levels even during speech activity. On the other hand, the continuous update of the noise estimate (independently in the sub-bands) is susceptible to erroneously capture speech energy. This, however, leads inevitably to speech deterioration in a subsequent noise reduction process. Fischer and Stahl (1999) investigated a spectral subtraction noise reduction algorithm with a continuous noise spectrum updating scheme. They found that the corruption of the noise estimate by speech is too large to be further considered and conclude that voice activity detection plays an important role and cannot be fully omitted. Recently, Nemer *et al.* (1999) proposed to use the kurtosis (fourth-order statistics) of the noisy signal to continuously estimate speech and noise energies. The examples presented used noisy speech signals with positive signal-to-noise ratios and yield promising results, but further research is required to extend these results to negative signal-to-noise ratios and different classes of noise, respectively.

Most authors reporting on noise reduction refer to speech pause detection when dealing with the problem of noise estimation. As Hirsch (1993) pointed out, "this is a very difficult and ultimately unsolved problem for realistic situations with a varying noise level". A lot of studies thus evade the problem by using an ideal speech pause detection using the clean speech signal or by using only short test signals with an initial noise-only period for

noise estimation without the need for updating the noise spectrum estimate. In some applications like audio restoration (e.g., restoration of old gramophone recordings) the noise estimation indeed can often be done "manually" off-line. However, other applications like noise reduction for mobile communication and for digital hearing aids require automatic updating of the noise spectrum estimate. Most authors agree that voice activity or speech pause detectors, respectively, are a very sensitive and often limiting part of systems for the reduction of additive noise in speech (Dendrinos and Bakamidis, 1994; Sovka and Pollák, 1995).

Various procedures for speech pause detection have been described in the literature so far. Kang and Fransen (1989) proposed a very simple scheme. Whenever the low-pass band energy (in the frequency range from 0 to 1 kHz) of a current signal frame is below a specific fraction of the low-pass band dynamic range as scanned in the past frames, the frame is used for updating the noise spectrum estimate. Obviously, this procedure has strong limitations. It will only work with higher signal-to-noise ratios and will fail in noises with prominently low frequencies. A more elaborate algorithm using adaptive energy thresholds was proposed by van Gerven and Xie (1997). Elberling *et al.* (1993) used the so-called synchro method for spectral estimation of the background noise. This procedure makes use of the specific characteristic of voiced speech sounds, i.e. that the energy is confined to pitch-harmonic frequencies. Based on successive multiplication of the envelopes from neighbouring pairs of band-pass signals, followed by a summation over all resulting signal-products, a global measure of energy synchronization is obtained which is then used to classify the time frames of the input signal into those dominated by speech (high synchronization) and those not dominated by speech (low synchronization). This patent application is reported to work successfully in signal-to-noise ratios ranging from +9 to −9 dB with various noises. However, an increase of wrong speech pause decisions with decreasing SNR is reported. Sheikhzadeh *et al.* (1995) proposed a pause detection algorithm based on an auto-correlation voicing detection which was performed on the enhanced signal (i.e., after the noise reduction rather than on the noisy signal). Although extensive testing is mentioned, no performance results are presented. However, the authors state that the algorithm is not supposed to work well below signal-to-noise ratios of 0 dB. Dendrinos and Bakamidis (1994) presented an algorithm for determining the starting and ending points of speech segments in coloured-noise environments through singular value decomposition based on some thresholds which have been determined experimentally. Good performance was proved for SNRs higher than 0 dB. However, the complexity of the algorithm makes a real-time implementation difficult. Recently, El-Maleh and Kabal (1997) performed a comparative study of three voice activity detection (VAD) algorithms: the VAD used in the GSM cellular system (Srinivasan and Gersho, 1993), the VAD used in the enhanced variable rate codec (EVRC) of the

North American CDMA-based PCS and cellular systems (TIA, 1996), and a third-oder statistics based VAD (Rangoussi and Carayannis, 1995). Unfortunately, the authors did not investigate false-alarm rates and hit rates systematically but present only some noisy waveforms with the respective VAD decisions. However, the EVRC VAD is reported to show consistent superiority over the other VADs. Davídek *et al.* (1996) implemented a speech activity detector using cepstral coefficients for use in a real-time noise cancellation system. However, a comprehensive evaluation of the detector itself is not given. Abdallah *et al.* (1997) introduced a local entropic criterion for speech signal detection. Very good performance down to SNRs of $-20\,\mathrm{dB}$ is reported. However, only white noise was tested so far. McKinley and Whipple (1997) suggested a model based speech pause detection algorithm which is claimed to be robust for low SNRs. The speech pause detection problem is formulated into a decision theory framework. However, this algorithm requires extensive training of a Hidden Markov Model with the set of speech prototypes to be encountered. Itoh and Mizushima (1997) proposed a speech/non-speech identification based on four different parameters. The first is the maximum value of the auto-correlation function of the LPC residual signal, which represents the degree of the periodicity of the signal waveform. Second is a spectral slope parameter, third is a reflection coefficient which itself is computed from some PARCOR coefficients, and fourth is the signal energy. For each of the parameters, Itoh and Mizushima (1997) used empirically determined thresholds for a speech/stationary noise/non-stationary noise decision. It seems, however, that the decision for non-stationary noise is made only on the basis of the spectral slope parameter. Unfortunately, the proposed algorithm was not tested in low SNR situations.

Irrespective of the actual kind of speech pause detector used, a comprehensive and fair evaluation should include its hit rate as well as its false-alarm rate using different noises with a large variety of signal-to-noise ratios. These measures reveal most of an algorithm's capabilities and deficiencies. For an application in noise reduction, the problem is that a speech pause detection algorithm with a high false-alarm rate results in remarkably deteriorated speech after the noise reduction. On the other hand, a speech pause detection algorithm that finds too few of the actual speech pauses results in worse reduction of the noise. Hence, noise estimation is a very sensitive stage in the noise reduction process.

The algorithm for speech pause detection that will be described in the next section dynamically tracks the minima in the signal's temporal power envelope as well as in its low- and high-pass frequency band power envelopes. After a number of threshold comparisons, a frame-by-frame decision is made on the presence of a speech pause. This approach was motivated by the work of Festen *et al.* (1993), who used the minima in the signal envelope for estimating the noise level in a speech-plus-noise signal to control an AGC (automatic gain control) algorithm for hearing aids. The proposed

algorithm can be regarded as an extension of the simple scheme proposed by Kang and Fransen (1989). In order to assess its applicability to real-time noise reduction for practical applications (see above), both the hit rate and false-alarm rate are evaluated for a large range of SNRs and different types of noise and compared to a voice activity detector (VAD) algorithm recommended by the International Telecommunication Union (ITU, 1996a).

2.2 Algorithm

The speech pause detection algorithm calculates the signal's temporal power envelope $E(p)$ by summing up the squares of the spectral components of the input signal in each short-time frame p:

$$E(p) = \sum_k |X(p, \omega_k)|^2 \tag{2.1}$$

Here, $X(p, \omega_k)$ denotes the spectral component of the noisy input signal at frequency ω_k at time frame p. In addition, a low-pass band power envelope and a high-pass band power envelope are calculated:

$$E_{\mathrm{LP}}(p) = \sum_l |X(p, \omega_l)|^2 \tag{2.2}$$

$$E_{\mathrm{HP}}(p) = \sum_m |X(p, \omega_m)|^2, \tag{2.3}$$

where l runs over all spectral components up to the cut-off frequency, and m runs over the remaining spectral components. In order to slightly smooth the envelopes, $E(p)$, $E_{\mathrm{LP}}(p)$ and $E_{\mathrm{HP}}(p)$ are averaged over a few frames by a recursive low-pass filter of first order with a release time constant τ_E; no smoothing is performed in case of an increase in energy (i.e., attack time zero) to avoid smearing over onsets. The algorithm tracks the minimum value and the maximum value of each envelope and uses these for the speech pause decision as described by the following scheme:

1. After an assumed 200 ms initial phase of noise only the minimum and maximum values are set as follows:

$$\begin{aligned} E_{\min}(p) &\equiv E(p) & \qquad E_{\max}(p) &\equiv E(p) \\ E_{\mathrm{LP,min}}(p) &\equiv E_{\mathrm{LP}}(p) & \qquad E_{\mathrm{LP,max}}(p) &\equiv E_{\mathrm{LP}}(p) \\ E_{\mathrm{HP,min}}(p) &\equiv E_{\mathrm{HP}}(p) & \qquad E_{\mathrm{HP,max}}(p) &\equiv E_{\mathrm{HP}}(p) \end{aligned} \tag{2.4}$$

   This guarantees that the minimum envelope values correspond roughly with the noise energy at the beginning.

2. The minimum and maximum values are updated for each of the three envelopes in the following manner:

- If the current envelope value is larger than the maximum value for the corresponding envelope, then the maximum value is set to the current value. Otherwise, the maximum value slowly decays. This is done by a recursive low-pass filter of first order with a release time constant τ_{decay}, which takes as input the current envelope value.
- If the current envelope value is smaller than the minimum value for the corresponding envelope, then the minimum value is set to the current value. Otherwise, the minimum value is slowly raised. This is done by a recursive low-pass filter of first order with attack time constant τ_{raise}, which takes as input the current envelope value.

3. The differences between the maximum and the minimum values are calculated for each envelope:

$$\begin{aligned} \Delta(p) &= E_{\max}(p) - E_{\min}(p) \\ \Delta_{\text{LP}}(p) &= E_{\text{LP,max}}(p) - E_{\text{LP,min}}(p) \\ \Delta_{\text{HP}}(p) &= E_{\text{HP,max}}(p) - E_{\text{HP,min}}(p) \end{aligned} \tag{2.5}$$

4. Three different criteria are introduced of which only one has to be true for making the decision that target speech is not present in the actual frame: a) the speech pause decision can be made because of a low signal dynamic in both the low-pass and the high-pass band (*Dyn Speech Pause*); b) the decision can be based on the low-pass band information (*LP Speech Pause*); and c) it can be made upon the high-band information (*HP Speech Pause*). These decision criteria are derived as follows:

   (a) If Δ_{LP} is smaller than some threshold η and also $\Delta_{\text{HP}} < \eta$ then it is assumed that only noise is present due to the very small dynamic range of the signal. ($\Rightarrow$ *Dyn Speech Pause*)

   (b) If (a) is not true, it is checked whether Δ_{LP} is bigger than η (otherwise the dynamic range in the low-pass band is very small and it should not receive too much attention $\Rightarrow$ *no LP Speech Pause*). Now, if the difference between the current $E_{\text{LP}}(p)$ and $E_{\text{LP,min}}(p)$ of the low-pass band envelope is smaller than some fraction *pc* of Δ_{LP} (which means that the actual envelope is near its minimum), a closer look at the high-pass band is necessary to support a speech pause decision:

      - *Case 1:* Δ_{HP} *of the high-pass band is smaller than threshold* η*.*
        In this case no additional information can be obtained from the high-pass band because of its small dynamic range. Now,

if at least $E(p)$ (the signal's envelope) lies in the lower half of its dynamic range (i.e. in the lower half between $E_{\min}(p)$ and $E_{\max}(p)$) the current frame can be assumed to be a speech pause because of the closeness of the low-pass band energy to its minimum value ($\Rightarrow$ *LP Speech Pause*) otherwise, however, there is not enough support for a speech pause decision ($\Rightarrow$ *no LP Speech Pause*).

- *Case 2:* Δ_{HP} *is bigger than two times the threshold* η*.*
  In this case, there is enough dynamic range to pay attention to the high-pass band. Thus, it is demanded that the difference between the current $E_{\mathrm{HP}}(p)$ and $E_{\mathrm{HP,min}}(p)$ of the high-pass envelope is smaller than two times the fraction pc of Δ_{HP} to support the small envelope value in the low-pass band. Then a noise-only frame is assumed ($\Rightarrow$ *LP Speech Pause*). This demand is not as strict as that for the low-pass band, to account for the case that the disturbing noise has a rather high-frequency characteristic. But if this condition is not fulfilled, speech may be present in the actual frame ($\Rightarrow$ *no LP Speech Pause*).
- *Case 3:* Δ_{HP} *is smaller than two times the threshold* η*, but bigger than* η*.*
  In this case, which is not as clear as Case 2, it is only demanded that $E_{\mathrm{HP}}(p)$ (the high-pass envelope) lies in the lower half of its dynamic range to support the small envelope value in the low-pass band. Then it is assumed that target speech is absent ($\Rightarrow$ *LP Speech Pause*). However, if this condition is not fulfilled, speech may be present in the actual frame ($\Rightarrow$ *no LP Speech Pause).*

(c) Condition (b) accounts for the case that the disturbing noise has a rather high-frequency characteristic, hence the speech pause decision should mainly be made upon the information in the low-pass band. To account also for the case that it has a rather low-frequency characteristic, the same conditions as under condition (b) have to be checked but now with reverse roles of the low-pass and the high-pass bands to determine whether target speech is absent (*HP Speech Pause*).

Figure 2.1 gives a flowchart of the proposed speech pause detection algorithm.

Due to its flexible design this novel approach for speech pause detection can easily be adjusted to obtain a rather low false-alarm rate by adapting the main parameters η and pc. Generally, a low false-alarm rate is desirable to reduce speech distortions in the subsequent noise reduction process. However, this also results in a reduced hit rate.

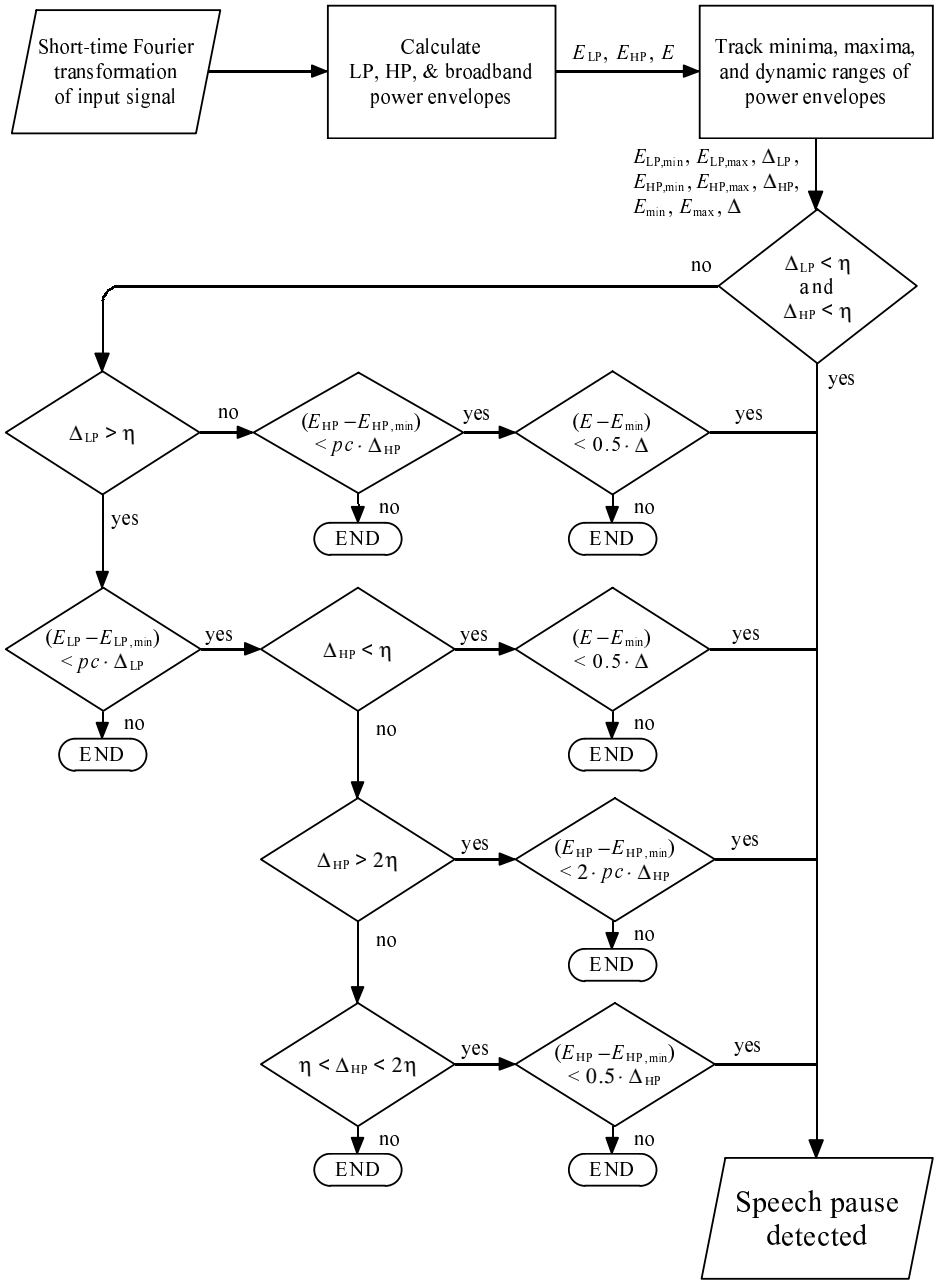


Figure 2.1: *Flowchart of the proposed speech pause detection algorithm operating on a single time frame. See text for details.*

During the development of the algorithm noisy signals generated from various different noise types and speech signals at several signal-to-noise ratios were used for performance verification. Finally, the following values were chosen for the free parameters: The input signal is digitized with a sampling frequency of 22050 Hz and partitioned in Hann-windowed segments of length 8 ms with 4 ms overlap. These segments are padded with zeros and a 256-point FFT is performed. This framework is compatible with most single-microphone noise reduction algorithms which can thus easily be integrated. The cut-off frequency between low-pass and high-pass band was set to 2 kHz, motivated by the fact that excluding speech frequencies above 1.9 kHz has a roughly similar effect on speech intelligibility as excluding those below this value (Jones, 1983). The time constant τ_E for the envelope smoothing was set to 32 ms, τ_{decay} and τ_{raise} were both set to 3 s. The threshold η was set to 5 dB and the fraction pc was set to 0.1.

2.3 Examples

To illustrate the speech pause detection scheme, Figures 2.4 to 2.11 show some detection examples using a target sentence of approximately 5 s length mixed with different noises.

Figures 2.4 to 2.6 show an example with noise from inside a running car. The bar at the bottom of the figures shows the real speech pauses which were determined manually. For comparison, the waveform of the clean sentence is displayed in Figure 2.2, the mixed signal with a signal-to-noise ratio of −5 dB is displayed in Figure 2.3. The speech pause decisions of the algorithm are displayed in the upper three bars. The distinct bars give additional information about the reason for the speech pause decision. The first bar shows a symbol whenever a speech pause is detected due to a small dynamic range of the signal in the low-pass band as well as in the high-pass band, and generally in the initial noise estimation phase (the first 200 ms). The second bar shows a symbol whenever a speech pause is detected on the basis of the low-pass band information. Finally, a symbol in the third bar means that the decision was based on the high-pass band information.

The car noise example shows that considering the signal's broadband power envelope only is not sufficient to obtain a reliable speech pause detection (cf. Figure 2.4). In this case, the signal's broadband envelope as well as the low-pass band envelope (Figure 2.5) are strongly disturbed by the noise. However, the high-pass envelope (Figure 2.6) is "clean enough" for making speech pause decisions. Actually, the third bar in the figures shows that the decision is mainly based on the high-pass information.

Figures 2.7 to 2.9 show an example, where the sentence is mixed with the noise of a drilling machine at +5 dB SNR. This noise makes it impossible to get reliable speech pause information from the high-pass channel, but in

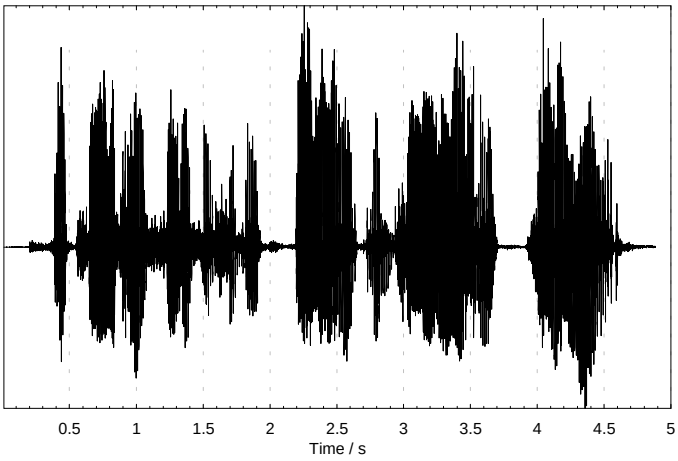


Figure 2.2: *Waveform of the sentence "I played in a theatre festival, honoring the German writer Heiner Müller."*

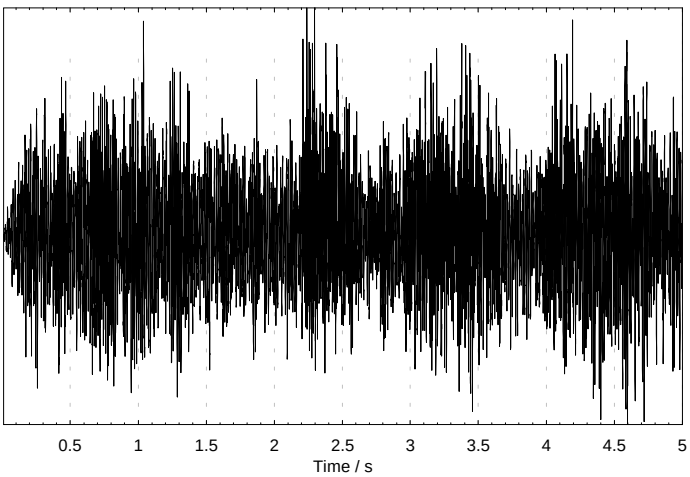


Figure 2.3: *Waveform of the sentence displayed in Fig. 2.2 mixed with car noise at -5 dB SNR.*

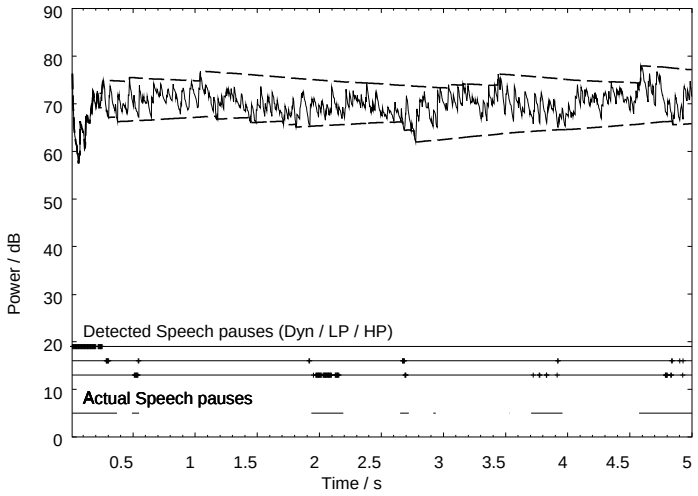


Figure 2.4: *Power envelope (solid curve) of the sentence displayed in Fig. 2.2 when mixed with car noise at -5 dB SNR together with detected and actual speech pauses. The dashed curves display $E_{\min}$ and $E_{\max}$, respectively.*

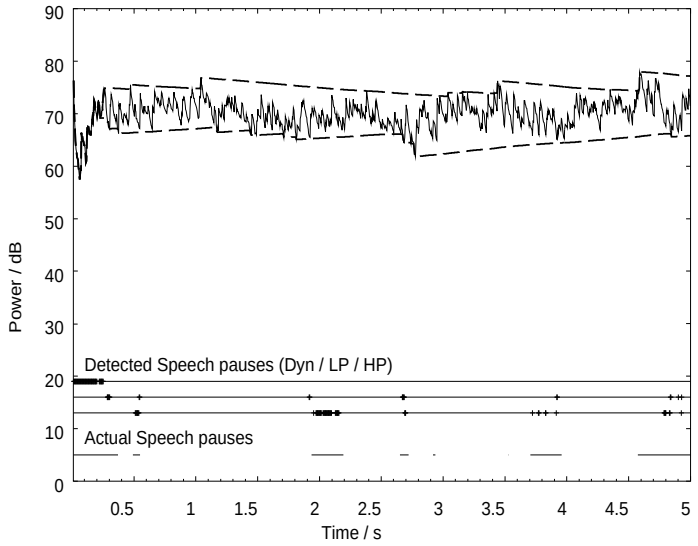


Figure 2.5: *Low-pass band power envelope (solid curve) of the sentence displayed in Fig. 2.2 when mixed with car noise at -5 dB SNR together with detected and actual speech pauses. The dashed curves display $E_{\mathrm{LP,min}}$ and $E_{\mathrm{LP,max}}$, respectively.*

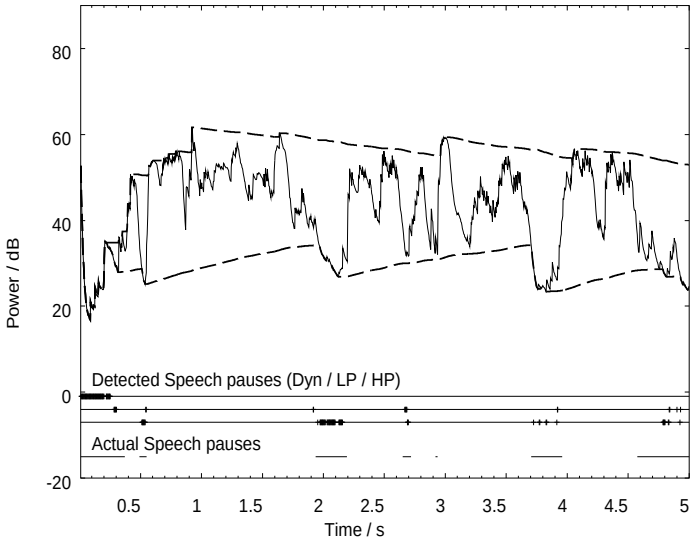


Figure 2.6: *High-pass band power envelope (solid curve) of the sentence displayed in Fig. 2.2 when mixed with car noise at −5 dB SNR together with detected and actual speech pauses. The dashed curves display $E_{\mathrm{HP,min}}$ and $E_{\mathrm{HP,max}}$, respectively.*

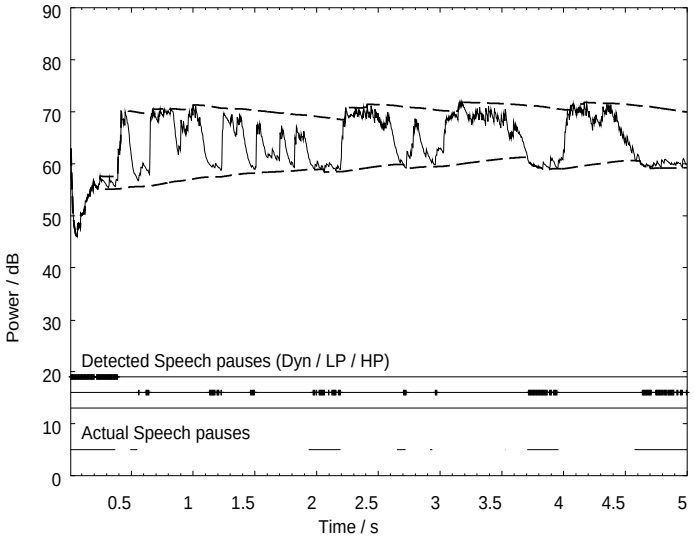


Figure 2.7: *Power envelope (solid curve) of the sentence displayed in Fig. 2.2 when mixed with drilling machine noise at +5 dB SNR together with detected and actual speech pauses. The dashed curves display E_{min} and E_{max}, respectively.*

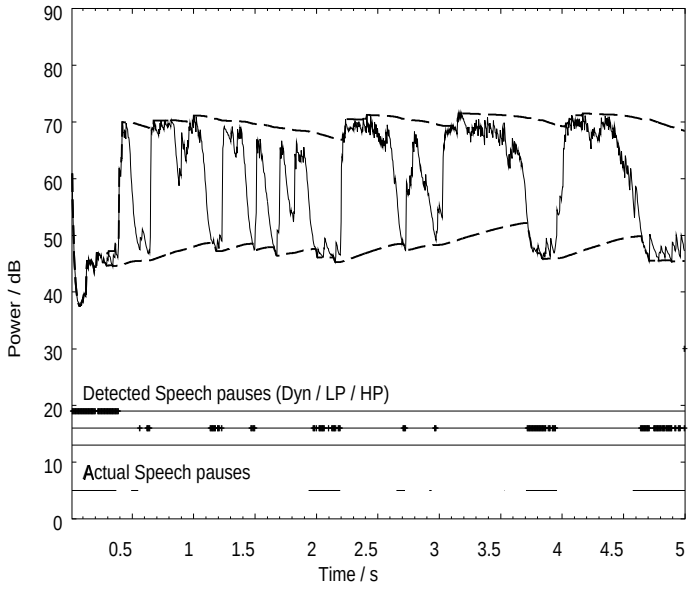


Figure 2.8: *Low-pass band power envelope (solid curve) of the sentence displayed in Fig. 2.2 when mixed with drilling machine noise at +5 dB SNR together with detected and actual speech pauses. The dashed curves display* $E_{\mathrm{LP,min}}$ *and* $E_{\mathrm{LP,max}}$*, respectively.*

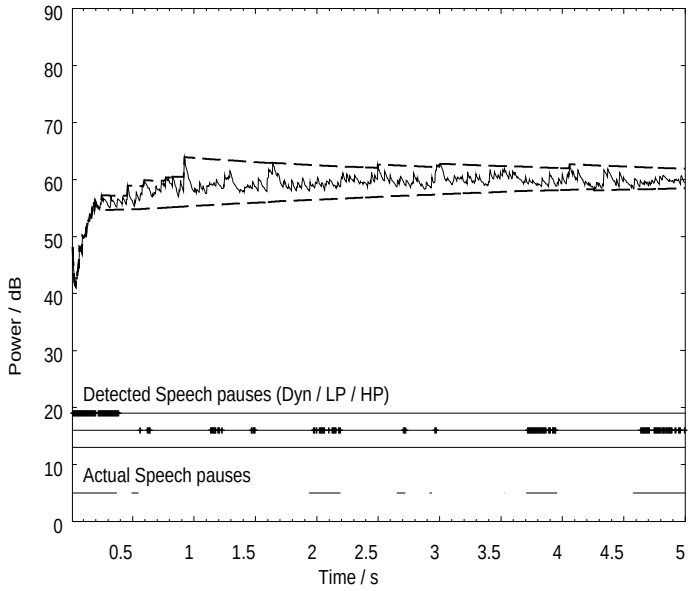


Figure 2.9: *High-pass band power envelope (solid curve) of the sentence displayed in Fig. 2.2 when mixed with drilling machine noise at +5 dB SNR together with detected and actual speech pauses. The dashed curves display* $E_{\mathrm{HP,min}}$ *and* $E_{\mathrm{HP,max}}$*, respectively.*

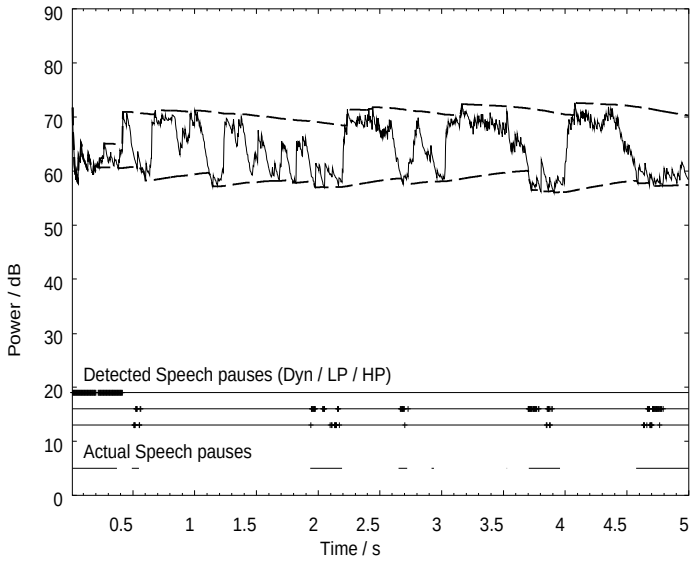


Figure 2.10: *Low-pass band power envelope (solid curve) of the sentence displayed in Fig. 2.2 when mixed with restaurant noise at +5 dB SNR together with detected and actual speech pauses. The dashed curves display* $E_{\mathrm{LP,min}}$ *and* $E_{\mathrm{LP,max}}$*, respectively.*

this case the low-pass band information can be used. Comparison with the lowest bar in the figures (the "true" speech pauses) shows that a good speech pause detection is obtained. Although the algorithm wrongly considers the time frames around 0.6 s ("p" from "played"), 1.2 s ("th" from "theatre") and around 1.5 s ("f" from "festival") as noise, these speech parts actually sound very similar to equally short segments of the drill noise. Hence, these wrong decisions are assumed to have no adverse effects on the speech quality when used for noise estimation in a noise reduction algorithm.

Figures 2.10 and 2.11 show an example with restaurant noise, which is neither mainly low-frequency nor high-frequency in its characteristics. As can be seen at the second and third bar in the figures, the speech pause detection, indeed, is sometimes based on the low-pass band information and sometimes on the high-pass information. In combination, a good speech pause detection performance is obtained.

2.4 Comparison with G.729 VAD algorithm

In 1996 the International Telecommunication Union (ITU) standardized a voice activity detector (VAD) algorithm as its Recommendation G.729 Annex B (ITU, 1996a). The VAD algorithm makes a voice activity decision every 10 ms based on differential parameters of the full-band energy, the low-

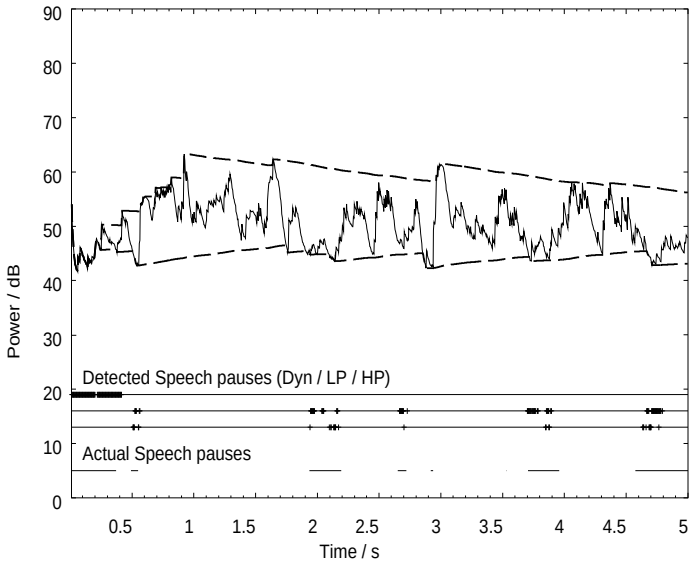


Figure 2.11: *High-pass band power envelope (solid curve) of the sentence displayed in Fig. 2.2 when mixed with restaurant noise at +5 dB SNR together with detected and actual speech pauses. The dashed curves display* $E_{\mathrm{HP,min}}$ *and* $E_{\mathrm{HP,max}}$*, respectively.*

pass band energy, the zero-crossing rate and a spectral distortion measure. These are obtained at each frame as differences between each parameter and its respective long-term average. The output of the VAD module is either 1 or 0, indicating the presence or absence of voice activity, respectively. Several publications compared their own algorithms with the G.729 VAD so far (Stegmann and Schröder, 1997; Sohn *et al.*, 1999).

2.4.1 Procedure

A female reading of a short story (41 s length) from the German PhonDat database (Draxler, 1995) was used to test the performance of the proposed algorithm versus the G.729 algorithm. The speech signal was mixed with a car noise, a multi-talker babble noise, an aircraft engine noise, and a factory noise, respectively, which were taken from the NOISEX-92 database (Steeneken and Geurtsen, 1988). Signal-to-noise ratios from −10 dB to +20 dB were employed. False-alarm rates (i.e., the fraction of all real speech frames that were erroneously detected as speech pauses) and hit rates (i.e., the fraction of all real speech pauses that were correctly detected as speech pauses) were determined in each noise condition for both the proposed algorithm and the G.729 algorithm. For the calculation of the false-alarm rate as well as the hit rate, the "real" speech frames and "real" speech pauses were determined using the G.729 VAD algorithm on the clean speech signal.

Using the G.729 itself as reference takes into consideration that no simple rule exists even for determining pauses in clean speech. Since the G.729 algorithm is recommended by the ITU, it can be taken for granted that it works well for clean speech. Note, that in the comparative test with the proposed new algorithm this may give an advantage for the G.729 algorithm, as it defines the "clean" standard.

Finally, both algorithms are compared in terms of receiver operating characteristics (ROC).[1]

2.4.2 Results

The detection results are shown in Figures 2.12 and 2.13. The upper panels show the false-alarm rate, the lower panels present the hit rate of both algorithms.

The comparison with the G.729 Annex B algorithm shows that the proposed speech pause detection algorithm yields a clearly lower false-alarm rate in each of the four different noises over the entire range of signal-to-noise ratios that were tested (cf. Figures 2.12 and 2.13). On the other hand, fewer speech pauses are actually detected than with the G.729 algorithm.

The false-alarm rates are lowest in car noise, followed by the multi-talker babble noise, the factory noise, and the aircraft engine noise. However, a principal difference between the algorithms is observed: While the proposed algorithm keeps the false-alarm rate and the hit rate almost constant with changing SNR, the performance of the G.729 algorithm strongly depends on the SNR – the lower the SNR, the larger the false-alarm rate as well as the hit rate.

In terms of receiver operating characteristics (ROC), the working point of the G.729 algorithm shifts up and to the right in ROC space with decreasing SNR, while the working point of the proposed algorithm stays nearly at the same place in ROC space. In general, the false-alarm rates can be decreased by changing threshold criteria in the algorithm's decision rules. This is, of course, connected with a decrease of the hit rates. Whether the proposed algorithm is generally "better" than the G.729 algorithm can be examined by comparing them in ROC space (in terms of discriminability, i.e. the area under the ROC curve). Figures 2.14, 2.15, and 2.16 show ROC curves of the proposed algorithm using car noise, babble noise, and aircraft noise, respectively. The left panels were obtained at signal-to-noise ratios of $-10\,\mathrm{dB}$; for the right panels SNRs of $+10\,\mathrm{dB}$ were used. The curves were generated by varying the threshold η in the decision rule of the proposed algorithm (cf. Section 2.2) from 1 to 25 dB in 1-dB steps.

Since in all noise conditions the G.729 algorithm falls below the ROC curve of the proposed algorithm, it may be concluded that the discriminability is better with the proposed speech pause detection algorithm.

Additionally, in Figure 2.16a the ROC curve was determined for the pro-

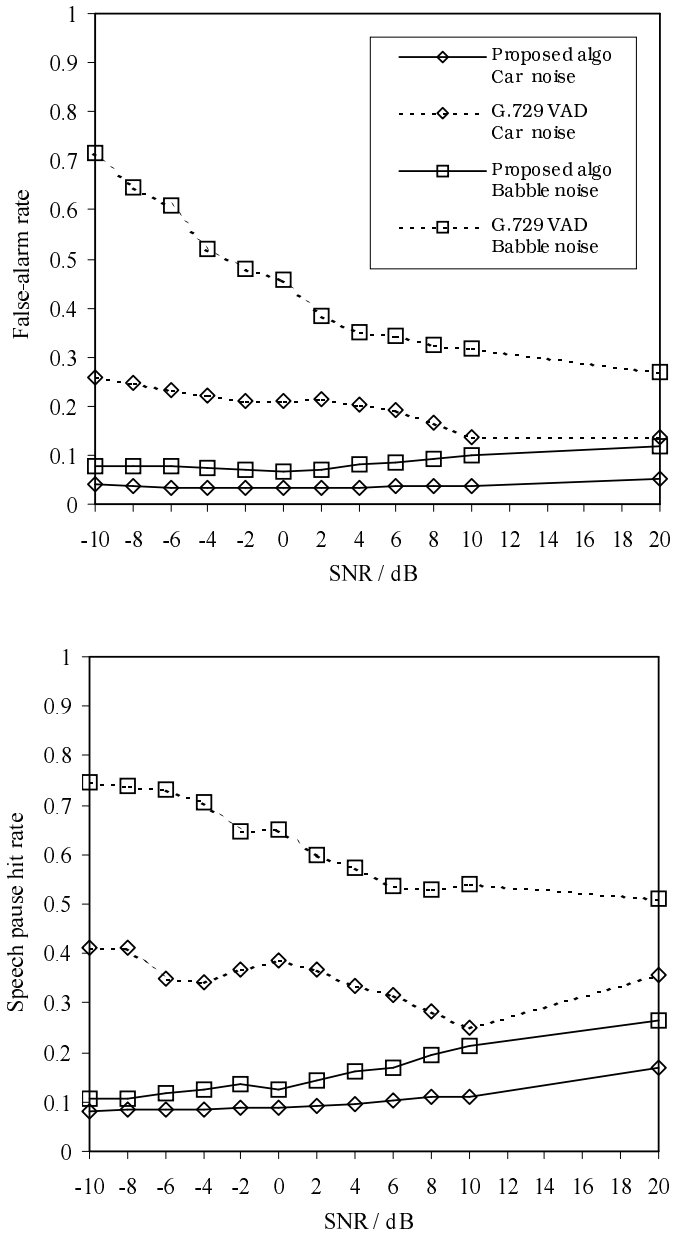


Figure 2.12: *Speech pause detection performance of the proposed algorithm and the G.729 VAD algorithm in car noise and multi-talker babble noise with signal-to-noise ratios ranging from −10 to +20 dB. The upper panel shows the false-alarm rates and the lower panel shows the hit rates with the respective algorithms.*

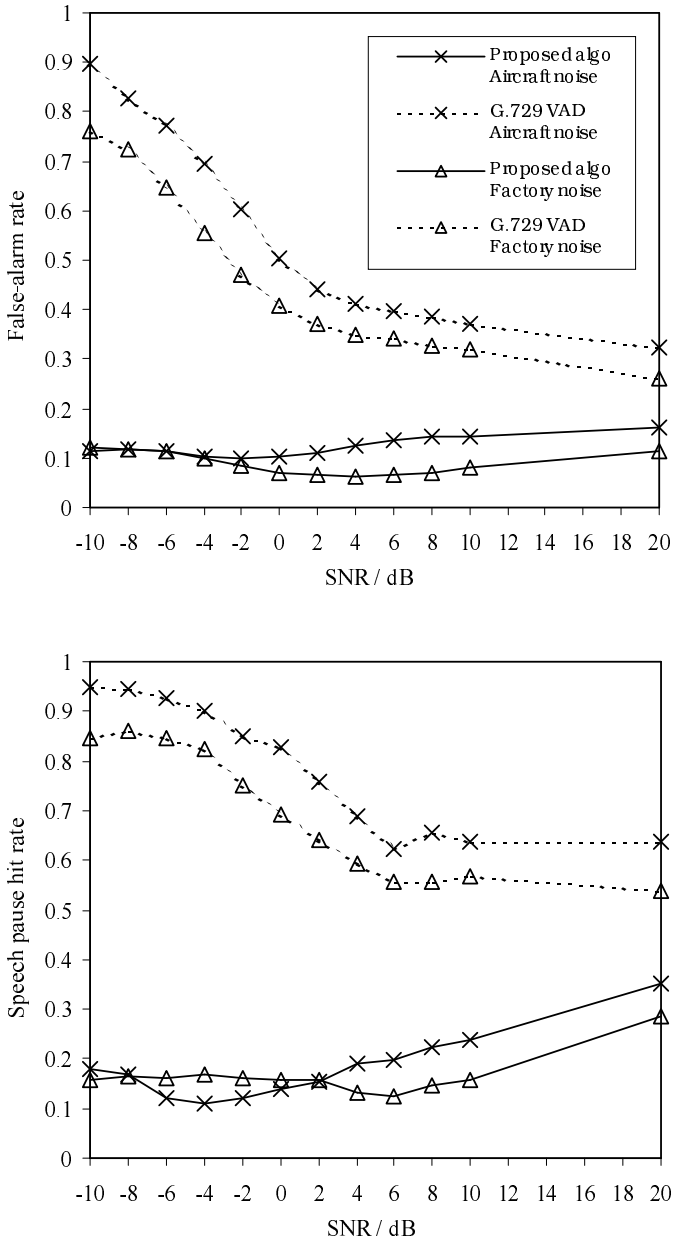


Figure 2.13: *Speech pause detection performance of the proposed algorithm and the G.729 VAD algorithm in aircraft engine and factory noise with signal-to-noise ratios ranging from −10 to +20 dB. The upper panel shows the false-alarm rates and the lower panel shows the hit rates with the respective algorithms.*

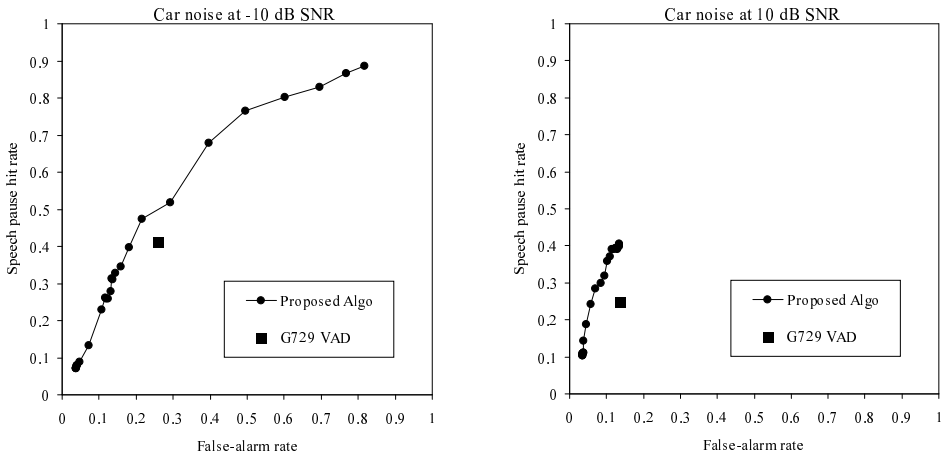


Figure 2.14: *ROC curve of the proposed algorithm using car noise at −10 dB SNR (left panel) and +10 dB SNR (right panel). For comparison, the performance of the G.729 VAD algorithm is also indicated.*

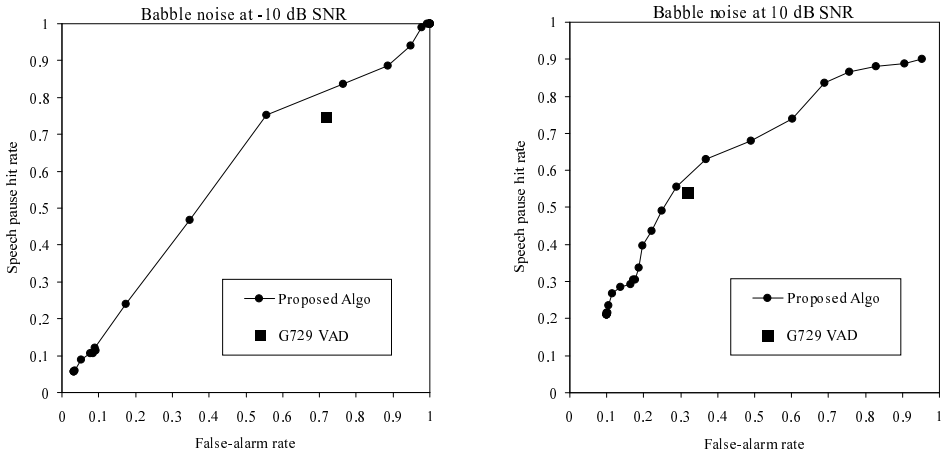
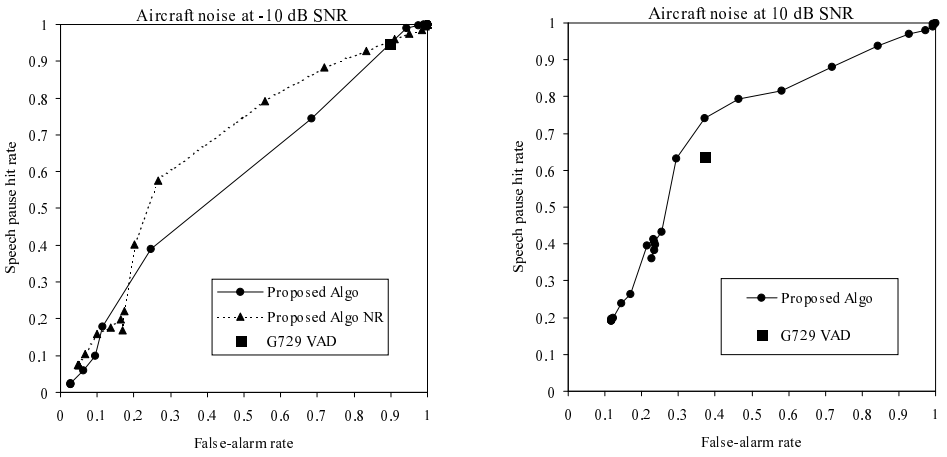


Figure 2.15: *ROC curve of the proposed algorithm using babble noise at −10 dB SNR (left panel) and +10 dB SNR (right panel). For comparison, the performance of the G.729 VAD algorithm is also indicated.*



(a) (b)

Figure 2.16: *ROC curve of the proposed algorithm using aircraft noise at −10 dB SNR (left panel) and +10 dB SNR (right panel). For comparison, the performance of the G.729 VAD algorithm is also indicated.*

posed algorithm using a noise-reduced signal as input for the speech pause detection (by employing the single-microphone noise reduction algorithm from Ephraim and Malah, 1984, on a frame-by-frame basis) instead of the noisy signal. The detected speech pauses are in turn used to adjust the noise spectrum estimate for the noise reduction. Although this leads to a recursive design of the signal flow, no stability problems were observed for a wide range of input signals and SNRs.

This modified algorithm is denoted as 'Proposed Algo NR'. Actually, the discriminability of the speech pause detection algorithm is further increased by this modification as can be seen at the larger area under the ROC curve (cf. Figure 2.16a).

2.4.3 Discussion

In a noise estimation application for noise reduction algorithms it is generally proposed to operate the speech pause detection at rather low hit rates to keep the false-alarm rate low. Large false-alarm rates in the speech pause detection lead to wrong noise spectrum estimates which include significant speech parts and hence cause artifacts in a subsequent noise reduction process. In fact, the proposed speech pause detection algorithm maintains a low false-alarm rate over a wide range of signal-to-noise ratios while the hit rate decreases only slightly at poorer SNRs. Hence, the algorithm keeps a relatively fixed position in ROC space over a wide range of SNRs. In contrast to the proposed algorithm, the algorithm of the ITU Recommendation

G.729 yields very large false-alarm rates (but also larger hit rates) at low SNRs.

Obviously, the G.729 was not designed to detect the true speech pauses in adverse noise conditions. In conditions where the speech is hardly noticeable, the G.729 VAD algorithm rather decides to classify this situation as speech-free (i.e., a kind of extended speech pause). Since this behaviour is inherent in the algorithmic design of the G.729 scheme, it cannot be overcome by global changes of its threshold parameters. In a noise reduction application, this behaviour probably makes it impossible for a noise reduction algorithm to "retrieve" the speech signal, if the whole signal is classified as noise. As the proposed algorithm detects speech pauses by tracking envelope minima, its behaviour at very poor SNRs differs here. It still decides for speech pauses only when energy minima occur.

The threshold parameters in the proposed speech pause detection algorithm were determined empirically to obtain low false-alarm rates for a wide range of input signals and SNRs. By this, speech deteriorations due to wrong noise spectrum estimates (i.e., including speech energy) in any subsequent noise reduction processing are minimized. However, low false-alarm rates are connected with lower hit rates which could also lead to signal deteriorations for certain types of strongly fluctuating noises. If the noise is strongly fluctuating in its characteristics between speech pauses, a noise estimate determined only when speech is absent is not sufficient to ensure effective noise reduction. For such conditions, noise reduction schemes have to be employed which exploit other features (for example separation in space between noise and target source, cf. Chapter 5 and Wittkop, 2000), or a running noise estimate has to be determined from the noisy signal and not only during speech pauses.

Apart from that, low hit rates in the proposed algorithm do not necessarily mean that some speech pause intervals are not detected at all, but rather that several frames *during* speech pauses are not detected as such (see for example Figure 2.4). For the adjustment of a noise spectrum estimate, the proposed algorithm can hence be employed at rather low hit rates to obtain low false-alarm rates and still detects at least some frames during most speech pauses.

It might seem strange that the false-alarm rates of the proposed algorithm increase slightly for *better* SNRs, but this is due to the fact that the G.729 defines the clean reference. Very soft consonant parts (with insignificant low energy) are classified as speech pause by the proposed algorithm. However, these parts are classified as speech by the G.729 algorithm.

2.5 Conclusions

The proposed speech pause detection algorithm maintains a low and approximately constant false-alarm rate over a wide range of signal-to-noise ratios. The hit rate decreases only slightly at poorer SNRs.

Since the proposed speech pause detection algorithm was shown to be superior to the G.729 VAD algorithm in terms of discriminability (area under the ROC curve) in speech with noise, it should be preferred in applications where noise disturbances may occur.

The performance can be further enhanced if the algorithm is combined with the single-microphone noise reduction algorithm proposed by Ephraim and Malah (1984) and the noise reduced signal is employed for the speech pause detection.

The relatively low complexity of the algorithm should allow an immediate application in, for example, digital hearing aids or cellular phones. The delay time due to the signal processing is below 10 ms.

Chapter 3

A review of the Ephraim-Malah noise reduction algorithms

An overview of the developments in the last decades concerning single-microphone noise reduction algorithms is given with the main focus on the schemes originally proposed by Ephraim and Malah (1984). One commonly reported artifact of most schemes based on conventional 'spectral subtraction', the musical tones phenomenon, counteracts expected benefits of the noise reduction processing since it is perceived as strongly disturbing the sound quality. Many attempts have been made so far to tackle this main drawback. Among others, especially the noise reduction algorithms proposed by Ephraim and Malah (1984, 1985) have been reported to not suffer from the musical tones artifact (Cappé, 1994). In addition, the implementation complexity of these schemes is relatively low. In contrast to most noise reduction algorithms based on Hidden Markov Models, the Ephraim-Malah schemes are not restricted in use to a previously learnt set of noises. Therefore, these single-microphone noise reduction algorithms are proposed for use in digital hearing aids. Some important features of the algorithms showing their capabilities are reviewed and discussed.

3.1 Introduction

Many studies have shown that noise evokes major difficulties for hearing-impaired subjects, even for persons with low to moderate hearing losses (Weiss and Neuman, 1993). According to the Working Group on Communication Aids for the Hearing-Impaired (1991), it is one of the most common complaints made by hearing-aid users that speech in noise, or speech in a reverberant room, is particularly difficult to understand.

However, in their overview on noise reduction in hearing aids, Weiss and

Neuman (1993) conclude that no hearing aid can be capable of extracting completely noise-free speech functions from a single-microphone input signal and the best that can be hoped for is that intelligibility improvements that are achieved through noise reduction are greater than intelligibility reductions that result from the loss or distortion of speech components due to the processing. In their comprehensive review of noise reduction in the 1970s, Lim and Oppenheim (1979) also point out that while many of the enhancement systems reduce the apparent background noise, many of them actually reduce intelligibility. Weiss and Neuman (1993) remark that only multi-microphone methods have been shown to be capable of improving speech intelligibility for a range of acoustic environments and noises so far. However, one advantage of single-microphone noise reduction procedures, compared to multi-microphone methods, is their robustness against the number of noise sources and the level of reverberation.

Another fact stresses the importance of investigating single-microphone noise reduction schemes for their applicability in hearing aids: Due to cosmetic reasons and due to the presumed rejection by the customers, most hearing aid manufacturers are not willing to implement multi-microphone noise reduction methods that need larger distances between the microphones than a few millimeters, even though the efficiency usually increases with distance between microphones.

3.2 Literature overview

3.2.1 Spectral subtraction

Most single-microphone noise reduction algorithms[2] proposed in the last decades are based on "spectral subtraction", which, according to Malca *et al.* (1996), has become "almost standard in noise reduction". In its simplest form, a noisy signal is overlap-partitioned in short time frames of some milliseconds which are transformed to the frequency domain by, for example, a Fast Fourier Transform. An estimated noise magnitude spectrum which is usually updated in speech pauses is subtracted from each noisy magnitude spectrum. The noise-reduced spectra are transformed back to the time domain using the unchanged phase of the noisy signal and overlap-added to give the noise-reduced output signal. The only limited importance of the phase in speech enhancement has been experimentally demonstrated by Lim and Wang (1982) and Vary (1985).[3]

Although spectral subtraction reduces the background noise, it does not seem to improve speech intelligibility. Niederjohn *et al.* (1987) conclude their overview on spectral subtraction noise reduction in claiming that it is probably not possible to enhance speech intelligibility in noise with this technique. They think that some information related to the speech signal must be extracted and used to enhance speech intelligibility. Actually, Heide

(1994) reports that spectral subtraction together with an enhancement of resonant formants provides a small but significant improvement in speech intelligibility when used in aircraft noise as front end for a linear predictor voice encoder.

Besides unsatisfactory results in terms of speech intelligibility improvement, another problem with almost all noise reduction algorithms based on spectral subtraction is the perceived sound quality: The algorithms are reported to produce artifacts in the residual noise, showing very unnatural disturbances (Boll, 1979; Preuss, 1979; Berouti *et al.*, 1979), which are due to the stochastic fluctuations in the spectral magnitudes of the noise signal. Whenever any current spectral magnitude of the noise exceeds the average noise estimate for the respective frequency, some noise energy is left after the subtraction at that spectral bin. This leads to stochastically distributed spectral peaks (i.e., tones) in the residual noise which is thus often called "musical noise" or "musical tones".

A few years after the description of these problems, different types of minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimators were proposed by Ephraim and Malah (1983, 1984, 1985). These algorithms are reported to yield a significant noise reduction while eliminating the musical noise phenomenon (Cappé, 1994).

3.2.2 HMM-based systems

The Ephraim-Malah algorithms (Ephraim and Malah, 1984, 1985) are derived from the assumption that speech and noise spectral components can be modeled as statistically independent Gaussian random variables. Doblinger (1995) called these algorithms the "best known spectral amplitude estimators of the noisy speech signal". However, since the Gaussian assumptions were not believed to be really fulfilled (Scalart *et al.*, 1996), many later developments in the 1980s and 1990s rather tried to circumvent assuming any specific distribution for speech (Boll, 1992). It was suggested to train Hidden Markov Models (HMM) on clean speech and/or the noise signals that are to be expected (Ephraim, 1992). However, the HMM-based noise reduction systems are inherently relying on the type of training data (Sheikhzadeh *et al.*, 1995). Hence, they work best with the trained type of noise but often worse with other types of noise. Boll (1992) concludes his review on noise reduction schemes in the 1980s in stating that better performance of noise reduction algorithms in machine speech recognition has come at the expense of more complex models supported by greater and greater computing requirements. Therefore, most of these algorithms are not suitable for hearing aid applications yet. Recently, Sameti *et al.* (1998) proposed a HMM-based enhancement system which is reported to have a computational complexity similar to that of spectral subtraction with a significantly superior performance. They evaluated their system with test sentences cor-

rupted by three types of noise at signal-to-noise ratios of 0, 5 and 10 dB. Five normal-hearing subjects were asked to rate the sound quality on a 5-point scale ranging from "bad" (Score 1) to "excellent" (Score 5). The proposed noise reduction system consistently outperformed the spectral subtraction system by one score on average. However, the results are still somewhat disappointing since the algorithm is only rated "poor" (Score 2) in multi-talker noise at 0 dB SNR and "fair" at 10 dB SNR. The helicopter noise and the white noise are rated "fair" at 0 dB SNR and "good" at 10 dB SNR. The ratings for the unprocessed signals are missing in the publication, but Sameti *et al.* admit that in all cases some listeners preferred the unprocessed signal over the enhanced one.

3.2.3 Usage of psychoacoustical properties

Parallel to the HMM-based developments, various modifications of the basic spectral subtraction noise reduction rule and also of the Ephraim-Malah noise reduction scheme were suggested in the literature. The mainstream in noise reduction in the 1990s can probably be characterized as "modified spectral subtraction using psychoacoustic criteria". In fact, a reduction of the musical tones phenomenon can already be obtained by applying the noise reduction to sub-bands derived from the frequency groups of the human auditory system instead of applying it to each frequency component of the FFT (Hirsch and Ehrlicher, 1995). Actually, this was already found by Peterson and Boll in 1981 and also used by Singh and Sridharan (1998). By this, the variance and the error in the noise estimate decreases. Averaged over six listeners, Singh and Sridharan (1998) found an improvement of 0.4 point on a 5-point quality scale, compared to spectral subtraction with linear frequency scale. Bodin and Villemoes (1997) proposed a rule for choosing the most favorable time-frequency decomposition using wavelet packets for spectral subtraction. Similarly, Nishimura *et al.* (1998) used a wavelet transform for the spectral subtraction technique. They performed speech intelligibility tests but did not find significant differences compared to standard spectral subtraction.

Virag (1995, 1999) proposed another spectral subtraction based noise reduction algorithm which considers masking properties of the human auditory system to reduce musical noise artifacts and speech distortions. This algorithm was compared to more simple spectral subtraction rules and was found to be superior with respect to the Itakura-Saito distortion measure, the Articulation Index, machine speech recognition, and subjective preference. Subjective rating results for the *unprocessed* signal, however, are not reported.

Tsoukalas *et al.* (1997a) proposed a noise reduction system which is similar to a noise reduction rule called "Wiener filtering". The Wiener filter minimizes the mean-squared error of best time domain fit to the speech wave-

form. A commonly used implementation of the Wiener filtering rule takes into consideration the power spectra of the noisy signal and an estimated noise power spectrum. It shows the same artifacts as spectral subtraction (i.e., "musical tones"). The algorithm developed by Tsoukalas *et al.* replaces the power spectra in the Wiener filter rule by their corresponding psychoacoustic representations derived from a psychoacoustic model. This system is reported to not suffer from musical noise artifacts. However, it was tested only down to +20 dB SNR as the primary application was supposed to be the restoration of audio recordings. In a further publication, Tsoukalas *et al.* (1997b) proposed another psychoacoustically motivated noise reduction algorithm based on the concept of the audible noise spectrum. This algorithm only modifies selective frequency components detected as containing audible noise, and thus reducing speech distortions. Speech intelligibility was assessed using a sentence test and a rhyme test with 20 subjects. In fact, small improvements in speech intelligibility were reported with this algorithm.

Haulick *et al.* (1997) proposed a post-processing method for spectral subtraction algorithms which is based on auditory masking thresholds to suppress musical noise. Informal listening tests confirmed that musical noise was actually reduced by this method, resulting in an output similar to that of the Ephraim-Malah algorithm.

3.2.4 The "rediscovery" of the Ephraim-Malah algorithms

Several modifications with respect to the Ephraim-Malah algorithm were proposed in the literature of the 1990s. Scalart *et al.* (1996) points out that the Ephraim-Malah algorithms have recently received much attention by many researchers for speech enhancement in the context of mobile hands-free radio communications. Moreover, Valiere *et al.* (1990) used the Ephraim-Malah algorithm to suppress the surface noise of old recordings. Applying a filter bank with bandwidths proportional to the frequency reduced transient distortions. It is reported that the algorithm efficiently reduces the noise without the creation of musical tones. Gülzow *et al.* (1998) proposed to use a wavelet transformation or another nonuniform filterbank for the Ephraim-Malah algorithm. They found out that, according to informal listening tests, this produces a more pleasant sound and a more natural sounding speech. Recently, Soon *et al.* (1998) adopted the Ephraim-Malah scheme to the discrete cosine transform. This is reported to result in stronger noise reduction compared to the original scheme, but the residual noise sounds less uniform.

Kleinschmidt *et al.* (1999) employed the Ephraim-Malah algorithms as preprocessor for automatic speech recognition and reported better recognition rates in noisy speech than without noise reduction.

Scalart and Vieira Filho (1996) suspect that the commonly reported

"good behaviour" of the Ephraim-Malah algorithm is caused by the decision-directed estimation approach for the *a priori* signal-to-noise ratio (cf. Section 3.3) and does not rely on the Gaussian assumptions of the suppression rule. Hence, they propose to include the concept of *a priori* SNR in classical speech enhancement schemes as Wiener filtering and spectral subtraction. Akbari Azirani *et al.* (1996) support this suggestion. They adopted the concepts of uncertainty of signal presence and decision-directed estimation of *a priori* signal-to-noise ratio as proposed by Ephraim and Malah (1984) to Wiener filtering noise reduction. Listening tests were performed with 30 subjects applying car noise at 0 dB SNR. On a 5-point quality scale, the proposed modified Wiener filter got a mean opinion score of 3.35 (which is between "fair" and "good"). The Ephraim-Malah algorithm received a score of 3.32 which is probably not significantly different from the proposed algorithm. This indicates that in fact the estimation approach of the *a priori* SNR is the main determinant in the Ephraim-Malah scheme. Classical magnitude spectral subtraction only received a mean opinion score of 2.87.

3.3 A closer look at the Ephraim-Malah schemes

The literature review revealed that especially the single-microphone noise reduction algorithms proposed by Ephraim and Malah (1983, 1984, 1985) have recently received much attention by different researchers. Due to good sound quality and only limited processing artifacts they appear to be suited for application in digital hearing aids. Hence, some details of these algorithms and a discussion of their behaviour will be provided in this section.

3.3.1 The suppression rule

In 1980, McAulay and Malpass (1980) introduced a concept similar to what was used later by Ephraim and Malah (1984). Both describe single-microphone noise reduction systems in which the gain is determined by two values: An *a posteriori* signal-to-noise ratio (SNR) and an *a priori* SNR. By this, McAulay and Malpass' Soft-Decision Noise Suppression Filter introduced an additional parameter which takes into account the uncertainty of speech presence in the noisy signal observations which offers the possibility to compromise between noise reduction and signal distortion.

From a theoretical point of view, the spectral subtraction algorithm is derived from an optimal variance estimator (in the maximum likelihood sense), and the Wiener filter algorithm is derived from the optimal minimum mean-square error (MMSE) signal spectral estimator. Thus, both are not optimal *spectral amplitude* estimators. Ephraim and Malah (1984) derived a MMSE short-time spectral amplitude estimator which is based on modelling speech and noise spectral components as statistically independent Gaussian random variables.

In a performance evaluation, Ephraim and Malah (1984) found out that the main difference between the enhanced speech signal of McAulay and Malpass' algorithm and their own algorithm lies in the nature of the residual noise. With Ephraim and Malahs' algorithm the residual noise is colorless, while musical residual noise results from the McAulay-Malpass algorithm.

First proposed in Ephraim and Malah (1983), the MMSE short-time spectral amplitude (STSA) estimator was further elaborated in Ephraim and Malah (1984). The following notation is adopted from Cappé (1994). The time and frequency indices p and ω_k are omitted to shorten the notation. The spectral gain $G(p, \omega_k)$ that is applied to each short-time spectrum value $X(p, \omega_k)$ is given by

$$G = \frac{\sqrt{\pi}}{2} \sqrt{\left(\frac{1}{1+R_{\text{post}}}\right)\left(\frac{R_{\text{prio}}}{1+R_{\text{prio}}}\right)} \cdot \mathrm{M}\left[(1+R_{\text{post}})\left(\frac{R_{\text{prio}}}{1+R_{\text{prio}}}\right)\right], \quad (3.1)$$

where M stands for the function

$$\mathrm{M}[\theta] = \exp\left(-\frac{\theta}{2}\right)\left[(1+\theta)\ \mathrm{I}_0\left(\frac{\theta}{2}\right) + \theta\ \mathrm{I}_1\left(\frac{\theta}{2}\right)\right], \quad (3.2)$$

where I_0 and I_1 are the modified Bessel functions of 0th and 1st order, respectively.[4] In the following, the focus shall be on the illustration of the suppression rule.

The gain depends on the two parameters R_{post} and R_{prio} which have to be calculated in each short-time frame p and for each spectral component ω_k. Following McAulay and Malpass (1980) and Ephraim and Malah (1984) these two parameters can be interpreted as the *a posteriori* signal-to-noise ratio R_{post} and the *a priori* SNR R_{prio}.[5]

R_{post} is given by

$$R_{\text{post}}(p, \omega_k) = \frac{|X(p, \omega_k)|^2}{\upsilon(\omega_k)} - 1, \quad (3.3)$$

where $\upsilon(\omega_k)$ denotes the noise power at frequency ω_k. Though the formulation of the noise suppression rule given here, following Cappé (1994), differs from the original formulation by Ephraim and Malah (1984), both are mathematically identical. According to Equation 3.3 the *a posteriori* SNR R_{post} is an estimate of the SNR calculated from the data in the current short-time frame. The *a priori* SNR R_{prio} is defined as follows:

$$R_{\text{prio}}(p, \omega_k) = (1-\alpha)\ \mathrm{P}\left[R_{\text{post}}(p, \omega_k)\right] + \alpha\ \frac{|G(p-1, \omega_k)\ X(p-1, \omega_k)|^2}{\upsilon(\omega_k)}, \quad (3.4)$$

where $\mathrm{P}[x] = x$ if $x \geq 0$, and $\mathrm{P}[x] = 0$ otherwise. This definition of the *a priori* SNR follows the "decision-directed" approach proposed by Ephraim and Malah (1984). It is obvious that the numerator of the second term in

Equation 3.4 denotes the spectral power of the noise-reduced output in the last short-time frame and thus corresponds to an estimate of the SNR in the last frame with index $p-1$. Hence, R_{prio} is an estimate of the SNR that takes into account the current short-time frame with weight $(1-\alpha)$ and the noise reduced previous frame with weight α. Based on simulations, both Ephraim and Malah (1984) and Cappé (1994) set the parameter α to 0.98.

Cappé (1994) showed that the *a priori* SNR R_{prio} is the dominant parameter in the Ephraim-Malah algorithm: Strong attenuation is obtained only if R_{prio} is low, and little attenuation is obtained only if R_{prio} is high. The *a posteriori* SNR acts as a correction parameter whose influence is limited to the case where R_{prio} is low. This correction is somewhat counter-intuitive: The larger R_{post}, the stronger the attenuation (cf. Figure 3.2a). However, this reduces the musical tones artifact: For low values of the *a priori* SNR concurrent with high values of the *a posteriori* SNR, a larger attenuation is assigned. Thus, values of the spectrum higher than the average noise level are "pulled down". By this, the algorithm avoids the appearance of local bursts of musical noise whenever the noise exceeds its average characteristics.

The musical noise phenomenon is further reduced by the following features of the *a priori* SNR:

1. When R_{post} stays below or is sufficiently close to 0 dB, the *a priori* SNR corresponds to a highly smoothed version of R_{post} over successive short-time frames.

2. When R_{post}, on the other hand, is much larger than 0 dB, the *a priori* SNR essentially follows R_{post} with a delay of one frame.

Figure 3.1 illustrates this behaviour with a section of a speech plus noise signal. The smoothness of the *a priori* SNR helps to reduce the musical noise effect. When only noise is present, the *a posteriori* SNR is low (theoretically, it should be $-\infty$ dB in average) which corresponds to Case 1 above. As can be seen in Figure 3.1, the *a priori* SNR has a significantly smaller variance in this case. Since the attenuation of the Ephraim-Malah algorithm depends mainly on the *a priori* SNR, the attenuation does not show large variations over successive frames, thus the musical noise is again reduced. Compared to other noise suppression rules which reduce the musical noise by averaging the short-time spectrum or the calculated gain over successive frames, one advantage of the Ephraim-Malah algorithm lies in the non-linearity of the averaging process. When the signal level is well above the noise level, the *a priori* SNR becomes almost equivalent to the *a posteriori* SNR with one frame delay, thus R_{prio} is no longer a smoothed SNR estimate, which is important to not deteriorate speech which is rather non-stationary.

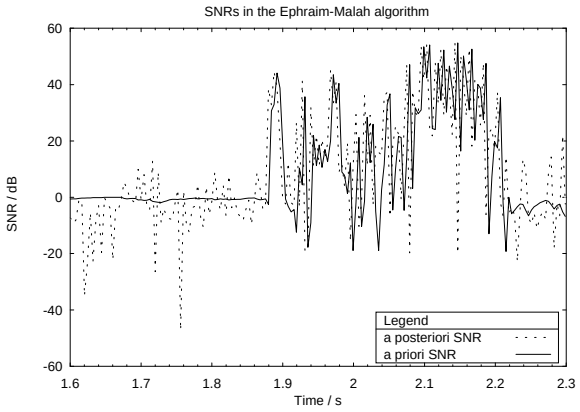


Figure 3.1: *Behaviour of the* a posteriori *SNR (dashed curve) and the* a priori *SNR (solid curve) displayed for one frequency band. Until* $t = 1.88\,\mathrm{s}$ *only noise is active in the input signal. Then a voice starts speaking (mixed with the noise; the overall SNR is 15 dB) until* $t = 2.22\,\mathrm{s}$*, then again only noise is present.*

3.3.2 Modifications of the suppression rule

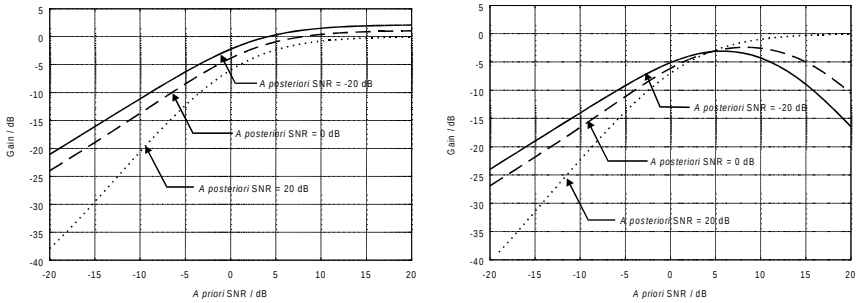
In addition to the above described suppression rule, Ephraim and Malah (1984) derived a modified MMSE amplitude estimator "under uncertainty of signal presence". The motivation for the modification was given by the fact that the target signal (i.e., speech) is absent in the noisy signal quite frequently (speech pauses) and appears with only insignificant energy in some noisy spectral components when the speech is of voiced type. Hence, Ephraim and Malah suggested a model in which a statistically independent random appearance of the signal in the noisy spectral components is assumed. When combining this model with the MMSE amplitude estimator the resulting algorithm has essentially one more parameter. This parameter q_k determines the probability of signal absence in the kth spectral component. Although theoretically this parameter could be set individually for each spectral component, Ephraim and Malah found that a global value of 0.2 gives good results. The spectral gain $G_{\mathrm{USP}}(p, \omega_k)$ of the MMSE amplitude estimator under uncertainty of signal presence (USP) that is applied to each short-time spectrum value $X(p, \omega_k)$ is given by

$$G_{\mathrm{USP}} = \frac{\Lambda}{1+\Lambda} \cdot G \Big|_{R_{\mathrm{prio,USP}}} \quad , \tag{3.5}$$

where $R_{\mathrm{prio,USP}}$ is the modified *a priori* SNR

$$R_{\mathrm{prio,USP}} = (1 - q_k) \cdot R_{\mathrm{prio}} \tag{3.6}$$

and G is the spectral gain of the original Ephraim-Malah algorithm. The time and frequency indices p and ω_k are again omitted for reasons of com-



(a) (b)

Figure 3.2: *Gain values for different* a priori *and* a posteriori *SNR values in the original Ephraim-Malah estimator (left panel), and in the Ephraim-Malah estimator under uncertainty of signal presence with* $q_k = 0.2$ *(right panel).*

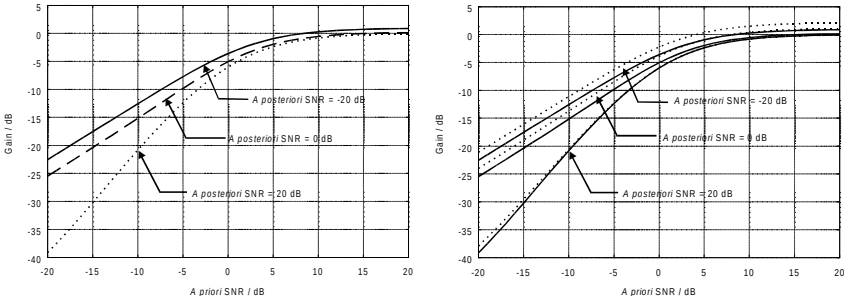
pactness. Λ is given by

$$\Lambda = \frac{(1-q_k)}{q_k} \frac{1}{1+R_{\text{prio,USP}}} \exp\left[\frac{R_{\text{prio,USP}}}{1+R_{\text{prio,USP}}}(1+R_{\text{post}})\right]. \qquad (3.7)$$

The main difference to the original MMSE amplitude estimator is its behaviour for high *a priori* signal-to-noise ratios. Figure 3.2 shows the gain values for different *a priori* and *a posteriori* SNR values of the original Ephraim-Malah algorithm (left panel) and under uncertainty of signal presence ($q_k = 0.2$; right panel). The decrease in gain as the *a posteriori* SNR decreases while the *a priori* SNR is high, is in contrast to the increase in gain in the original Ephraim-Malah algorithm for the same SNR values. This behaviour is caused by favoring the hypothesis of signal absence in such a situation.

Ephraim and Malah (1984) also compared the two amplitude estimators with respect to sound quality. They found that with a slightly increased weighting factor in the *a priori* SNR calculation ($\alpha = 0.99$ instead of 0.98) the estimator under uncertainty of signal presence results in a further reduction of the colorless residual noise, with negligible additional distortions in the enhanced speech signal. In the limit case of $q_k = 0$, the amplitude estimator under uncertainty of signal presence becomes the original Ephraim-Malah amplitude estimator. This can easily be seen as $\Lambda/(1+\Lambda)$ in Equation 3.5 then equals unity.

A further modification of the algorithm minimizes the mean-square error of the *log*-spectra (Ephraim and Malah, 1985). This estimator was motivated by the fact that a distortion measure which is based on the mean-square error (MSE) of the log-spectra was reported to be more suitable and subjectively meaningful for speech processing than the MSE of the spectra themselves (Gray *et al.*, 1980).



(a) (b)

Figure 3.3: *Gain values for different* a priori *and* a posteriori *SNR values in the Ephraim-Malah log-spectral amplitude estimator (left panel). The log-spectral estimator always gives a stronger attenuation (solid curves in the right panel) than the original Ephraim-Malah estimator (dotted curves in the right panel).*

The spectral gain $G_{\log}$ of the MMSE log-spectral amplitude estimator is given by

$$G_{\log} = \frac{R_{\text{prio}}}{1 + R_{\text{prio}}} \exp\left[\frac{1}{2}\int_{\kappa_k}^{\infty} \frac{e^{-t}}{t}\, dt\right] , \tag{3.8}$$

where κ_k is defined by

$$\kappa_k = \frac{R_{\text{prio}}}{1 + R_{\text{prio}}} \left(1 + R_{\text{post}}\right) . \tag{3.9}$$

R_{prio} and R_{post} are defined as in the original MMSE amplitude estimator. The integral in Equation 3.8 is the exponential integral $E_1(x)$.[6]

Figure 3.3 shows the parametric gain curves of the MMSE log-spectral amplitude estimator (left panel). For easier comparison, the right panel shows the gain curves of the log-spectral estimator together with the curves of the original MMSE amplitude estimator. It is interesting to note that the new gain function (Equation 3.8) always gives a lower gain than the original MMSE amplitude estimator and thus provides higher attenuation. Ephraim and Malah (1985) reported that the log-spectral estimator results in much less residual noise compared to the original estimator, while no difference in the speech itself was noticed. The residual noise, however, sounds a little less uniform in the log-spectral estimator. They also found that the enhanced speech obtained by the MMSE amplitude estimator under uncertainty of signal presence and that obtained by the log-spectral estimator sound very similar, except that in the latter one the residual noise sounds a little less uniform.

3.4 Conclusions

Concluding this short review of single-microphone noise reduction algorithms with a special focus on the schemes proposed by Ephraim and Malah (1984, 1985), it may be emphasized that the Ephraim-Malah algorithms seem to be well suited for hearing aid applications. This conclusion is supported by three important facts: First, the musical tones artifact is reduced to a high degree resulting in a better sound quality compared to conventional spectral subtraction. Secondly, the computational complexity is relatively low which facilitates an implementation in the next generation of digital hearing aids. Last but not least, the Ephraim-Malah schemes are not restricted to a previously learnt set of noises in contrast to HMM-based noise reduction algorithms.

The gain function of the algorithm has to be pre-calculated for a range of *a posteriori* and *a priori* signal-to-noise ratios and to be stored in look-up tables to make an implementation of the Ephraim-Malah noise reduction schemes in digital hearing aids possible, since the online calculation of the Bessel functions or the exponential integral is not feasible with the available computing power in today's hearing aids. By this, the complexity of the algorithms reduces to that of conventional spectral subtraction. This is an important advantage of the Ephraim-Malah noise reduction schemes over most other algorithms which were proposed to reduce the musical tones artifact, which often require the calculation of complex (psychoacoustical) models in real time.

An important prerequisite for the use of the Ephraim-Malah noise reduction schemes in hearing aids is their combination with an automatic procedure to update the noise spectrum estimate. This update is proposed to be done during speech pauses which might be detected with appropriate algorithms. One such algorithm with low computational complexity and a relatively low false-alarm rate was proposed in Chapter 2.

Chapter 4

Noise reduction schemes for digital hearing aids: I. Listening effort and speech intelligibility

Subjective methods for the evaluation of benefits from different noise reduction schemes were proposed and tested with six normal-hearing and six hearing-impaired subjects. With respect to speech intelligibility, no improvement was found with the Ephraim-Malah single-microphone noise reduction algorithms (Ephraim and Malah, 1984), compared to the unprocessed signal. However, benefits with respect to reductions in listening effort were found by applying a newly developed listening effort test.

In addition, a binaural directional filter and dereverberation algorithm was considered (Wittkop, 2000). Reverberant binaural recordings obtained with a dummy head were used to test the algorithms with hearing-impaired subjects. In addition, small but significant improvements in speech reception thresholds (SRT) were found with the binaural algorithm. Small benefits regarding "ease of listening" were found in speech-shaped noise with the monaural Ephraim-Malah algorithm as well as with the binaural algorithm.

The results indicate that conventional SRT tests and tests of listening effort appear to measure different aspects of the effect of noise reduction schemes in speech perception. Also, the binaural noise suppression appears to be more effective in enhancing speech intelligibility than the single-microphone algorithms.

4.1 Introduction

Noise evokes major listening difficulties in hearing-impaired subjects, even in persons with low to moderate hearing losses (Weiss and Neuman, 1993).

These difficulties are often experienced as a real handicap especially at the working place and during social activities. They are connected with decreased speech intelligibility and with an often enormously increased effort to understand speech in noise. Noise reduction schemes in digital hearing aids may help to overcome these deficiencies by increasing the signal-to-noise ratio. They aim at reducing speech reception thresholds, i.e. increasing speech intelligibility, lower the listening effort and improve the perceived quality of the acoustic environment.

However, the literature on noise reduction indicates that generally no or hardly any improvements with single-microphone noise reduction are found regarding speech intelligibility if a speech signal is degraded by wideband noise (for a review see the book from Studebaker and Hochberg, 1993, and Chapter 3). On the other hand, in general, listening to the algorithms confirms that the noise *is* actually reduced. A potential benefit of such algorithms is supposed to be an increased "ease of listening" which is assumed to be connected with less listening effort. Indeed, fatigue and increased effort when listening in noise is a common complaint of hearing-impaired subjects. Even normal-hearing subjects have this complaint after a workday's noise exposure. However, speech recognition tests did not reflect this fatigue (Ivarsson and Arlinger, 1993). The fatigue may well be related to non-auditory functions, involving concentration, attention et cetera.[7] In 1950, Wyatt recounts in his autobiography that a great deal of time and effort was devoted, unsuccessfully, as it turned out, to the search for a measure of fatigue. Actually, the problem of assessing fatigue is still not solved satisfactorily.

Downs and Crum (1978) found longer durations in a probe reaction-time task during an auditory learning task when presenting competing speech. They related this increase in reaction times to an increase in learning effort. An effect of the competing speech on learning performance was actually not found. Gatehouse (1994) suggested a "sentence verification test" in which response times are measured to assess listening effort. This test had already been applied by Baer *et al.* (1993) to evaluate a spectral contrast enhancement technique. They found effects for the response times that were twice as large as for intelligibility scores and statistically more robust. Measuring response times in an intelligibility test was proposed already by Hecker *et al.* (1966). He suggested that they provide an independent measure and can increase the sensitivity of conventional tests.

A different approach is used by Hoeks and Levelt (1993). They use pupillary dilation as a measure of the level of mental effort needed in an attentional task. However, to obtain stable (and reliable) results, the measurement of response times as well as the measurement of pupillary dilation need averaging over numerous trials since the trial-to-trial variability is large. Moreover, these physical effects, though more "objective" than a subjective self-assessment, are generally too small in higher signal-to-noise ratios, i.e.

the sensitivity of these tests is rather restricted if the measurement time is fixed. In addition, other factors that cannot be controlled effectively, such as boredom, strongly influence response times (Davies *et al.*, 1983).

Dillon and Lovegrove (1993) (referring back to Lim and Oppenheim, 1979) point out that long-term listening at a reduced fatigue level may lead to a long-term gain in intelligibility if without a noise reduction system the listener tires more quickly. This connects listening effort with speech intelligibility on a larger time scale. If it holds true, listening effort measurements may be used to estimate long-term speech intelligibility.

For these reasons a novel procedure is introduced here which is based on a subjective self-assessment. This procedure is supposed to be sensitive for the effects of noise on speech concerning listening effort at arbitrary signal-to-noise ratios.

In addition, conventional speech reception threshold (SRT) tests were also employed to compare listening effort and speech intelligibility measures with the same subjects and the same set of algorithms and test signals. Two types of noise reduction algorithms were employed: The single-microphone noise reduction algorithms proposed by Ephraim and Malah (1984, 1985; see Chapter 3) and the binaural directional filter and dereverberation algorithm by Wittkop (2000).

Noise reduction is expected to yield benefits even for normal-hearing listeners and not only for hearing-impaired subjects. However, an evaluation using only normal-hearing listeners may be misleading because results with normal-hearing subjects can differ significantly from those with hearing-impaired subjects. Levitt *et al.* (1993), for example, report that normal-hearing subjects showed a significant decrement while hearing-impaired subjects showed a significant improvement in consonant recognition using a Wiener filtering noise reduction algorithm. Hygge *et al.* (1992) found that hearing-impaired and normal-hearing persons differ substantially in how they are affected by background noise when trying to comprehend foreground speech. For this reason, all measurements which will be reported in the next sections were primarily carried out with hearing-impaired subjects. To investigate whether any systematic differences exist between normal-hearing and hearing-impaired listeners with respect to the noise reduction processing, the first experiment was additionally performed with six normal-hearing listeners.

4.2 Algorithms

The literature review in Chapter 3 revealed the high potential of the single-microphone noise reduction schemes proposed by Ephraim and Malah (1984, 1985) for the use in digital hearing aids. These algorithms overcome the annoying "muscial tones" artifact of conventional schemes based on spectral

subtraction while keeping relatively low computational complexity. Moreover, the Ephraim-Malah schemes are not restricted to a previously learnt set of noises as many HMM-based noise reduction algorithms are. Nevertheless, they yield strong reductions of different (stationary) background noises. One advantage of single-microphone noise reduction procedures, compared to multi-microphone methods, is their robustness against the number of noise sources and the level of reverberation (at least as long as the noise is stationary). In addition, the performance of most multi-microphone noise reduction schemes improves with increasing distance between their microphones. Hence, they are hardly applicable in a standard hearing aid, or, alternatively they need a (wireless) bidirectional communication between a left-ear and a right-ear hearing aid, a yet unsolved technical problem.

Since in most single-microphone algorithms an update of the noise spectrum estimate can only be performed during speech pauses, such schemes are only capable of reducing noises which do not change too drastically between two speech pauses. Moreover, only multi-microphone algorithms may provide improvements in speech intelligibility over a wide range of – especially wide-band – noises as is indicated by the literature on noise reduction reviewed in Chapter 3.

The two-microphone directional filter and dereverberation algorithm developed by Peissig (1993) was shown to yield significant improvements in speech intelligibility (Kollmeier *et al.*, 1993). The directional filtering stage of this algorithm attenuates lateral sound sources while passing through sounds from the front. The dereverberation stage reduces diffuse noise and reverberation. This algorithm was further elaborated and improved by Wittkop *et al.* (1999) and Wittkop (2000). The main objective was the preservation of a high signal quality in the processed signal. Because of its potential benefits with respect to speech intelligibility, this binaural noise reduction algorithm was considered in addition to the single-microphone Ephraim-Malah schemes in the present experiments.

Three different single-microphone noise reduction algorithms were employed in the first experiment (cf. Chapter 3): The minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator (Eq. 7 in Ephraim and Malah, 1984; called E7 in the following), a modified estimator under uncertainty of signal presence (Eq. 30 in Ephraim and Malah, 1984; called E30 in the following), and the MMSE log-spectra estimator (Ephraim and Malah, 1985; called EL in the following). The implementations of these algorithms use the decision-directed approach for estimating the a priori signal-to-noise ratio (Eq. 51 in Ephraim and Malah, 1984). An important prerequisite for the application of these algorithms in hearing aids is their combination with an automatic procedure to update the noise spectrum estimate since the acoustic environment is supposed to change over time. Hence, the speech pause detection algorithm proposed in Chapter 2 with low computational complexity and a relatively low false-alarm rate was

employed. The noise spectrum estimate is updated during detected speech pauses.

In a second experiment, the single-microphone algorithm E7 and the two-microphone (binaural) directional filter and dereverberation algorithm by Wittkop *et al.* (1999) and Wittkop (2000) are considered (called DD in the following). The main principle of the directional filter is the comparison of the current interaural level and phase differences in each frequency group (according to the Bark scale; Zwicker and Terhardt, 1980) with mean values measured for particular sound incidence directions in a previous reference measurement. If the current level and phase differences are in the range of the values obtained for a certain angle range of frontal incidence directions, then the signal is also assumed to be emitted from a frontal sound source. In this case, magnitude gain factors of unity are applied. Otherwise, the signal is assumed to be emitted from a lateral sound source, and thus low magnitude gain factors are applied. Due to ambiguities of the level and phase differences, sound sources located in the median plane can not be distinguished by the directional filter. Additionally, two basic acoustical configurations are considered by Wittkop's algorithm for which optimal noise reduction processing strategies are applied. First, a situation with one target sound source and some additional diffuse noise, and secondly a situation with one target sound source and one interfering sound source, which are clearly separated in their spatial location, i.e., mainly in their azimuthal location. In order to classify the acoustical environment, a "degree of diffusiveness" measure is calculated concurrently which controls the different processing stages. This control mechanism was introduced because the directional filtering considerably deteriorates the signal if active in the presence of multiple noise sound sources or diffuse noise.

Moreover, the sequential processing of the binaural algorithm DD and the single-microphone algorithm E7 is considered in Experiment 2 (denoted as DDE7). While the processing by DD is assumed to reduce reverberation and diffuse noise as well as distinct noise sources separated in space, the processing by E7 is assumed to further reduce stationary components of the background noise.

4.3 Subjects

Six normal-hearing (three male and three female students aging from 21 to 29 years) participated in the evaluation of the single-microphone noise reduction algorithms. They had no prior experience in psychoacoustic measurements, no history of hearing problems and pure tone thresholds less than 10 dB HL for at least seven of the nine audiometric frequencies between 125 Hz and 8 kHz. Moreover, six subjects with moderately bilateral sensorineural hearing losses (three males and three females aging from 23 to

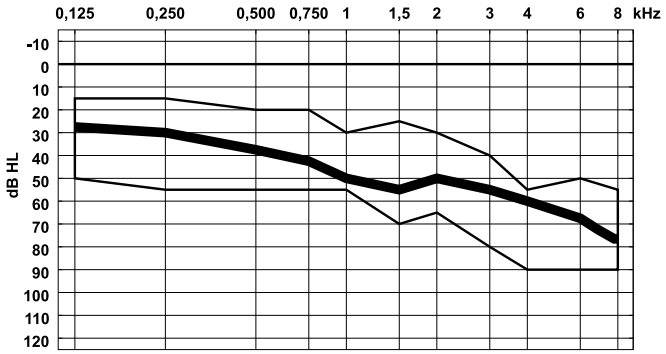


Figure 4.1: *Median audiogram of the hearing-impaired subjects. The solid line indicates the median hearing loss of the left and right ears of the six hearing-impaired subjects. The shaded area shows the total range.*

78 years; cf. Table A.1 in Appendix A) participated in both the evaluation of the single-microphone as well as the binaural noise reduction algorithms. For the hearing-impaired subjects, Figure 4.1 gives a visual impression of the median and the range of the hearing losses. They were experienced listeners from other psychoacoustic measurements.

4.4 Measurement setup

The subjects were seated in a sound-insulated booth. All experiments were performed computer controlled. In the listening effort measurements, the subjects were allowed to switch among algorithms with a handheld touchscreen response box. In the speech intelligibility measurements, the experimenter was also seated in the booth and operated the computer by using the touchscreen response box.

To approximately compensate for the hearing loss of the hearing-impaired subjects at intermediate levels, a third-octave band equalization (range of ± 16 dB in each band) was employed for each ear independently (28-Band Graphic Equalizer TC1128X from T.C.Electronic). The required gain in each frequency band was determined from the audiogram of each subject using the one-half gain rule, i.e. target gain at each frequency is simply the subject's audiometric threshold multiplied by 0.5 (Lybarger, 1944, 1978).[8]

In the listening effort measurements, the unprocessed signal (i.e, without noise reduction) and the signals processed by the three noise reduction algorithms under consideration (cf. Table 4.1) were recorded synchronous to each other on four tracks of an Alesis ADAT 8 Track Professional Digital Audio Recorder. For playback, the output of the ADAT Recorder is

amplified (equalizers together with Sony STR-D315 stereo amplifier) and presented to the subjects via headphones (Sennheiser HD 25).

In the speech intelligibility measurements, the additional amplification was provided by an audiometric amplifier and a Sennheiser HDA 200 headphone was used.[9]

Table 4.1 gives an overview of the noise signals used in the different measurements.

Table 4.1: *Noise signals used in the measurements.*

Noise signal		Listening effort		Speech intelligibility	
		Experiment 1 (UN,E7,EL,E30)	Experiment 2 (UN,E7,DD,DDE7)	Experiment 1 (UN,E7,EL,E30)	Experiment 2 (UN,E7,DD,DDE7)
Monaural recordings					
Drill	Drilling machine	× (−5 dB SNR)		×	
Caf_m	Cafeteria noise	× (0 dB SNR)		×†	
Binaural recordings					
Ssn_{60}	Speech-shaped noise (from 60°)		× (5 dB SNR)		×
Caf_b	Cafeteria noise (diffuse)		× (5 dB SNR)		

† Only UN and the best noise reduction algorithm from the drill noise measurement

In each experiment, a stationary noise as well as a fluctuating noise were employed.[10] In the second experiment binaural recordings were applied. The target speech signals and the speech-shaped noise signal were recorded with a Head Acoustics dummy head in a seminar room with reverberation time $T_{60} = 0.6\,\text{s}$. The speech-shaped noise was presented from 60° incidence direction and 1 m distance to the dummy head. The target speech was always presented from the front. The signal-to-noise ratios were set by measuring long-term RMS values in the right channel and attenuating or amplifying the speech signal accordingly. This channel of the binaurally processed signals was presented diotically to the subjects.

In both experiments, the signals were presented diotically to the subjects via headphones. This ensures that the subjects' own binaural "noise reduction" capabilities (which might be quite different among hearing-impaired subjects) are circumvented. The presentation level was individually adjusted so that perception was "loud but still comfortable" to guarantee that most signal parts were audible for the subject.

4.5 Statistical methods

The individual results as well as median values[11] over all subjects and median absolute deviations (MAD)[12] will be reported. The MAD is defined

as

$$\begin{aligned} \mathrm{MAD}(x_1, \ldots, x_N) &= 1.4826 \cdot \mathrm{median}(|x_1 - \mathrm{median}(x_1, \ldots, x_N)|, \ldots, \\ &\quad |x_N - \mathrm{median}(x_1, \ldots, x_N)|)\,, \end{aligned} \tag{4.1}$$

where $x_1, \ldots, x_N$ is the respective data set. The constant 1.4826 ensures that the MAD approximates the standard deviation σ if the data have a Gaussian bell-shaped distribution.[13]

A Friedman two-way analysis of variance by ranks test[14] is applied to the experimental results to find out whether there are significant differences among the algorithms. In all experiments, a difference is only regarded as significant if the P-value of the Friedman chi-square statistics (χ_r^2 with df degrees of freedom) is below $\alpha = 0.05$, otherwise it is regarded as not significant.[15]

A test for the overall concordance among the subjects' ratings is Kendall's W coefficient of concordance. Kendall's W is a normalization of the Friedman statistic, hence the χ^2 and df values are the same. Table 4.2 gives a classification of the W values.

Table 4.2: *Classification of Kendall's W coefficient of concordance. Although these divisions are clearly arbitrary, they do provide useful "benchmarks" for the discussion of concordance among subjects. This subdivision is adapted from a classification of the Kappa statistic by Landis and Koch (1977).*

W Statistic	Strength of Concordance
0.00–0.17	Poor
0.18–0.33	Slight
0.34–0.50	Fair
0.51–0.67	Moderate
0.68–0.83	Substantial
0.84–1.00	Almost Perfect

If the Friedman test indicates significant differences among algorithms, a closer look at the data is worthwhile. For this, Dunn's post test for multiple comparisons is employed.[16]

4.6 Listening effort

4.6.1 Procedure

Different local radio newscasts that were at least two years old have been re-recorded in a radio studio, spoken by a professional male newscaster. The newscasts were put together to give blocks of 2 1/2 minutes and were mixed afterwards with different noises at different signal-to-noise ratios (see Table 4.1). The noisy newscasts were then processed by the noise reduction algorithms under consideration.

During the measurement session, at first a part of a radio newscast is presented to the subject (seated in a sound-insulated booth) to adjust the overall gain so that the overall loudness impression is at the top border of the comfortable loudness range. Thereafter, the subject is requested to listen to a newscast of 2 1/2 min duration, mixed with background noise. The task is to listen carefully to the news and to repeat afterwards as much news as can be remembered (Step 0). The repetition of the news will be recorded with a dictaphone by the experimenter. However, the content of the subject's report of the newscast is not evaluated. According to Jones (1983), cognition experiments revealed that the overall number of words reported is not reduced in noise, but they are produced in a fashion which is both less coherent and organized. In this test, the repetition procedure only serves to put stress on the subject and to force the subject to really listen carefully (thereby producing probably an additional fatigue effect). For this reason news were chosen that were at least two years old and only locally relevant, so that the subjects didn't know them beforehand. Step 0 is introduced for the subject to exercise the listening task and to realize the annoyance of the background noise.

In the next step (Step 1), the subject has to listen again to the same newscast, but has now the opportunity to switch among the different programs. This is done by the subject with a handheld touchscreen response box showing four "buttons" to switch among the different algorithms. The subject is instructed to try all four programs and to judge them according to the 5-point listening effort scale recommended by the ITU (1996b)[17] (Table 4.3). This scale is listed on a form sheet which the subject is required to fill out after listening. The numerical scores are not visible to the subject. They are only relevant for the evaluation of the responses.

Table 4.3: *Listening effort scale as recommended by the International Telecommunication Union (ITU, 1996b).*

Effort required to understand the news	Score
Complete relaxation possible; no effort required	5
Attention necessary; no appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No meaning understood with any feasible effort	1

In addition, the subject has the opportunity to give comments on the different programs orally or in writing. Step 1 was introduced to let the subject get accustomed to the available algorithms and to try them without pressure.

In the following step (Step 2), the subject is requested to listen to a further and yet unknown newscast and to use a new form for judging the programs (the experimenter removes the first one). The task is to repeat

again as much news as can be remembered. This task is meant to put pressure on the subject in order to really assess listening effort associated with the algorithms and not primarily sound quality. The instructions are slightly different than before: The subject can switch among the four programs at will, but is also allowed to stay at a program, if listening is easier with that certain program. After the repetition of the news, the subject fills out the judgment form again.

Depending on the question under study, the procedure used in Step 2 can further be repeated with a different noise and/or signal-to-noise ratio and/or speaker. In the present studies, a female newscaster and cafeteria noise were chosen as Step 3.

4.6.2 Results

Experiment 1

The normal-hearing subjects' results of the listening effort test (Table 4.4) show a median improvement of one point with algorithm EL in drill noise compared to no noise reduction (UN) in Steps 1 and 2, whereas an improvement with algorithm E7 is only found in Step 1, and no difference between E30 and UN is found. In cafeteria noise, all three noise reduction algorithms were judged worse than UN, in median. The difference between UN and E7, however, is only small (0.5 point).

Table 4.4: *Listening effort test results of Experiment 1 with the normal-hearing subjects. Shown are the scores according to the listening-effort scale for the UN (unprocessed), E7, EL, and E30 algorithms, respectively. Given are also the median values and the median absolute deviations (MAD). A high score corresponds with low listening effort. Noise conditions were drilling machine noise at −5 dB SNR and cafeteria noise at 0 dB SNR.* "Without repetition" *indicates that subjects didn't have to repeat the news after listening.* "With repetition" *means that subjects had to repeat the remembered news.*

Normal-Hearing Subject	Step 1: Drill Without Repetition				Step 2: Drill With Repetition				Step 3: Cafeteria With Repetition			
	UN	E7	EL	E30	UN	E7	EL	E30	UN	E7	EL	E30
AA	2	4	4	4	2	3	4	3	2	4	4	4
FJ	3	3	3	3	3	3	3	3	4	4	2	2
GI	3	4	4	3	4	3	3	2	4	2	2	3
MI	5	5	4	3	4	4	4	3	4	3	2	1
MS	4	4	3	3	2	3	4	3	3	3	2	2
RE	3	4	4	4	3	3	4	3	3	3	2	2
Median	3.0	4.0	4.0	3.0	3.0	3.0	4.0	3.0	3.5	3.0	2.0	2.0
MAD	0.7	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.7	0.7	0.0	0.7

The hearing-impaired subjects seem to benefit more from the noise reduction processing than the normal hearing (Table 4.5). Algorithm EL is

Table 4.5: *Listening effort test results of Experiment 1 with the hearing-impaired subjects.*

Hearing-Impaired Subject	Step 1: Drill Without Repetition				Step 2: Drill With Repetition				Step 3: Cafeteria With Repetition			
	UN	E7	EL	E30	UN	E7	EL	E30	UN	E7	EL	E30
BD	2	4	4	4	2	4	4	3	2	1	1	1
GM	1	4	5	3	1	3	5	2	5	4	1	2
HM	1	3	5	4	1	3	5	4	1	3	2	1
KF	3	4	5	5	3	3	4	5	4	3	2	2
KR	5	5	5	5	3	4	5	3	2	3	3	1
WH	2	4	5	2	3	5	4	3	2	4	2	1
Median	2.0	4.0	5.0	4.0	2.5	3.5	4.5	3.0	2.0	3.0	2.0	1.0
MAD	1.5	0.0	0.0	1.5	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.0

judged three and two points better than the unprocessed condition UN in drill noise in Steps 1 and 2, respectively. Algorithm E7 improves listening effort by two and one points, and E30 by two and a half point in Steps 1 and 2, respectively. In cafeteria noise, the median results indicate better performance with E7 compared to UN (one point improvement), no differences between EL and UN, and worst performance with E30.

A look at Table 4.5 reveals that every hearing-impaired subject reported less listening effort with at least one of the noise reduction algorithms compared to no noise reduction in drill noise. However, the subjects show only little concordance in the judgments of the different Ephraim-Malah algorithms. While in Step 2 subjects GM, HM, and KR experience the least listening effort with algorithm EL, subject KF does so with algorithm E30, subject WH with algorithm E7, and subject BD is indecisive between algorithms E7 and EL.

The concordance between subjects is generally lower in Step 3 (cafeteria noise) than in Steps 1 and 2. Subjects HM, KR, and WH still benefit from the noise reduction, while subjects BD, GM, and KF report less listening effort without noise reduction.

For the normal-hearing subjects' results in Step 1, the Friedman test gives $\chi_r^2 = 4.297$, $df = 3$, $P = 0.231$ and $W = 0.239$. This indicates no significant differences among algorithms and a slight agreement among subjects. For Step 2: $\chi_r^2 = 6.077$, $df = 3$, $P = 0.108$ and $W = 0.338$ (no significant differences, but fair agreement among subjects); the same for Step 3 ($\chi_r^2 = 7.041$, $df = 3$, $P = 0.071$ and $W = 0.391$).

The hearing-impaired subjects' results, however, show significant differences among algorithms. For Step 1, the Friedman test gives $\chi_r^2 = 11.659$, $df = 3$, $P = 0.009$ (moderate concordance among subjects, $W = 0.648$). Dunn's post test reveals that the improvement concerning listening effort with the Ephraim-Malah algorithm EL compared to the unprocessed signal UN is statistically significant. The median judgment using the unprocessed

signal is "considerable effort required", whereas with the EL algorithm the judgment is in median "complete relaxation is possible, no effort required". Other differences were not found to be significant. For Step 2: $\chi_r^2 = 11.089$, $df = 3$, $P = 0.011$ (moderate concordance among subjects, $W = 0.616$). According to Dunn's post test, again the only significant difference in listening effort is between UN and EL. For Step 3: $\chi_r^2 = 7.788$, $df = 3$, approximate $P = 0.051$, exact $P < 0.05$[18] Dunn's post test shows that according to the rank sum differences only the difference between E7 and E30 can be significant, i.e. E30 performs significantly worse in the cafeteria than E7 concerning listening effort. The concordance among subjects is fair ($W = 0.433$).

The statistical analysis can be summarized as follows: No significant differences between the algorithms are found with the normal-hearing subjects. In case of the hearing-impaired subjects, algorithm EL is significantly better than no noise reduction (UN) with respect to listening effort in drill noise, and algorithm E30 is significantly worse than E7 in the cafeteria noise.

Experiment 2

Only the hearing-impaired subjects participated in Experiment 2. From the three different Ephraim-Malah algorithms which were used in Experiment 1, E7 was chosen for Experiment 2, since this algorithm performed better than EL and E30 in the adverse cafeteria noise condition.

Table 4.6: *Listening effort test results of Experiment 2 with the hearing-impaired subjects using algorithms UN (unprocessed), E7, DD, and DDE7. Noise conditions were speech-shaped noise and cafeteria noise at 5 dB SNR.*

Hearing-Impaired Subject	Step 1: Speech-Shaped Without Repetition				Step 2: Speech-Shaped With Repetition				Step 3: Cafeteria With Repetition			
	UN	E7	DD	DDE7	UN	E7	DD	DDE7	UN	E7	DD	DDE7
BD	3	4	4	5	3	4	4	3	3	2	3	2
GM	5	3	3	3	5	2	2	2	3	2	2	2
HM	5	4	3	1	5	3	4	1	4	2	5	3
KF	5	4	3	4	3	4	3	2	4	3	2	2
KR	4	5	3	3	4	5	3	3	4	4	5	3
WH	5	3	2	3	4	4	5	4	4	5	4	4
Median	5.0	4.0	3.0	3.0	4.0	4	3.5	2.5	4.0	2.5	3.5	2.5
MAD	0.0	0.7	0.0	0.7	1.5	0.7	0.7	0.7	0.0	0.7	2.2	0.7

The data suggests that *in median* the noise reduction algorithms rather *increase* listening effort instead of decreasing it (Table 4.6). In the speech-shaped noise, algorithm E7 is judged best from all noise reduction algorithms. In median, it is one point worse than UN in Step 1, but no difference to UN is found in Step 2. In the cafeteria noise (Step 3), algorithm DD is only slightly worse than UN.

A closer look at the data in Table 4.6 reveals that in Step 2, the poor median performance of the noise reduction algorithms is only due to subjects GM and HM who have strong opinions on the algorithm without noise reduction (UN). These two subjects reported that no listening effort is necessary without noise reduction but increased effort was required with the other algorithms. Nevertheless, four from six of the subjects (BD, KF, KR, and WH) still benefit from one or another noise reduction algorithm in the speech-shaped noise (Step 2). Subjects KF and KR experienced the least listening effort with algorithm E7, subject WH with algorithm DD, and subject BD is indecisive between E7 and DD. Two subjects judged the listening effort being similar with DDE7 and UN, but in four subjects listening effort increased with the combined algorithm DDE7 compared to no noise reduction.

In the cafeteria noise (Step 3), two of the six subjects experienced least listening effort with algorithm DD, two subjects with no noise reduction (UN), one subject with E7, and one subject was indecisive between DD and UN. With regard to the noise reduction, three subjects had least listening effort with DD, two with E7, and one subject was indecisive between all three algorithms.

None of the differences between algorithms UN, E7, DD, and DDE7 (Table 4.6) is statistically significant due to a low concordance among the subjects. For Step 1: $\chi_r^2 = 6.288$, $df = 3$, $P = 0.098$, $W = 0.349$. For Step 2: $\chi_r^2 = 6.063$, $df = 3$, $P = 0.109$, $W = 0.337$. Step 3: $\chi_r^2 = 5.750$, $df = 3$, $P = 0.124$, $W = 0.319$.

4.6.3 Discussion

The listening effort test introduced here could be administered successfully to assess the differences among algorithms for a number of normal-hearing and hearing-impaired subjects. The main characteristics of the test are the rating of listening effort (according to ITU recommendation) and the stepwise evaluation procedure with subjects that forces them to really concentrate on the listening effort. Anecdotally, the judgments of subject KR indicate that a repetition task as included in Step 2 is sensible (cf. Table 4.5): In Step 1, subject KR reported that no effort is required with each of the algorithms. In Step 2, however, the same subject differentiated well between the four algorithms. In addition to some accustomization to the test situation, this might reflect the influence of the serious listening task in Step 2, which is missing in Step 1.

Since a formal evaluation of the test procedure is still open, it is not proven yet whether subjects are able to differentiate *reliably* between algorithms in this test. Some subjects reported that they actually experienced differences between algorithms but that these differences did not cause them to assign different categories to the algorithms since the steps between the

verbal categories were too big. Hence, it might be advisable for future studies to use an extended listening effort scale with more categories or some subdivisions in between. Actually, Humes *et al.* (1997) who developed a similar listening effort test as proposed here, required a magnitude-estimate of listening effort on a 0–100 scale from their subjects in a clinical study. They used 10-sentence encyclopedia-style passages in cafeteria noise and babble backgrounds to assess the benefit of different hearing aid processing schemes. However, the test proposed by Humes *et al.* does not include a task that actually requires effort from the subjects.

Instead of using more categories in the rating procedure, paired comparisons could be applied as an alternative. As will be shown in Chapter 5, paired comparisons are very sensitive even to small differences between algorithms. In general, this procedure is superior to category rating because of context and range effects and different application of the scale by the subjects in a category rating procedure (see for example Johnson and Mullally, 1969). Hence, a paired comparison experiment is supposed to give significant results even with a small number of subjects (provided differences between algorithms exist), where category methods fail, or at least need a lot more subjects to give significant results (Bech, 1987). Moreover, scale values on a difference scale level can be derived from the data obtained with paired comparisons (Bradley and Terry, 1952; cf. Chapter 5). However, a major disadvantage of paired comparisons is longer measurement time.

As the proposed listening effort test exposes the subject to an effortful listening task, it is expected that the subject is actually *able* to judge the listening effort afterwards. At least, it is supposed that the danger of giving judgments strongly influenced by other perceptual dimensions as, e.g., "pleasantness of the sound", "perceived artifacts" etc., is reduced compared to an experiment in which the subject is asked to judge the listening effort by just listening (probably with hardly any effort) to a short sound sample of some seconds duration. It is believed to be important to let the subject really experience a situation which definitely causes or affects the phenomenon to be judged, i.e. listening effort. Actually, this is the main difference between the proposed test and most other tests proposed in the literature so far (e.g., Humes *et al.*, 1997). However, strictly speaking, it is not proved nor guaranteed that the proposed test actually assesses "listening effort".[19]

A significant improvement for hearing-impaired subjects concerning listening effort was found with the Ephraim-Malah algorithm EL compared to no noise reduction when used in drilling machine noise. In the fluctuating cafeteria noise, no significant effects were found. Generally, the differences among algorithms were more pronounced for the hearing-impaired subjects than for the normal-hearing subjects.

Moreover, the results indicate that the Ephraim-Malah noise reduction algorithms produce more artifacts when applied in the fluctuating cafeteria

noise than in the stationary drilling noise. These artifacts are obviously more prominent with algorithms EL and E30 than with E7 and seem to strongly counteract possible reductions in listening effort which were expected due to signal-to-noise ratio improvements. In Experiment 2, algorithm DDE7 (the combination of E7 and DD) is characterized rather by increased artifacts than by fruitful synergy effects when compared to the results of E7 and DD in isolation.

Since improvements in terms of listening effort were found with algorithm E7 in Experiment 1 but not in Experiment 2, the worse performance of E7 in the latter experiment can probably attributed to the different noise conditions. The noise in Steps 1 and 2 was changed from drilling noise in Experiment 1 to a speech-shaped noise in Experiment 2, which more effectively masks the target speech. Moreover, the signals were deteriorated by reverberation in the second experiment.

As in Experiment 1, it is again observed in Experiment 2 that in Step 1 some subjects (KF and WH) experienced no listening effort without noise reduction and didn't benefit from the noise reduction, but did so in Step 2, where actually more concentration was required to fulfill the listening task. This is further evidence for the need of a really strenuous listening task to be able to judge effort. However, subjects GM and HM did *not* revise their positive judgments of algorithm UN from Step 1 to Step 2.

To conclude, even with the coarse listening effort scale as recommended by the ITU (1996b) and with only six hearing-impaired subjects, a significant difference in listening effort between the single-microphone noise reduction algorithm EL and the algorithm without noise reduction was found in Experiment 1. Therefore, the proposed listening effort test is supposed to be an adequate tool to assess the benefits of noise reduction techniques with regard to listening effort when carefully listening to a target speaker in background noise over a longer period of time. However, this proposed subjective measurement method has not yet been formally evaluated. Even if a formal proof of the validity of the test (i.e., that it measures what it is supposed to) can probably not be provided in principal, future research has to address the reliability of the proposed listening effort test (i.e., determining if and how precise subjects can reproduce their judgments in re-tests). In future experiments, it might also be worthwhile to investigate the potential correlation between the subjective preference judgments concerning listening effort with the durations a subject chose to listen to each algorithm in the course of the experiment.

4.7 Speech intelligibility

Measuring speech intelligibility is indispensable for any comprehensive hearing aid study. Speech intelligibility measurements are often regarded as "ob-

jective". Strictly speaking, however, word recognition tests are *not* objective, since objective measurements are those that do not require a response from the patient (Fabry and Schum, 1994). Nevertheless, Fabry and Schum argue that a word recognition test can be regarded as an almost objective technique, since the responses provided by the listener are generally automatic, requiring a minimal amount of cognitive evaluation, i.e. the listener is not called on to provide a studied evaluation of the auditory signal, although this might be less true for a sentence test than for a single-word test.

4.7.1 Procedure

In Experiment 1, the Göttingen sentence test (Kollmeier and Wesselkamp, 1997) was applied to determine the speech reception thresholds (SRT) for the different algorithms, i.e. the signal-to-noise ratio at which 50 % of the speech is correctly understood. The sentences of the Göttingen sentence test, which were spoken by an unschooled male speaker, are combined to lists of 10 sentences each. In Experiment 1, one test list served for determining the speech intelligibility at one fixed signal-to-noise ratio (SNR). For a reliable estimation of the psychometric function (speech intelligibility versus SNR) three points of this function were determined individually for each condition by measuring at three different SNRs. All test sentences were mixed with drilling machine noise and cafeteria noise, respectively, at SNRs from -26 to $+16$ dB and were processed offline with the noise reduction algorithms under consideration. The overall gain was adjusted in a preliminary test run for each subject so that the overall loudness impression was at the top border of the comfortable loudness range. The subjects were seated in a sound-insulated booth together with the experimenter. The task of the subject was to repeat what was heard. The correct sentence is displayed to the experimenter on a handheld touchscreen response box. The experimenter marks the words that were not heard or wrongly repeated by the subject. Then the next sentence is presented. At least three points (three different SNRs) were measured to determine the psychometric function for each subject for each algorithm. Table 4.1 shows the conditions that were tested. Both the order of the test lists and the order of the algorithms were chosen at random. A logistic function was fitted to the measured data using a maximum likelihood method. In this way, it was possible to determine the speech reception threshold (SRT) and the slope of the psychometric function.

Recently, another German sentence test, the Oldenburg sentence test, was developed (Wagener *et al.*, 1998; Wagener *et al.*, 1999). Because of the limited set of sentences in the Göttingen sentence test which restricts the possible number of conditions that can be measured without learning effect, the Oldenburg sentence test was applied in Experiment 2, which employs syntactically correct but semantically nonsense sentences. Sentences (and test lists, respectively) can be used repeatedly during the measurement be-

cause of their nonsense character. Test lists consisting of 20 sentences were employed.

4.7.2 Results

Experiment 1

Tables 4.7 to 4.10 show the speech reception thresholds and the slopes of the psychometric functions for the normal-hearing subjects and the hearing-impaired subjects obtained in drill noise and in cafeteria noise with the different algorithms. Given are also median values over all subjects with respective median absolute deviations (MAD).

Table 4.7: *Sentence test results of Experiment 1 with the normal-hearing subjects in drill noise. Shown are the speech reception thresholds (SRT) and the slopes s of the fitted psychometric functions for the UN (unprocessed), E7, EL, and E30 algorithms in drilling machine noise. Given are also the median values and the median absolute deviations (MAD). A lower SRT corresponds to better speech intelligibility.*

Normal-Hearing Subject	UN		E7		EL		E30	
	SRT dB	s dB^{-1}	SRT dB	s dB^{-1}	SRT dB	s dB^{-1}	SRT dB	s dB^{-1}
AA	−21.5	0.06	−17.1	0.11	−16.7	0.10	−14.8	0.07
FJ	−22.3	0.13	−21.1	0.09	−15.2	0.06	−17.7	0.04
GI	−19.8	0.06	−21.0	0.10	−16.7	0.12	−20.2	0.04
MI	−21.6	0.07	−20.1	0.12	−19.9	0.04	−18.7	0.14
MS	−19.3	0.07	−18.3	0.10	−22.5	0.10	−19.3	0.10
RE	−21.5	0.04	−19.9	0.15	−18.2	0.08	−16.0	0.11
Median	−21.5	0.07	−20.0	0.11	−17.5	0.09	−18.2	0.09
MAD	0.7	0.01	1.6	0.01	2.2	0.03	2.3	0.05

In median the noise reduction processing seems to decrease speech intelligibility, rather than to increase it. From the three different Ephraim-Malah algorithms, E7 in median yields the best intelligibility scores for the normal-hearing and for the hearing-impaired subjects.

Normal-hearing subject GI actually performs better with E7 than without noise reduction. Subject MS's SRT is 3.5 dB better with EL than without noise reduction (Table 4.7). The other four normal-hearing subjects, however, obtain better speech intelligibility without noise reduction.

Half of the normal-hearing subjects perform better with the noise reduction in the cafeteria noise, the other half performs better without. Again, subject MS benefits most from the noise reduction with a 1.5 dB better SRT than with UN (Table 4.8).

Surprisingly, the normal-hearing subjects benefit more from the noise reduction than the hearing-impaired subjects in terms of speech intelligibility.

Table 4.8: *Sentence test results of Experiment 1 with the normal-hearing subjects in cafeteria noise.* E *denotes the Ephraim-Malah algorithm (E7, EL, or E30) with which each subject performed best in the drill noise. A lower SRT corresponds to better speech intelligibility.*

Normal-Hearing Subject	UN		E	
	SRT dB	s dB^{-1}	SRT dB	s dB^{-1}
AA	−3.9	0.16	−2.2	0.11
FJ	−2.6	0.1	−3.4	0.15
GI	−4	0.14	−3.8	0.09
MI	−3.9	0.12	−4.1	0.23
MS	−2.7	0.15	−4.2	0.11
RE	−4.1	0.12	−2.7	0.22
Median	−3.9	0.13	−3.6	0.13
MAD	0.2	0.02	0.8	0.04

Table 4.9: *Sentence test results of Experiment 1 with the hearing-impaired subjects in drill noise. A lower SRT corresponds to better speech intelligibility.*

Hearing-Impaired Subject	UN		E7		EL		E30	
	SRT dB	s dB^{-1}	SRT dB	s dB^{-1}	SRT dB	s dB^{-1}	SRT dB	s dB^{-1}
BD	−10.3	0.03	−10.2	0.03	−7.8	0.05	−7.9	0.10
GM	−18.0	0.10	−19.2	0.14	−17.8	0.17	−14.3	0.14
HM	−21.5	0.08	−18.2	0.17	−18.0	0.10	−16.1	0.07
KF	−23.0	0.10	−20.1	0.05	−21.7	0.07	−20.1	0.05
KR	−19.3	0.07	−19.1	0.08	−16.5	0.08	−16.0	0.11
WH	−18.6	0.05	−18.4	0.09	−13.7	0.07	−14.1	0.06
Median	−19.0	0.08	−18.8	0.09	−17.2	0.08	−15.2	0.09
MAD	2.6	0.04	0.7	0.07	3.2	0.02	1.5	0.04

Only the hearing-impaired subject GM obtains a better speech intelligibility in drill noise with noise reduction processing (E7) than with UN (Table 4.9). The hearing-impaired subjects perform worse with algorithm EL and worst with E30.

In cafeteria noise, the speech reception thresholds of the hearing-impaired subjects are in median 1.8 dB worse using the noise reduction algorithm than without noise reduction processing (Table 4.10). Only subject WH shows no difference between the Ephraim-Malah algorithm and UN.

A statistical analysis of the data reveals that only few differences are actually statistically significant due to the low concordance among subjects.

For the normal-hearing subjects in drill noise, the Friedman test gives $\chi_r^2 = 6.458$, $df = 3$, $P = 0.091$, $W = 0.359$ (no significant differences between the four algorithms; fair concordance between subjects).

Only two algorithms were tested with the cafeteria noise. Hence, Wil-

Table 4.10: *Sentence test results of Experiment 1 with the hearing-impaired subjects in cafeteria noise.* E *denotes the Ephraim-Malah algorithm (E7, EL, or E30) with which each subject performed best in the drill noise. A lower SRT corresponds to better speech intelligibility.*

Hearing-Impaired Subject	UN		E	
	SRT dB	s dB^{-1}	SRT dB	s dB^{-1}
BD	2.0	0.04	3.4	0.13
GM	−2.2	0.15	−0.1	0.11
HM	−4.6	0.12	−2.4	0.24
KF	−4.0	0.19	−2.3	0.09
KR	−3.7	0.20	−2.1	0.16
WH	−0.3	0.17	−0.3	0.19
Median	−3.0	0.16	−1.2	0.15
MAD	2.0	0.05	1.6	0.06

coxon's matched pairs signed rank test was applied to test for significant differences. It yields $Z = -0.210$, $P = 0.833$ for the normal-hearing subjects, which is far from any statistical significance.

For the hearing-impaired subjects in drill noise, the Friedman test gives $\chi_r^2 = 13.068$, $df = 3$, $P = 0.004$, which is significant. According to Dunn's post test, algorithms EL and E30 are significantly worse than UN. All other differences are not significant. Overall, there is a substantial concordance between subjects ($W = 0.726$). In the cafeteria noise, the Wilcoxon test yields $T = 0$, $Z = -2.023$ with an asymptotic $P = 0.043$, which is significant.[20] The hearing-impaired subjects thus obtain significantly worse SRTs with the Ephraim-Malah algorithm than without noise reduction in the cafeteria noise.

Experiment 2

The results of the speech intelligibility measurements in Experiment 2 are given in Table 4.11. In this experiment, a speech-shaped noise is employed.

Two of the six hearing-impaired subjects obtained better SRTs with algorithm E7 than with UN, three performed worse with E7, and one reaches almost the same SRTs with E7 and UN.

All subjects obtained better SRTs with the directional filter and dereverberation algorithm DD than with UN. The median improvement is 1.1 dB, the maximum improvement is found with subject WH (2.6 dB). The difference might seem small, but in relation to the very steep slope of the psychometric function it corresponds to a difference in number of recognized test words of the order of 13 %, for subject WH even of about 26 %.

Only subject GM performed better with algorithm E7 than with DD, but her SRT differences are relatively small.

Table 4.11: *Sentence test results of Experiment 2 with the hearing-impaired subjects in speech-shaped noise using algorithms UN, E7, DD, and DDE7. A lower SRT corresponds to better speech intelligibility.*

Hearing-Impaired Subject	UN		E7		DD		DDE7	
	SRT dB	s dB^{-1}	SRT dB	s dB^{-1}	SRT dB	s dB^{-1}	SRT dB	s dB^{-1}
BD	1.40	0.07	0.30	0.10	0.00	0.11	0.60	0.08
GM	−1.00	0.08	−1.90	0.12	−1.40	0.16	1.00	0.08
HM	−4.70	0.12	−3.40	0.13	−5.30	0.12	−4.30	0.09
KF	−3.00	0.14	−2.80	0.13	−3.90	0.11	−3.70	0.10
KR	−2.50	0.12	−0.70	0.16	−3.60	0.15	−0.90	0.09
WH	0.50	0.08	−0.70	0.09	−2.10	0.13	−0.80	0.07
Median	−1.75	0.10	−1.30	0.13	−2.85	0.13	−0.85	0.09
MAD	2.59	0.03	1.56	0.02	1.85	0.02	2.45	0.01

Four of the six subjects performed slightly better with the combined algorithm DDE7 than with E7. However, all subjects obtained better SRTs with algorithm DD than with DDE7. Hence, with respect to speech intelligibility, a combination of algorithms DD and E7 is not advantageous, compared to using DD alone.

The Friedman test gives $\chi_r^2 = 8.600$, $df = 3$ with an asymptotic $P = 0.035$ and an exact $P = 0.029$. The differences between UN and DD, and between E7 and DD are found to be statistically significant.[21] The differences between UN and algorithm E7, as well as between UN and DDE7 are statistically insignificant.

To summarize, with the binaural algorithm DD a small but significant improvement of speech reception thresholds compared to no noise reduction (UN) is obtained.

4.7.3 Discussion

Although the differences between the single-microphone noise reduction algorithms are not found to be significant, the noise reduction processing seems to decrease speech intelligibility, rather than to increase it. From the three different Ephraim-Malah algorithms, E7 in median yields the best intelligibility scores for the normal-hearing and for the hearing-impaired subjects. It should also be noted that in median this algorithm does not obtain worse speech reception thresholds than no noise reduction (UN) with normal-hearing and hearing-impaired listeners in drill noise as well as in the reverberant speech-shaped noise condition. This indicates that the processing artifacts are limited with the Ephraim-Malah algorithm, even though many of the enhancement systems known from the literature actually reduce intelligibility (Lim and Oppenheim, 1979).

In cafeteria noise, however, the speech intelligibility for the hearing-

impaired subjects is decreased by the Ephraim-Malah noise reduction processing. Probably, this can be attributed to additional speech distortions which are introduced by the processing due to the fluctuating character of the cafeteria babble noise. The assumption of the stationarity of the noise between speech pauses is strongly violated here.

The SRTs obtained in drill noise in Experiment 1 are generally quite low compared to those that are obtained with speech-shaped noise in Experiment 2. This can be attributed to the fact that the drill noise has significant frequency components even above the typical speech range, and that the calculation of the signal-to-noise ratio covers the whole frequency range (0–11 kHz).

Clear improvements in terms of speech intelligibility are found with the binaural noise reduction algorithm DD in Experiment 2. These would probably not have been found if the signals had been presented dichotically (which, however, is a more realistic condition) instead of diotically. But through the diotic presentation the subject's own binaural processing capabilities are by-passed and the potential of the noise reduction processing itself is tested. This is motivated by the fact that the binaural processing capabilities vary considerably among hearing-impaired subjects and that the respective loss can not be predicted by their audiograms or other psychoacoustical parameters (Kinkel *et al.*, 1991, 1992; Holube, 1993). Hence, the binaural system of some hearing-impaired subjects will be more effective than the directional filter and dereverberation algorithm DD. Although the amount of noise reduction can be increased, early experiments and field tests with the algorithm have shown that subjects prefer less noise reduction in favor of better overall sound quality, i.e. less artifacts (Wittkop, 2000).

Finally, it should be considered that the failure of showing benefits with respect to speech intelligibility using the single-microphone noise reduction algorithms could be due to missing acclimatization to the algorithms. Gatehouse (1992) found that benefits from providing a particular frequency shaping to hearing-impaired subjects did not emerge immediately, but over a time course of at least 6–12 weeks. He concludes that the existence of perceptual acclimatization effects call into question short-term methods of hearing aid evaluation. Punch and Parker (1981) point out that already "Carhart (1946) recommended that the prospective hearing aid user be allowed to spend a substantial amount of time in individual and group listening activities prior to the recommendation of a specific instrument."

Since the whole potential of the noise reduction processing (especially concerning long-term speech intelligibility) cannot be assessed by a short-term laboratory evaluation, an implementation in a wearable digital hearing aid device is advisable for carrying out an evaluation in the field. The discussed results from Gatehouse (1992) strongly support this demand. It is assumed that the relatively low complexity of the algorithms allows an

application in digital hearing aids in the near future. For that, a combination with a dynamic compression algorithm should be considered.

4.8 Conclusions

Due to its design, which involves a strenuous listening task, the listening effort test proposed here is believed to actually assess listening effort and not merely subjective preference in terms of better sound quality. Therefore, the proposed test is recommended for evaluations of noise reduction algorithms in general.

Although a significant amount of noise reduction without the disturbing "musical tones" artifact is obtained with the Ephraim-Malah algorithms for various noise conditions, an increase in speech intelligibility was not found. However, this is in line with the results of most publications on single-microphone noise reduction schemes. At least, the Ephraim-Malah algorithm E7 did not worsen the speech intelligibility, which is not self-evident as the speech reception thresholds are quite low, i.e. in a region where single-microphone noise reduction techniques show significant processing artifacts. But this stresses the importance of assessing other subjective criteria which characterize the effectiveness of noise reduction algorithms. Listening effort can definitely be regarded as one of these and was assessed by the test introduced here. Indeed, significant benefits with respect to listening effort were found for algorithm EL compared to UN (no noise reduction), although this algorithm obtained significantly worse speech reception thresholds than UN.

Chapter 5

Noise reduction schemes for digital hearing aids: II. Subjective quality assessment based on paired comparisons

The Ephraim-Malah single-microphone noise reduction algorithms (Ephraim and Malah, 1984, 1985), a binaural directional filter and dereverberation algorithm (Wittkop, 2000), and their combination were evaluated with six normal-hearing and six hearing-impaired subjects. Paired comparisons were applied with respect to different subjective criteria (overall preference, naturalness of the speech, strength of noise reduction, and speech intelligibility). The Bradley-Terry model (Bradley and Terry, 1952) was used to analyze the data.

Noise reduction was generally preferred over no processing in machinery noise. While the monaural Ephraim-Malah algorithm was preferred in higher signal-to-noise ratios (SNR) in cafeteria noise as well as in speech-shaped noise, the binaural directional filter and dereverberation algorithm was preferred at lower SNR. However, the combined algorithm was not able to merge both benefits but was found at an intermediate position, indicating that a more sophisticated combination or alternatively an intelligent switching algorithm is needed.

The Bradley-Terry scale values from the paired comparison judgments concerning speech intelligibility show perfect concordance with the results of the "objective" sentence test measurements reported in Chapter 4, indicating that subjects are well able to judge *speech intelligibility by paired comparisons.*

5.1 Introduction

Several studies showed that subjects were well able to differentiate between hearing aids using subjective judgments of intelligibility or quality when no differences were found with "objective" speech intelligibility tests (Cox and McDaniel, 1984; Studebaker *et al.*, 1982; Tecca and Goldstein, 1984; cf. the overview in Kuk *et al.*, 1990 and Kuk, 1994). Moreover, it was shown that subjective judgments are as reliable or even more reliable than word-recognition tests (Punch and Parker, 1981; Studebaker *et al.*, 1982; Tecca and Goldstein, 1984). This observation raised the possibility that subjects who showed no improvement in word-recognition scores may nevertheless benefit subjectively from such aids. Kuk *et al.* (1990) conclude that a noise reduction hearing aid should be considered effective if a patient reports improvement in speech recognition *or* subjective judgment, or both. Moreover, Kuk *et al.* stress the need for the development of reliable measurement tools to reflect the benefits of noise reduction circuits.

According to Kuk (1994), Zerlin (1962) was the first who proposed a manageable method to use paired comparisons for the assessment of hearing aids.[22] Today, digital hearing aids allow to store separate programs, i.e. different processing strategies, and thus facilitate the use of paired comparisons as a clinical tool.

Kuk (1994) even concludes that the reliability of paired comparison is as good as, if not better than, speech recognition. Kuk and Tyler (1990) found that hearing-impaired subjects could well differentiate among various subjective criteria. Using several criteria is regarded as being useful for the evaluation of non-linear and other types of signal processing hearing aids. In the field of speech coding, naturalness and noisiness were reported to be two attributes that determine the two main factors (dimensions) when judging sentences processed by different speech coders (Halka and Heute, 1992).

Another aspect was raised by Studebaker (1982). He suggested that judgment of small differences between stimuli is easier when performed in direct comparison than in isolated judgments. In a paired comparison experiment, the subject does not judge one stimulus on a given scale, but judges which of two stimuli exhibits more of the property under question; a task which is much simpler for the subject. Lukas (1991) also argues in favor of paired comparisons because the alternative, category judgments, are highly susceptible to context and range effects, anchor stimuli, reference objects, different application of the scale by the subjects, etc. (cf. Johnson and Mullaly, 1969). In addition, Lukas argues that category rating requires a stable idea of the meaning of the categories. Such a reference system, however, can at best be expected from a trained and experienced subject. These problems cannot be solved by using magnitude estimation instead, where subjects are asked to give their subjective judgments directly as "numbers" without using any verbal categories. Lukas warns that the danger arises of confounding

the judgments of the subject (which are only the empirical basis for the scaling) with the result of the scaling (which is a theory-based estimation of numerical values). Unfortunately, the numbers assigned by the subjects are often taken directly as numerical values, but without justification as to which numerical operations are really empirically meaningful.

For the analysis of paired comparison data a well-founded theory exists. One frequently used analysis model is the Bradley-Terry (BT) model.[23] Proposed by Bradley and Terry (1952) the model is strongly related to the choice axiom of Luce (1959) and hence also called Bradley-Terry-Luce (BTL) model by some authors (e.g., Colonius, 1980; Koehler and Ridpath, 1982; Tutz, 1986; Lukas, 1991). The model was developed independently by Zermelo (1929), Bradley and Terry (1952), and Ford (1957), however, based on different practical applications, namely the evaluation of chess players, taste testing experiments, and league competitions, respectively. Overviews over the method of paired comparisons are given by Bradley (1976) and David (1988). A comprehensive bibliography is given in Davidson and Farquhar (1976).[24]

After having empirically tested the theoretical premises by checking the goodness of fit, the BT model provides a scaling of the paired comparison data based on a ratio scale or a difference scale, respectively. Hence, distances between algorithms are represented meaningfully and it is easy to graphically describe the relative positions of the ratings of each algorithm.

To conclude, the method of paired comparisons and the application of the Bradley-Terry model for scaling the data can be regarded as an adequate tool for the assessment of different subjective aspects of the processing quality of hearing aid algorithms.

In the following experiments the method of paired comparisons is used to evaluate different classes of noise reduction algorithms with respect to different subjective criteria (overall preference, naturalness of the speech, strength of noise reduction, and speech intelligibility).

5.2 Algorithms

Three different single-microphone noise reduction algorithms were employed in the first experiment (cf. Chapter 3): The minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator (Eq. 7 in Ephraim and Malah, 1984; denoted as E7 in the following), a modified estimator under uncertainty of signal presence (Eq. 30 in Ephraim and Malah, 1984; denoted as E30 in the following), and the MMSE log-spectra estimator (Ephraim and Malah, 1985; denoted as EL in the following). The implementations of these algorithms use the decision-directed approach for estimating the a priori signal-to-noise ratio (Eq. 51 in Ephraim and Malah, 1984). These algorithms were combined with the speech pause detection algorithm pro-

posed in Chapter 2. This allows for an update of the noise spectrum estimate during detected speech pauses.

In addition, the two-microphone (binaural) directional filter and dereverberation algorithm by Wittkop *et al.* (1999) and Wittkop (2000) is considered in a second experiment (denoted as DD in the following; cf. Chapter 4). Furthermore, the sequential processing of the binaural algorithm DD and the single-microphone algorithm E7 is considered (denoted as DDE7 in the following). While the processing by DD is assumed to reduce reverberation and diffuse noise as well as distinct noise sources separated in space, the post-processing by E7 is assumed to further reduce stationary noise parts.

5.3 Subjects

Six normal-hearing (three male and three female students aging from 21 to 29 years) participated in the evaluation of the single-microphone noise reduction algorithms. They had no prior experience in psychoacoustic measurements, no history of hearing problems and pure tone thresholds less than 10 dB HL for at least seven of the nine audiometric frequencies between 125 Hz and 8 kHz. Moreover, six subjects with moderately bilateral sensorineural hearing losses (three males and three females aging from 23 to 78 years; cf. Table A.1 in Appendix A) participated in both the evaluation of the single-microphone as well as the binaural noise reduction algorithms. They were experienced listeners from other psychoacoustic measurements. The median and the range of their hearing losses were given in Figure 4.1 of Chapter 4.

5.4 Measurement setup

The subjects were seated in a sound-insulated booth. The experiments were performed computer controlled. The subjects were allowed to switch between two algorithms with a handheld touchscreen response box. The experimenter was also seated in the booth and noted the preferences of the subjects.

To approximately compensate for the hearing loss of the hearing-impaired subjects at intermediate levels, a third-octave band equalization (range of ± 16 dB in each band) was employed for each ear independently (28-Band Graphic Equalizer TC1128X from T.C.Electronic). The required gain in each frequency band was determined from the audiogram of each subject using the one-half gain rule, i.e. target gain at each frequency is simply the subject's audiometric threshold multiplied by 0.5 (Lybarger, 1944, 1978).[25]

From the equalizers, the signals were fed into a Sony STR-D315 stereo amplifier and finally via a Sennheiser HD 25 headphone to the subject.

For the second experiment binaural recordings were employed (cf. Table 5.1). The target speech signals as well as the noise signals were recorded with a Head Acoustics dummy head. The industrial noise and the speech-shaped noise were presented from 60° incidence direction and 1 m distance to the dummy head in a seminar room with reverberation time $T_{60} = 0.6\,\mathrm{s}$. Target speech was always presented from the front. The cafeteria noise was recorded with the dummy head in a crowded cafeteria. The signal-to-noise ratios were set by measuring long-term RMS values in the right channel and attenuating or amplifying the speech signal accordingly. This channel of the binaurally processed signals was presented diotically to the subjects.

In both experiments, the diotic signal presentation ensures that the subjects' own binaural "noise reduction" capabilities (which might be quite different among hearing-impaired subjects) are circumvented. The presentation level was individually adjusted so that perception was "loud but still comfortable" to guarantee that most signal parts were audible for the subject.

5.5 Procedure

A complete paired comparison experiment (Round Robin tournament) of all algorithms (including no noise reduction) was performed with regard to different criteria. In Experiment 1, the algorithms UN (unprocessed), E7, EL, and E30 were compared to each other. The following questions were asked one after another:

1. Which of the two programs do you like more?
2. With which of the two programs does the speech sound more natural?
3. Which of the two programs shows less background noise?

In Experiment 2, the algorithms UN, E7, DD, and DDE7 were compared. Here, the subjects were additionally asked to compare the algorithms with respect to speech intelligibility.

Although the subjects were not explicitly asked for the perceived *sound quality*, the above criteria are of course strongly connected with the multi-dimensional phenomenon "quality". The first criterion (overall preference) can be interpreted as an indicator of the acceptance of the respective algorithm. Furthermore, naturalness and noisiness were reported to be two attributes that determine the two main factors when judging speech quality (Halka and Heute, 1992).

Table 5.1 gives an overview of the noise signals employed in the experiments. Stationary noise signals (drilling machine noise, speech-shaped noise) as well as fluctuating noise signals (cafeteria babble, industrial noise) at different signal-to-noise ratios were used. In each condition, four algorithms

were compared to each other. The six comparisons for each noise condition were performed as one measurement block. The comparisons were randomly arranged in each of these blocks. The target speech signal was always the same short sentence (in German language) of approximately 4 s duration, chosen from the news recordings that were produced for the listening effort test (cf. Chapter 4). A male voice was chosen because Punch (1978) reported that quality judgments of hearing-impaired subjects were more reliable with a male than with a female voice.

Table 5.1: *Noise signals used for the paired comparisons.*

Noise signal		Experiment 1 (UN,E7,EL,E30)	Experiment 2 (UN,E7,DD,DDE7)
Monaural recordings			
D–5	Drilling machine	×	
D+5	(−5 and +5 dB SNR)	×	
C–5	Cafeteria noise	×	
C+5	(−5 and +5 dB SNR)	×	
Binaural recordings			
Ind0	Industrial noise		×
Ind10	(0 and 10 dB SNR)		×
Caf0	Cafeteria noise		×
Caf10	(0 and 10 dB SNR)		×
Ssn0	Speech-shaped noise		×
Ssn10	(0 and 10 dB SNR)		×

The subject was allowed to switch between the two programs under consideration at will by use of a handheld touchscreen response box, and listen to the programs as often as desired. The judgments of the subject were given orally and were written down by the experimenter. Both were seated in a sound-treated booth.

The subjects had to state a preference for one of the two presented algorithms depending on the respective criterion. In Experiment 1, ties (i.e., equal judgments) between algorithms were permitted. For the BT analysis, however, the ties were equally distributed to the tied algorithms. In Experiment 2, ties were not permitted, since allowing for ties in Experiment 1 encouraged "lazy" judgments and, as David (1988) remarks, introduces difficulties, since some judges may declare ties more readily than others.

The fitting of a Bradley-Terry model to the paired comparison data, which results in difference scale values for each algorithm, is described in detail in Appendix B.

5.6 Results

The Bradley-Terry model gives a good fit to the empirical data in all conditions of both Experiment 1 and 2 with the only exception of the overall

preference judgments for industrial noise at 10 dB SNR in Experiment 2. The detailed test statistics are given in Appendix C.

The results of Experiment 1 are shown in Figures 5.1 to 5.4. Given are the scale values according to the Bradley-Terry model for the three criteria that were asked for: Overall preference (O), naturalness of the speech (N), and reduction of the background noise (R).[26]

Since the Bradley-Terry model yields scale values on a difference scale, there is one free parameter for the absolute position of the values. This was chosen in a way that algorithm UN (i.e., no noise reduction) was always assigned to zero. Hence, negative scale values mean that an algorithm is judged worse than no noise reduction. Differences between scale values from different measurements (i.e., different noise conditions), even from different studies, are comparable with each other (if the same objects are compared to each other). The absolute values, however, are not comparable. In contrast to an interval scale, the *range* of a difference scale is also of significance (Gediga, 1998). A small range, as for example in the overall preference judgments (O) in Figure 5.1, indicates that the algorithms were judged much more alike than if they encompass a large range as is the case in the noise reduction judgments (R). For this reason, a scale axis is also given in the figures.

Figures 5.1 and 5.2 show the results of the six normal-hearing subjects for the algorithms in drill noise and cafeteria noise at two different signal-to-noise ratios, respectively.

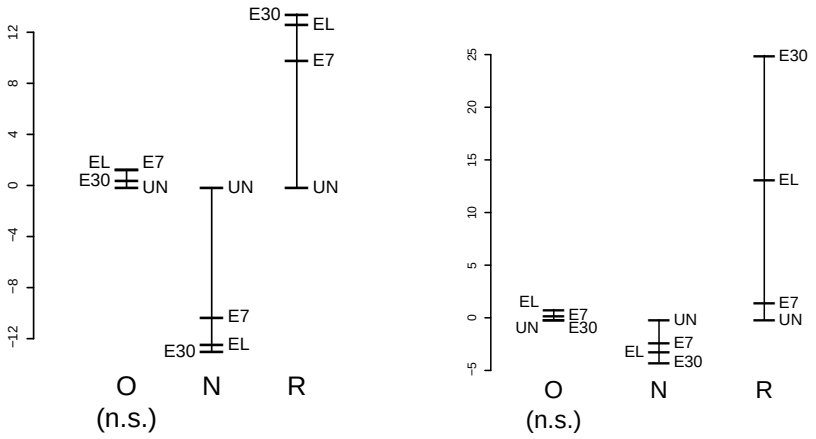
Figures 5.3 and 5.4 present the results of the six hearing-impaired subjects.

The noise reduction algorithms are generally preferred over no noise reduction processing in case of the drill noise. This preference is a little bit stronger (and statistically significant) in case of the hearing-impaired subjects than in case of the normal-hearing subjects.

The drilling machine noise is reduced most with the Ephraim-Malah algorithm E30, followed by EL, and least with algorithm E7. This was found with the normal-hearing (Figure 5.1) as well as with the hearing impaired subjects (Figure 5.3). The range of the scale values indicates that the amount of noise reduction is larger at the higher SNR.

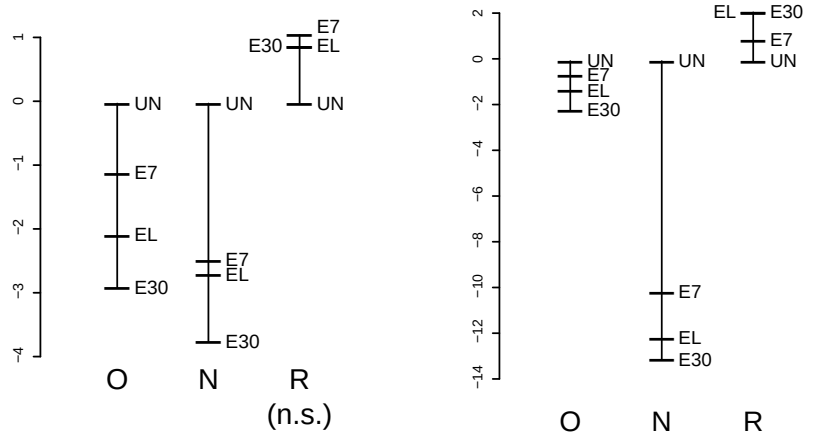
Apparently, the distortions of the target speech increase together with the amount of noise reduction as is obvious from the normal-hearing subjects' scale values for the naturalness of the speech. Probably, the negative effects of the noise reduction on the naturalness of the target speech are the reason for the only small overall preference for the noise reduction algorithms over UN in case of the normal-hearing subjects. EL and E7 are little preferred over E30, which has (with its highest noise reduction) the strongest negative impact on speech naturalness.

However, the hearing-impaired subjects judge the naturalness of the



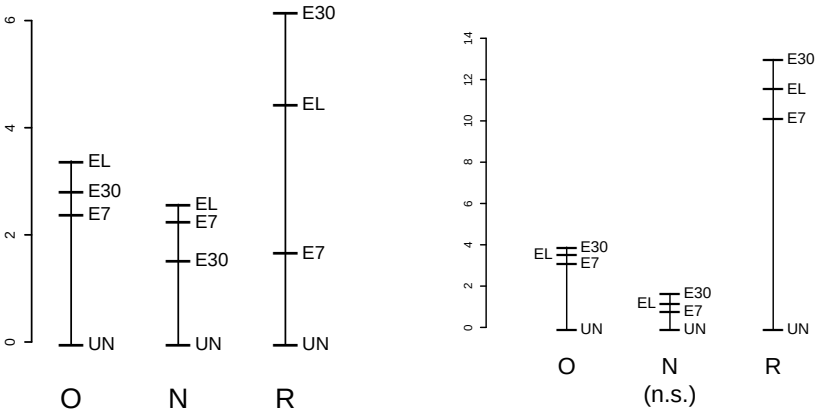
(a) Drill noise −5 dB SNR (b) Drill noise +5 dB SNR

Figure 5.1: *Normal-hearing subjects' results of the paired comparisons in Experiment 1 with drill noise. Plotted are the Bradley-Terry scale values for the different algorithms. Algorithm UN is arbitrarily set to zero. Negative values thus indicate worse performance than without noise reduction. The bar denoted* O *shows the results concerning the overall preference,* N *naturalness of the speech, and* R *reduction of the background noise.* (n.s.) *means that the differences among algorithms were not found to be significant.*



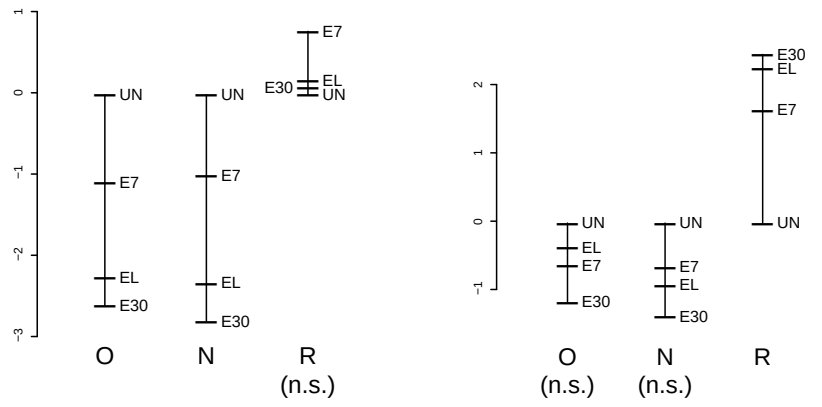
(a) Cafeteria −5 dB SNR (b) Cafeteria +5 dB SNR

Figure 5.2: *Normal-hearing subjects' results of the paired comparisons in Experiment 1 with cafeteria noise. Notation as in Figure 5.1.*



(a) Drill noise −5 dB SNR (b) Drill noise +5 dB SNR

Figure 5.3: *Hearing-impaired subjects' results of the paired comparisons in Experiment 1 with drill noise. Notation as in Figure 5.1.*



(a) Cafeteria −5 dB SNR (b) Cafeteria +5 dB SNR

Figure 5.4: *Hearing-impaired subjects' results of the paired comparisons in Experiment 1 with cafeteria noise. Notation as in Figure 5.1.*

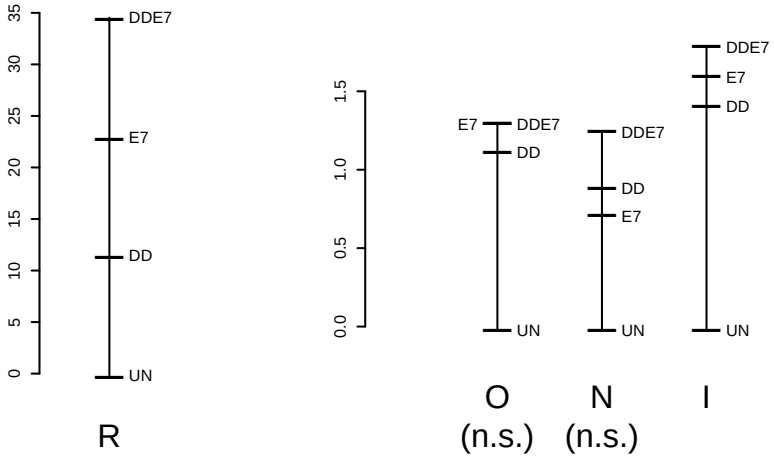
speech being better with the noise reduction algorithms than without. This reveals a difference between normal-hearing and hearing-impaired subjects.

The results of the paired comparisons in cafeteria noise show a high concordance between normal-hearing and hearing-impaired subjects (Figures 5.2 and 5.4). Here, the only prominent difference between these two groups of subjects is the range of the scale values for the naturalness of the speech. The normal-hearing subjects perceived larger differences, especially between UN and the noise reduction algorithms, than the hearing-impaired subjects did. The speech distortions increase from E7 over EL to E30. In general, the amount of perceived noise reduction provided by the Ephraim-Malah algorithms is much smaller than in the drill noise, but the order is the same in the cafeteria noise at the higher SNR. At the lower SNR, the order is reversed: E7 provides the strongest noise reduction, followed by EL and E30.

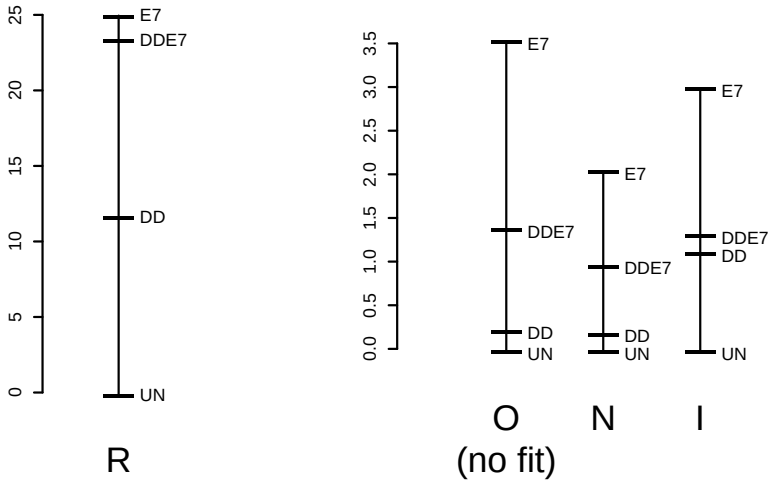
The results for Experiment 2 are given in Figure 5.5 (industrial noise), Figure 5.6 (cafeteria noise), and Figure 5.7 (speech-shaped noise). In addition to the three criteria of Experiment 1, the subjects were also asked to judge the intelligibility of the speech. The Bradley-Terry scale values of these paired comparisons are given in the bars denoted with "I".

The differences between algorithms are strongest concerning the reduction of the background noise (Figures 5.5 to 5.7). Here, the range of the scale values is so much larger than for the other criteria, that two graphs with distinct scale axes are given in the figures to resolve differences between algorithms concerning overall preference, naturalness of the speech, and intelligibility. With respect to the amount of noise reduction, the order of algorithms with increasing noise reduction is UN, DD, E7, and DDE7, apart from one small exception where E7 provides slightly more suppression (industrial noise at 10 dB SNR). The differences between algorithms with respect to the overall preference are generally very small. A plain difference is only found in the industrial noise at 10 dB SNR where algorithm E7 is clearly preferred over all other algorithms. However, the Bradley-Terry model only yields a poor fit to the data due to inconsistent judgments of subject HM.[27] With subject HM excluded, however, the model fits the data. Then, E7 is still clearly preferred over the other algorithms but no difference is found anymore between algorithm DD and UN. However, at the smaller SNR (0 dB SNR) the difference between UN and DD is more distinct.

With respect to overall preference, naturalness of the speech and intelligibility, DD seems to perform generally better at the lower SNR than at the higher SNR. In the cafeteria at 0 dB SNR, algorithm DD is preferred over UN while E7 is worse than UN. The differences are, however, not significant. At 10 dB SNR, DD is barely preferred over UN but now E7 is preferred. Similarly in the speech-shaped noise at 0 dB SNR, DD is little preferred over UN which in turn is little preferred over E7, but at 10 dB

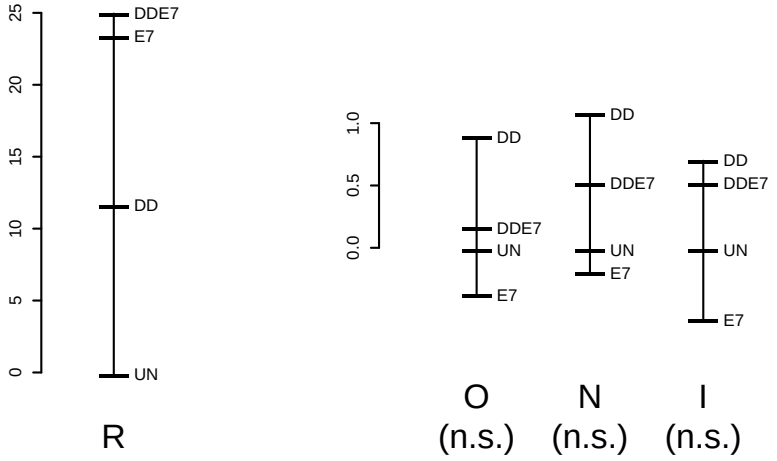


(a) Industry 0 dB SNR

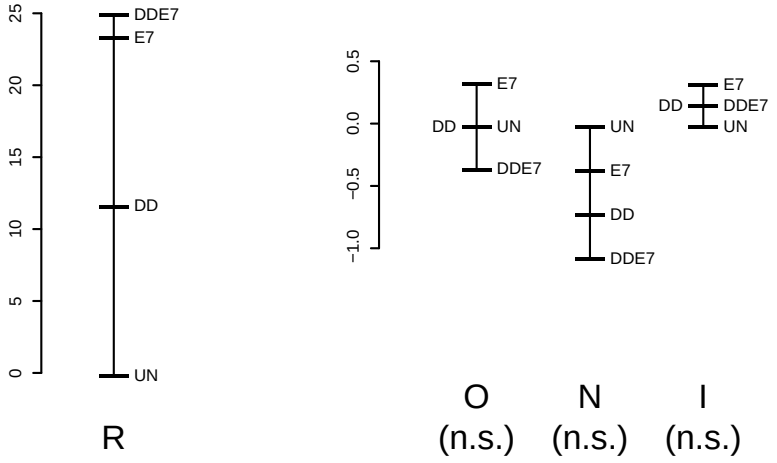


(b) Industry 10 dB SNR

Figure 5.5: *Hearing-impaired subjects' results of the paired comparisons in Experiment 2 with industrial noise. Plotted are the Bradley-Terry scale values for the different algorithms. Algorithm UN is arbitrarily set to zero. Negative values thus indicate worse performance than without noise reduction. The bar denoted* R *shows the results concerning the reduction of the background noise. Due to their much smaller scale values, the other results are plotted in an extra graph. The bar denoted* O *shows the results concerning the overall preference,* N *naturalness of the speech, and* I *intelligibility of the speech.* (n.s.) *means that the differences among algorithms were not found to be significant.* (no fit) *means that the BT model fit was below the predefined* α*-level, indicating a poor fit.*

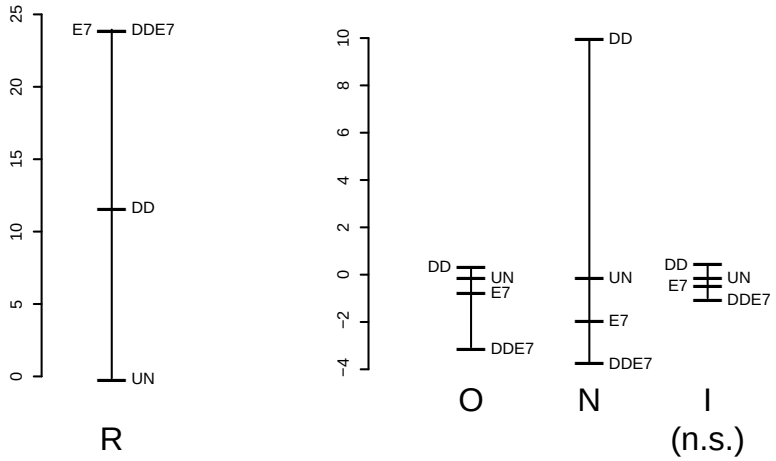


(a) Cafeteria 0 dB SNR

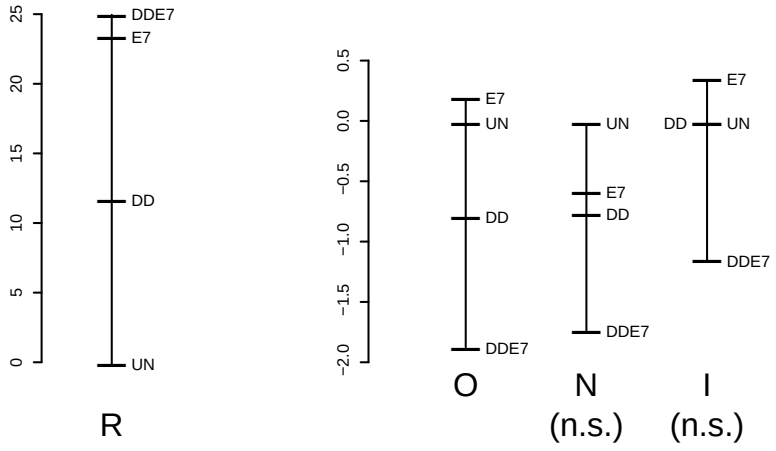


(b) Cafeteria 10 dB SNR

Figure 5.6: *Hearing-impaired subjects' results of the paired comparisons in Experiment 2 with cafeteria noise. Notation as in Figure 5.5.*



(a) Speech-shaped noise0 dB SNR



(b) Speech-shaped noise10 dB SNR

Figure 5.7: *Hearing-impaired subjects' results of the paired comparisons in Experiment 2 with speech-shaped noise. Notation as in Figure 5.5.*

SNR the order is reversed: E7 preferred over UN, which is preferred over DD.

In the industrial noise, the speech intelligibility judgments are completely in line with the judgments of the amount of noise reduction (with regard to the order of algorithms; the actual differences are much smaller in the intelligibility judgments), i.e. the stronger the noise reduction, the more intelligible the speech. In the cafeteria noise at 0 dB SNR, however, the intelligibility judgments are in line with the naturalness of the speech. In particular, the subjects recognize that processing artifacts of algorithm E7 distort the speech, leading to an intelligibility loss. Though all noise reduction algorithms are judged worse than UN with respect to the naturalness of the speech at the higher SNR of 10 dB, the subjects here see an intelligibility gain by the noise reduction processing. The speech is judged most intelligible with algorithm E7, which is also preferred overall. Though algorithm DDE7 is not preferred over UN (probably due to the low speech naturalness), the subjects acknowledge a better speech intelligibility with DDE7.

5.7 Discussion

In Chapter 4, speech intelligibility measurements employing the speech-shaped noise were reported for Experiment 2 (Table 4.11). It is of interest to compare the intelligibility *judgments* of the present study with the actual measurements of the speech reception thresholds performed with the Oldenburg sentence test. The median speech reception threshold (SRT) for algorithm UN was found to be -1.75 dB. Since the paired comparisons were performed near this SRT (at a SNR of 0 dB), the results are assumed to be comparable. In the sentence test, the SRT of E7 is in median 0.45 dB worse than UN, DDE7 is 0.9 dB worse than UN, and DD is 1.1 dB better than UN. This same order was also found in the paired comparison judgments of the intelligibility (Figure 5.7a). The Bradley-Terry scale values even reflect that the difference between UN and E7 is only about half that of UN and DD, and UN and DDE7.

This perfect agreement between judgments and measurements of the speech intelligibility underlines that subjects are well able to subjectively *judge* intelligibility, showing high correlation to "objective" word recognition scores, a fact also found by other researchers (Speaks *et al.*, 1972; Punch and Parker, 1981; Cox and McDaniel, 1984; Cox *et al.*, 1991; Rankovic and Levy, 1997; Wesselkamp and Kollmeier, 1993; Kollmeier and Wesselkamp, 1997). Note, however, that this comparison between subjective judgments and the objective speech reception thresholds is only possible because the Bradley-Terry model is used to obtain difference scale values from the paired comparisons data.

In the overall preference judgments, speech distortions due to the processing seem to counterbalance the positive effects of the well-perceived reduction of the noise. Obviously, normal-hearing subjects were able to disregard the background noise and concentrate on the target speech when judging "speech naturalness". Thus, they noticed distortions due to the noise reduction processing. In case of the drill noise, the hearing-impaired subjects, however, seem not to have noticed these distortions and hence look for other cues, since they judged the speech naturalness as being better *with* the noise reduction. Interviews with the subjects confirmed that the *presence of background noise* was perceived as being "unnatural". Hence, the speech was judged as least natural with algorithm UN (i.e., without noise reduction).

In most conditions in Experiment 1, algorithm E30 was found to provide the strongest noise reduction, followed by EL and E7. However, in cafeteria noise at the lower SNR, E7 provides the strongest noise reduction, followed by EL and E30. This change in order is probably connected with the non-stationary characteristic of the cafeteria noise. Since the cafeteria noise itself is fluctuating and irregular, it can be assumed that at this low signal-to-noise ratio of −5 dB, which is even below the speech-reception threshold (the SRT is in median −3.9 dB for the normal hearing and −3 dB for the hearing-impaired subjects; cf. Chapter 4), the additional distortions by the noise reduction processing contribute to the perceived amount of noise.

As the amount of noise reduction provided by the algorithms is much smaller in the cafeteria noise than in the drill noise, the overall preference judgments in the cafeteria seem to be strongly dominated by the naturalness of the speech. Hence, no noise reduction (UN) is preferred over E7, EL, and strongest over E30. A slight disagreement is found at +5 dB SNR where the hearing-impaired subjects prefer EL over E7.

The results for Experiment 2 (Figures 5.5 to 5.7) further support the assumption that the overall preference is a function mainly of the naturalness of the speech and of the amount of noise reduction. In the industrial noise at 0 dB SNR, for example, the speech is perceived more natural with algorithm DD than with E7. However, since E7 provides considerably more noise reduction, E7 is finally preferred over DD. In the cafeteria noise as well as in the speech-shaped noise at 10 dB SNR, the speech is perceived most natural without noise reduction (i.e., with algorithm UN), followed by algorithm E7. Again, algorithm E7 provides significant noise reduction and hence is finally preferred over UN. In the cafeteria noise as well as in the speech-shaped noise at 0 dB SNR, the speech is perceived most natural with algorithm DD. Although algorithm E7 yields much more noise reduction, DD is finally preferred since the naturalness of the speech is worse with E7 than without noise reduction UN.

It is assumed that the relatively low complexity of the algorithms allows an application in digital hearing aids in the near future. The delay time

due to the signal processing is below 10 ms. One cosmetic obstacle for the application of the directional filter and dereverberation algorithm, however, might be that it needs binaural input, i.e. microphones in both ears. A wireless bidirectional communication between a left-ear and a right-ear hearing aid is a yet unsolved technical problem. The Ephraim-Malah noise reduction algorithms, however, can be employed in a single hearing aid.

5.8 Conclusions

The application of the method of paired comparisons with respect to different quality aspects (i.e., overall preference, naturalness of the speech, strength of noise reduction, and speech intelligibility) in combination with the Bradley-Terry scaling model provided a consistent and comprehensive assessment of algorithmic performance that coincides very well with other assessment methods (e.g. measurement of speech reception thresholds). The advantage of paired comparisons in contrast to category rating is its ease for the subjects and elimination of judgment bias.

The results of the paired comparisons show that noise reduction processing is worthwhile in all of the different noises that were investigated. The Ephraim-Malah single-microphone noise reduction algorithms can be recommended for use in rather stationary noises. They fail, however, in strongly fluctuating noises (cafeteria babble) where the binaural directional filter and dereverberation algorithm may be used, particularly at lower SNRs.

The combined algorithm DDE7 was not able to merge both benefits but was found at an intermediate position. It is concluded that it is not appropriate to use both noise reduction schemes at the same time. These findings stress the importance of developing a more sophisticated combination of the noise reduction algorithms with an intelligent control algorithm.

It is striking that the hearing-impaired subjects did not perceive any distortions in the speech due to the processing of the noise reduction algorithms in drilling noise in Experiment 1. The naturalness of the speech was judged better with the Ephraim-Malah algorithms than without noise reduction, contrary to the normal-hearing subjects. This confirms once more that tests with hearing-impaired subjects should be performed when hearing-aid applications are considered, even in the case of noise reduction algorithms which are commonly believed to yield the same positive effects for normal-hearing listeners.

Unfortunately, the algorithms with the largest amount of noise reduction show also the strongest speech distortions in unfavorable noise situations as can be seen with the cafeteria noise at -5 dB SNR. In such situations the naturalness of the speech is best preserved by the Ephraim-Malah algorithm E7.

Future studies have to consider a combination with algorithms which

are able to classify the noise characteristics and the acoustic environment to automatically adjust parameters for optimal noise reduction performance. Recently, promising candidates for this task were suggested (Ostendorf *et al.*, 1997; Ostendorf *et al.*, 1998; Tchorz, 2000). These algorithms are based on an analysis of modulation frequencies.

To conclude, the single-microphone algorithms are very effective in the reduction of the background noise, but they also lead to a loss or distortion of speech components at lower signal-to-noise ratios. Hence, future research has to address the reduction of these distortions in adverse noise conditions.

Chapter 6

Predicting the subjective quality of noise reduction algorithms for hearing aids

Most noise reduction schemes proposed in the literature so far that aim at enhancing speech in a noisy background were evaluated using objective quality measures. Since listener tests were mostly not included, conclusions might be invalid due to the inappropriateness of some objective measures for the specific kind of distortions introduced by noise reduction algorithms. Moreover, if the noise reduction schemes are intended for use in digital hearing aids, the objective measures have to be validated with quality judgments from hearing-impaired subjects.

In the present study, different objective speech quality measures were applied to the same test signals that were judged by six hearing-impaired subjects. Single-microphone noise reduction algorithms proposed by Ephraim and Malah (1984, 1985) and a binaural noise reduction algorithm (directional filter and dereverberation) proposed by Wittkop (2000) were employed. The paired comparisons carried out by the subjects were analyzed with the Bradley-Terry model (Bradley and Terry, 1952) resulting in difference scale values for each algorithm. The Log-Area-Ratio (LAR) objective measure shows the highest correlation with overall subjective preference, while the PMF measure (Hansen and Kollmeier, 2000) corresponds best with the subjectively perceived amount of noise suppression.

A potential application of objective measures is the optimization of noise reduction algorithms by effectively evaluating the often large parameter space.

6.1 Introduction

Following Halka and Heute (1992), a "measure" is regarded as *objective* if it is computable, in contrast to a *subjective* measure which is always based on a human perception test. Several different objective prediction methods for the transmission quality of low-bit rate speech coding algorithms (speech codecs) have been proposed in the literature so far. Some of these employ a quantitative processing model of the auditory system to estimate the deviations between the distorted signal and a corresponding reference signal in a perceptually relevant domain. The "perceptual speech-quality measure" (PSQM) introduced by Beerends and Stemerdink (1994) was obtained by modifying the previously introduced "perceptual audio quality measure" PAQM (Beerends and Stemerdink, 1992). While the PAQM was developed for evaluating Hi-Fi equipment and was shown to be superior to other measures as for example the NMR (noise-to-mask ratio), the PSQM was optimized for measuring the quality of *speech* codecs. In 1996, the PSQM was standardized by the International Telecommunication Union as Recommendation P.861 (ITU, 1996c). A detailed description and discussion of both the PAQM and PSQM quality measures is given by Beerends (1998). Another measure using a more elaborated psychoacoustical model is the objective speech quality measure q_c, denoted as PMF in the following (M. Hansen and Kollmeier, 2000).

Both PSQM and PMF show good performance in speech quality prediction for different speech quality test data bases (Beerends and Stemerdink, 1994; Hansen and Kollmeier, 2000). In general, the speech test data bases consist of a number of test signals with corresponding subjective mean opinion scores (MOS), which are obtained by averaging the judgments on a five-point scale ranging from "bad" (Score 1) to "excellent" (Score 5) from several subjects for each signal.

A different approach was proposed by J. Hansen and Pellom (1998) to evaluate speech enhancement algorithms by using several "technical" distance measures: the Itakura-Saito Distortion Measure (IS), the Log-Likelihood Ratio Measure (LLR), the Log-Area-Ratio Measure (LAR), the Segmental Signal-to-Noise Ratio Measure (SSNR) and the Weighted Spectral Slope Measure (WSS). The measures each calculate a distance between the clean speech signal (reference input) and either the noisy speech signal itself or the signal processed by a noise reduction algorithm, respectively. Although J. Hansen and Pellom (1998) stress the importance of applying subjective tests next to objective quality evaluation, and recommend a pairwise preference test, they do not present any subjective results to validate the objective measures.

Most applications of objective quality measures so far were concerned with the evaluation of speech codecs rather than noise reduction systems. According to Gustafsson *et al.* (1996), there is little experience in the use

of objective measures for the evaluation of speech enhancement systems so far. Although a lot of publications actually used objective measures for the evaluation of noise reduction algorithms, this statement honestly reflects the fact that the tradeoff between maximizing "the noise reduction while keeping the audible distortions of the speech signal at an acceptable level [...] makes it hardly possible to describe the quality of a speech enhancement system with a single figure." (Gustafsson *et al.*, 1996). Moreover, it is not clear a priori which objective measure assesses which subjective dimension.

In this chapter, therefore the predictive power of several quality measures is investigated with respect to the subjective noise reduction effect for hearing-impaired listeners.

Most studies on objective speech quality measures reported in the literature so far considered correlations between the predictions of the respective measures with mean-opinion scores (MOS), which are the mean of subjective quality ratings on a five-point absolute category rating scale from a large number of subjects (Quackenbush *et al.*, 1988). However, Studebaker (1982) suggested that judgment of small differences between stimuli is easier when performed in direct comparison than in isolated judgments. In addition, Lukas (1991) argues that a major problem of category judgments is that they are highly susceptible to context and range effects, anchor stimuli, reference objects, different application of the scale by the subjects, etc. (cf. Johnson and Mullaly, 1969).

In a paired comparison experiment, the subject does not judge one stimulus on a given scale, but judges which of two stimuli exhibits more of the property under question. The task for the subject is much simpler, and context effects are strongly reduced. Hence, this method was employed here. One frequently used scaling procedure for paired comparisons is the Bradley-Terry (BT) model (Bradley and Terry, 1952). More details on using this method to obtain difference scale values from paired comparisons of noise reduction algorithms are given in Appendix B.

6.2 Subjective preference data sets

Data from two experiments on noise reduction algorithms with hearing-impaired subjects are considered here.

Experiment 1 (Marzinzik and Kollmeier, 1999, 2000; cf. Chapters 4 and 5) evaluates three different single-microphone noise reduction algorithms with six moderately sensorineural hearing-impaired subjects: The minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator (Eq. 7 in Ephraim and Malah, 1984; denoted as E7), a modified estimator under uncertainty of signal presence (Eq. 30 in Ephraim and Malah, 1984; denoted as E30), and the MMSE log-spectra estimator (Ephraim and Malah, 1985; denoted as EL; cf. Chapter 3 for more details on the algorithms).

A complete paired comparison test (Round Robin tournament) of the algorithms E7, EL, and E30, and the unprocessed condition UN were performed. The individual hearing loss was compensated using third-octave band equalizers supplying an amplification in each frequency band determined from the audiogram of each subject using the one-half gain rule, i.e. target gain at each frequency is simply the subject's audiometric threshold multiplied by 0.5 (Fabry and Schum, 1994; Lybarger, 1944; Lybarger, 1978).

Different noise signals were employed for the measurements (cf. Table 6.1). The target speech signal was always the same short sentence (in German language) of approximately 4 s duration spoken by a male talker. The subjects were asked to give three preference judgments for each condition with regard to a) overall preference, b) naturalness of the speech, and c) suppression of the background noise.[28]

Experiment 2 used a similar setup and test protocol as in Experiment 1 but employed different noise reduction algorithms and different speech and noise signals (Marzinzik *et al.*, 1999; cf. Chapters 4 and 5). A binaural directional filter and dereverberation algorithm (Wittkop *et al.*, 1999; Wittkop, 2000; denoted as DD) was evaluated in Experiment 2 as well as algorithm E7 from Experiment 1, and the sequential processing of DD and E7 (denoted as DE). The unprocessed condition is again denoted as UN. The different noise conditions are given in Table 6.1.

Table 6.1: *Noise conditions used in the experiments.*

Noise signal		Experiment 1 (UN,E7,EL,E30)	Experiment 2 (UN,E7,DD,DE)
Monaural recordings			
Caf-5	Cafeteria noise	×	
Caf5	(−5 and +5 dB SNR)	×	
Drill-5	Drilling machine	×	
Drill5	(−5 and +5 dB SNR)	×	
Binaural recordings			
Caf0	Cafeteria noise		×
Caf10	(0 and 10 dB SNR)		×
Ind0	Industrial noise		×
Ind10	(0 and 10 dB SNR)		×
Ssn0	Speech-shaped noise		×
Ssn10	(0 and 10 dB SNR)		×

The Bradley-Terry model was used to calculate difference scale values for each algorithm using the preference data from the paired comparison experiments (cf. Chapter 5).

6.3 Implementation of objective measures

The perceptual speech-quality measure (PSQM) was employed as standardized by the International Telecommunication Union (ITU, 1996c). The processing includes a frequency transformation (critical band rates) with level compression. Silent intervals are excluded when the final quality measure is derived from the differences between the representations of the clean reference and the (processed) noisy signal.

The PMF measure employs a more elaborated psychoacoustical model (Hansen, 1998). Each frequency channel output of the incorporated model is time-averaged in frames of 20 ms with 50 % overlap. The final objective speech quality measure is calculated as the overall correlation coefficient between the representations of the reference and the test signal. A value of 1.0 therefore reflects identical representations. Lower values indicate stronger distortions.

J. Hansen and Pellom (1998) intended to provide a common evaluation test platform for developers of speech enhancement algorithms. MATLAB source codes of several speech quality measures were made available via WWW[29]. These implementations were employed in the present study. The Itakura-Saito Distortion Measure (IS), the Log-Likelihood Ratio Measure (LLR), and the Log-Area-Ratio Measure (LAR) are based on the assessment of dissimilarity of linear prediction (LP) coefficients between the clean and the (processed) noisy signals. More details on these objective measures are given by Quackenbush *et al.* (1988). The Segmental Signal-to-Noise Ratio Measure (SSNR) is a frame-based estimation of the signal-to-noise ratio where SNRs above 35 dB are replaced with 35 dB, since these do not reflect large perceptual differences. Similarly, a lower threshold of −10 dB is chosen. The Weighted Spectral Slope Measure (WSS) calculates a weighted difference between the spectral slopes in each of 36 overlapping filter bands with increasing bandwidth. For the calculation of each of these measures, the signals are sampled with 8 kHz and segmented in frames of 30 ms with 3/4 overlap, i.e. a window skip of 7.5 ms. The respective objective measure is calculated for each frame of the input signal and finally the median over all frames is taken. Hansen and Pellom (1998) suggested to alternatively take the average over all frames by ignoring 5 % of the largest values (corresponding to larger distortions), because the mean over *all* frames is typically biased by a few frames in the tail of the quality measure distribution. However, taking the median was preferred here and serves the same purpose.

6.4 Correlating subjective data and objective measures

6.4.1 Procedure

The psychoacoustically motivated speech quality measures PMF (M. Hansen and Kollmeier, 2000) and PSQM (Beerends and Stemerdink, 1994), as well as the more "technical" quality measures proposed by J. Hansen and Pellom (1998) – the Itakura-Saito Distortion Measure IS, the Log-Likelihood Ratio Measure LLR, the Log-Area-Ratio Measure LAR, the Segmental Signal-to-Noise Ratio Measure SSNR, and the Weighted Spectral Slope Measure WSS – were applied to the same test signals that were used in the subjective tests described in Section 6.2.

Since the Bradley-Terry scaling procedure was applied to the subjective paired comparisons data, difference scale values were obtained for each algorithm which in the present study allow to use Pearson's correlation coefficient ρ and scatter plots to determine the relationship between the subjective data and the objective measures.

6.4.2 Results

Tables 6.2 and 6.3 give Pearson's correlation coefficients ρ between the subjective preference and objective measures data of Experiments 1 and 2, respectively.[30] The values obtained with the objective measures were rescaled so that the unprocessed signals UN are always zero, and larger values indicate better performance. Scatter plots are given for those quality measures which are highest correlated with the respective subjective dimension.[31]

Table 6.2: *Pearson's correlation coefficients ρ between transformed objective measures and subjective data from Experiment 1.*

Subjective Criterion	Objective Measure						
	PMF	PSQM	LAR	LLR	IS	WSS	SSNR
Noise Suppression	0.90	0.65	0.48	0.56	−0.50	0.28	0.71
Speech Naturalness	0.43	0.85	0.92	0.88	−0.06	0.78	0.68
Overall Preference	0.67	0.87	0.87	0.86	−0.21	0.68	0.77

A look at Table 6.2 shows that the objective measure PMF has the highest correlation with the subjective "noise suppression" judgments in Experiment 1, whereas the LAR measure has the highest correlation with "speech naturalness" and "overall preference" judgments.

Pearson's correlation between the objective measure PMF and the subjective data with criterion "noise suppression" is 0.9. A look at Figure 6.1 confirms this very high correlation.

Table 6.3: *Pearson's correlation coefficients ρ between transformed objective measures and subjective data from Experiment 2.*

Subjective Criterion	Objective Measure						
	PMF	PSQM	LAR	LLR	IS	WSS	SSNR
Noise Suppression	0.54	0.70	−0.14	0.07	−0.52	−0.78	0.44
Speech Naturalness*	−0.11	0.31	0.73	0.65	0.45	0.26	0.21
Overall Preference	−0.04	0.41	0.70	0.64	0.37	0.23	0.25

* After removing one single outlier (the subjective judgment of algorithm DD in speech-shaped noise at 0 dB SNR). Including the outlier the correlations are: 0.15, 0.14, 0.46, 0.40, 0.37, 0.26, 0.05.

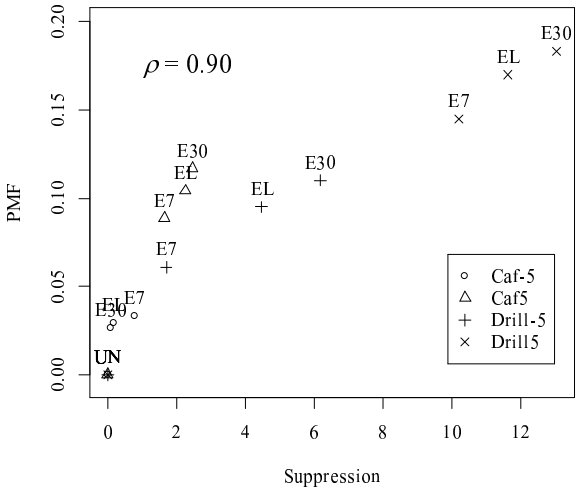


Figure 6.1: *Scatter plot of the objective measure PMF vs. the subjective data (Bradley-Terry scale values) with the criterion "noise suppression" in Experiment 1. Noise conditions are abbreviated as denoted in Table 6.1.*

Not only that the PMF measure is able to give the correct ranking of the algorithms for all different conditions (drilling machine noise and cafeteria noise, both for −5 and +5 dB SNR), even the amount of noise suppression in the different noise conditions as perceived by the subjects is very well reflected.

For the criterion "speech naturalness", the correlation between the LAR measure and the subjective data is the largest (0.92, see Table 6.2). While the rankings and distances are very well predicted for cafeteria noise (the noise suppression algorithms are judged worse than unprocessed, here), the LAR measure fails to give correct rankings of the algorithms for the drilling noise conditions (see Figure 6.2).

However, the LAR measure correctly predicts that the speech naturalness of the noise suppressed signals with drilling noise is generally judged better than unprocessed. It seems that the reversed ranking of algorithms

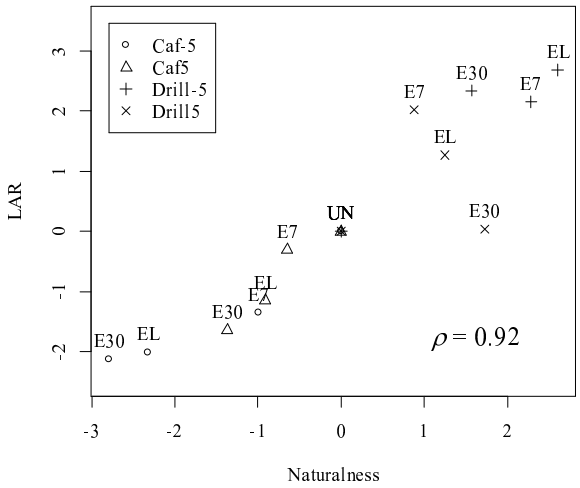


Figure 6.2: *Scatter plot of the objective measure LAR vs. the subjective data (Bradley-Terry scale values) with the criterion "speech naturalness" in Experiment 1. Noise conditions are abbreviated as denoted in Table 6.1.*

E7, EL, and E30 for drilling noise at 5 dB SNR is due to a change of criterion by the subjects. In this noise condition the subjects judged the speech the more natural, the more noise was suppressed. This gives rise to the assumption that above a certain signal-to-noise ratio the subjects rate the simple presence of noise as disturbing the "speech naturalness" and small distortions (artifacts) of the speech itself due to noise reduction processing are not weighed as high as the influence of the noise. This criterion change, however, is not reflected by the objective measures.

The LAR measure is also highest correlated with the "overall preference" judgments from the subjects in Experiment 1. Although nominally the PSQM measure has the same high correlation (0.87), a look at Figures 6.3 and 6.4 reveals the superiority of the LAR measure.

An almost perfectly linear relation between LAR predictions and subjective data for cafeteria noise and drilling noise at signal-to-noise ratios of -5 dB is obtained. In the better SNR condition ($+5$ dB), however, the same observation can be made as for the naturalness judgments: The subjects seem to have changed their criterion which results in a reversed rank order of algorithms. Probably, this process of switching criteria is beginning to occur in the cafeteria noise condition at $+5$ dB SNR: The judgments of the noise reduction algorithms are better than at -5 dB SNR (though still worse than unprocessed) but now EL (with more noise suppression) has already outperformed E7.

A look at Table 6.3 shows that in Experiment 2 (as in Experiment 1) the LAR objective measure has the highest correlation with the subjective data with regard to speech naturalness and overall preference. This is a strong

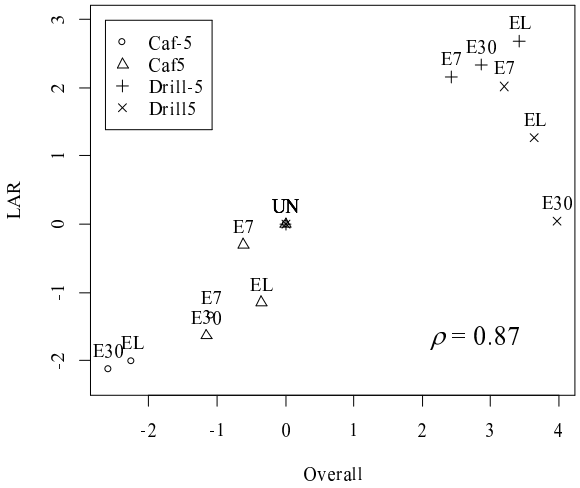


Figure 6.3: *Scatter plot of the objective measure LAR vs. the subjective data (Bradley-Terry scale values) with the criterion "overall preference" in Experiment 1. Noise conditions are abbreviated as denoted in Table 6.1.*

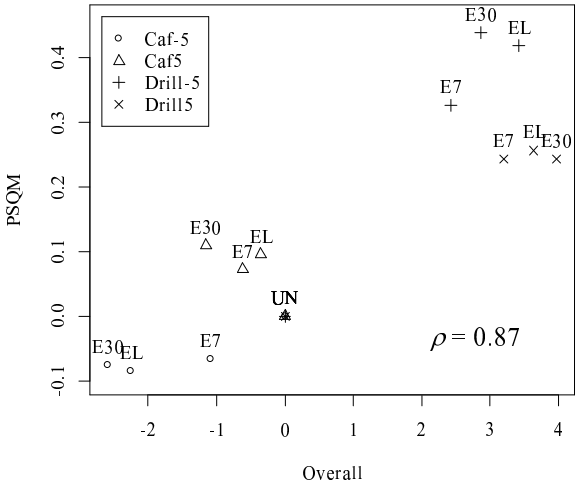


Figure 6.4: *Scatter plot of the objective measure PSQM vs. the subjective data (Bradley-Terry scale values) with the criterion "overall preference" in Experiment 1. Noise conditions are abbreviated as denoted in Table 6.1.*

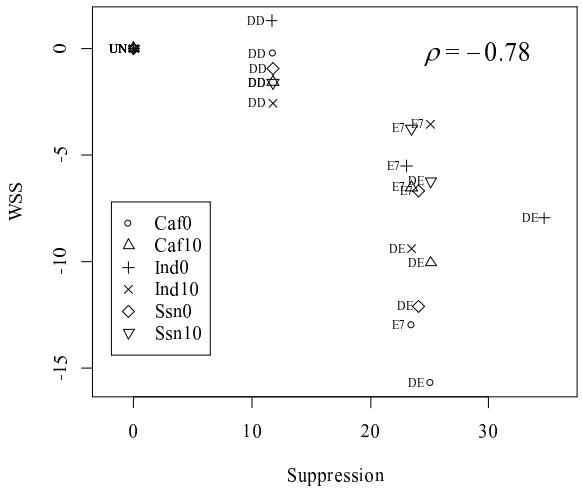


Figure 6.5: *Scatter plot of the objective measure WSS vs. the subjective data (Bradley-Terry scale values) with the criterion "noise suppression" in Experiment 2. Noise conditions are abbreviated as denoted in Table 6.1.*

indicator of the robustness of the LAR measure since different noise reduction algorithms with different noise signals at different signal-to-noise ratios were tested in Experiment 2. Moreover, the recordings of Experiment 2 were made in a room with a reverberation time of $T_{60} = 0.6\,\mathrm{s}$, whereas in Experiment 1 the signals featured almost no reverberation.

Figures 6.5 to 6.7 give the best-performing objective measures with regard to "noise suppression" (cf. Table 6.3) as a function of the subjective data for different noise conditions and algorithms (scatter plots).

As is obvious from the figures, the subjective paired comparison experiment did not reveal substantial differences among different noise conditions (essentially, only three different subjective values were obtained). This is probably an artifact of the experimental design (there were no inter-noise comparisons), which in this case was not overcome by the Bradley-Terry scaling because the subjects showed a very high concordance in their preferences indicating very distinct differences in the amount of noise suppression obtained with the different algorithms. However, the objective measures do discriminate between the different noise conditions, which is also obvious from the figures. Altogether, this results in vertical stripes in the respective scatter plots. The highest correlations between the subjective "noise suppression" scaling and objective measures (cf. Table 6.3) are found for WSS (negative correlation, see Figure 6.5) and PSQM (see Figure 6.6), followed by PMF (see Figure 6.7). The good performance of WSS to predict noise suppression in this experiment is somewhat surprising since this measure showed only a low correlation with the subjective noise suppression data of Experiment 1. Therefore, the evidence is not sufficient to recommend the WSS measure for predicting the amount of subjectively perceived noise

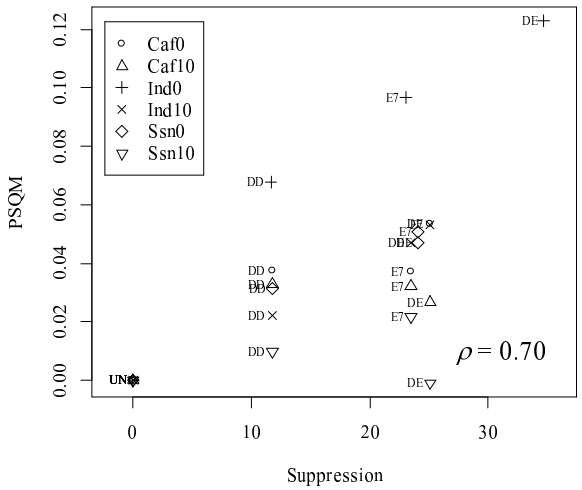


Figure 6.6: *Scatter plot of the objective measure PSQM vs. the subjective data (Bradley-Terry scale values) with the criterion "noise suppression" in Experiment 2. Noise conditions are abbreviated as denoted in Table 6.1.*

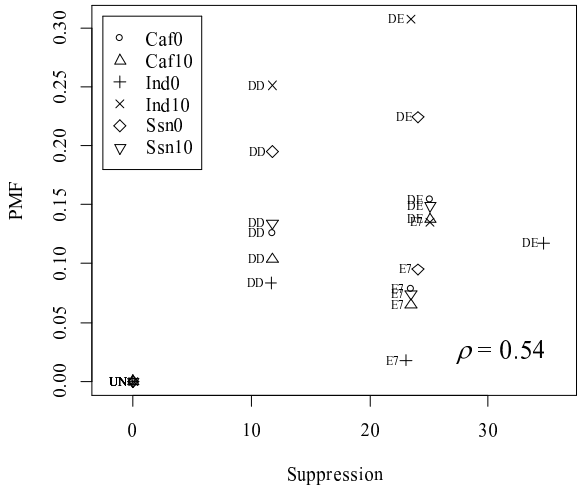


Figure 6.7: *Scatter plot of the objective measure PMF vs. the subjective data (Bradley-Terry scale values) with the criterion "noise suppression" in Experiment 2. Noise conditions are abbreviated as denoted in Table 6.1.*

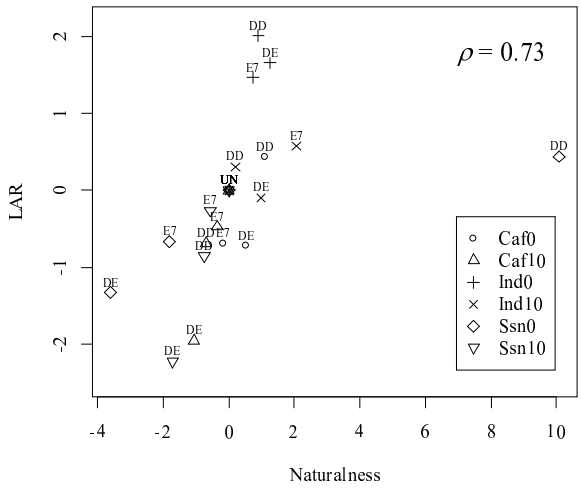


Figure 6.8: *Scatter plot of the objective measure LAR vs. the subjective data (Bradley-Terry scale values) with the criterion "speech naturalness" in Experiment 2. Noise conditions are abbreviated as denoted in Table 6.1.*

suppression. However, as in Experiment 1 there is again a significant correlation between the PMF measure and the subjective noise suppression data. It is striking that the rank order of algorithms UN, DD and DE is perfectly predicted by PMF (see Figure 6.7), but algorithm E7 is systematically underrated. By excluding algorithm E7, the correlation between PMF and the subjective noise suppression data increases to 0.71 (in this case PSQM also increases to 0.73 and WSS to −0.82). The difference between the PSQM measure and PMF observed in the correlation coefficient is only small and is only due to one noise condition (Ind0). A comparison of the scatter plots in Figures 6.7 and 6.6 shows no general advantage of PSQM over PMF.

A scatter plot of the objective measure LAR vs. the subjective data for the criterion "naturalness of speech" is given in Figure 6.8.

For three of the six noise conditions, the rank order of the algorithms is exactly predicted by LAR. In all six noise conditions the rank order of algorithms UN, E7 and DD (and to a high degree even the distances) are correctly predicted (algorithm DD was rated exceptionally high in the speech-shaped noise at 0 dB SNR which is therefore considered as an "outlier"). Algorithm DE, however, is underrated in the cafeteria noise at 0 dB SNR and in the industrial noise, leading to a wrong rank prediction for this algorithm in these noise conditions.

Although the correlations between the LAR measure and the subjective "overall preference" ratings are quite high (0.70), the scatter plot between both quantities (Figure 6.9) reveals that there are some more discrepancies in the rank order of the algorithms derived from the objective and the subjective data, respectively.

On the basis of this experiment it cannot be decided whether the discrep-

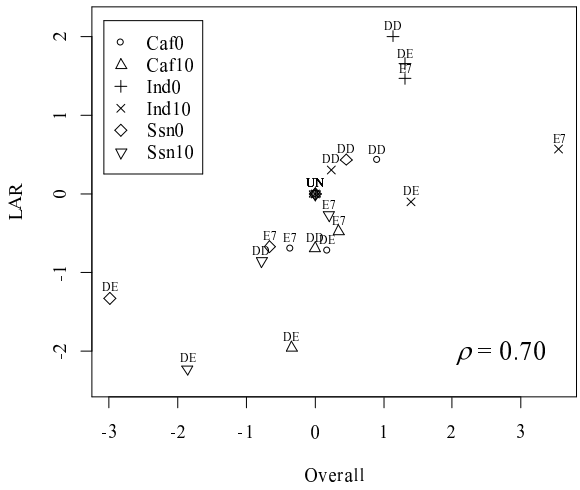


Figure 6.9: *Scatter plot of the objective measure LAR vs. the subjective data (Bradley-Terry scale values) with the criterion "overall preference" in Experiment 2. Noise conditions are abbreviated as denoted in Table 6.1.*

ancies are due to the inadequacy of the LAR objective measure to predict overall preference or whether they are due to too high standard errors in the subjective data.

6.4.3 Discussion

The observed change of judgment criterion by the subjects with increasing signal-to-noise ratio in drill noise (found in the "speech naturalness" and "overall preference" judgments in Experiment 1) indicate that probably only a composite measure will be able to reflect such effects. For higher SNRs the overall judgments were found to be strongly correlated with the amount of noise reduction which is best reflected by the PMF measure, while in more adverse noise conditions the overall preference is more connected with signal distortions which appears to be better reflected in the LAR measure. Hence, some kind of noise classification seems to be necessary to switch between different quality measures to give a final composite quality measure. However, more subjective data sets are required to validate the observed subjective effects and to develop a new measure composed of some basic measures. Another fact complicates this issue even more: Experiment 1 was also carried out with normal-hearing subjects (cf. Chapter 5), and it was found that the normal-hearing listeners did not show the change in judgment criterion as discussed for the hearing-impaired subjects. In case of the normal-hearing subjects, the order of algorithms with increasing perceived naturalness in drill noise was reported to be E30, EL, and E7 for both -5 dB and $+5$ dB SNR (Chapter 5). This correlates with decreasing noise reduction, and hence less processing artifacts. Although in this case the rank order

of noise reduction algorithms is correctly predicted by the LAR measure, it fails in predicting that the unprocessed signal UN is still perceived the most natural by the normal-hearing subjects in drill noise at +5 dB SNR. The clear differences in the subjective preferences between normal-hearing and hearing-impaired subjects indicate the importance of validating objective measures with the group for which predictions are aimed at.

6.4.4 Comparison with literature results

Gannot *et al.* (1997) evaluated the MMSE log-spectra estimator EL (Ephraim and Malah, 1985) by means of the Itakura-Saito distortion measure (IS). They found a slight degradation of the speech quality with the EL algorithm compared to no noise reduction over a range of signal-to-noise ratios from −10 dB to +10 dB. With respect to the stationary characteristics, the drill noise employed in Experiment 1 (cf. Section 6.4.2) is assumed to be comparable to the computer fan noise which was employed by Gannot *et al.* (1997). The findings of the present study for algorithm EL using the objective Itakura-Saito measure are in line with the results reported by Gannot *et al.* (1997). However, a look at Table 6.2 reveals that the IS measure does not correlate very well with the *subjective* judgments. This stresses again the importance of validating an objective quality measure for a specific application and, unfortunately, emphasizes the doubtful value of several noise reduction evaluations reported in the literature so far. However, most studies on noise reduction did not use objective quality measures for evaluation only, but considered also (informal) subjective testing, and often (with an increasing tendency) machine speech recognition tests.

Gannot *et al.* (1997) also compared the EL noise reduction algorithm with their own algorithm using the IS measure. As a result, EL performed slightly worse than no noise reduction but their own algorithm performed clearly better. The improvement was found to be strongest at the lowest SNR. There, the IS measure improves from about 2.1 to 1.2 with one sentence and from about 3.5 to 2.2 with another sentence. In the present study using the drill noise at −5 dB SNR (Experiment 1), the IS measure indicates an only small improvement from 4.4 to 4.3 for the EL algorithm, but a large improvement from 4.4 to 2.6 for the E7 algorithm. Hence, it is assumed that algorithm E7 still outperforms the algorithms proposed by Gannot *et al.* (1997). Unfortunately, these authors did not consider this algorithm for comparison in their study.

Meyer and Simmer (1997) used the LAR objective measure to evaluate a new noise reduction scheme. They compared their own algorithm with Wiener filtering and spectral subtraction noise reduction schemes. The LAR measure was chosen by these authors because Quackenbush *et al.* (1988) showed that this measure has the highest correlation with subjective quality measurement, a finding which is supported by the present study. Car noise

was chosen for the evaluation at SNRs ranging from 0 to 12 dB. Below 8 dB SNR, the new proposed algorithm was found to be better than both Wiener filtering and spectral subtraction. According to the LAR measure, all three algorithms provide better speech quality than no noise reduction. The largest improvement was found at about 3 dB SNR. Here, the LAR measure improves by 0.6 from about 2.7 to 2.1. In the present study, the E7 noise reduction algorithm improves the LAR measure by about 2.1 (from 11.5 to 9.4) compared to no noise reduction using drill noise at −5 dB SNR, by about 2.0 (from 9.2 to 7.2) at +5 dB SNR, and by about 1.4 (from 7.3 to 5.9) compared to no noise reduction using industry noise at 0 dB SNR and by about 0.6 (from 4.6 to 4.0) at 10 dB SNR. Although the same trend is found in both studies, the higher absolute values of the LAR measure found in the present study are due to the stronger impact of the chosen noises on the speech than is the case with the car noise used by Meyer and Simmer.

6.5 Using objective measures to optimize noise reduction schemes

The LAR objective measure was shown to have the highest correlation with subjective "overall preference" in the experiments considered in this study (cf. Section 6.4.2).

In the following it is demonstrated how to use this objective measure to aid in parameter optimizations in noise reduction algorithms.

6.5.1 Procedure

Slight modifications are introduced in the E7 noise reduction scheme. The noise reduction is applied to sub-bands derived from the frequency groups of the human auditory system instead of applying it to each frequency component of the FFT. This modified frequency spacing results in the speech power being distributed more uniformly across the sub-bands and is reported in the literature to result in better overall quality in adverse noise conditions (cf. Chapter 3).

Two additional signal processing parameters are introduced with the objective to further increase perceived quality. First is a first-order recursive low-pass filter with recursive factor G_r to smooth the gain factors over frequency. Secondly, the gain factors are smoothed over time using a first-order recursive low-pass filter with time constant τ.

For the optimization of the two parameters, the drill noise and the cafeteria noise from Experiment 1 (cf. Section 6.4.2) are employed at SNRs of −5 and +5 dB. The signals are processed by the modified noise reduction scheme for a wide range of parameter settings, and the LAR measure is determined for each.

6.5.2 Results

First of all, Table 6.4 reports the transformed LAR objective measure data for the algorithms E7, EL, E30, and no noise reduction UN from Experiment 1 (cf. Section 6.4.2; higher values indicate better performance). In addition, the LAR values for the modified E7 algorithm with frequency group spacing (denoted as E7MOD) are given in the last row.

Table 6.4: *Transformed LAR objective quality measure data for no noise reduction (UN), noise reduction algorithms E7, EL, E30, and modified E7 algorithm (E7MOD) using the setup of Experiment 1. Higher values indicate better performance.*

Algorithm	Noise, SNR			
	Drill		Cafeteria	
	–5 dB	+5 dB	–5 dB	+5 dB
UN	0.00	0.00	0.00	0.00
E7	1.99	2.20	−1.42	−0.37
EL	2.58	1.29	−2.13	−1.24
E30	2.18	−0.02	−2.27	−1.72
E7MOD	2.22	2.30	−0.88	−0.07

According to the LAR measure, this modification increases the quality of the E7 algorithm in all noise conditions. However, the noise reduction processing still results in worse (predicted) quality in the cafeteria noise than no noise reduction (UN).

Smoothing the gain factors over frequency increases the quality predictions of the LAR measure in the cafeteria noise as is shown in Table 6.5.

Table 6.5: *Transformed LAR objective quality measure data for the modified noise reduction algorithm with smoothing the gain factors over frequency using the setup of Experiment 1. Higher values indicate better performance. The gain factors are smoothed over frequency using a first-order recursive low-pass filter with recursive factor G_r.*

G_r	Noise, SNR			
	Drill		Cafeteria	
	–5 dB	+5 dB	–5 dB	+5 dB
0.0	2.22	2.30	−0.88	−0.07
0.4	2.29	2.36	−0.66	0.02
0.5	2.29	2.37	−0.58	0.03
0.6	2.29	2.36	−0.51	0.03
0.7	2.25	2.31	−0.43	0.01
0.8	2.11	2.14	−0.37	−0.07
0.9	1.59	1.54	−0.36	−0.24

An optimum for the higher signal-to-noise ratio is reached with a factor of $G_r = 0.6$, while the predicted quality at the lower SNR increases further with

even stronger smoothing. However, the predicted quality of the processing in drill noise decreases, probably due to less noise reduction caused by the gain smoothing.

As Table 6.6 shows, further improvements in the cafeteria noise are reached if the gain factors are additionally smoothed over time.

Table 6.6: *Transformed LAR objective quality measure data for the modified noise reduction algorithm with smoothing the gain factors over time using the setup of Experiment 1. Higher values indicate better performance. The gain factors are smoothed over time using a first-order recursive low-pass filter with time constant τ.*

τ	Noise, SNR			
	Drill		Cafeteria	
ms	–5 dB	+5 dB	–5 dB	+5 dB
0	2.29	2.36	−0.51	0.03
10	2.26	2.30	−0.47	0.07
20	2.22	2.27	−0.40	0.11
25	2.21	2.26	−0.37	0.13
30	2.19	2.24	−0.35	0.13
35	2.18	2.23	−0.33	0.13
40	2.16	2.22	−0.31	0.13
50	2.13	2.20	−0.29	0.12
100	2.03	2.10	−0.27	0.04

For the fixed frequency smoothing factor $G_r = 0.6$, the predicted quality in the cafeteria noise at the lower SNR still increases, the stronger the smoothing. For the higher SNR, however, a maximum is reached with a time constant of about 30 ms. On the other hand, the predicted quality in the drill noise still drops further. Although the noise reduction algorithm again performs worse than no processing in the cafeteria noise at −5 dB SNR, it does perform better than no noise reduction at the higher signal-to-noise ratio with the additional smoothing of the gain factors over time. But on the other hand, these modifications clearly decrease the effective noise reduction.

6.5.3 Discussion

The presented example shows how objective measures can be used to optimize signal processing parameters in noise reduction applications. Although a maximum in predicted quality was found for specific parameter settings in cafeteria noise at +5 dB SNR, the same settings do not maximize predicted quality in the other noise conditions. Nevertheless, the LAR predictions can aid in finding a compromise in parameter settings based on the performance in different noise conditions.

Objective measures are definitely a comfortable tool to aid researchers in

"sampling" the often large parameter space for the development of noise reduction algorithms. If checked with informal listening, this parameter space sampling can lead to a preselection of noise reduction algorithms which are worthwhile a comprehensive subjective evaluation as suggested, for example, in Chapters 4 and 5.

6.6 Conclusions

Different aspects of the noise reduction processing seem to be reflected in different objective measures.

The PMF and LAR objective quality measures in particular have shown to successfully reflect different subjective results and hence are promising candidates for future "objective" evaluations of noise reduction algorithms.

The PMF objective measure (Hansen and Kollmeier, 2000) was found to almost perfectly predict the subjectively perceived amount of noise suppression for different single-microphone noise reduction algorithms in different noise conditions without reverberation (Experiment 1), but correspondence was poorer in Experiment 2 for binaural noise reduction algorithms and reverberant test signals. The LAR objective measure was superior to all other measures that were under investigation in predicting "naturalness of speech" and "overall preference".

The example in Section 6.5 shows that objective quality measures can be used to optimize signal processing parameters.

Chapter 7

Summary and conclusions

The aim of the current study was to improve the construction and the assessment methods of noise reduction schemes for future digital hearing aids.

The speech pause detection algorithm proposed in Chapter 2 detects speech pauses by tracking minima in a noisy signal's power envelope, specifically in its low-pass and high-pass power envelopes. It maintains a low false-alarm rate over a wide range of signal-to-noise ratios. This facilitates its application for noise estimation in noise reduction algorithms. Large false-alarm rates in the speech pause detection would lead to wrong noise spectrum estimates which include significant speech parts and hence cause artifacts in a subsequent noise reduction process. The proposed scheme maintains a relatively fixed position in ROC (receiver operating characteristic) space as opposed to the standardized algorithm of ITU G.729 Annex B, which yields very large false-alarm rates (together with large hit rates) at low SNRs.

Chapter 3 showed that the musical noise phenomenon, one widely reported artifact of most single-microphone noise reduction schemes based on spectral subtraction, can to a high degree be overcome by the Ephraim-Malah noise reduction algorithms (Ephraim and Malah, 1984, 1985). If combined with the procedure for automatically adjusting the noise spectrum estimate during speech pauses (Chapter 2), a self-adaptive noise reduction scheme is obtained.

Comprehensive evaluations of the Ephraim-Malah noise reduction algorithms with hearing-impaired subjects showed that besides better "sound quality" (Chapter 5), most obvious benefits are reductions in the mental effort needed to listen to speech in noise and hence in listener fatigue over longer periods of time (Chapter 4). To assess this feature, a new listening effort test was developed. Due to its design, which involves a strenuous listening task, the proposed listening effort test is believed to actually assess listening effort and not merely subjective preference in terms of better sound quality. Therefore, the test is recommended for evaluations of noise reduc-

tion algorithms in general. To further increase the sensitivity of the test, its experimental design could be changed from a category rating procedure to paired comparisons, whenever feasible in appropriate measurement time. For the analysis of the paired comparison data the well-established Bradley-Terry model is proposed, which was successfully applied in Chapter 5 to scale subjective quality data.

Although a significant amount of noise reduction is obtained with the Ephraim-Malah algorithms for various noise conditions, an increase in speech intelligibility measured with a sentence test was not found. Only the binaural directional filter and dereverberation algorithm (Wittkop, 2000) was found to provide speech intelligibility improvements. This, however, is in agreement with Weiss and Neuman (1993) who remark that only *multi*-microphone methods have been shown to be capable of improving speech intelligibility for a range of acoustic environments and (especially wideband) noises so far.

On the other hand, differences in terms of listening effort were found for different algorithms which did not show up in word recognition scores. These findings indicate that conventional speech recognition tests and tests of listening effort appear to measure different aspects of the effect of noise reduction schemes in speech perception.

The results of the paired comparisons presented in Chapter 5 show that noise reduction is worthwhile in all of the different noises that were investigated. The Ephraim-Malah single-microphone noise reduction algorithms can be recommended for use in rather stationary noises (drilling machine as well as other industrial noise). They fail, however, in strongly fluctuating noises (cafeteria babble) where the binaural directional filter and dereverberation algorithm may be used, particularly at lower SNRs.

The combined algorithm (binaural processing followed by Ephraim-Malah processing) was not able to merge the benefits of both schemes. It is concluded that it is not appropriate to use both noise reduction schemes at the same time. These findings stress the importance of developing a more sophisticated combination of the noise reduction algorithms with an intelligent control algorithm.

It is striking that the hearing-impaired subjects did not perceive any distortions in the speech due to the processing of the noise reduction algorithms in drilling noise: The "naturalness of the speech" was judged better with the Ephraim-Malah algorithms than without noise reduction processing. The opposite was found with the normal-hearing subjects. This confirms once more that tests with hearing-impaired subjects should be performed when hearing-aid applications are considered, even in the case of noise reduction algorithms which are commonly believed to yield positive effects even for normal-hearing listeners.

Unfortunately, the algorithms with the largest amount of noise reduction (EL and E30) show also the strongest speech distortions in unfavorable noise

situations (cafeteria noise at −5 dB SNR). In such situations the naturalness of the speech is best preserved by the Ephraim-Malah algorithm E7 which, however, yields the lowest amount of noise reduction in the more stationary noises.

Future studies should consider a combination with algorithms which are able to classify the noise characteristics and the acoustic environment to automatically adjust parameters for optimal noise reduction performance and/or switch among different noise reduction schemes. Recently, promising candidates for this task were suggested (Ostendorf *et al.*, 1997; Ostendorf *et al.*, 1998; Tchorz, 2000). These algorithms are based on an analysis of modulation frequencies.

The Bradley-Terry scale values from the paired comparison judgments concerning speech intelligibility showed perfect concordance with the results of the "objective" sentence test measurements, indicating that subjects are well able to *judge* speech intelligibility by paired comparisons.

Since the whole potential of the noise reduction processing, especially concerning long-term speech intelligibility, cannot be assessed by a short-term laboratory evaluation, an implementation in a wearable digital hearing aid device is advisable for carrying out an evaluation in the field. This demand is supported by findings from Gatehouse (1992) who observed that benefits from providing a particular frequency shaping to hearing-impaired subjects did not emerge immediately, but over a time course of at least 6–12 weeks. He concludes that the existence of perceptual acclimatization effects call into question short-term methods of hearing aid evaluation. In particular, benefits may well emerge over time even if no benefits were found in laboratory evaluations.

It is assumed that the relatively low complexity of the proposed algorithms allows an application in digital hearing aids in the near future. A combination with a dynamic compression algorithm should be considered to guarantee audibility of low-level speech parts and to protect the subject from uncomfortable high signal levels (Marzinzik *et al.*, 1999a).

Finally, in Chapter 6 the predictive power of several "objective" speech quality measures was investigated with respect to the subjective noise reduction effect for hearing-impaired listeners. Particularly the PMF and LAR objective quality measures have shown to successfully reflect different subjective results.

The PMF objective measure (Hansen and Kollmeier, 2000) was found to almost perfectly predict the subjectively perceived amount of noise suppression for the Ephraim-Malah noise reduction algorithms in different noise conditions. The log-area-ratio measure (LAR) was superior to all other measures that were under investigation in predicting "naturalness of speech" and "overall preference".

Obvious differences in the subjective preferences between normal-hearing

and hearing-impaired subjects stress the importance of validating objective measures with the group for which predictions are aimed at.

As demonstrated, objective measures can be employed to assess the often large parameter space in the development of noise reduction algorithms aiming at a preselection of noise reduction algorithms and certain parameter settings, respectively, which are worthwhile a comprehensive subjective evaluation.

However, some puzzling effects and discrepancies between the "objective" predictions and the subjective perception emphasize the need to still carry out subjective evaluations in future hearing aid studies.

Taken together, an attempt was made here to improve both the assessment methods and the available algorithms for noise reduction in hearing aids. In particular, the whole development chain from the construction of algorithms, subjective assessment of algorithmic performance by normal-hearing and hearing-impaired listeners as well as objective assessment methods has been considered. It is hoped that these methods might be used in the future to provide further benefit to hearing-impaired patients from "intelligent" digital hearing aids.

Appendix A

Audiograms of the hearing-impaired subjects

Table A.1: *Audiograms of the sensorineural hearing-impaired subjects. In column* Ear, *'r' denotes right and 'l' left ear. The pure tone thresholds were obtained with an Interacoustics–Audiometer DA 930. The conductive components of the hearing losses were less than 10 dB.*

Subject	Ear	Sex	Age	Hearing Loss (dB HL)								
				Frequency (Hz)								
				125	250	500	1 k	2 k	3 k	4 k	6 k	8 k
BD	r	m	74	35	45	50	50	30	40	55	65	70
	l			50	55	55	55	45	70	70	80	75
GM	r	f	72	35	35	40	50	65	70	90	90	90
	l			45	45	45	45	60	80	90	90	90
HM	r	f	23	15	25	45	55	55	45	60	65	70
	l			15	20	35	50	50	55	60	70	80
KF	r	m	66	15	20	30	55	60	60	60	65	80
	l			25	30	35	50	55	55	70	50	55
KR	r	f	76	20	25	30	35	40	40	60	70	65
	l			30	30	20	30	30	60	75	75	75
WH	r	m	78	15	15	20	30	50	50	55	50	80
	l			40	50	45	45	45	45	60	65	80

Appendix B

Fitting Bradley-Terry models

In the paired comparison experiments, the subjects ("judges") were asked to state preference within pairs of algorithms ("items") with respect to a certain attribute. The Bradley-Terry model for binary paired comparisons assumes that the properties of the items with respect to the given attribute can be expressed by values on a linear scale. If these values are denoted by θ_i, the probability that algorithm i is preferred to algorithm j is assumed to be

$$p(i|i,j) = \frac{\exp(\theta_i)}{\exp(\theta_i) + \exp(\theta_j)}. \tag{B.1}$$

The unknown scale values θ_i appear as parameters of a probability model. From B.1 follows that

$$\log\left(\frac{p(i|i,j)}{p(j|i,j)}\right) = \theta_i - \theta_j. \tag{B.2}$$

Since the system of equations B.2 contains only differences of scale values, one scale value can be chosen freely. In the present studies, the reference algorithm UN was always set to the scale value zero.

The Bradley-Terry model is essentially heuristic. Whether the model is appropriate for a given data set can be checked by means of a goodness-of-fit test. Two strong assumptions are connected with Equation B.1 (Gediga, 1998): First, irrelevant alternatives, i.e. items with probability 0, must not have an influence on the preference probabilities of the other items. Second, too many circularities would violate the model, i.e. the preference of item i over item j, and of item j over k, *but* of item k over item i. Such circularities might occur due to a change of the judgment criterion by the judge. Of course, the items are no longer representable on a single linear scale if this happens too often in a given data set.

Several ways of fitting a Bradley-Terry (BT) model have been suggested in the literature so far. For small sample sizes (i.e., a small number of items that are to compared and a small number of judges), Bradley and Terry (1952) and Bradley (1954b) provided tables with the values for the item ratings which were calculated by means of maximum likelihood estimation. Tutz (1986) proposed to use a representation of the BT model as a *linear model*. For the solution of the model, Tutz prefers the method of Grizzle *et al.* (1969) based on a weighted least squares estimation instead of a method based on maximum likelihood estimation. Duineveld *et al.* (2000) used a representation of the Bradley-Terry model as a *log linear model* which can easily be solved with modern statistical software. The log linear model representation was already suggested by Fienberg and Larntz (1976). Gediga (1998) shows how to use *logistic regression* to fit a Bradley-Terry model with the SPSS software package.

Critchlow and Fligner (1991) have shown that the *generalized linear model* (GLM) provides a natural framework for modeling a variety of paired comparison experiments

and describe the fitting of a Bradley-Terry model by use of the GLIM computer package which uses an iteratively reweighted least squares algorithm to obtain maximum likelihood estimates, and also provides likelihood ratio test statistics for hypotheses of interest. The same can be obtained using the software package "R" (Ihaka and Gentleman, 1996; Hornik, 2000). It is very similar to the widespread S language which is best known through its commercial S-PLUS implementation. R has a home page at `http://www.r-project.org/`. It is free software and an official part of the GNU project ("GNU S"). Source code as well as binaries for many computer platforms are available. The fitting of a Bradley-Terry model with this software is presented below.

According to Larntz (1978), several statistics are commonly used to judge the goodness of fit for counted data models (including the BT model), and some statisticians follow the practice of reporting two or more statistics. The null hypothesis is that the model fits the data. The usual chi-squared statistic (Pearson statistic) is defined by

$$X^2 \quad = \quad \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \tag{B.3}$$

An alternative statistic is the likelihood ratio statistic, also called residual deviance

$$G^2 \quad = \quad 2 \cdot \sum_{\text{all cells}} \text{observed} \cdot \ln(\text{observed}/\text{expected}) \tag{B.4}$$

The Pearson statistic was suggested as a goodness-of-fit test for the BT model by Bradley (1954a). However, Bradley notes that difficulties are encountered when observed cell frequencies are small. Koehler and Ridpath (1982), for example, report both statistics for the fit of the Bradley-Terry model to basketball results. For large sample sizes, both X^2 and G^2 are approximately distributed as central χ^2 random variables with the same degrees of freedom when the model is correct, i.e. under the null hypothesis. Since only six subjects participated in the paired comparisons in the studies considered in this thesis, the large sample χ^2 approximation could be unreliable. In small samples, however, the X^2 statistic is still closer to the χ^2 distribution than is G^2 (Lawal, 1984). Larntz (1978) concludes that a P-value based on the asymptotic chi-squared approximation is "on average" about right for the Pearson statistic, but is understated for the likelihood ratio statistic when there are small cell expectations, i.e. X^2 has Type I error rates closest to the nominal levels based on the asymptotic chi-squared approximation while G^2 yields too many rejections under the null distribution. TenVergert *et al.* (1993) recommend to consider both statistics. Fienberg and Larntz (1976) regard a P-value above 0.05 for G^2 as reasonably good. Bäuml (1991) chose an α-level of 0.1. According to Gediga (1998), however, the α-level should be set at least to 0.2 in any test in which the null hypothesis carries the research question.

In the experiments reported in this thesis, an α-level of 0.1 was chosen for the G^2 statistic and an α-level of 0.2 for the Pearson X^2 statistic. Different α-levels were chosen because of the described discrepancies between these both statistics.

If the Bradley-Terry model does not fit the paired comparisons data, this may essentially be due to two reasons:

1. It is possible that the subjects do not judge according to a common criterion. This problem can be illustrated with the following example. Consider the comparison of four bicycles with regard to overall preference. Some subjects may base their preference on the number of gears, others perhaps on how comfortable the saddle is. Since there is no common criterion, the Bradley-Terry model may result in a bad fit. Kendall and Babington Smith defined a coefficient of agreement u to test agreement among subjects (Kendall and Gibbons, 1990; David, 1988). Exact significance levels for the coefficient of agreement may for example be found in Table 10D of Kendall and Gibbons (1990). Alternatively to the coefficient of agreement u, Cochran's Q is often used to test for agreement among subjects in paired comparisons. Due to the

too small number of subjects in this thesis, however, the chi square approximation for Q is not valid here (Siegel, 1956).

2. It may be that some subjects do not respond consistently, i.e. there are too many inconsistencies in the paired comparisons, called "circular triads". This can be assessed by a coefficient of consistence ζ (Kendall and Gibbons, 1990), which was also introduced by Kendall and Babington Smith.

A good fit of the Bradley-Terry model to the data means that the assumptions of the model are met. In a second step, it is checked whether significant differences between algorithms exist. An approximate test is obtained by taking the difference between the null deviance and the residual deviance of the model fit which is approximately distributed as χ^2 for large samples. For small sample sizes as is the case here, Bradley and Terry (1952) and Bradley (1954b) provided tables based on the exact distribution to test for differences among algorithms. The null hypothesis is that no differences among algorithms exist. Tables for the respective B_1 statistic with exact significance levels for the case of six subjects and four algorithms are included in Bradley and Terry (1952). However, it is only possible to check for *any* significant differences among algorithms, not for pairwise significant differences. Calculating standard errors for the scale values could probably provide this information. Generally, standard errors from asymptotic theory (assuming large samples) are calculated automatically by most procedures to fit a Bradley-Terry model. Bootstrapping, a nonparametric way to obtain standard errors if the distribution of values is unknown, results in somewhat larger standard errors. Teebagy and Chatterjee (1989) found that asymptotic estimates underestimate while bootstrap estimates overestimate the real standard errors. Hence, if standard errors are of interest both procedures should be applied to obtain lower and upper boundaries for the real standard errors. However, it is abstained from reporting standard errors in this thesis (note, that other publications using the Bradley-Terry model neither report standard errors of the scale values). Independent of *statistical* significance of small differences between algorithms, the Bradley-Terry scale values nevertheless best represent (in a least squares sense) the paired comparison data of the participating subjects.

The statistics for the BT models fitted to the paired comparison data of Experiments 1 and 2 (cf. Chapter 5) are given in Tables C.1 to C.3. An interesting result of the investigations is that the approximate χ^2 deviance test for differences among algorithms leads to the same conclusions as the exact test in all conditions of both experiments. Hence, the tedious work with the tables from the exact distribution (Bradley and Terry, 1952; Bradley, 1954b) can be replaced by the approximate χ^2 deviance test (which can automatically be performed by the fitting routines), even at such small sample sizes of only six subjects.

```
#####################################################################
# Fit a Bradley-Terry model for paired comparison data using "R" #
#####################################################################
# Type this script manually in the R terminal or gui,            #
# or call this script in R if saved as "BT.R" with the command:  #
# source("BT.R", print.eval=TRUE)                                 #
#####################################################################

# Maximum number of iterations
glm.control(maxit=1000)

# Example with 6 judges comparing 4 items:
#
#                   Number of Preferences
# Pair (i,j)          n(i|i,j) n(j|i,j)
```

```
# (1,2) (E7,DD)          6         0
# (1,3) (E7,DDE7)        5         1
# (1,0) (E7,UN)          5         1
# (2,3) (DD,DDE7)        3         3
# (2,0) (DD,UN)          5         1
# (3,0) (DDE7,UN)        5         1

# Set up the data
#################
wins   <- c(6,5,5,3,5,5)
losses <- c(0,1,1,3,1,1)
wl <- cbind(wins,losses)

# Set up the design matrix.
# (Item UN is supposed to be the
# "reference" with scale value 0)
#################################
E7   <- c(1,1,1,0,0,0)
DD   <- c(-1,0,0,1,1,0)
DDE7 <- c(0,-1,0,-1,0,1)
UN   <- c(0,0,-1,0,-1,-1)

# Calculate the Bradley-Terry model
# using a generalized linear model
###################################
bt <- glm(wl~E7+DD+DDE7+UN-1,family=binomial)
summary(bt)

# Show model statistics with asymptotic P-values
################################################
noquote("G^2=")
Gsq<-deviance(bt)
Gsq
noquote("Significance of G^2:")
Gp<-pchisq(deviance(bt),df.residual(bt),lower.tail=FALSE)
Gp
noquote("-----------------------------------------------")
noquote("X^2=")
Xsq<-sum(residuals(bt,type="pearson")^2)
Xsq
noquote("Significance of X^2:")
Xp<-pchisq(Xsq,df.residual(bt),lower.tail=FALSE)
Xp
noquote("-----------------------------------------------")
noquote("Chi-square of differences between treatments:")
delta<-bt$null-bt$deviance
delta
noquote("Significance of differences between treatments:")
deltap<-pchisq(bt$null-bt$deviance,df.residual(bt),lower.tail=FALSE)
deltap
noquote("-----------------------------------------------")
noquote("Only to test the appropriateness")
noquote("of the large sample approximation:")
noquote("Rank sum values for exact significance")
```

```
noquote("to look up in Bradley & Terry Tables:")
E7<-sum(wins[1:3])+2*sum(losses[1:3])
DD<-sum(wins[4:5])+losses[1]+2*(sum(losses[4:5])+wins[1])
DDE7<-wins[6]+losses[4]+losses[2]+2*(losses[6]+wins[4]+wins[2])
UN<-losses[3]+losses[5]+losses[6]+2*(wins[3]+wins[5]+wins[6])
sort(cbind(UN,E7,DD,DDE7))
noquote("*********************************************")
```

Above example yields following output if called in "R":

```
R : Copyright 2000, The R Development Core Team
Version 1.1.0  (June 15, 2000)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type    "?license" or "?licence" for distribution details.

R is a collaborative project with many contributors.
Type    "?contributors" for a list.

Type    "demo()" for some demos, "help()" for on-line help, or
        "help.start()" for a HTML browser interface to help.
Type    "q()" to quit R.

>source("BT.R", print.eval=TRUE)

Call:
glm(formula = wl ~ E7 + DD + DDE7 + UN - 1, family = binomial)

Deviance Residuals:
[1]   1.29584  -0.06962  -1.09828   0.25627   0.46589   0.25869

Coefficients:
     Estimate Std. Error z value Pr(>|z|)
E7     3.0208     0.9405   3.212  0.00132 **
DD     1.1250     0.7254   1.551  0.12092
DDE7   1.3345     0.7382   1.808  0.07064 .
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19.9619  on 6  degrees of freedom
Residual deviance:  3.2399  on 3  degrees of freedom
AIC: 18.859

Number of Fisher Scoring iterations: 4

[1] G^2=
[1] 3.239922
[1] Significance of G^2:
[1] 0.3560923
[1] ---------------------------------------------
[1] X^2=
```

```
[1] 3.188103
[1] Significance of X^2:
[1] 0.3635227
[1] ---------------------------------------------
[1] Chi-square of differences between treatments:
[1] 16.72197
[1] Significance of differences between treatments:
[1] 0.0008061529
[1] ---------------------------------------------
[1] Only to test the appropriateness
[1] of the large sample approximation:
[1] Rank sum values for exact significance
[1] to look up in Bradley & Terry Tables:
[1] 20 27 28 33
[1] *********************************************
>
```

Hence, the difference scale values for this specific example are UN = 0, E7 = 3.0208, DD = 1.1250, and DDE7 = 1.3345. The goodness-of-fit statistics are $X^2 = 3.1881$ with $P = 0.3635$, and $G^2 = 3.2399$ with $P = 0.3561$, which prove the appropriateness of the Bradley-Terry model for this set of paired comparisons data. According to the difference between null and residual deviance which is 16.72 ($P = 0.0008$) there are significant differences between algorithms.

The following script shows how to obtain bootstrap standard errors using the R software package.

```
##############################################################################
# Apply bootstrapping for obtaining standard errors of BT scale values.      #
##############################################################################
# Type this script manually in the R terminal or gui,                        #
# or call this script in R if saved as "BOOTBT.R" with the command:          #
# source("BOOTBT.R", print.eval=TRUE)                                        #
##############################################################################

# Maximum number of iterations
glm.control(maxit=1000)
options(warn=-1)

# The add-on package "boot" is necessary for bootstrapping
library(boot)

# Example with 6 judges comparing 4 items.

# In the following table, a "1" means that the first
# algorithm was preferred, a "0" means that the second
# algorithm was preferred by the respective judge.

# Pairwise    | Judge
# comparisons | 1 2 3 4 5 6
# ------------------------
# (E7,DD)     | 1 1 1 1 1 1
# (E7,DDE7)   | 1 1 0 1 1 1
# (E7,UN)     | 1 1 0 1 1 1
# (DD,DDE7)   | 0 0 1 1 0 1
```

```
# (DD,UN)     | 1 1 1 1 1 0
# (DDE7,UN)   | 1 1 1 1 1 0

# Set up the data
#################
Judge   <- c(1, 2, 3, 4, 5, 6)
E7.DD   <- c(1, 1, 1, 1, 1, 1)
E7.DDE7 <- c(1, 1, 0, 1, 1, 1)
E7.UN   <- c(1, 1, 0, 1, 1, 1)
DD.DDE7 <- c(0, 0, 1, 1, 0, 1)
DD.UN   <- c(1, 1, 1, 1, 1, 0)
DDE7.UN <- c(1, 1, 1, 1, 1, 0)
pc.data <- data.frame(Judge,E7.DD,E7.DDE7,E7.UN,DD.DDE7,DD.UN,DDE7.UN)
rm(Judge,E7.DD,E7.DDE7,E7.UN,DD.DDE7,DD.UN,DDE7.UN)
attach(pc.data)

# Set up the bootstrap function
###############################
bt.boot.fun<-function(dat,inds)
{
   assign(".inds",inds,envir=.GlobalEnv)
   wins<-c(sum(E7.DD[.inds]),sum(E7.DDE7[.inds]),sum(E7.UN[.inds]),
           sum(DD.DDE7[.inds]),sum(DD.UN[.inds]),sum(DDE7.UN[.inds]))
   losses<-length(Judge)-wins
   # Set up the design matrix.
   # (Item UN is supposed to be the
   # "reference" with scale value 0)
   ################################
   E7   <- c(1,1,1,0,0,0)
   DD   <- c(-1,0,0,1,1,0)
   DDE7 <- c(0,-1,0,-1,0,1)
   UN   <- c(0,0,-1,0,-1,-1)
   wl<-cbind(wins,losses)
   # Calculate the Bradley-Terry model
   # using a generalized linear model
   ###################################
   bt<-glm(wl~E7+DD+DDE7+UN-1,family=binomial)
   remove(".inds",envir=.GlobalEnv)
   c(bt$coef)
}

# Run bootstrap with 200 replications
#####################################
bt.boot<-boot(pc.data,bt.boot.fun,R=200)
bt.boot
```

This example might result in following output:

```
> source("BOOTBT.R", print.eval=TRUE)

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = pc.data, statistic = bt.boot.fun, R = 200)
```

```
Bootstrap Statistics :
    original   bias    std. error
t1* 3.020817 4.461506    7.398908
t2* 1.125016 1.018128    3.072738
t3* 1.334454 1.226167    3.829435
WARNING: All values of t4* are NA
>
```

This indicates the following bootstrap standard errors for the original BT scale values of the different algorithms: 7.399 for E7, 3.073 for DD, and 3.829 for DDE7. As this example shows, the bootstrap standard errors are very high. These standard errors suggest that no significant differences exist between the four algorithms in this case. However, in this experiment the large bootstrap standard errors are caused by the small number of judges. Actually, the scale values depend sensitively on the choice of judges, here. Nevertheless, the scale values for the different algorithms best represent the paired comparisons data of the participating subjects.

Appendix C

BT test statistics of Experiments 1 and 2

Table C.1: *Bradley-Terry model statistics for the paired comparisons of the normal-hearing subjects in Experiment 1. Conditions are abbreviated as denoted in Table 5.1;* O*: Overall preference,* N*: Naturalness of the speech,* R*: Reduction of the noise.*

Condition		Goodness of fit				Differences among algorithms			
		Pearson		Likelihood ratio		Deviance test		Bradley's exact test	
		X^2	asy. P	G^2	asy. P	χ^2	asy. P	B_1	exact P
D–5	O	1.592	0.6612	1.656	0.6467	7.145	0.0674	9.14-9.39	0.0590-0.1069
	N	0.096	0.9923	0.093	0.9926	35.291	<0.0001	2.98-3.27	<0.0001
	R	0.257	0.9679	0.405	0.9393	38.422	<0.0001	1.66-3.01	<0.0001
D+5	O	1.338	0.7203	1.315	0.7257	3.519	0.3183	9.94-10.17	0.2923-0.4382
	N	0.617	0.8926	0.701	0.8729	23.603	<0.0001	5.46-5.90	0.0001
	R	0.000	1.0000	0.001	1.0000	44.499	<0.0001	1.17	<0.0001
C–5	O	0.578	0.9015	0.874	0.8316	17.084	0.0007	6.64-7.54	0.0004-0.0023
	N	1.991	0.5743	2.303	0.5119	21.002	0.0001	6.20-6.30	0.0002
	R	1.164	0.7618	1.183	0.7571	3.800	0.2839	10.01	0.3264
C+5	O	0.833	0.8417	0.825	0.8435	11.276	0.0103	8.18	0.0098
	N	0.532	0.9119	0.813	0.8463	36.049	<0.0001	3.01	<0.0001
	R	2.063	0.5595	2.185	0.5348	13.406	0.0038	7.93	0.0055

Table C.2: *Bradley-Terry model statistics for the paired comparisons of the hearing-impaired subjects in Experiment 1. Notation as in Table C.1.*

Condition		Goodness of fit				Differences among algorithms			
		Pearson		Likelihood ratio		Deviance test		Bradley's exact test	
		X^2	asy. P	G^2	asy. P	χ^2	asy. P	B_1	exact P
D–5	O	1.219	0.7485	1.603	0.6588	22.425	0.0001	5.67-6.20	0.0001-0.0002
	N	2.478	0.4792	2.850	0.4153	15.048	0.0018	7.32-7.77	0.0018-0.0037
	R	0.229	0.9727	0.369	0.9466	35.282	<0.0001	2.35-3.69	<0.0001
D+5	O	0.239	0.9711	0.239	0.9710	27.131	<0.0001	4.71-5.12	<0.0001
	N	2.496	0.4760	2.445	0.4854	7.759	0.0513	9.05-9.21	0.0536-0.0718
	R	0.151	0.9851	0.142	0.9864	35.509	<0.0001	2.43-3.68	<0.0001
C–5	O	0.508	0.9172	0.543	0.9093	17.223	0.0006	6.85-7.28	0.0006-0.015
	N	0.342	0.9519	0.357	0.9490	17.577	0.0005	6.03-7.83	0.0001-0.0048
	R	0.248	0.9695	0.249	0.9693	2.125	0.5470	10.32-10.39	0.5826-0.6081
C+5	O	1.784	0.6184	1.977	0.5771	4.593	0.2042	9.70-9.94	0.1888-0.2923
	N	0.358	0.9488	0.356	0.9492	5.139	0.1619	9.54-9.86	0.1455-0.2382
	R	0.966	0.8095	1.376	0.7111	14.123	0.0027	7.77	0.0037

Table C.3: *Bradley-Terry model statistics for the paired comparisons of the hearing-impaired subjects in Experiment 2. Conditions are abbreviated as denoted in Table 5.1;* R*: Reduction of the noise,* O*: Overall preference,* N*: Naturalness of the speech,* I*: Intelligibility of the speech.*

Condition		Goodness of fit				Differences among algorithms			
		Pearson		Likelihood ratio		Deviance test		Bradley's exact test	
		X^2	asy. P	G^2	asy. P	χ^2	asy. P	B_1	exact P
Ind0	R	0.000	1.0000	0.001	1.0000	49.906	<0.0001	0.000	<0.0001
	O	3.710	0.2945	5.032	0.1695	6.004	0.1114	9.533	0.1328
	N	1.110	0.7748	1.119	0.7724	4.510	0.2114	9.858	0.2382
	I	0.301	0.9599	0.303	0.9595	9.110	0.0279	8.859	0.0383
Ind10	R	0.000	1.0000	0.001	1.0000	44.499	<0.0001	1.174	<0.0001
	O	5.638	0.1306	4.295	0.2313	21.753	0.0001	6.114	0.0001
	N	1.512	0.6794	1.591	0.6615	11.413	0.0097	8.359	0.0152
	I	3.188	0.3635	3.240	0.3561	16.722	0.0008	7.206	0.0013
Caf0	R	0.000	1.0000	0.000	1.0000	42.268	<0.0001	1.659	<0.0001
	O	3.165	0.3669	3.351	0.3406	4.510	0.2114	9.858	0.2382
	N	1.106	0.7756	1.092	0.7791	5.217	0.1566	9.704	0.1888
	I	3.104	0.3759	3.323	0.3444	5.217	0.1566	9.704	0.1888
Caf10	R	0.000	1.0000	0.001	1.0000	44.499	<0.0001	1.174	<0.0001
	O	1.474	0.6882	1.562	0.6681	1.349	0.7175	10.544	0.7502
	N	1.441	0.6959	1.522	0.6772	3.428	0.3302	10.093	0.3669
	I	1.675	0.6425	1.704	0.6359	0.334	0.9535	10.764	0.9919
Ssn0	R	0.000	1.0000	0.000	1.0000	41.588	<0.0001	1.806	<0.0001
	O	3.765	0.2879	3.869	0.2760	19.948	0.0002	6.505	0.0003
	N	0.228	0.9729	0.381	0.9441	38.712	<0.0001	2.431	<0.0001
	I	0.348	0.9507	0.349	0.9506	5.960	0.1136	9.543	0.1455
Ssn10	R	0.000	1.0000	0.001	1.0000	44.499	<0.0001	1.174	<0.0001
	O	0.876	0.8312	0.911	0.8227	11.413	0.0097	8.359	0.0152
	N	2.520	0.4718	2.608	0.4561	7.485	0.0580	9.212	0.0718
	I	0.814	0.8462	0.811	0.8468	6.370	0.0949	9.454	0.1175

Appendix D

Objective measures data

Table D.1: *Objective quality measures data from Experiment 1 (without re-scaling).*

Objective Measure	Algorithm	Noise, SNR			
		Drill		Cafeteria	
		–5 dB	+5 dB	–5 dB	+5 dB
PMF	UN	0.2079	0.3011	0.2587	0.5192
	E7	0.2687	0.4458	0.2920	0.6080
	EL	0.3030	0.4710	0.2879	0.6234
	E30	0.3181	0.4844	0.2850	0.6361
PSQM	UN	0.6115	0.3884	0.4317	0.3731
	E7	0.2855	0.1448	0.4982	0.3003
	EL	0.1928	0.1321	0.5163	0.2771
	E30	0.1726	0.1451	0.5076	0.2634
LAR	UN	11.5064	9.2200	7.7968	6.5827
	E7	9.3510	7.1998	9.1404	6.8857
	EL	8.8278	7.9517	9.8112	7.7309
	E30	9.1679	9.1813	9.9233	8.2197
LLR	UN	2.5228	1.5862	1.7427	1.1940
	E7	1.8355	1.0659	1.7475	1.0531
	EL	1.7461	1.1553	1.8797	1.1873
	E30	1.8346	1.3244	1.9096	1.2609
IS	UN	4.4302	2.7344	4.1035	2.6635
	E7	2.5856	2.3974	3.3215	2.7269
	EL	4.2899	7.2097	4.4961	9.3055
	E30	10.9475	27.2709	4.9813	23.3697
WSS	UN	45.3000	31.8938	63.2885	46.3178
	E7	51.1321	36.4713	91.3016	65.5325
	EL	50.8525	37.4865	104.1227	74.0975
	E30	52.9305	39.4636	107.8562	77.5392
SSNR	UN	−4.4884	1.7042	−6.4054	−1.8111
	E7	1.5187	6.8755	−4.7819	0.8904
	EL	3.4555	8.2815	−4.6538	1.5513
	E30	3.1648	7.2015	−4.7948	1.5331

Table D.2: *Objective quality measures data from Experiment 2 (without re-scaling).*

Objective Measure	Algorithm	Noise, SNR					
		Cafeteria		Speech-shaped		Industry	
		0 dB	10 dB	0 dB	10 dB	0 dB	10 dB
PMF	UN	0.4811	0.7364	0.5251	0.7952	0.2523	0.4993
	E7	0.5595	0.8011	0.6200	0.8695	0.2702	0.6346
	DD	0.6066	0.8401	0.7203	0.9295	0.3359	0.7505
	DE	0.6359	0.8738	0.7493	0.9444	0.3699	0.8066
PSQM	UN	0.2801	0.1308	0.2136	0.0829	0.2965	0.1367
	E7	0.2428	0.0986	0.1628	0.0610	0.1997	0.0836
	DD	0.2426	0.0979	0.1822	0.0731	0.2287	0.1143
	DE	0.2265	0.1040	0.1665	0.0837	0.1736	0.0897
LAR	UN	5.0260	3.1155	4.5451	2.6135	7.3434	4.5615
	E7	5.7155	3.5922	5.2132	2.8807	5.8797	3.9923
	DD	4.5984	3.8017	4.1088	3.4640	5.3372	4.2543
	DE	5.7471	5.0683	5.8738	4.8400	5.6882	4.6590
LLR	UN	0.6823	0.2846	0.4989	0.1818	1.0742	0.4413
	E7	0.7585	0.2833	0.5631	0.1656	0.7545	0.3273
	DD	0.5132	0.2970	0.3982	0.2292	0.6824	0.4122
	DE	0.7302	0.4598	0.6090	0.3862	0.6793	0.4577
IS	UN	1.7811	0.9033	1.4730	0.6911	2.7478	1.2751
	E7	2.8954	1.6143	2.7028	1.6953	1.6569	2.4275
	DD	1.2343	0.6942	1.0011	0.5369	1.7832	0.8127
	DE	5.0175	2.8453	5.4824	2.4582	2.8297	4.2154
WSS	UN	46.8849	22.7723	47.0356	21.7986	40.5950	20.1103
	E7	59.8665	29.2939	53.7248	25.5622	46.1204	23.6618
	DD	47.1383	24.3565	47.9913	23.4008	39.3048	22.6758
	DE	62.5807	32.8239	59.1484	28.0388	48.5680	29.5260
SSNR	UN	−1.8496	6.0069	−2.1467	5.7745	−3.1683	6.0370
	E7	0.3086	6.8655	0.6989	5.3984	2.0512	7.7422
	DD	−0.3682	5.7926	−1.6780	4.6465	−0.5505	5.9779
	DE	1.1162	4.5183	0.7382	4.2712	1.7924	5.3571

Notes

1 According to Egan (1975), the receiver operating characteristic (ROC) is a function which summarizes the possible performances of an observer faced with the task of detecting a signal in noise. In general, the ROC is given as a plot of the hit rate versus the false-alarm rate which is obtained by modifying the decision criterion. In the present study, the signal to be detected is a "speech pause" occurring in a noisy speech signal.

2 The focus here is on *noise reduction* which means that the algorithms considered here focus on reducing additive noise in a speech-plus-noise signal – as opposed to *speech enhancement* which focuses on the enhancement of typical speech cues by for example amplifying consonants, increasing consonant durations, spectral contrast enhancement etc. (Montgomery and Edge, 1988; Revoile and Holden-Pitt, 1993; Baer *et al.*, 1993; Franck *et al.*, 1999). However, many publications do not use these terms as strictly.

3 Vary (1985) reported that if the actual phase of clean speech is replaced by zero-phase, then the resynthesized speech sounds completely voiced and monotonous. If the phase is randomly chosen uniformly distributed between $\pm\pi$ a rough and completely unvoiced speech is obtained. If noise is added to the actual phase, nothing is to be recognized below a certain threshold. Above that noise level some roughness is perceived. Vary concludes that noise suppression can be achieved by estimating just the spectral magnitude and by leaving the noisy phase as it is.

4 Numerical algorithms to calculate the modified Bessel functions are for example given in Press *et al.* (1992).

5 Actually, this definition of R_{post} following Cappé (1994) is slightly different from that originally given by Ephraim and Malah (1984). However, the gain formulae are the same and the modification is only made for the purpose of easier interpretation.

6 The exponential integral $E_1(x)$ is, for example, described in Abramowitz and Stegun (1964). It can be calculated numerically by, e.g., the procedure `expint` in Press *et al.* (1992).

7 In conventional speech intelligibility tests, the rate of correctly repeated words – isolated or in sentences – presented in noise at a certain level with certain signal-to-noise ratios is determined. As most subjects will try not to disappoint the experimenter, it can be expected that they will do their very best to overcome any fatigue caused by increased listening effort in noise. This is known as the Hawthorne effect (Roethlisberger and Dickson, 1939). Hence, a conventional speech intelligibility test is supposed to be inadequate to assess listening effort.
Figure D.1 shows an optical illustration of noise effects. Its discussion might help to understand the acoustical case. Figure D.1a gives an example of how noise disturbs a target signal. The text is blurred and the noise suppression tries to bring it back into focus. Most probably, a reading test with subjects will not reveal any significant differences in the error rate between blurred and the noise-reduced text as the subjects are able to counteract the blurring of the text by more concentration. However, after having read a long text, the subjects with the blurred text are generally more

(a) High signal-to-noise ratio (b) Low signal-to-noise ratio

Figure D.1: *Optical illustration of noise effects.*

tired than others having read a clean text. But how can this fatigue be assessed? If the subjects have to perform a reading test after a preceding "fatiguing" phase, again, the subjects are most probably able to counteract the fatigue by more concentration. As most subjects will try not to disappoint the experimenter, it can be expected that they will do their very best to overcome the fatigue (Hawthorne effect; Roethlisberger and Dickson, 1939). Hence, a conventional reading test (and in the acoustical case this would mean a conventional speech intelligibility test) seems not to be adequate to assess this fatiguing effect of the noise. It is therefore suggested to let the subjects read two long texts (blurred and clean) and let them self-assess their fatigue (or alternatively the effort that was needed or the "ease of reading"). This procedure leads to the listening effort test presented in this thesis. A different approach is based on the conjecture that response times are longer with the noisy signal compared to the clean signal. With the noise from the optical example of Figure D.1a, however, it seems more likely that a presumed lengthening of the reading time is too small and perhaps even compensated by more concentration to be reliably detected. On the other hand, if the "signal-to-noise ratio" gets worse so that whole signal parts actually get lost, longer response times are probable. The reader can check this with Figure D.1b.

8 Most threshold-based approaches which are in use today are modifications of the simple one-half gain rule (Fabry and Schum, 1994). The actual frequency response of the equalizers was verified with a spectrum analyzer from Stanford Research Systems, Model SR780.

9 The equipment was calibrated using a Brüel & Kjær Measuring Amplifier Type 2610 and a Brüel & Kjær Artificial Ear Type 4153.

10 In the first experiment, noise samples recorded from a drilling machine were used since these represent a class of technical noises which are often rather stationary. In addition, performance in a cafeteria noise with babble in the background was tested. This noise is strongly fluctuating and represents a noisy environment which is typical of social gatherings.

11 Most statistical calculations and graphics in this chapter were performed with "R" (Ihaka and Gentleman, 1996; Hornik, 2000). The R software is very similar to the S language which is best known through its commercial S-PLUS implementation. R has a home page at `http://www.r-project.org/`. It is free software and an official part of the GNU project ("GNU S"). Source code as well as binaries for many computer platforms are available.

12 The median absolute deviation (MAD) is used for describing variation. It has the advantage over the ordinary range or standard deviation that it is not sensitive to extreme outliers and that it does not assume a Gaussian distribution of the data as the standard deviation does. Compared to the interquartile range (IQR) the MAD is even more robust. Furthermore, the calculation of the median absolute deviation is straightforward, whereas different statistical software packages seem not to be in agreement on how to calculate quartiles. Remarkably enough, the SPSS software

obviously uses an approach different from, e.g., R and Excel. While the latter two programs yield quartiles which give an IQR of 0.75 for algorithm UN (i.e., IQR of {2,3,3,5,4,3}, see Table 4.4), SPSS calculates quartiles which result in an IQR value of 1.5. MATLAB gives an IQR of 1.0. These discrepancies possibly make it hard to compare interquartile ranges from different studies (at least if small sample sizes are used, as is often the case in psychoacoustics), since each of these programs is widespread in use.

13 The MAD function as used here is available in the R software package. Care has to be taken as some programs use the acronym MAD for *mean* absolute deviation instead of the *median* absolute deviation.

14 The Friedman two-way analysis of variance by ranks test is also known under the names Friedman rank sum test or just Friedman test.

15 According to Motulsky (1995), many statisticians avoid terms like "very significant" or "extremely significant" and think that the word significant should never be prefaced by an adjective, since once an α-level is set, a result is just either statistically significant or not.

16 In general, Wilcoxon's matched pairs signed rank test can be used to find out which algorithms differ significantly from each other. However, it is not appropriate to repeatedly use a Wilcoxon test with the same significance level as if only two algorithms were tested. To compare various pairs of algorithms a correction for multiple comparisons has to be applied (Wright, 1992). R provides the function `pairwise.wilcox.test` for this purpose. Since most corrections are too conservative, Dunn's post test for multiple comparisons is performed here instead of a Wilcoxon test. This test is available for example in the InStat software. A free demo (which is not limited in the statistical calculations) is available at `http://www.graphpad.com`.

17 Following German translations were used for the measurements: Völlige Entspannung ist möglich, keine Anstrengung erforderlich (5); Aufmerksamkeit ist erforderlich, aber keine nennenswerte Anstrengung nötig (4); Mäßige Anstrengung ist erforderlich (3); Beträchtliche Anstrengung ist erforderlich (2); Trotz größter Anstrengung ist die Bedeutung unverständlich (1).

18 The exact P-values for the Friedman test can for example be found in Table A–22 from Marascuilo and McSweeney (1977).

19 In principal, it is not even *possible* to formally prove that a test actually assesses listening effort, as listening effort is assumed to be mainly a mental phenomenon (as opposed to a physical fatigue, which could be assessed by muscular measurements). A kind of proof could only be established by showing high correlation with any "objective" measure which is again *believed* to be strongly correlated with the (mental) listening effort. One example is pupillary dilation (Hoeks and Levelt, 1993), another is the counting of "errors" in subsequent or parallel mental tasks. But these measures again have to be validated by subjective judgments, if a correlation with *perceived* effort is wanted. Otherwise, any functional definition is only *postulated* to measure listening effort. The same problems apply to the similar concept of fatigue. Trenchantly, Muscio (1921) concluded that it is not possible to device an acceptable test of fatigue because there exist no observable criteria for fatigue, other than those provided by the test itself, against which the test might be validated. Holding (1983) points out that "feeling tired does not necessarily correlate with physiological impairment, nor with reduced efficiency in work output or other kinds of human performance. As a result, the research literature dealing with attempts to find objective tests for fatigue contains many disappointing outcomes." Whatever route is taken, finally one ends up at asking subjects for their opinion.

20 The exact P-value is also below 0.05. Exact P-values for the Wilcoxon test can for example be found in Table G from Siegel (1956).

21 In fact, Dunn's post test does not find the significant differences, although the Friedman test claims that they exist. Since the rank sum differences are greatest between

UN and DD as well as between E7 and DD, these are supposed to be the significant differences. In this experiment, Dunn's post test requires a difference of at least 11.8 to be significant at the 5-percent level. Actually, above differences are 11.0.

22 One requirement for paired comparisons is the ability to rapidly switch among various hearing aids. Zerlin (1962) recorded speech processed by different hearing aids on two separate tracks of a magnetic tape. Listeners switched between the two tracks of the tape and indicated preference for one of the two hearing aids. However, this was primarily a research tool and was impractical for clinical use.

23 Stern (1992) compared several models for the analysis of paired comparison data including the Bradley-Terry model and the well-known Thurstone-Mosteller model and found that all provide adequate and almost identical fits to the data.

24 A wide range of applications of the Bradley-Terry model is established in the literature. It was used for consumer preference and taste testing experiments (Bradley, 1953; Lukas, 1991; Duineveld *et al.*, 2000), preferences among political candidates (Bäuml, 1991), and sports league competitions (Koehler and Ridpath, 1982). In these areas, the Bradley-Terry model has become very popular. Although paired comparison experimental designs are quite common in acoustics research, too, the Bradley-Terry model has seldom be used for the analysis of the data, so far. Kousgaard (1987) applied the BT model to the analysis of loudspeaker listening tests. Pressnitzer and McAdams (1997) applied it to construct a roughness scale from subjective paired comparison data.

25 Most threshold-based approaches which are in use today are modifications of the simple one-half gain rule (Fabry and Schum, 1994). The actual frequency response of the equalizers was verified with a spectrum analyzer from Stanford Research Systems, Model SR780.

26 A remark on the number of test subjects ("judges") is necessary here. The model is generally not restricted to a minimum number of test subjects. Bradley and Terry (1952) actually provided tables for small sample sizes down to just *one* test subject. Of course, sample sizes are often much larger when considering marketing issues, taste testing (Bradley, 1953; Duineveld *et al.*, 2000) or sports league competitions (Koehler and Ridpath, 1982). However, no smallest sample size exists for which the Bradley-Terry model could reliably be applied. Strictly speaking, however, the scale values reflect only the judgments of the participating subjects: If the goodness-of-fit test confirms that the Bradley-Terry model can be applied, the scale values *best* reflect differences between treatments (algorithms) in a least squares sense according to the underlying paired comparison data. In fact, the order of algorithms derived from the BT scale values is always the same as that derived from just counting wins and losses in the paired comparisons. Of course, this increases the confidence in this method since nothing obscure is supposed to happen by the scaling procedure. Punch (1978) showed that intrasubject reliability of paired comparison preference judgments of aided speech is acceptably high. Hence, the only problematic point is that of generalizing the results obtained with a few subjects to the "whole universe", which is, on the other hand, not a specific problem of paired comparisons and the Bradley-Terry model, but of all subjective measurements with a small number of test persons in general.

27 The model statistics are given in Table C.3 in Appendix C. In general, there are two possible explanations why the Bradley-Terry model may not fit the data: First, it is possible that the subjects do not judge according to a common criterion. As subjects seem to base their overall preference on the naturalness of the speech and on the amount of noise reduction, overall preference may not be a one-dimensional criterion, with the consequence that the fits of the model are often poorer for these data. However, in this case (industrial noise at 10 dB SNR) the coefficient of agreement u (cf. Appendix B) amounts to 0.49 ($\Sigma = 67$) which indicates a significant agreement between subjects ($P = 0.00048$).

A second explanation for a poor model fit may be that some subjects do not respond consistently, i.e. there are too many inconsistencies in the paired comparisons. In the present case, the coefficients of consistence (cf. Appendix B) amount to $\zeta = 1$ for all subjects except for subject HM, i.e. no circular triads occured except for HM which showed two circular triads yielding $\zeta = 0$. In fact, if subject HM is excluded from the analysis, the Bradley-Terry model yields a good fit to the data with $X^2 = 1.311$, $P = 0.726$ and $G^2 = 1.699$, $P = 0.6371$. The scale values are then as follows: UN = 0, DD = 0, DDE7 = 2.2, and E7 = 13.5. The rank order is essentially the same as with HM included, but now the difference between E7 and UN is much larger. According to the difference between null and residual deviance which is 28.16 ($P < 0.0001$) there are significant differences between algorithms. This is also confirmed by Bradley's exact test which yields $B_1 = 2.917$ with $P < 0.0001$.

28 Naturalness and noisiness were reported to be two attributes that determine the two main factors (dimensions) when judging sentences processed by different speech coders (Halka and Heute, 1992).

29 `http://cslr.colorado.edu/rspl/rspl_software.html`

30 The absolute value of the correlation coefficient can be judged according to following commonly used classification. $|\rho| < 0.20$: slight correlation, almost no relationship; $0.21 - 0.40$: low correlation, only a small relationship; $0.41 - 0.70$: moderate correlation, substantial relationship; $0.71 - 0.90$: high correlation, strong relationship; $|\rho| > 0.90$: very high correlation, very strong relationship. The sign of the correlation coefficient indicates the direction of the relationship.

31 Discussing the use of the speech transmission index (STI) developed by Steeneken and Houtgast (1980) as an extension of the articulation index (AI) of French and Steinberg (1947), Schmidt-Nielsen (1987) points out that the correspondence between the STI and listener tests looks fairly good only when a very wide range of speech intelligibilities is tested. He emphasizes that the fit is simply not good enough for many of the most useful applications in a smaller range due to wide scatter with several rank order reversals. This points to a general problem: A wide range of conditions often leads easily to good overall correlation coefficients. Just one extreme point at one end or another of the tested range could increase the linear correlation drastically. This fact stresses the importance of looking at scatter plots and not to trust correlation coefficients alone.

References

Abdallah, I., Montrésor, S., and Baudry, M. (1997). Speech signal detection in noisy environment using a local entropic criterion. *Proceedings of the 5th European Conference on Speech Communication and Technology, EUROSPEECH '97, Rhodes, Greece*.

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions*, volume 55 of *Applied Mathematics Series*. Dover Publications, New York. Reprinted 1968.

Akbari Azirani, A., Le Bouquin Jeannès, R., and Faucon, G. (1996). Speech enhancement using a Wiener filtering under signal presence uncertainty. *Signal Processing VIII, Theories and Applications. Proceedings of EUSIPCO-96, Vol.II*, Trieste, Italy, LINT, 971–974.

Baer, T., Moore, B. C. J., and Gatehouse, S. (1993). Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times. *Journal of Rehabilitation Research* **30** (1), 49–72.

Bäuml, K.-H. (1991). Präferenzen zwischen politischen Kandidaten: Versuch einer Repräsentation durch BTL-Modell, Präferenzbäume und Eliminierung-nach-Aspekten. *Z. Psychol.* **199**, 337–352.

Bech, S. (1987). Planning of listening test – choice of rating scale and test procedure. Bech, S. and Pedersen, O. J., editors, *Perception of Reproduced Sound*, 62–70. ISBN 87-982562-1-1.

Beerends, J. G. (1998). Audio quality determination based on perceptual measurement techniques. Kahrs, M. and Brandenburg, K., editors, *Applications of Digital Signal Processing to Audio and Acoustics*, Kluwer Academic Publishers, Boston, chapter 1, 1–38.

Beerends, J. G. and Stemerdink, J. A. (1992). A perceptual audio quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc.* **40** (12), 963–978.

Beerends, J. G. and Stemerdink, J. A. (1994). A perceptual speech-quality measure based on a psychoacoustic sound representation. *J. Audio. Eng. Soc.* **42** (3), 115–123.

Berouti, M., Schwartz, R., and Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. *International Conference on Acoustics, Speech, and Signal Processing 1979, Conference Proceedings*, New

York, IEEE, 208–211.

Bodin, P. and Villemoes, L. F. (1997). Spectral subtraction in the time-frequency domain using wavelet packets. *IEEE Workshop on Speech Coding for Telecommunications. Proceedings*, New York, IEEE, 47–48.

Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing* **ASSP-27** (2), 113–120.

Boll, S. F. (1992). Speech enhancement in the 1980s: Noise suppression with pattern matching. Furui, S. and Sondhi, M. M., editors, *Advances in Speech Signal Processing*, Marcel Dekker, New York, chapter 10, 309–325.

Bradley, R. A. (1953). Some statistical methods in taste testing and quality evaluation. *Biometrics* **9**, 22–38.

Bradley, R. A. (1954a). Incomplete block rank analysis: On the appropriateness of the model for a method of paired comparisons. *Biometrics* **10**, 375–390.

Bradley, R. A. (1954b). The rank analysis of incomplete block designs. II. Additional tables for the method of paired comparisons. *Biometrika* **41**, 502–537.

Bradley, R. A. (1976). Science, statistics, and paired comparisons. *Biometrics* **32**, 213–232.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs, I. The method of pair comparisons. *Biometrika* **39**, 324–345.

Cappé, O. (1994). Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing* **2** (2), 345–349.

Carhart, R. (1946). Selection of hearing aids. *Archives of Otolaryngology* **44**, 1–18.

Colonius, H. (1980). Representation and uniqueness of the Bradley-Terry-Luce model for pair comparisons. *British Journal of Mathematical and Statistical Psychology* **33**, 99–103.

Cox, R. M., Alexander, G. C., and Rivera, I. M. (1991). Comparison of objective and subjective measures of speech intelligibility in elderly hearing-impaired listeners. *Journal of Speech and Hearing Research* **34** (4), 904–915.

Cox, R. M. and McDaniel, D. M. (1984). Intelligibility ratings of continuous discourse: Application to hearing aid selection. *Journal of the Acoustical Society of America* **76** (3), 758–766.

Critchlow, D. E. and Fligner, M. A. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika* **56** (3), 517–533.

David, H. A. (1988). *The Method of Paired Comparisons.* Griffin, 2nd edition.

Davídek, V., Šika, J., and Štusák, J. (1996). Noise cancellation system on TMS320C31. *Proceedings of the First European DSP Education and*

Research Conference, Paris, ESIEE, 134–138.

Davidson, R. R. and Farquhar, P. H. (1976). A bibliography on the method of paired comparisons. *Biometrics* **32**, 241–252.

Davies, D. R., Shackleton, V. J., and R., P. (1983). Monotony and boredom. Hockey, R., editor, *Stress and Fatigue in Human Performance*, John Wiley & Sons, New York, chapter 1, 1–32.

Dendrinos, M. and Bakamidis, S. (1994). Voice activity detection in coloured-noise environment through singular value decomposition. *Proceedings of the 5th International Conference on Signal Processing Applications and Technology, Vol.1*, Waltham, MA, USA, DSP Associates, 137–141.

Dillon, H. and Lovegrove, R. (1993). Single-microphone noise reduction systems for hearing aids: A review and an evaluation. Studebaker, G. A. and Hochberg, I., editors, *Acoustical Factors Affecting Hearing Aid Performance*, Allyn and Bacon, chapter 20, 353–372.

Doblinger, G. (1995). Computationally efficient speech enhancement by spectral minima tracking in subbands. *Proceedings of the 4th European Conference on Speech Communication and Technology EUROSPEECH '95. Madrid, Spain, September 1995.*, ESCA, 1513–1516.

Downs, D. W. and Crum, M. A. (1978). Processing demands during auditory learning under degraded listening conditions. *Journal of Speech and Hearing Research* **21** (4), 702–714.

Draxler, C. (1995). Introduction to the Verbmobil-PhonDat Database of Spoken German. *Proceedings of the Third International Conference on the Practical Application of Prolog*, Paris, 201–212.

Duineveld, C. A. A., Arents, P., and King, B. M. (2000). Log-linear modelling of paired comparison data from consumer tests. *Food Quality and Preference* **11**, 63–70.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. Academic Press, New York.

El-Maleh, K. and Kabal, P. (1997). Comparison of voice activity detection algorithms for wireless personal communications systems. *Proceedings of the CCECE '97 Canadian Conference on Electrical and Computer Engineering. Vol.2*, New York, NY, USA, IEEE, 470–473.

Elberling, C., Ludvigsen, C., and Keidser, G. (1993). The design and testing of a noise reduction algorithm based on spectral subtraction. *Scand. Audiol.* **Suppl. 38**, 39–49.

Ephraim, Y. (1992). A Bayesian estimation approach for speech enhancement using Hidden Markov Models. *IEEE Transactions on Signal Processing* **40** (4), 725–735.

Ephraim, Y. and Malah, D. (1983). Speech enhancement using optimal non-linear spectral amplitude estimation. *International Conference on Acoustics, Speech, and Signal Processing 1983, Conference Proceedings*, New York, NY, USA, IEEE, 1118–1121.

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

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

Fabry, D. A. and Schum, D. J. (1994). The role of subjective measurement techniques in hearing aid fittings. Valente, M., editor, *Strategies for Selecting and Verifying Hearing Aid Fittings*, Thieme Medical Publishers, New York, 136–155.

Fienberg, S. E. and Larntz, K. (1976). Log linear representation for paired and multiple comparisons models. *Biometrika* **63**, 245–254.

Fischer, A. and Stahl, V. (1999). On improvement measures for spectral subtraction applied to robust automatic speech recognition in car environments. *Proceedings of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions. Tampere, Finland, May 1999*, 75–78.

Ford, L. R., J. (1957). Solution of a ranking problem from binary comparisons. *Amer. Math. Monthly* **64** (8), 28–33.

Franck, B. A. M., Van Krefeld-Bos, S. G. M., and Dreschler, W. A. (1999). Evaluation of spectral enhancement in hearing aids, combined with phonemic compression. *J. Acoust. Soc. Am.* **106** (3), 1452–1464.

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

Gannot, S., Burshtein, D., and Weinstein, E. (1997). Iterative-batch and sequential algorithms for single microphone speech enhancement. *International Conference on Acoustics, Speech, and Signal Processing 1997, Conference Proceedings*, Los Alamitos, CA, USA, IEEE.

Gatehouse, S. (1992). The time course and magnitude of perceptual acclimatization to frequency responses: Evidence from monaural fitting of hearing aids. *J. Acoust. Soc. Am.* **92** (3), 1258–1268.

Gatehouse, S. (1994). Components and determinants of hearing aid benefit. *Ear and Hearing* **15**, 30–49.

Gediga, G. (1998). *Skalierung: Eine Einführung in die Methodik zur Entwicklung von Test- und Meßinstrumenten in den Verhaltenswissenschaften*, volume 5 of *Osnabrücker Schriften zur Psychologie*. LIT, Münster, Germany.

Gerven, S. V. and Xie, F. (1997). A comparative study of speech detection methods. *Proceedings of the 5th European Conference on Speech Communication and Technology, EUROSPEECH '97, Rhodes, Greece*.

Gray, R. M., Buzo, A., Gray Jr., A. H., and Matsuyama, Y. (1980). Distortion measures for speech processing. *IEEE Trans. Acoust., Speech, Signal Processing* **28**, 367–376.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969). Analysis of categor-

ical data by linear models. *Biometrics* **25**, 489–504.

Gülzow, T., Engelsberg, A., and Heute, U. (1998). Comparison of a discrete wavelet transformation and a nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement. *Signal Processing* **64**, 5–19.

Gustafsson, S., Martin, R., and Vary, P. (1996). On the optimization of speech enhancement systems using instrumental measures. *Workshop on Quality Assessment in Speech, Audio and Image Communication, Darmstadt, Germany, March, 11th–13th, 1996*, ITG Informationstechnische Gesellschaft, EURASIP European Association for Signal Processing, 36–40.

Halka, U. and Heute, U. (1992). A new approach to objective quality-measures based on attribute-matching. *Speech Communication* **11** (1), 15–30.

Hansen, J. and Pellom, B. (1998). An effective quality evaluation protocol for speech enhancement algorithms. *Proceedings ICSLP '98, Sydney, Australia.*

Hansen, M. (1998). *Assessment and Prediction of Speech Transmission Quality with an Auditory Processing Model.* PhD thesis, Universität Oldenburg, Oldenburg.

Hansen, M. and Kollmeier, B. (2000). Objective modeling of speech quality with a psychoacoustically validated auditory model. *J. Audio. Eng. Soc.* **48** (5), 395–409.

Haulick, T., Linhard, K., and Schrögmeier, P. (1997). Residual noise suppression using psychoacoustic criteria. *Proceedings of the 5th European Conference on Speech Communication and Technology, EUROSPEECH '97, Rhodes, Greece, Vol.3*, 1395–1398.

Hecker, M. H. L., Stevens, K. N., and Williams, C. E. (1966). Measurements of reaction time in intelligibility tests. *Journal of the Acoustical Society of America* **39** (6), 1188–1189.

Heide, D. (1994). Encoded speech intelligibility improvement in the F/A-18 noise environment using spectral subtraction preprocessing. *Proceedings of the 5th International Conference on Signal Processing Applications and Technology, Vol.2*, Waltham, MA, USA, DSP Associates, 1535–1540.

Hirsch, H. G. (1993). Estimation of noise spectrum and its application to SNR-estimation and speech enhancement. Technical Report TR-93-012, International Computer Science Institute, Berkeley, California, USA.

Hirsch, H. G. and Ehrlicher, C. (1995). Noise estimation techniques for robust speech recognition. *International Conference on Acoustics, Speech, and Signal Processing 1995, Conference Proceedings Vol.1*, New York, NY, USA, IEEE, 153–156.

Hoeks, B. and Levelt, W. J. M. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers* **25** (1), 16–26.

Holding, D. H. (1983). Fatigue. Hockey, R., editor, *Stress and Fatigue in*

Human Performance, John Wiley & Sons, New York, chapter 6, 145–167.

Holube, I. and Kollmeier, B. (1993). A perception model to predict speech intelligibility in impaired listeners using psychoacoustical parameters. Schick, A., editor, *Contributions to Psychological Acoustics, 6th Oldenburg Symposium*, Oldenburg, BIS Oldenburg, 557–566.

Hornik, K. (2000). The R FAQ. `http://www.ci.tuwien.ac.at/~hornik/R/`.

Humes, L. E., Christensen, L. A., Bess, F. H., and Hedley-Williams, A. (1997). A comparison of the benefit provided by well-fit linear hearing aids and instruments with automatic reductions of low-frequency gain. *Journal of Speech, Language, and Hearing Research* **40**, 666–685.

Hygge, S., Rönnberg, J., Larsby, B., and Arlinger, S. (1992). Normal-hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds. *Journal of Speech and Hearing Research* **35**, 208–215.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5** (3), 299–314.

Itoh, K. and Mizushima, M. (1997). Environmental noise reduction based on speech/non-speech identification for hearing aids. *International Conference on Acoustics, Speech, and Signal Processing 1997, Conference Proceedings*, Los Alamitos, CA, USA, IEEE, 419–422.

ITU (1996a). *ITU-T Recommendation G.729 – Annex B: A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70.* International Telecommunication Union.

ITU (1996b). *ITU-T Recommendation P.800: Methods for Subjective Determination of Transmission Quality.* International Telecommunication Union.

ITU (1996c). *ITU-T Recommendation P.861: Objective quality measurement of telephone-band (300-3400 Hz) speech codecs.* International Telecommunication Union.

Ivarsson, U. S. and Arlinger, S. D. (1993). Speech recognition in noise before and after a work-day's noise exposure. *Scand. Audiol.* **23**, 159–163.

Johnson, D. M. and Mullally, C. R. (1969). Correlation-and-regression model for category judgments. *Psychological Review* **76** (2), 205–215.

Jones, D. M. (1983). Noise. Hockey, R., editor, *Stress and Fatigue in Human Performance*, John Wiley & Sons, New York, chapter 3, 61–95.

Kang, G. S. and Fransen, L. J. (1989). Quality improvement of LPC-processed noisy speech by using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **37** (6), 930–942.

Kendall, M. and Gibbons, J. D. (1990). *Rank Correlation Methods.* Edward Arnold, London, 5th edition.

Kinkel, M. and Kollmeier, B. (1992). Binaurales Hören bei Normal- und Schwerhörigen II: Analyse der Ergebnisse. *Audiologische Akustik* **31**, 22–

33.

Kinkel, M., Kollmeier, B., and Holube, I. (1991). Binaurales Hören bei Normal- und Schwerhörigen I: Methoden und Ergebnisse. *Audiologische Akustik* **30**, 192–201.

Kleinschmidt, M., Marzinzik, M., and Kollmeier, B. (1999). Combining monaural noise reduction algorithms and perceptive preprocessing for robust speech recognition. Dau, T., Hohmann, V., and Kollmeier, B., editors, *Psychophysics, Physiology and Models of Hearing*, World Scientific Publishing, Singapore, 267–270. ISBN 981-02-3741-3.

Koehler, K. J. and Ridpath, H. (1982). An application of a biased version of the Bradley-Terry-Luce model to professional basketball results. *Journal of Mathematical Psychology* **25**, 187–205.

Kollmeier, B., Peissig, J., and Hohmann, V. (1993). Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain. *Scand. Audiol.* **Suppl. 38**, 28–38.

Kollmeier, B. and Wesselkamp, M. (1997). Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *J. Acoust. Soc. Am.* **102** (4), 2412–2421.

Kousgaard, N. (1987). The application of binary paired comparisons to listening tests. Bech, S. and Pedersen, O. J., editors, *Perception of Reproduced Sound*, 71–80. ISBN 87-982562-1-1.

Kuk, F. and Tyler, R. (1990). Relationship between consonant recognition and subjective ratings of hearing aids. *Br. J. Audiol.* **24**, 171–177.

Kuk, F. K. (1994). Use of paired comparisons in hearing aid fittings. Valente, M., editor, *Strategies for Selecting and Verifying Hearing Aid Fittings*, Thieme Medical Publishers, New York, chapter 6, 108–135.

Kuk, F. K., Tyler, R. S., and Mims, L. (1990). Subjective ratings of noise-reduction hearing aids. *Scand. Audiol.* **19**, 237–244.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association* **73**, 253–263.

Lawal, H. B. (1984). Comparisons of the X^2, Y^2, Freeman-Tukey and William's Improved G^2 test statistics in small samples of one-way multinomials. *Biometrika* **71**, 415–418.

Levitt, H., Bakke, M., Kates, J., Neuman, A., Schwander, T., and Weiss, M. (1993). Signal processing for hearing impairment. *Scand. Audiol.* **Suppl. 38**, 7–19.

Lim, J. S. and Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE* **67** (12), 1586–1604.

Lim, J. S. and Wang, D. Y. (1982). The unimportance of phase in speech enhancement. *IEEE Transactions on Acoustics, Speech and Signal Processing* **ASSP-30**, 679–681.

Luce, R. D. (1959). *Individual Choice Behaviour: A Theoretical Analysis*. Wiley, New York.

Lukas, J. (1991). BTL-Skalierung verschiedener Geschmacksqualitäten von Sekt. *Zeitschrift für experimentelle und angewandte Psychologie* **38** (4), 605–619.

Lybarger, S. (1944). Method of fitting hearing aids. U.S. Patent Application S.N. 543,278.

Lybarger, S. (1978). Selective amplification – a review and evaluation. *J. Am. Audio. Soc.* **3** (6), 258–266.

Malca, Y., Wulich, D., Ramponi, G., Sicuranza, G. L., Carrato, S., and Marsi, S. (1996). Improved spectral subtraction for speech enhancement. *Signal Processing VIII, Theories and Applications. Proceedings of EUSIPCO-96, Vol.2*, Trieste, Italy, LINT, 975–978.

Marascuilo, L. A. and McSweeney, M. (1977). *Nonparametric and Distribution-Free Methods for the Social Sciences*. Brooks/Cole, Monterey, California.

Martin, R. (1993). An efficient algorithm to estimate the instantaneous SNR of speech signals. *Proceedings EUROSPECH '93, Vol.1*, ESCA.

Martin, R. (1994). Spectral subtraction based on minimum statistics. Holt, M. J. J., Cowan, C. F. N., Grant, P. M., and Sandham, W. A., editors, *Signal Processing VII, Theories and Applications. Proceedings of EUSIPCO-94, Vol.I*, Lausanne, Switzerland, European Association for Signal Processing.

Marzinzik, M., Hohmann, V., Appell, J.-E., and Kollmeier, B. (1999a). Dynamic compression algorithms: Laboratory evaluation with hearing-impaired subjects. *Zeitschrift für Audiologie* **38** (1), 16–25.

Marzinzik, M. and Kollmeier, B. (1999). Development and evaluation of single-microphone noise reduction algorithms for digital hearing aids. Dau, T., Hohmann, V., and Kollmeier, B., editors, *Psychophysics, Physiology and Models of Hearing*, World Scientific Publishing, Singapore, 279–282. ISBN 981-02-3741-3.

Marzinzik, M. and Kollmeier, B. (2000). Quality assessment of noise reduction for digital hearing aids: Measurements and predictions. *Fortschritte der Akustik – DAGA 2000*, Oldenburg, DEGA e.V.

Marzinzik, M., Wittkop, T., and Kollmeier, B. (1999b). Combination of monaural and binaural noise suppression algorithms and its use for the hearing impaired. *J. Acoust. Soc. Am.* **105** (2, Pt. 2), 1211.

McAulay, R. J. and Malpass, M. L. (1980). Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Processing* **28**, 137–145.

McKinley, B. L. and Whipple, G. H. (1997). Model based speech pause detection. *International Conference on Acoustics, Speech, and Signal Processing 1997, Conference Proceedings*, Los Alamitos, CA, USA, IEEE, 1179–1182.

Meyer, J. and Simmer, K. U. (1997). Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction. *International Conference on Acoustics, Speech, and Signal Processing 1997, Conference Proceedings*, Los Alamitos, CA, USA, IEEE.

Montgomery, A. A. and Edge, R. A. (1988). Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults. *Journal of Speech and Hearing Research* **31** (3), 386–393.

Motulsky, H. J. (1995). *Intuitive Biostatistics*. Oxford University Press.

Muscio, B. (1921). Is a fatigue test possible? *British Journal of Psychology* **12**, 31–46. Cited in: Holding (1983).

Nemer, E., Goubran, R., and Mahmoud, S. (1999). SNR estimation of speech signals using subbands and fourth-order statistics. *IEEE Signal Processing Letters* **6** (7), 171–174.

Niederjohn, R. J., Lee, P.-J., and Josse, F. (1987). Factors related to spectral subtraction for speech in noise enhancement. *Proceedings of IECON '87*, New York, IEEE, 985–996.

Nishimura, R., Asano, F., Suzuki, Y., and Sone, T. (1998). Speech enhancement using spectral subtraction with wavelet transform. *Electronics and Communications in Japan* **81** (1), 24–31.

Ostendorf, M., Hohmann, V., and Kollmeier, B. (1997). Empirische Klassifizierung verschiedener akustischer Signale und Sprache mittels einer Modulationsfrequenzanalyse. *Fortschritte der Akustik – DAGA 97*, Oldenburg, Germany, DEGA.

Ostendorf, M., Hohmann, V., and Kollmeier, B. (1998). Klassifikation von akustischen Signalen basierend auf der Analyse von Modulationsspektren zur Anwendung in digitalen Hörgeräten. *Fortschritte der Akustik – DAGA 98*, Oldenburg, Germany, DEGA, 402–403.

Paul, D. B. (1981). The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **ASSP-29** (4), 786–794.

Peissig, J. (1993). *Binaurale Hörgerätestrategien in Störschallsituationen*. VDI-Verlag, Düsseldorf.

Peterson, T. L. and Boll, S. F. (1981). Acoustic noise suppression in the context of a perceptual model. *International Conference on Acoustics, Speech, and Signal Processing 1981, Conference Proceedings, Vol.3*, New York, IEEE, 1086–1088.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C – The Art of Scientific Computing*. Cambridge University Press.

Pressnitzer, D. and McAdams, S. (1997). Influence of phase effects on roughness modeling. *International Computer Music Conference ICMC 97, Thessaloniki, Greece*.

Preuss, R. D. (1979). A frequency domain noise cancelling preprocessor for narrowband speech communication systems. *International Conference on*

Acoustics, Speech, and Signal Processing 1979, Conference Proceedings, New York, IEEE, 212–215.

Punch, J. and Parker, C. (1981). Pairwise listener preferences in hearing aid evaluation. *J. Speech Hear. Res.* **24**, 366–374.

Punch, J. L. (1978). Quality judgments of hearing aid-processed speech and music by normal and otopathologic listeners. *Journal of the American Audiology Society* **3** (4), 179–188.

Quackenbush, S. R., Barnwell, T. P., and Clements, M. A. (1988). *Objective Measures of Speech Quality*. Prentice Hall, New Jersey.

Rangoussi, M. and Carayannis, G. (1995). Higher order statistics based Gaussianity test applied to on-line speech processing. *Proceedings of the IEEE Asilomar Conference*, IEEE, 303–307.

Rankovic, C. M. and Levy, R. M. (1997). Estimating articulation scores. *J. Acoust. Soc. Am.* **102** (6), 3754–3761.

Revoile, S. G. and Holden-Pitt, L. D. (1993). Some acoustic enhancements of speech and their effect on consonant identification by the hearing impaired. Studebaker, G. A. and Hochberg, I., editors, *Acoustical Factors Affecting Hearing Aid Performance*, Allyn and Bacon, chapter 21, 373–385.

Roethlisberger, F. J. and Dickson, W. J. (1939). *Management, and the Worker*. Harvard University Press, Cambridge, MA. Cited in: Jones (1983).

Scalart, P. and Vieira Filho, J. (1996). Speech enhancement based on a priori signal to noise estimation. *International Conference on Acoustics, Speech, and Signal Processing 1996, Conference Proceedings*, New York, IEEE, 629–632.

Scalart, P., Vieira Filho, J., and Geraldo Chiquito, J. (1996). On speech enhancement algorithms based on MMSE estimation. *Signal Processing VIII, Theories and Applications. Proceedings of EUSIPCO-96, Vol.I*, Trieste, Italy, LINT, 471–474.

Schmidt-Nielsen, A. (1987). Comments on the use of physical measures to assess speech intelligibility. *J. Acoust. Soc. Am.* **81** (6), 1985–1987.

Sheikhzadeh, H., Brennan, R. L., and Sameti, H. (1995). Real-time implementation of HMM-based MMSE algorithm for speech enhancement in hearing aid applications. *International Conference on Acoustics, Speech, and Signal Processing 1995, Conference Proceedings Vol.1*, New York, NY, USA, IEEE, 808–811.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

Singh, L. and Sridharan, S. (1998). Speech enhancement using critical band spectral subtraction. *Proceedings ICSLP '98, Sydney, Australia*, 2827–2830.

Sohn, J., Kim, N. S., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters* **6** (1), 1–3.

Soon, I. Y., Koh, S. N., and Yeo, C. K. (1998). Noisy speech enhancement using discrete cosine transform. *Speech Communication* **24**, 249–257.

Sovka, P. and Pollák, P. (1995). The study of speech/pause detectors for speech enhancement methods. *Proceedings of the 4th European Conference on Speech Communication and Technology EUROSPEECH '95. Madrid, Spain, September 1995.*, ESCA, 1575–1578.

Speaks, C., Parker, B., Harris, C., and Kuhl, P. (1972). Intelligibility of connected discourse. *J. Speech Hear. Res.* **15**, 590–602.

Srinivasan, K. and Gersho, A. (1993). Voice activity detection for cellular networks. *Proceedings of the IEEE Speech Coding Workshop*, IEEE, 85–86.

Steeneken, H. J. M. and Geurtsen, F. W. M. (1988). Description of the RSG.10 noise database. Technical Report IZF 1988-3, TNO Institute for Perception, Soesterberg, The Netherlands.

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

Stegmann, J. and Schröder, G. (1997). Robust voice-activity detection based on the wavelet transform. *Proceedings of the 1997 IEEE Workshop on Speech Coding for Telecommunications*, New York, NY, USA, IEEE, 99–100.

Stern, H. (1992). Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences* **23** (1), 103–117. ISSN 0165-4896.

Studebaker, G. (1982). Hearing aid selection: An overview. Studebaker, G. and Bess, F., editors, *The Vanderbilt Hearing-Aid Report: State of the Art Research Needs*, Upper Darby, PA, 147–155. Cited in: Kuk (1994).

Studebaker, G., Bisset, J., Van Ort, D., and Hoffnung, S. (1982). Paired comparison judgments of relative intelligibility in noise. *J. Acoust. Soc. Am.* **72**, 80–92.

Studebaker, G. A. and Hochberg, I. (1993). *Acoustical Factors Affecting Hearing Aid Performance*. Allyn and Bacon.

Tchorz, J. (2000). *Auditory-Based Signal Processing for Noise Suppression and Robust Speech Recognition*. PhD thesis, Carl von Ossietzky Universität Oldenburg, Germany.

Tecca, J. and Goldstein, D. (1984). Effect of low frequency hearing aid response on four measures of speech perception. *Ear and Hearing* **5**, 22–29.

Teebagy, N. and Chatterjee, S. (1989). Inference in a binary response model with applications to data analysis. *Decision Sciences* **20**, 393–403.

TenVergert, E., Gillespie, M., and Kingma, J. (1993). Testing the assumptions and interpreting the results of the Rasch model using log-linear procedures in SPSS. *Behavior Research Methods, Instruments, & Computers* **25** (3), 350–359.

TIA (1996). Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems. Document PN-3292.

Tsoukalas, D. E., Mourjopoulos, J. N., and Kokkinakis, G. (1997a). Perceptual filters for audio signal enhancement. *J. Audio Eng. Soc.* **45** (1/2), 22–36.

Tsoukalas, D. E., Mourjopoulos, J. N., and Kokkinakis, G. (1997b). Speech enhancement based on audible noise suppression. *IEEE Transactions on Speech and Audio Processing* **5** (6), 497–514.

Tutz, G. (1986). Bradley-Terry-Luce models with an ordered response. *Journal of Mathematical Psychology* **30**, 306–316.

Vary, P. (1985). Noise suppression by spectral magnitude estimation – Mechanism and theoretical limits. *Signal Processing* **8**, 387–400.

Virag, N. (1999). Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on Speech and Audio Processing* **7** (2), 126–137.

Wagener, K., Kühnel, V., Brand, T., and Kollmeier, B. (1998). Entwicklung und Evaluation eines Sprachverständlichkeitstests für die deutsche Sprache. *Fortschritte der Akustik - DAGA 98*, Oldenburg, DEGA, 322–323. ISBN 3-9804568-3-8.

Wagener, K., Kühnel, V., and Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache. *Z. Audiol.* **38** (1, 2, 3).

Weiss, M. and Neuman, A. C. (1993). Noise reduction in hearing aids. Studebaker, G. A. and Hochberg, I., editors, *Acoustical Factors Affecting Hearing Aid Performance*, Allyn and Bacon, chapter 19, 337–352.

Wesselkamp, M. and Kollmeier, B. (1993). Vergleich von gemessener und subjektiv skalierter Sprachverständlichkeit mit einem optimierten Satztest. *Fortschritte der Akustik*, DAGA'93, Bad Honnef, DPG-Kongreß-GmbH, 1064–1067.

Wittkop, T. (2000). *Two-Channel Noise Reduction Algorithms Motivated by Models of Binaural Interaction.* PhD thesis, Carl von Ossietzky Universität Oldenburg, Germany. To be submitted.

Wittkop, T., Hohmann, V., and Kollmeier, B. (1999). Noise reduction strategies employing interaural parameters. *ACUSTICA · Acta Acustica* **85**, 285.

Working Group on Communication Aids for the Hearing-Impaired (1991). Speech-perception aids for hearing-impaired people: Current status and needed research. *J. Acoust. Soc. Am.* **90** (2), 637–685.

Wright, S. P. (1992). Adjusted p-values for simultaneous inference. *Biometrics* **48**, 1005–1013.

Wyatt, S. (1950). An autobiography. *Occupational Psychology* **24**, 65–74. Cited in: Davies *et al.* (1983), p. 5.

Zerlin, S. (1962). A new approach to hearing-aid selection. *Journal of Speech and Hearing Research* **5** (4), 370–376.

Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Math. Zeit.* **29**, 436–460.

Zwicker, E. and Terhardt, E. (1980). Analytical expressions for critical-band

rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.* **68** (5), 1523–1525.

Erklärung

Hiermit erkläre ich, daß ich die vorliegende Arbeit selbständig verfaßt und keine anderen als die angegebenen Hilfsmittel benutzt habe.

Oldenburg, den 20. November 2000

Mark Marzinzik

Danksagung

Die vorliegende Dissertation entstand während meiner Tätigkeit als wissenschaftlicher Angestellter in der Arbeitsgruppe Medizinische Physik an der Carl von Ossietzky Universität Oldenburg, gefördert durch die Europäische Union im Rahmen des Projektes “SPACE” (DE 3012).

Dem Leiter der Arbeitsgruppe, Herrn Prof. Dr. Dr. Birger Kollmeier, möchte ich herzlich danken für die Anregung zu dieser Dissertation, für die ausgezeichneten Arbeitsbedingungen und insbesondere für die sehr gute Betreuung und Unterstützung während meiner ganzen Arbeit.

Herrn Prof. Dr. Volker Mellert danke ich für die freundliche Übernahme des Korreferats.

Der Firma GN ReSound danke ich dafür, daß sie durch einen Forschungsauftrag wesentliche Impulse für Teile der vorliegenden Arbeit gegeben hat.

Allen Kollegen gilt mein Dank für die hervorragende Arbeitsatmosphäre und zahlreiche anregende Diskussionen. Mein ganz besonderer Dank gilt dabei Jürgen Tchorz und Dr. Thomas Brand, die mit mir über Jahre in einem Büro “gesessen” haben.

Erwähnen möchte ich auch noch Thomas Wittkop und Michael Kleinschmidt, mit denen ich enger zusammenarbeiten durfte. Uns verbindet nicht zuletzt das gemeinsame Interesse an den Möglichkeiten der Störgeräuschunterdrückung.

Schließlich gilt mein herzlicher Dank meinen Eltern, die mir überhaupt erst das Studium der Physik ermöglicht haben, sowie meiner Frau Kerstin, die mich so manches Mal ermutigt hat und viel Verständnis für diese Arbeit aufgebracht hat. Ihr widme ich diese Dissertation.

Oldenburg, im November 2000

Mark Marzinzik

Lebenslauf

Am 28. Juni 1970 wurde ich, Mark Marzinzik, als erstes Kind von Günter und Brunhild Marzinzik, geb. Bödeker, in Bremen mit deutscher Staatsangehörigkeit geboren.

Von 1976–1980 besuchte ich die Grundschule in Leeste, von 1980–1982 die dortige Orientierungsstufe und von 1982–1989 das Gymnasium der KGS Weyhe, an dem ich meine Schulbildung mit dem Abitur abschloß.

Meinen Wehrdienst leistete ich von Juni 1989 bis August 1990 in Goslar und Delmenhorst ab.

Das Physikstudium nahm ich im Oktober 1990 an der Universität Oldenburg auf, von der ich im Oktober 1992 das Vordiplom erhielt. Meine Diplomarbeit über "Dynamikkompression für Hörgeräte" fertigte ich in der Arbeitsgruppe Medizinische Physik unter der Leitung von Prof. Dr. Dr. Birger Kollmeier an. Im Juni 1996 schloß ich mein Physikstudium mit der Diplomprüfung ab.

Von Juli 1996 bis Juni 2000 arbeitete ich als wissenschaftlicher Angestellter in der Arbeitsgruppe Medizinische Physik. Dort fertigte ich unter der Anleitung von Prof. Dr. Dr. Birger Kollmeier die vorliegende Dissertation an und arbeitete mit an einem europäischen Projekt zur Entwicklung und Evaluation von Signalverarbeitungsstrategien für Schwerhörige.

Für den Abschluß der Dissertation erhielt ich von Juli bis Dezember 2000 ein Promotionsstipendium des Europäischen Graduiertenkollegs "Neurosensorik", gefördert durch die Deutsche Forschungsgemeinschaft.